الجمهورية العربية السورية
جامعة دمشق
كلية الاقتصاد
قسم الإحصاء التطبيقي



استخدام تقانة اكتشاف المعرفة من قواعد البيانات في دراسة بيانات مخابر التحاليل الطبية

(حالة عملية)

Using Knowledge Discovery in Databases
To Study Clinical Laboratories Data

بحث علمي أعد لنيل درجة الدكتوراه في الإحصاء التطبيقي
اختصاص قواعد بيانات

إعداد

المعيدة أنمار نافع عرابي

إشراف

الدكتورة ميساء حمشو

أستاذ مساعد في كلية الاقتصاد بجامعة دمشق

العام الدراسي 2010-2011

استخدام تقانة اكتشاف المعرفة من قواعد البيانات
في دراسة بيانات مخابر التحاليل الطبية
(حالة عملية)



بحث علمي أعد لنيل درجة الدكتوراه
في الإحصاء التطبيقي
اختصاص قواعد بيانات

إعداد

المعيدة أنمار نافع عرابي

إشراف

الدكتورة ميساء حمشو

أستاذ مساعد في كلية الاقتصاد بجامعة دمشق

العام الدراسي 2010-2011

أعضاء لجنة الحكم

د. راكـــان رزوق

الأستاذ في قسم هندسة البرمجيات
كلية الهندسة المعلوماتية – جامعة دمشق
الاختصاص: قواعد بيانات

د. شـــفيق عربش

الأستاذ في قسم الإحصاء التطبيقي
كلية الاقتصاد – جامعة دمشق
الاختصاص: انفورماتيك

د. أحمد صقر أحمد

الأستاذ في قسم هندسة النظم والشبكات الحاسوبية
كلية الهندسة المعلوماتية – جامعة تشرين
الاختصاص: نظم نقل المعلومات

د. ميســـاء حمشو

الأستاذ المساعد في قسم الإحصاء التطبيقي
كلية الاقتصاد – جامعة دمشق
الاختصاص: المعالجة الآلية للمعلومات

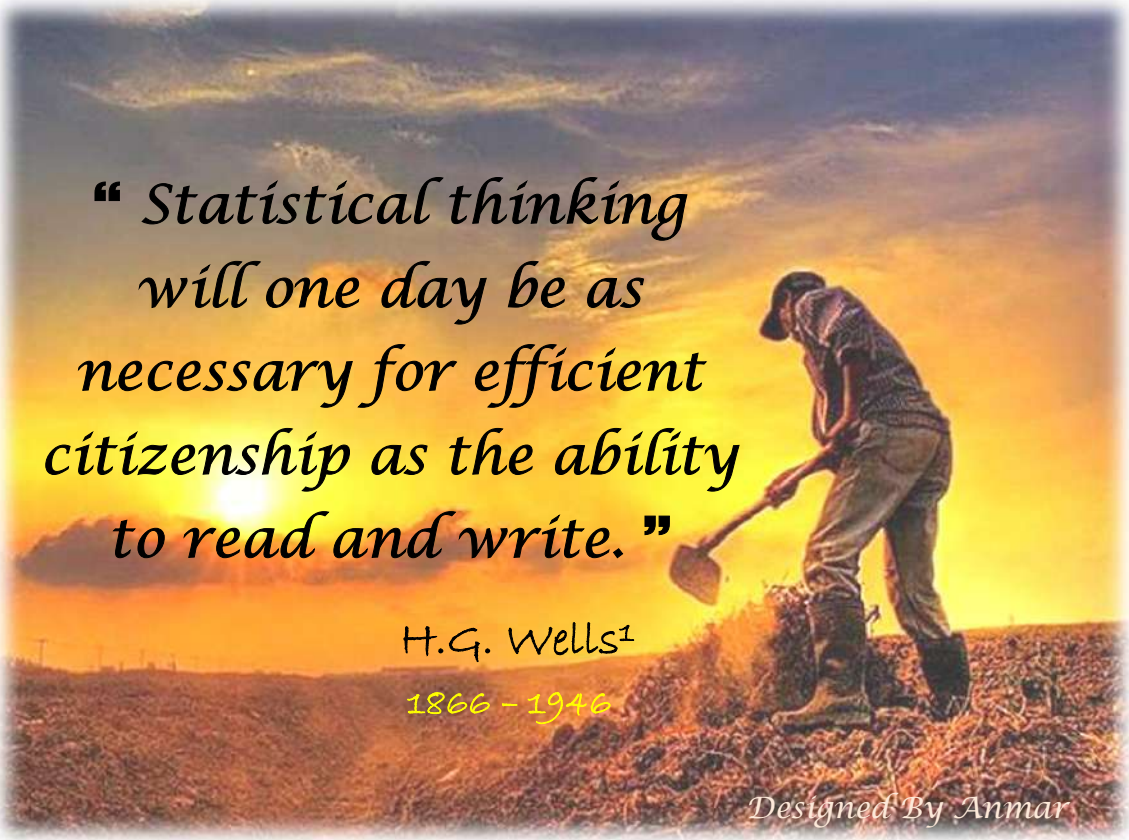
د. عثمـــان نقـــار

المدرس في قسم الإحصاء التطبيقي
كلية الاقتصاد – جامعة دمشق
الاختصاص: المعالجة الآلية للمعلومات

بسم الله الرحمن الرحيم

"وكان فضل الله عليك عظيماً"

صدق الله العظيم



[1] *Herbert George Wells* was an English author, now best known for his work in the science fiction genre. He was also a prolific writer in many other genres, including contemporary novels, history, politics and social commentary, even writing text books. H.G. Wells's best known works are THE TIME MACHINE (1895), one of the first modern science fiction stories, THE INVISIBLE MAN (1897), and THE WAR OF THE WORLDS (1898).

باقة إهداء إلى



﴿ شكر وامتنان ﴾

إلى كل من دعمني وزودني بالمعلومات المفيدة التي أغنت البحث

وأخص بالشكر

الدكتورة ميساء حمشو

وأعضاء مجلس قسم الإحصاء التطبيقي

وأعضاء لجنة الحكم

﴿﴾

فهرس المحتويات

الفصل التمهيدي	1
مقدمة	2
كلمات مفتاحية	3
أهمية البحث	3
مشكلة البحث	4
أهداف البحث	4
منهجية البحث	5
مجتمع البحث	5
لمحة عامة عن فصول البحث	5
دراسات سابقة	7
الفصل الأول: مفاهيم أساسية ومقارنات في مجال اكتشاف المعرفة من قواعد البيانات	11
مقدمة	12
1•1 المفاهيم الأساسية	13
1•1•1 اكتشاف المعرفة من قواعد البيانات (Knowledge Discovery in Database (KDD))	13
2•1•1 التنقيب في البيانات (Data Mining)	16
1•2•1•1 أنواع التنقيب في البيانات	17
2•2•1•1 أنماط خوارزميات التنقيب في البيانات	18
3•1•1 ذكاء الأعمال (Business Intelligence (BI))	20
4•1•1 قواعد البيانات العلاقاتية	20
5•1•1 مستودع البيانات (Data Warehouse)	21
6•1•1 مكعب البيانات (Data Cube)	23
7•1•1 المعالجة التحليلية الآنية (Online Analytical Processing (OLAP))	26
8•1•1 تعليم الآلة (Machine Learning)	27
9•1•1 العرض المرئي (Visualization)	28
2•1 مقارنات في مجال اكتشاف المعرفة من قواعد البيانات	28
1•2•1 مقارنة بين قاعدة البيانات (Databases) ومستودع البيانات (Data Warehouse)	28
2•2•1 مقارنة بين معالجة المبادلات الآنية (OLTP) والمعالجة التحليلية الآنية (OLAP)	29
3•2•1 مقارنة بين لغة الاستعلام المهيكلة (Structural Query Language (SQL)) ومكعب البيانات (Data Cube)	29

4•2•1 مقارنة بين التنقيب في البيانات (Data Mining) والمعالجة التحليلية الآنية للبيانات (OLAP)	30
5•2•1 مقارنة التنقيب في البيانات (Data Mining) والإحصاء (Statistics)	31
الفصل الثاني: أبرز مجالات تطبيق تقانة اكتشاف المعرفة من قواعد البيانات	33
مقدمة	34
1•2 مجالات الأعمال (Business Fields)	36
1•1•2 الاتصالات (Telecommunications)	36
2•1•2 التنقيب في استخدام شبكة الانترنت (Web Usage Mining (WUM))	39
3•1•2 التنقيب في بيانات الأسواق المركزية	40
4•1•2 استخدام التنقيب في البيانات في فهم الزبائن وتوقع مجالات اهتمامهم	42
5•1•2 استخدام التنقيب في البيانات في الإعلانات الموجهة لمستخدمي البريد الالكتروني	43
6•1•2 استخدام تقانة التنقيب في صناعة السياحة	44
7•1•2 استخدام التنقيب في البيانات في شركات التأمين	45
8•1•2 استخدام التنقيب في البيانات في التعليم العالي	45
2•2 مجالات الأبحاث العلمية	46
1•2•2 التنقيب في البيانات الطبية	46
2•2•2 استخدام التنقيب في البيانات في الأبحاث الجينية	53
3•2•2 التنقيب في مضمون المؤلفات والكتابة الالكترونية	51
4•2•2 استخدام التنقيب في البيانات في علم الفلك	55
3•2 مجالات أخرى للتنقيب في البيانات	57
1•3•2 استخدام التنقيب في البيانات في مراقبة الأشخاص عبر الانترنت وفي توقع جماعات الإرهاب	57
2•3•2 استخدام التنقيب في الصور (Image Mining) في محرك البحث Google	58
الفصل الثالث: مراحل عملية اكتشاف المعرفة من قواعد البيانات	64
مقدمة	65
1•3 مرحلة استيعاب مجال المسألة (Understanding of the problem domain)	68
1•1•3 تعريف حدود المسألة والمجال الذي تطبق فيه	68
2•1•3 تحديد المستفيدين المستهدفين من عملية اكتشاف المعرفة	69
3•1•3 تحديد الأدوات البرمجية الأساسية اللازمة لانجاز مراحل اكتشاف المعرفة	69
2•3 مرحلة استيعاب البيانات (Understanding of the Data)	70
3•3 مرحلة تهيئة البيانات (Data Preparation)	72
1•3•3 انتقاء الحقول التي تفيد في انجاز مراحل اكتشاف المعرفة	72
2•3•3 تنظيف وتنسيق البيانات وإنشاء بيانات تركيب جديدة	73

1•2•3•3 تنظيف وتنسيق البيانات (Data Cleaning and Formatting)	73
1•1•2•3•3 التعامل مع القيم المفقودة (Missing Values)	74
2•1•2•3•3 التعامل مع البيانات المشوشة (Noisy Data)	75
3•1•2•3•3 تنسيق البيانات (Data Formatting)	76
2•2•3•3 إنشاء بيانات تركيب جديدة (New Constructed data)	76
3•3•3 بناء مستودع بيانات جديد (New Data Warehouse)	77
4•3•3 المعالجة التحليلية الآنية للبيانات (OLAP)	79
4•3 مرحلة التنقيب في البيانات	80
1•4•3 إنشاء مشروع التنقيب وتحديد مصدر البيانات	81
2•4•3 بناء هيكل التنقيب	81
3•4•3 تصميم نماذج التنقيب	83
1•3•4•3 اختيار حقول البيانات ومهامها في النموذج	83
2•3•4•3 تخصيص نموذج التنقيب بمجموعة جزئية من البيانات	84
3•3•4•3 اختيار خوارزمية التنقيب	84
4•3•4•3 تعديل محددات خوارزمية النموذج (Parameters) في حالات الضرورة	89
4•4•3 استعراض مخرجات نماذج التنقيب	90
5•4•3 استخدام اكسل في عمليات التنقيب في البيانات	90
5•3 مرحلة تقييم المعرفة المكتشفة (Evaluation of the Knowledge Discovered)	91
6•3 مرحلة استخدام المعرفة المكتشفة (Use of the Discovered Knowledge)	92
7•3 التغذية الراجعة (Feed Back)	93
الفصل الرابع: التعريف بمصدر بيانات مخبر التحاليل الطبية	96
مقدمة	97
1•4 مفاهيم أساسية	97
1•1•4 مخبر التحاليل الطبية	97
2•1•4 بيانات مخبر التحاليل الطبية	98
3•1•4 السجل الطبي الالكتروني (Electronic Medical Record (EMR))	98
4•1•4 نظام المعلومات الطبية في المشفى (Hospital Information System (HIS))	99
5•1•4 سجل الصحة الالكتروني (Electronic Health Record (HER))	100
6•1•4 تقانة المعلومات الصحية (Health Information Technology (HIT))	101
2•4 آلية تجميع وتخزين البيانات في مخبر التحاليل الطبية	102
3•4 الأبحاث العلمية المستندة إلى بيانات التحاليل المخبرية الطبية	104

1•3•4 الأبحاث المخبرية التقليدية	105
1•1•3•4 مراحل انجاز الأبحاث المخبرية الطبية التقليدية	105
2•1•3•4 أمثلة عن الأبحاث المخبرية التقليدية	106
2•3•4 الأبحاث المخبرية المستندة إلى التطورات التقنية في مجال مخابر التحاليل الطبية	108
1•2•3•4 الأبحاث المخبرية المرافقة لتطور الأجهزة والتقنيات المخبرية	108
2•2•3•4 الأبحاث المخبرية المرافقة لتراكم البيانات في الحواسيب	109
3•2•3•4 الأبحاث المخبرية المرافقة لتقانة اكتشاف المعرفة	110
الفصل الخامس: مراحل اكتشاف المعرفة من قواعد بيانات مخبر التحاليل الطبية	113
تمهيد	114
المبحث الأول: مرحلة استيعاب مجال مسألة اكتشاف المعرفة من بيانات مخبر التحاليل الطبية	116
مقدمة	116
1•1•5 تعريف حدود المسألة والمجال الذي تطبق فيه	116
1•1•1•5 مجال تطبيق تقانة اكتشاف المعرفة في دراسة بيانات مخبر التحاليل الطبية	117
1•1•1•1•5 واقع الأتمتة في مخبر التحاليل الطبية المدروس	118
2•1•1•1•5 واجهات البرنامج الحاسوبي المستخدم في المخبر المدروس	119
2•1•1•5 صياغة مسائل اكتشاف المعرفة من بيانات مخبر التحاليل الطبية المتاح	125
2•1•5 تحديد المستفيدين المستهدفين	126
3•1•5 تحديد الأدوات البرمجية التي ستُستخدَم في جميع المراحل	127
المبحث الثاني: مرحلة استيعاب بيانات مخبر التحاليل الطبية والبيانات المرتبطة بها	129
مقدمة	129
1•2•5 وصف مصادر البيانات التي تم الحصول عليها	129
1•1•2•5 قواعد بيانات مخبر التحاليل الطبية	130
2•1•2•5 قاعدة بيانات مكتب القبول في المشفى	131
3•1•2•5 قاعدة بيانات السجلات الطبية في المشفى	134
2•2•5 وصف جداول البيانات من حيث المضمون والحجم	135
1•2•2•5 قواعد بيانات مخبر التحاليل الطبية	135
2•2•2•5 قاعدة بيانات مكتب القبول	139
3•2•2•5 قاعدة بيانات مكتب السجلات الطبية	140
3•2•5 التعرف على بنى الجداول في قواعد البيانات المدروسة	140
1•3•2•5 التعرف على بنى الجداول في قاعدة بيانات مخبر التحاليل الطبية	141
2•3•2•5 التعرف على بنى الجداول في قاعدة بيانات مكتب القبول	152

3•3•2•5 التعرف على بنى الجداول في قاعدة بيانات قسم السجلات الطبية	155
المبحث الثالث: مرحلة تهيئة بيانات مخبر التحاليل الطبية	161
مقدمة	161
1•3•5 انتقاء الحقول التي تفيد في انجاز مراحل اكتشاف المعرفة	162
2•3•5 تنظيف وتنسيق البيانات وإنشاء بيانات تركيب جديدة	165
1•2•3•5 تنظيف وتنسيق البيانات وإنشاء بيانات تركيب جديدة في قاعدة بيانات مخبر التحاليل الطبية	166
2•2•3•5 تنظيف بيانات القبول في المشفى	177
3•2•3•5 تنظيف بيانات مكتب السجلات الطبية	183
4•2•3•5 دمج قواعد البيانات الثلاثة ضمن قاعدة بيانات واحدة في برنامج أكسس	184
3•3•5 بناء مستودع البيانات الجديد (Clinical Laboratory Study) في نظام (Microsoft SQL Server 2008)	186
1•3•3•5 تجهيز البيانات المراد نقلها	186
2•3•3•5 إنشاء بنية مستودع البيانات الجديد في SQL Server	191
3•3•3•5 إجراءات سحب البيانات من قاعدة البيانات إلى مستودع البيانات	192
4•3•3•5 تعريف العلاقات بين الجداول في مستودع البيانات	197
5•3•3•5 تطوير مستودع البيانات (Clinical Laboratory Study)	199
4•3•5 المعالجة التحليلية الآنية للبيانات ((On-line Analytical Processing (OLAP))	202
1•4•3•5 إنشاء مشروع المعالجة التحليلية الآنية	203
2•4•3•5 تعريف مصدر البيانات	203
3•4•3•5 تعريف مقاييس جديدة في مرأى مصدر البيانات	205
4•4•3•5 تعريف أبعاد المكعب	205
5•4•3•5 تعريف مكعب البيانات	207
6•4•3•5 تدقيق وبناء وتفعيل مكعب بيانات التحاليل الطبية	207
7•4•3•5 استطلاع البيانات (Explore Data)	209
8•4•3•5 بناء تراتبات هرمية (Hierarchy) للأبعاد	211
9•4•3•5 إضافة مقاييس جديدة داخل مكعب البيانات	212
10•4•3•5 تصفح مكعب البيانات في برنامج (Microsoft Visual Studio)	214
11•4•3•5 تصفح مكعب البيانات في برنامج (Microsoft Excel 2007)	222
المبحث الرابع: مرحلة التنقيب في بيانات مخبر التحاليل الطبية	228
مقدمة	228
1•4•5 مشروع التنقيب في البيانات الشخصية والمرضية لمرضى المشفى	229

5•1•4•1 إنشاء المشروع وتحديد مصدر البيانات	230
5•1•4•2 بناء هيكل التنقيب	230
5•1•4•3 تصميم واستعراض نماذج التنقيب التي تهدف إلى اكتشاف أسباب الوفاة الأكثر تأثراً بالخاصيات المختلفة للمرضى	232
5•1•4•3•1 نموذج أشجار القرار (DecisionTrees_DeathReason)	234
5•1•4•3•2 نموذج قواعد الاقتران (AssociationRules_DeathReason)	234
5•1•4•3•3 نموذج التجزئة العنقودية (Clustering_DeathReason)	238
5•1•4•4 تصميم واستعراض نماذج التنقيب الخاصة بالتنبؤ بحالة الخروج من المشفى	244
5•1•4•4•1 نموذج نايف بايز (NaiveBayes_OutStae)	245
5•1•4•4•2 نموذج الشبكات العصبونية (NeuralNetworks_OutState)	247
5•1•4•4•3 نموذج الانحدار المنطقي (LogisticRegresion_OutState)	250
5•2•4 مشروع التنقيب في بيانات نتائج التحاليل الطبية المخبرية	253
5•2•4•1 إنشاء المشروع وتحديد مصدر البيانات	253
5•2•4•2 بناء هيكل التنقيب	255
5•2•4•3 تصميم واستعراض نماذج التنقيب في بيانات نتائج التحاليل الطبية المخبرية	255
5•2•4•3•1 نموذج أشجار القرار (DecisionTrees_ PredictTestResult)	257
5•2•4•3•2 نموذج قواعد الاقتران (AssociationRules_PredictTestResult)	262
5•2•4•3•3 نموذج نايف بايز (NaiveBayes_PredictTestResult)	265
5•2•4•3•4 نموذج خوارزمية الشبكات العصبونية (NeuralNetwork_PredictTestResult)	268
5•2•4•3•5 نموذج الانحدار المنطقي (LogisticRegrision_PredictTestResult)	270
5•2•4•3•6 نموذج التجزئة العنقودية (Clustering Model) باستخدام أداة اكتشاف الفئات في برنامج اكسل	271
5•2•4•4•7 نموذج تحليل عوامل التأثير الرئيسة (Analyze Key Influencers) في اكسل	278
المبحث الخامس: مرحلة تقييم المعرفة المكتشفة من بيانات مخبر التحاليل الطبية	282
مقدمة	282
5•5•1 تقييم قدرات نماذج مشروع التنقيب في البيانات الشخصية والمرضية للمرضى	282
5•5•1•1 اختيار مفردات التقييم	283
5•5•1•2 مخطط الارتقاء (Lift Chart) لنماذج توقع حالة خروج المرضى من المشفى	285
5•5•1•2•1 مخطط الارتقاء (Lift Chart) العام	285
5•5•1•2•2 مخطط الارتقاء المخصص لتقييم قدرة النماذج على التنبؤ بقيمة محددة	286
5•5•1•3 مصفوفة التصنيف (Classification Matrix) لنماذج التنبؤ بحالة الخروج من المشفى	291

4•1•5•5 مخطط الارتقاء (Lift Chart) للنماذج المتعلقة بدراسة أسباب وفاة المرضى	293
5•1•5•5 مصفوفة التصنيف لنماذج التنبؤ بأسباب الوفاة	295
2•5•5 تقييم قدرات نماذج مشروع التنقيب في نتائج التحاليل الطبية المخبرية	296
1•2•5•5 اختيار مفردات التقييم في نافذة المخطط البياني لدقة التنقيب	296
2•2•5•5 مخطط الارتقاء (Lift Chart) لنماذج توقع نتيجة تحليل تعداد كريات الدم البيضاء	297
1•2•2•5•5 مخطط الارتقاء (Lift Chart) العام	297
2•2•2•5•5 مخطط الارتقاء المخصص لتقييم قدرة النماذج على التنبؤ بالقيمة المرتفعة لتحليل تعداد كريات الدم البيضاء	298
3•2•5•5 مصفوفة التصنيف (Classification Matrix)	299
المبحث السادس: مرحلة استخدام المعرفة المكتشفة من بيانات مخبر التحاليل الطبية	302
مقدمة	302
1•6•5 تقرير نتائج عملية التنقيب	302
2•6•5 الاستعلام المباشر من نماذج التنقيب وتصدير النتائج	306
1•2•6•5 بناء الاستعلام من نماذج التنقيب	306
2•2•6•5 استعرض الاستعلام وتخزينه	307
3•2•6•5 استخدام الاستعلام الناتج المخزن في مستودع بيانات التحاليل الطبية	308
3•6•5 التنبؤ بالإيراد الشهري لمخبر التحاليل الطبية لعام 2009	309
4•6•5 استعمال المعرفة المكتشفة في تتبع تتالي الأسباب المرضية المؤدية إلى تكرار دخول المرضى إلى المشفى	313
5•6•5 تصميم أداة في اكسل للتنبؤ بالأشخاص المعرضين للوفاة باستخدام الصفات الشخصية والأسباب المرضية ونتائج التحاليل الطبية	321
6•6•5 اكتشاف المتغيرات الطبية الأكثر تأثيراً على تغير نتائج أحد التحاليل الطبية	325
7•6•5 استخدام تقانة التنقيب في البيانات في البحث عن الأسباب الكامنة وراء تحول المرضى إلى مشفى آخر	327
الخاتمة: النتائج والمقترحات	329
النتائج	330
المقترحات والأعمال المستقبلية	338
قائمة المصطلحات	341
المراجع	346

فهرس الأشكال البيانية

الترقيم	العنوان	الصفحة
الشكل (1-1)	تدفق البيانات عبر التقنيات المختلفة في تقانة اكتشاف المعرفة	16
الشكل (1-2)	مكعب بيانات ثلاثي الأبعاد لكميات مبيعات (حقائق) شركة ألبان وأجبان بحسب البعد الزمني والجغرافي والصنف	23
الشكل (1-3)	أمثلة عن بعض التراتبات الهرمية للبعد الزمني	24
الشكل (1-4)	المخطوطة النجمية لمكعب بيانات المبيعات	25
الشكل (1-5)	العمليات الأساسية في المعالجة التحليلية الآنية	27
الشكل (2-1)	مجالات تطبيق تقانة التنقيب في البيانات	35
الشكل (2-2)	نتائج تحليل سلوك استعمال الهاتف الخلوي باستخدام التجزئة العنقودية والتصنيف	38
الشكل (2-3)	التنقيب في بيانات الأسواق المركزية	42
الشكل (2-4)	التنقيب في البيانات الطبية	47
الشكل (2-5)	مثال عن استدعاء الصور الشعاعية المخزنة الأقرب إلى الصور الملتقطة	49
الشكل (2-6)	طريقة كتابة الشيفرة الوراثية في الحمض النووي	52
الشكل (2-7)	طريقة فهم النص باستخدام تقانة التنقيب في النصوص	54
الشكل (2-8)	صورة فضائية ملتقطة، واستخدام تقانة التنقيب في التعرف على مكوناتها	56
الشكل (2-9)	مثال عن استخدام محرك البحث Google لتقانة التنقيب في الصور لإيجاد صور مشابهة	59
الشكل (2-10)	مثال على التجزئة العنقودية للصور	60
الشكل (2-11)	مثال عن محدودية قدرات التنقيب في الصور	61
الشكل (2-12)	النتيجة الأولى للبحث عن صور لباص مدرسة	62
الشكل (2-13)	صور فوتوغرافية لباص مدرسة	62
الشكل (2-14)	رسومات غير ملونة لباص مدرسة	63
الشكل (3-1)	مراحل عملية اكتشاف المعرفة من قواعد البيانات	67
الشكل (3-2)	بعض الأسماء التجارية للأدوات البرمجية المستخدمة في عمليات التنقيب في البيانات	70
الشكل (3-3)	مكونات مشروع التنقيب في البيانات	82
الشكل (3-4)	مثال عن آلية عمل خوارزمية الشبكات العصبونية	88
الشكل (3-5)	قائمة محددات خوارزمية قواعد الاقتران	89
الشكل (3-6)	قائمتي التنقيب في البيانات المضافتين إلى برنامج اكسل (Excel)، ونافذة الربط مع مخدم (SQL Server) على مخدم الحاسوب المحلي	91
الشكل (4-1)	صور لبعض أدوات وأجهزة مخبر التحاليل الطبية	98
الشكل (4-2)	استخدام الرماز القضباني للتعرف على العينات في مخبر التحاليل الطبية	104
الشكل (5-1-1)	واجهة الاستقبال في برنامج الوافي للمخابر والتحاليل	119
الشكل (5-1-2)	الأبواب التي تصنف ضمنها التحاليل الطبية	120
الشكل (5-1-3)	واجهة الإعدادات في برنامج الوافي للمخابر والتحاليل	120
الشكل (5-1-4)	واجهة تعريف التحاليل في برنامج الوافي للمخابر والتحاليل	121

الترقيم	العنوان	الصفحة
الشكل (5-1-5)	بعض نماذج طباعة نتائج التحاليل الطبية (تحليل البول، تحليل الكريات البيض والصيغة، وتحليل الكرياتينين) في برنامج الوافي للمخابر والتحاليل	122
الشكل (5-1-6)	واجهة ميزانية المخبر في برنامج الوافي للمخابر والتحاليل	122
الشكل (5-1-7)	واجهة الموردين والمواد في برنامج الوافي للمخابر والتحاليل	123
الشكل (5-1-8)	واجهة طلب إحصاءات التحاليل في برنامج الوافي للمخابر والتحاليل	124
الشكل (5-1-9)	واجهة إحصاءات التحاليل في برنامج الوافي للمخابر والتحاليل	124
الشكل (5-1-10)	البرامج الحاسوبية المستخدمة في انجاز مراحل اكتشاف المعرفة من قواعد بيانات مخبر التحاليل الطبية	128
الشكل (5-2-1)	استمارة قبول المريض، واستمارة خروج المريض	132
الشكل (5-2-2)	النسخة الالكترونية لاستمارة قبول المرضى في المشفى والموجودة في برنامج قاعدة بيانات مكتب القبول	133
الشكل (5-2-3)	واجهة التقارير الإحصائية في برنامج مكتب القبول	134
الشكل (5-2-4)	جدول أبواب التحاليل الطبية المخبرية (Doors)، والجدول الفرعي لأسماء التحاليل المنتمية إلى باب التحاليل الكيمائية الحيوية (Biochemistry)	137
الشكل (5-2-5)	العلاقات بين الجداول في برنامج الوافي للتحاليل المخبرية	138
الشكل (5-2-6)	مخطط كيان-علاقة للجداول الرئيسة في قاعدة بيانات مخبر التحاليل المدروس	139
الشكل (5-2-7)	مخطط كيان-علاقة لقاعدة بيانات مكتب القبول	140
الشكل (5-2-8)	بنية جدول المرضى في قاعدة بيانات مخبر التحاليل الطبية	143
الشكل (5-2-9)	بنية جدول أسماء التحاليل المخبرية (TestsInfo)	144
الشكل (5-2-10)	مضمون الحقول في جدول أسماء التحاليل (TestInfo)	145
الشكل (5-2-11)	استعلام عن التحاليل التي لها شروط مدرجة في جدول أسماء التحاليل (TestInfo)	146
الشكل (5-2-12)	تعرفة التحاليل الطبية الصادرة من قبل وزارة الصحة حسب القرار / 26 ت تاريخ 11 آب 2008	146
الشكل (5-2-13)	بنية جدول القيم الطبيعية للتحاليل الطبية	148
الشكل (5-2-14)	جدول القيم الطبيعية للتحاليل الطبية، ويُظهر اختلاف هذه القيم بحسب الجنس وبحسب الفئة العمرية	148
الشكل (5-2-15)	بنية جدول تحاليل المريض (PatTest) في برنامج الوافي للتحاليل الطبية	150
الشكل (5-2-16)	القيم المخزنة في جدول تحاليل المرضى (PatTest)	150
الشكل (5-2-17)	أعداد التحاليل التي تم إجراؤها في المخبر مرتبة تنازلياً بحسب عدد المرات التي أجري فيها كل تحليل طبي للمرضى خلال كل من عامي 2007 و 2008	151
الشكل (5-2-18)	أمثلة عن بنية جداول التحاليل التي تتضمن عدة نتائج	152
الشكل (5-2-19)	بنية جدول (PatientData1) الموجود في قاعدة بيانات مكتب القبول	153
الشكل (5-2-20)	بنية جدول (PatientData2) الموجود في قاعدة بيانات مكتب القبول	154
الشكل (5-2-21)	استعلام عن طرق القبول في المشفى	154
الشكل (5-2-22)	استعلام عن الجهة المسددة لنفقات المشفى	155

الترقيم	العنوان	الصفحة
الشكل (5-2-23)	استعلام عن أسباب القبول في المشفى من قاعدة بيانات القبول	155
الشكل (5-2-24)	مكونات جدول بيانات المرضى من قاعدة بيانات السجلات الطبية	156
الشكل (5-2-25)	المدرج التكراري لعدد أيام الإقامة في المشفى بحسب بيانات مكتب السجلات الطبية، باستخدام برنامج SPSS	158
الشكل (5-2-26)	استعلام عن الحالات الصحية للمرضى عند خروجهم من المشفى	159
الشكل (5-2-27)	استعلام عن أسماء الأطباء المشرفين على المرضى وعدد المرضى لكل منهم بحسب قاعدة بيانات السجلات الطبية	159
الشكل (5-3-1)	تداخل الخطوات المتبعة أثناء مرحلة تحضير البيانات	162
الشكل (5-3-2)	جدول التحاليل الفرعية (SubTests)	168
الشكل (5-3-3)	العلاقات الترابطية في برنامج أكسس بين الجداول (PatTest2007) و(SubTests) و(TestsInfo)	169
الشكل (5-3-4)	تحويل كل سجل في جدول تحليل كريات الدم الحمراء (RBC) إلى ثلاثة سجلات في الجدول الجديد (PatTest2007)	170
الشكل (5-3-5)	جدول تحويل التعابير غير الرقمية لنتائج التحاليل الطبية إلى قيم رقمية	171
الشكل (5-3-6)	الجدول الجديد للقيم الطبيعية (SubTestsNormalRanges) ويتضمن القيم الطبيعية للتحاليل الفرعية	172
الشكل (5-3-7)	القيم الطبيعية لتحليل البروتين مقاسة بالملي-غرام في عام 2008، بينما كانت تقاس بالغرام في عام 2007	173
الشكل (5-3-8)	طريقة حذف الرموز غير المرغوبة من حقل النتائج في جدول تحاليل المرضى	174
الشكل (5-3-9)	الاستعلام عن مناطق ومدن إقامة المرضى قبل وبعد عمليات التنظيف	179
الشكل (5-3-10)	جدول مكان الإقامة الجديد وقد أضيف إليه حقل المناطق الجغرافية	179
الشكل (5-3-11)	استعلام عن الحالات المرضية التي تحتوي كلمة (رئة) قبل وبعد تنظيف البيانات	181
الشكل (5-3-12)	تقلص عدد الأسباب المرضية (من 9490 إلى 3501 سبب) بعد عمليات التنظيف	181
الشكل (5-3-13)	جدول الشعب الطبية وقد أضيف إليه حقل مجموعات الشعب الطبية	182
الشكل (5-3-14)	نتائج الاستعلام عن حالة خروج المرضى بعد تنظيف بيانات جدول السجلات الطبية	183
الشكل (5-3-15)	مخطط كيان-علاقة النظري لقاعدة البيانات الجديدة المستنتجة من قاعدة بيانات مخبر التحاليل الطبية وقاعدة بيانات مكتب القبول وقاعدة بيانات مكتب السجلات الطبية	185
الشكل (5-3-16)	تصميم جدول البعد الزمني باستخدام التوابع الموجودة في برنامج اكسل	187
الشكل (5-3-17)	النموذج الأولي لمكعب البيانات المخصص لاستعراض تكاليف وأعداد التحاليل الطبية بحسب الأبعاد المختلفة	189
الشكل (5-3-18)	التصميم الأولي لمكعب البيانات الخاص بالتحاليل الطبية، وبعض عمليات التقطيع والاستعلام من هذا المكعب	191
الشكل (5-3-19)	إنشاء بنية مستودع بيانات مخبر التحاليل الطبية (Clinical Laboratory Study)	192
الشكل (5-3-20)	إنشاء مشروع لنقل البيانات من قاعدة البيانات إلى مستودع البيانات باستخدام خدمات التكامل في برنامج (Visual Studio)	193

الترقيم	العنوان	الصفحة
الشكل (5-3-21)	إدراج مهمة لتدفق البيانات في برنامج (Microsoft Visual Studio)	193
الشكل (5-3-22)	تعريف الاتصال بمصدر البيانات المرغوب نقلها وتعريف مراحل التنفيذ باستخدام خدمات التكامل في برنامج (Visual Studio)	194
الشكل (5-3-23)	عملية تغيير أنماط بعض البيانات المنقولة من خلال إجراء تحويل أنماط البيانات	195
الشكل (5-3-24)	تعريف اسم جدول البيانات الجديد والحقول التي سيتكون منها، وذلك ضمن إجراء مكان قاعدة البيانات المقصودة (DB Destination)	195
الشكل (5-3-25)	مشهد تم التقاطه أثناء أداء عملية نقل سجلات البيانات من أكسس إلى SQL Server	196
الشكل (5-3-26)	تعريف المفتاح الأولي للجدول المنقول، وتعديل خصائص الحقول بحسب الضرورة	197
الشكل (5-3-27)	مخطط العلاقات الترابطية بين الجداول الأساسية في مستودع بيانات التحاليل الطبية	198
الشكل (5-3-28)	إنشاء دالة لتقييم نتائج التحاليل الطبية	200
الشكل (5-3-29)	مرأى (vResultEvaluation) الخاص بتقييم نتائج التحاليل الطبية	200
الشكل (5-3-30)	استخدام الاستعلام التصالبي في تحويل أسماء التحاليل الطبية إلى خاصيات	201
الشكل (5-3-31)	الجدول التصالبي الذي تحولت فيه أسماء التحاليل الطبية إلى خاصيات	202
الشكل (5-3-32)	البنية التأسيسية لمشروع خدمات المعالجة التحليلية الآنية لبيانات مخابر التحاليل الطبية في برنامج (Microsoft Visual Studio)	203
الشكل (5-3-33)	اختيار جداول مرأى مصدر البيانات في برنامج (Visual Studio)	204
الشكل (5-3-34)	مرأى بنية مصدر البيانات في مشروع المعالجة التحليلية الآنية لبيانات مخبر التحاليل الطبية المدروس	204
الشكل (5-3-35)	تعريف مقياس سعر التحليل وهو مقياس جديد مشتق من المقاييس الموجودة في جدول الحقيقة	205
الشكل (5-3-36)	اختيار الأنماط المناسبة للخاصيات في البعد الزمني أثناء إجراء تعريف هذا البعد	206
الشكل (5-3-37)	مكونات مكعب البيانات الخاص بدراسة بيانات مخبر التحاليل الطبية	207
الشكل (5-3-38)	تقارير نجاح معالجة مكعب البيانات أو أية تغيُّرات تطرأ عليه	208
الشكل (5-3-39)	مكعب بيانات التحاليل الطبية المخبرية المستخدم في عمليات المعالجة التحليلية الآنية	209
الشكل (5-3-40)	استعراض عينة من البيانات المسحوبة من جدول بيانات دخول المرضى إلى المشفى باستخدام ميزة استطلاع البيانات في هيكل مكعب البيانات	210
الشكل (5-3-41)	تصفح وسطي تكاليف التحاليل الطبية للمرضى بعمر السنتين وبحسب الجنس والأمراض باستخدام (Pivot Char)	210
الشكل (5-3-42)	التراتب الهرمي للتحاليل الطبية (مجموعات التحاليل، التحاليل الرئيسة، التحاليل الفرعية) والتراتب الهرمي لمكان إقامة المرضى (المنطقة الجغرافية في سورية، المدينة، المنطقة)	211
الشكل (5-3-43)	تطبيق العمليات الحسابية في إيجاد مقاييس جديدة داخل مكعب البيانات، وهو الوسط الحسابي لنتائج التحاليل الطبية للمرضى	213
الشكل (5-3-44)	إضافة مقاييس جديدة إلى مكعب البيانات تساوي المقاييس القديمة ولكنها تختلف عنها في الخصائص	213

الترقيم	العنوان	الصفحة
الشكل (5-3-45)	استعلام من مكعب بيانات التحاليل الطبية عن التوزع النسبي لأعداد التحاليل الطبية بحسب نوعها خلال أيام الأسبوع	215
الشكل (5-3-46)	استعلام من مكعب بيانات التحاليل الطبية عن التوزع النسبي لأعداد التحاليل الطبية اليومي بحسب مجموعات التحاليل	215
الشكل (5-3-47)	الاستعلام المتعدد الأبعاد عن وسطي نتائج التحاليل الطبية لأكثر التحاليل الطبية طلباً في الشعب الطبية التي تُعنى بالجراحة في المشفى، وموزعة حسب الجنس	217
الشكل (5-3-48)	الاستعلام من مكعب بيانات التحاليل الطبية عن مقاييس العمر (العمر الوسطي، العمر الأعلى، العمر الأدنى) وأعداد المرضى في كل من الجنسين وبحسب الأسباب المرضية	218
الشكل (5-3-49)	توزع أعداد المرضى القادمين إلى المشفى بحسب شعبة القبول والسبب المرضي وتوقيت القدوم إلى المشفى	219
الشكل (5-3-50)	التوزع النسبي لإيرادات المخبر من التحاليل الطبية على أرباع السنة وبحسب مجموعات التحاليل	220
الشكل (5-3-51)	التوزع النسبي لإيرادات المخبر من مجموعات التحاليل الطبية في الربع الثالث وبحسب السنوات	220
الشكل (5-3-52)	التوزع النسبي لإيرادات المخبر من التحاليل الطبية على أرباع السنة وبحسب أماكن إقامة المرضى	221
الشكل (5-3-53)	تعريف الاتصال في برنامج اكسل مع مكعب بيانات التحاليل الطبية الذي تم إنشاؤه سابقاً في برنامج (Viual Studio)، واختيار طريقة عرض تقرير البيانات	223
الشكل (5-3-54)	مكونات ورقة العمل الناتجة عن اتصال اكسل مع مكعب بيانات التحاليل الطبية في مشروع المعالجة التحليلية الآنية	223
الشكل (5-3-55)	مثال عن استخدام اكسل في استعراض مكعب بيانات التحاليل الطبية	224
الشكل (5-3-56)	الاستفادة من مخرجات المعالجة التحليلية الآنية في حساب التكاليف الوسطية للتحاليل الطبية التي يتحملها المرضى مصنفة بحسب المرض	225
الشكل (5-3-57)	الاستعلام متعدد الأبعاد عن تكاليف التحاليل الطبية التي دفعها المرضى للمشفى مصنفة شاقولياً بحسب العام والجنس، وأفقياً بحسب نوع المرض وأرقام أضابير المرضى الذين تعرضوا للإصابة بكل من هذه الأمراض	226
الشكل (5-4-1)	إجراءات بناء هيكل للتنقيب في بيانات التحاليل الطبية باستخدام البرنامج الحاسوبي (Microsoft Visual Studio)، وتقرير نجاح عملية التدقيق والبناء والتفعيل	231
الشكل (5-4-2)	هيكل التنقيب في بيانات مرضى مخبر التحاليل الطبية، ومضمون النوافذ المختلفة المتعلقة بعملية التنقيب	232
الشكل (5-4-3)	نافذة نماذج التنقيب في بيانات المرضى للتنبؤ بأسباب الوفاة، وقائمة انتقاء فقط البيانات المتعلقة بوفيات المرضى	233
الشكل (5-4-4)	نموذج أشجار القرار (DecisionTrees_DeathReason) للتعرف على خصائص المرضى المتوفين بالأسباب المختلفة للوفاة	234
الشكل (5-4-5)	نافذة قواعد الاقتران في نموذج قواعد الاقتران (AssociationRules_DeathReason)	236
الشكل (5-4-6)	نافذة شبكة التبعية في نموذج قواعد الاقتران (AssociationRules_DeathReason)	237

الترقيم	العنوان	الصفحة
الشكل (5-4-7)	نافذة مجموعات الاقتران في نموذج قواعد الاقتران (AssociationRules_DeathReason)	237
الشكل (5-4-8)	نافذة خصائص عقد العنقود (Cluster Profiles) في نموذج التجزئة العنقودية (Clustering_DeathReason)	239
الشكل (5-4-9)	نافذة الرسم البياني العنقودي (Cluster Diagram) في نموذج التجزئة العنقودية (Clustering_DeathReason)	240
الشكل (5-4-10)	نافذة الرسم البياني العنقودي (Cluster Diagram) في نموذج التجزئة العنقودية (Clustering_DeathReason)	240
الشكل (5-4-11)	نافذة صفات العقدة (Cluster Characteristics) في نموذج التجزئة العنقودية (Clustering_DeathReason)	242
الشكل (5-4-12)	نافذة تمييز العقد (Cluster Discrimination) في نموذج التجزئة العنقودية (Clustering_DeathReason)	244
الشكل (5-4-13)	نافذة شبكة التبعية (Dependency Network) في نموذج (NaiveBayes_OutStae)	246
الشكل (5-4-14)	نافذة اللمحة المختصرة للخاصيات (Attribute Profile) في نموذج (NaiveBayes_OutStae)	246
الشكل (5-4-15)	نافذة صفات الخاصيات (Attribute Characteristics) في نموذج (NaiveBayes_OutStae)	247
الشكل (5-4-16)	نافذة نتائج نموذج الشبكات العصبونية (NeuralNetworks_OutState) المتعلقة بتقييم تأثير كل من قيم خاصيات المرضى على حالة الخروج من المشفى للفئة العمرية 13-29 سنة، ومرتبة تنازلياً بحسب العلامات المحسوبة داخلياً لكل حالة	248
الشكل (5-4-17)	نافذة نتائج نموذج الشبكات العصبونية (NeuralNetworks_OutState) المتعلقة بتقييم تأثير كل من قيم خاصية "سبب قبول المرضى" على خاصية حالة الخروج من المشفى في الأعمار التي تزيد عن 65 سنة	249
الشكل (5-4-18)	لقطات من مرأى شجرة المحتويات الشاملة (Generic Content Tree View) لنموذج الشبكات العصبونية (NeuralNetworks_OutState) المتعلق بتوقع حالة الخروج من المشفى	250
الشكل (5-4-19)	نافذة نتائج نموذج الانحدار المنطقي (LogisticRegresion_OutState) المتعلقة بتقييم تأثير كل من قيم خاصيات المرضى على حالة الخروج من المشفى للفئة العمرية 13-29 سنة، ومرتبة تنازلياً بحسب العلامات المحسوبة داخلياً لكل حالة	251
الشكل (5-4-20)	نافذة نتائج نموذج الانحدار المنطقي (LogisticRegresion_OutState) المتعلقة بتقييم تأثير كل من قيم خاصية سبب قبول المرضى على خاصية حالة الخروج من المشفى في الأعمار التي تزيد عن 65 سنة	252
الشكل (5-4-21)	استعراض مرأى (ResultEvaluation_Crosstab) المستخدم في مشروع التنقيب في بيانات نتائج التحاليل الطبية	255
الشكل (5-4-22)	استخدام مكعب البيانات الخاص بالدراسة في التعرف على أكثر التحاليل الطبية طلباً في مخبر التحاليل الطبية التابع للمشفى المدروس	256

الترقيم	العنوان	الصفحة
الشكل (5-4-23)	استخدام مكعب البيانات الخاص بالدراسة في التعرف على أكثر التحاليل الطبية طلباً في مخبر التحاليل الطبية التابع للمشفى المدروس مصنفة بحسب الأمراض وبحسب الشعب الطبية	257
الشكل (5-4-24)	مخطط شجرة القرار الناتج عن نموذج التنقيب (DecisionTrees_PredictTestResult)، وقد استدعيت هذه الشجرة للتنبؤ أو استنتاج نتيجة تحليل الكريات البيضاء (WBC) استناداً إلى نتائج التحاليل الأخرى	259
الشكل (5-4-25)	أمثلة عن النتائج المتوقعة للتحليل الطبي (WBC) استناداً إلى نتائج التحاليل الطبية الأخرى والمستمدة من نموذج شجرة القرار (DecisionTrees_PredictTestResult)	260
الشكل (5-4-26)	نافذة شجرة القرار في نموذج (DecisionTrees_PredictTestResult) الخاصة بالتنبؤ بنتائج التحليل الطبي لكريات الدم الحمراء (RBC)	261
الشكل (5-4-27)	منظر عام لنافذة شبكة التبعية في نموذج التنقيب المستند إلى خوارزمية قواعد الاقتران (AssociationRules_PredictTestResult)، ومنظر جزء من هذه النافذة يُمثِّل علاقات اقتران نتائج التحاليل الطبية المتبادلة مع التحليل الطبي لتعداد كريات الدم البيضاء	262
الشكل (5-4-28)	نافذة القواعد في نموذج التنقيب المستند إلى خوارزمية قواعد الاقتران (AssociationRules_PredictTestResult)، مرتبة تنازلياً بحسب الاحتمال	264
الشكل (5-4-29)	نافذة القواعد في نموذج التنقيب المستند إلى خوارزمية قواعد الاقتران (AssociationRules_PredictTestResult)، مرتبة تنازلياً بحسب قوة القواعد	265
الشكل (5-4-30)	مثالين لنافذة شبكة التبعية في نموذج التنقيب (NaiveBayes_PredictTestResult)، الأول حول تحليل (GlucoseFasting) السكر الصيامي، والثاني حول تحليل تعداد الكريات البيضاء (WBC)	266
الشكل (5-4-31)	أمثلة لنافذة صفات الخاصيات (Attribute Characteristics) في نموذج (NaiveBayes_PredictTestResult)، ويتعلق المثال الأول بقيم التحاليل الطبية المرافقة لارتفاع تعداد الكريات البيضاء (WBC) واحتمال كل منها، والثاني بقيم التحاليل الطبية المرافقة لارتفاع تحليل السكر الصيامي (Glucose Fasting) واحتمال كل منها	267
الشكل (5-4-32)	نافذة نتائج نموذج التنقيب (NeuralNetwork_PredictTestResult) المتعلقة بتقييم تأثير كل من نتائج التحاليل الطبية على ارتفاع أو انخفاض تحليل تعداد كريات الدم البيضاء	269
الشكل (5-4-33)	الاستعلام من مستودع البيانات عن سجلات المرضى التي تحتوي على نتائج مرتفعة في تحليل (BH)	270
الشكل (5-4-34)	نافذة نموذج الانحدار المنطقي (LogisticRegrision_PredictTestResult) المتعلقة بتقييم تأثير كل من نتائج التحاليل الطبية على ارتفاع أو انخفاض تحليل تعداد كريات الدم البيضاء (WBC)	271
الشكل (5-4-35)	استدعاء أداة اكتشاف الفئات (Detect Categories) في اكسل واختيار الخاصيات المطلوبة لعملية التجزئة	272
الشكل (5-4-36)	نافذة تقرير الفئات المكتشفة باستخدام أداة اكتشاف الفئات في اكسل والتي طُبقت على بيانات المرضى الشخصية والمرضية ونتائج التحاليل الطبية	273

الترقيم	العنوان	الصفحة
الشكل (5-4-37)	مخرجات أداة اكتشاف الفئات بعد انتقاء التسميات المناسبة للفئات بحسب قيم الخصائص في كل منها	274
الشكل (5-4-38)	مخطط محوري مرفق مع تقرير فئات المرضى الناتج عن استخدام أداة التنقيب (Detect Category)، وقد جرى اختيار استعراض نتائج تحليل كريات الدم البيضاء (N, H ,L, Missing) في كل فئة من فئات المرضى	276
الشكل (5-4-39)	الحقلُ الذي أضافته أداة اكتشاف الفئات إلى جدول البيانات المنقب فيها، ويَعرض هذا الحقل اسم الفئة التي ينتمي إليها كل مريض	277
الشكل (5-4-40)	استدعاء أداة تحليل عوامل التأثير الرئيسة (Analyze Key Influencers) في برنامج اكسل للتنقيب في البيانات الشخصية والمرضية ونتائج التحاليل الطبية	278
الشكل (5-4-41)	تقرير العوامل المؤثرة على سبب القبول في المشفى باستخدام أداة تحليل عوامل التأثير الرئيسة في برنامج اكسل (Analyze Key Influencers)	280
الشكل (5-4-42)	أداة توليد تقارير المقارنة في نموذج تحليل عوامل التأثير الرئيسة (Discrimination based on Key Influencers)، وتقرير المقارنة بين تشمع الكبد والتهاب الكبد من ناحية العوامل المؤثرة على كل منهما	281
الشكل (5-5-1)	نافذة اختيار العناصر المرغوبة لاستدعاء أدوات اختبار دقة التنقيب	284
الشكل (5-5-2)	مخطط الارتقاء (Lift Chart) العام لتقييم قدرة بعض نماذج التنقيب المنشأة على توقع حالتي خروج المريض من المشفى (تحسن أو وفاة)	286
الشكل (5-5-3)	مخطط الارتقاء (Lift Chart) المخصص لاختبار قدرة نماذج التنقيب على التنبؤ بخروج المرضى بحالة وفاة	289
الشكل (5-5-4)	مخطط الارتقاء (Lift Chart) المخصص لاختبار قدرة نماذج التنقيب على التنبؤ بخروج المرضى بحالة تحسن، وذلك عند نسبة 96.08% من البيانات	290
الشكل (5-5-5)	نافذة مصفوفة التصنيف (Classification Matrix)، التي تُبيِّن مقارنة بين أعداد القيم المتنبأ بها الصحيحة منها والخاطئة في مجموعة بيانات الاختبار لكل نموذج من نماذج التنقيب الخاصة بالتنبؤ بحالة الخروج من المشفى (الشفاء أو الوفاة)	292
الشكل (5-5-6)	مخطط الارتقاء (Lift chart)، لنماذج التنقيب الثلاثة الخاصة بالتنبؤ بأسباب الوفاة لدى مرضى المشفى	294
الشكل (5-5-7)	نافذة مصفوفة التصنيف لنماذج التنقيب الخاصة بالتنبؤ بأسباب الوفاة انطلاقاً من البيانات الشخصية والمرضية لمرضى المشفى	295
الشكل (5-5-8)	نافذة انتقاء عناصر تقييم نماذج التنقيب الخاصة بمشروع التنقيب في نتائج التحاليل الطبية، وقد تم انتقاء تقييم قدرات النماذج في التنبؤ بنتائج تحليل كريات الدم البيضاء	297
الشكل (5-5-9)	مخطط الارتقاء العام لتقييم قدرة نماذج التنقيب على التنبؤ بنتائج التحليل الطبي (WBC)	298
الشكل (5-5-10)	مخطط الارتقاء المخصص لتقييم قدرة النماذج على التنبؤ بالقيمة المرتفعة لتحليل تعداد الكريات البيضاء	299
الشكل (5-5-11)	نافذة مصفوفة التصنيف الخاصة بعرض نتائج التنبؤات التي قامت بها كل من نماذج التنقيب الخاصة بالتنبؤ بنتائج التحليل الطبي لكريات الدم البيضاء (WBC) استناداً إلى نتائج التحاليل الطبية الأخرى المقترنة معه	300

الترقيم	العنوان	الصفحة
الشكل (5-6-1)	نافذة تصميم الاستعلامات عن تنبؤات نماذج التنقيب (Mining Model Prediction)، وقد تم الاستعلام على بيانات جدول بيانات المرضى الذي بُنيت عليه نماذج التنقيب في هذه الدراسة، وتحديداً المرضى الذين توفوا في المشفى	307
الشكل (5-6-2)	نافذة نتائج الاستعلام عن توقعات نموذج (NaïveBayes_OutState) حول بيانات المرضى المتوفين في جدول بيانات المرضى (PatientEnterProcess)، ونافذة الإجراء الخاص بتخزين جدول الاستعلام في مستودع بيانات التحاليل الطبية	308
الشكل (5-6-3)	نافذة برنامج (SQL Server Management Studio) ويظهر فيها انضمام جدول الاستعلام عن توقعات نموذج التنقيب (OutStatePredictionByNaiveByesModel) إلى مجموعة جداول مستودع بيانات التحاليل المخبرية، والاستعلام التلخيصي عن أعداد التوقعات الصحيحة والتوقعات الخاطئة لحالة وفاة المرضى	309
الشكل (5-6-4)	إنشاء مرأى في برنامج (SQL Server Management Studio) لحساب الإيرادات الشهرية للتحاليل الطبية	310
الشكل (5-6-5)	الاستيراد في برنامج اكسل لمرأى الإيرادات الشهرية للتحاليل الطبية الموجود في مستودع بيانات التحاليل الطبية	311
الشكل (5-6-6)	استدعاء إجراء التنبؤ (Forecast) في برنامج اكسل للتنبؤ بالإيراد الشهري لمخبر التحاليل الطبية خلال عام 2009	312
الشكل (5-6-7)	مخطط التنبؤ بالدخل الشهري لمخبر التحاليل الطبية خلال عام 2009	312
الشكل (5-6-8)	جدول تلخيصي لإظهار تكرار قبول المرضى في المشفى	314
الشكل (5-6-9)	خطوات تحليل سلة الأسباب المرضية التي تؤدي إلى تكرار قبول المرضى في المشفى باستخدام أداة التنقيب (Shopping Basket Analysis)	316
الشكل (5-6-10)	تتابع الأسباب المرضية المؤدية إلى الدخول إلى المشفى خلال الفترة المدروسة (2007-2008)، ويوجد بعض البيانات عن المرضى في عام 2005	317
الشكل (5-6-11)	استخدام أدوات الانتقاء في اكسل لاختيار جزء من الحزم الموجودة في جدول تتابع الأسباب المرضية والمتعلقة بمرض "خناق الصدر"	318
الشكل (5-6-12)	نسب الإصابات المرضية التي تلت التعرض إلى خناق الصدر عند المرضى الذين تكرر قبولهم في المشفى المدروس	319
الشكل (5-6-13)	استعراض حزم الأسباب المتتابعة التي تحتوي على مشكلة الآلام البطنية	321
الشكل (5-6-14)	استدعاء إجراء تصميم أداتي تنبؤ بالوفاة الأولى آلية تفاعلية والثانية ورقية	322
الشكل (5-6-15)	مثال واقعي عن استخدام أداة التنبؤ الآلي بحدوث الوفاة استناداً إلى الصفات الشخصية والمرضية ونتائج التحاليل الطبية للمرضى	223
الشكل (5-6-16)	مقاطع من الاستمارة الورقية لأداة التنبؤ بالوفاة الناتجة عن تطبيق إجراء التنقيب في البيانات (Predict Calculation)	324
الشكل (5-6-17)	تحليل العوامل المؤثرة على ارتفاع وانخفاض نتيجة التحليل الطبي (SGOT) باستخدام أداة (Analyze Key Influence)	326
الشكل (5-6-18)	شبكة التبعية في نموذج قواعد الاقتران (AssociationRules_OutState) الخاص بالتنبؤ بحالة خروج المرضى مع التأكيد على حالة تحويل المرضى إلى مشفى آخر	328

فهرس الجداول

الترقيم	العنوان	الصفحة
الجدول (1–1)	مقارنة معالجة المبادلات الآنية مع المعالجة التحليلية الآنية	29
الجدول (1–2)	أبرز الفروقات بين التنقيب في البيانات والإحصاء التقليدي	32
الجدول (4–1)	عناوين بعض رسائل الدراسات العليا في التشخيص المخبري والطب المخبري في جامعة دمشق	107
الجدول (5–2–1)	قواعد بيانات مخبر التحاليل الطبية المتاحة للدراسة وأحجام كل منها	130
الجدول (5–3–2)	أسماء الجداول والحقول المنتقاة لانجاز مراحل اكتشاف المعرفة	163
الجدول (5–6–1)	أبرز النتائج التي قدمتها نماذج التنقيب في مشاريع التنقيب التي بنيت في مرحلة التنقيب في البيانات	303

الفصل التمهيدي

الفصل التمهيدي

مقدمة

حظي حقل اكتشاف المعرفة من قواعد البيانات (Knowledge Discovery in Database) باهتمام عالمي كبير في الآونة الأخيرة بسبب التزايد الهائل للبيانات المتراكمة في الأنظمة الحاسوبية، وظهور الحاجة إلى تحويل مثل هذه البيانات ليس فقط إلى معلومات مفيدة كما في أنظمة قواعد البيانات التقليدية وإنما تحويلها إلى معرفة. وتصلح تقانة اكتشاف المعرفة للاستخدام في عددٍ كبيرٍ من المجالات مثل إدارة الأعمال، والاتصالات، وتحليل الأسواق، واستخدام الانترنت، والأبحاث والاستكشافات العلمية.. حيث أنها تساعد في كسب البصيرة والفهم العميق المخفي لسلوك الحالات المدروسة وارتباطها مع بعضها.

يُعتبر التنقيب في البيانات إحدى المراحل الأساسية في العملية الكلية لاكتشاف المعرفة من قواعد البيانات، وهذا ما تم الاتفاق عليه في المؤتمر الأول لاكتشاف المعرفة من قواعد البيانات الذي عُقد في مونتريال عام 1995. أما واقعياً فقد جرت العادة على اعتبار المفهومين مترادفين، ويتم في الكثير من الأحيان الاكتفاء باستخدام مفهوم التنقيب في البيانات للتعبير عن مجمل العملية. ويعود السبب في ذلك إلى أن مرحلة التنقيب في البيانات هي المرحلة الأساسية المميزة لعملية استخلاص المعرفة، والتي تلتقي فيها نتائج العمليات التقنية والتحليلية التي أنجزها أخصائيو التنقيب مع احتياجات المستفيدين الحقيقيين من المعرفة المكتشفة. وقد يسيء هذا الاستخدام غير الصحيح لمفهوم التنقيب إلى كلا المفهومين، ويؤدي إلى الوقوع في جملة من المشاكل المفاهيمية خصوصاً في بلادنا التي ما يزال تطبيق هذه التقانة فيها حديثاً ومحدوداً. وتبعاً لذلك سينطلق البحث الحالي في أساسه النظري من مفهوم تقانة اكتشاف

المعرفة، ثم سيتم التعرف على مكوناته وبعض المفاهيم المرتبطة به، وأبرز مجالات تطبيقه والمراحل التي يتكون منها.

يهتم الجانب التطبيقي من البحث الحالي بالبيانات المخزنة في مخابر التحاليل الطبية، والتي تتألف عادةً من نوعين أساسيين من البيانات هما البيانات الشخصية والمرضية للمرضى وبيانات نتائج التحاليل الطبية التي تُجرى لهم. وتبعاً لذلك سينطلق هذا البحث في أساسه التطبيقي من التعرف على مخبر التحاليل الطبية باعتباره مصدر الحصول على بيانات هذه الدراسة من جهة، وباعتباره يشكل أيضاً جزءاً من نظام المعلومات الطبية في المشافي العالمية، ويعمل على التزود من السجل الطبي الالكتروني، ويزوده ببيانات نتائج التحاليل الطبية. كما سيتم الاطلاع على بعض الأساليب المتبعة في انجاز الدراسات الطبية المحلية والعالمية التي تهتم ببيانات التحاليل الطبية. ثم تنطلق بعد ذلك الرحلة الطويلة والشاقَّة لاكتشاف المعرفة الكامنة في بيانات مخبر التحاليل الطبية المتاحة للدراسة، والتي ستجتاز عدداً من المحطات الضرورية لها قبل أن تصل في النهاية إلى غايتها.

كلمات مفتاحية

تقانة اكتشاف المعرفة، التنقيب في البيانات، مستودع البيانات، المعالجة التحليلية الآنية، مكعب البيانات، قواعد البيانات، مخبر التحاليل الطبية.

أهمية البحث

تُعتبر الدراسات المرتبطة بتقانة اكتشاف المعرفة من قواعد البيانات قليلة وحديثة العهد نسبياً، ولذلك فهي تحتوي على الكثير من النقاط البحثية التي ما تزال بحاجة إلى الرفد بالدراسات والتجارب التطبيقية، وتنبع أهمية هذا البحث من كونه يهتم بتطبيق هذه التقانة على حقل هام جداً من حقول الحياة والمجتمع والعلم وهو الحقل الطبي.

ويُستفاد من هذا البحث في كشف الحقائق المفيدة الكامنة في البيانات الالكترونية المخزنة في قواعد بيانات طبية، لاستخدامها بما فيه مصلحة صحة الإنسان من خلال رفد الأبحاث الطبية بالاكتشافات والمسائل البحثية ودعم عمليات اتخاذ القرارات الإدارية في المنشآت الصحية. ويعتبر هذا النوع من الدراسات إضافياً ومكملاً للأساليب الأخرى في دراسة وتلخيص البيانات كالأساليب الإحصائية وأساليب الاستعلام من قواعد البيانات، وليس بديلاً عنها.

مشكلة البحث

تدور المشكلة الأساسية في هذا البحث حول التعرف على مدى إمكانية وكيفية الاستفادة من تقانة اكتشاف المعرفة في دراسة بيانات مخابر التحاليل الطبية ضمن الخصوصيات والعيوب التي تتسم بها قواعد البيانات المحلية.

تختلف طريقة التعبير عن نتائج التحاليل الطبية (رقمية، رمزية) ومدلولاتها من تحليل طبي لآخر، ويقود ذلك إلى صعوبة التوصل إلى معرفة مفيدة من هذه البيانات مجتمعةً، وستجري المحاولة في هذا البحث لحل هذه المشكلة أيضاً.

أهداف البحث

يُعتبر البحث الحالي دراسة حالة عملية، ولذلك فهو يتضمن بعض الجوانب النظرية لتقانة اكتشاف المعرفة، والكثير من المكونات التطبيقية. وبشكل عام يسعى البحث الحالي إلى تحقيق الأهداف الآتية:

- التعريف بتقانة اكتشاف المعرفة وبالأعمال التي تستطيع أن تقوم بها في عدد من مجالات الأعمال والأبحاث العلمية وغيرها.
- تطبيق تقانة تحليلية وتوجيهية حديثة هي تقانة اكتشاف المعرفة في المجال الطبي، لكي تستفيد من مخرجاتها عدةُ جهات كمنظمات الصحة وإدارات المراكز الطبية والأطباء والباحثين وشركات التأمين الصحي.
- التعرف على قدرة البيانات الشخصية والمرضية ونتائج التحاليل الطبية المتوفرة على تقديم مواد أولية كافية ومناسبة لعمليات التنقيب.
- التحقق بالدرجة الأولى من قدرة أدوات التنقيب في البيانات على استخراج البديهيات التي تتصف بها البيانات المتاحة للدراسة، لأنها بيانات اجتمعت في قواعد البيانات بهدف التوثيق الالكتروني، وهذه هي المرة الأولى التي تُطبَّق عليها هذه الأدوات.
- محاولة التوصل إلى معارف جديدة ومفيدة كامنة في البيانات الضخمة المتراكمة في قواعد بيانات مخابر التحاليل الطبية.
- البحث عن استخدامات مفيدة للمعرفة المكتشفة من البيانات المتاحة.

منهجية البحث

يعتمد البحث الحالي في جوانبه النظرية على المنهج الوصفي، حيث جرى التزود بالمعرفة الضرورية لبناء هذه الجوانب من خلال الاطلاع على الكتب والأبحاث والمقالات الورقية والالكترونية المرتبطة بمواضيع الدراسة الحالية، وذلك من ناحية استخدام تقانة اكتشاف المعرفة، وناحية الأبحاث الطبية المتعلقة ببيانات مخابر التحاليل الطبية.

كما ويعتمد البحث في جانبه التطبيقي على المنهج التحليلي الاستقرائي، وذلك من خلال القيام بالتهيئة والتحليل والتنقيب في البيانات الشخصية والمرضية ونتائج التحاليل الطبية للمرضى الذين أجريت لهم تحاليل طبية في مخبر التحاليل الطبية المدروس، والتوصل أخيراً إلى أنماط معرفية جديدة ومفيدة.

مجتمع البحث

يستمد البحث الحالي بياناته من مخبر التحاليل الطبية ومن مكتبي القبول والسجلات الطبية التابعين لأحد المشافي الحكومية في دمشق خلال الفترة الواقعة بين عامي 2007 و2008.

لمحة عامة عن فصول البحث

يتألف البحث الحالي من جزء نظري وجزء تطبيقي هما:

الجزء الأول هو الجزء النظري والذي يتكون بدوره من ثلاثة فصول نظرية على النحو التالي:

- كُرِّس **الفصل الأول** للتعريف بالمفاهيم الأساسية المرتبطة بمجال تقانة اكتشاف المعرفة، ولإجراء المقارنات الضرورية بين بعض المفاهيم في هذا المجال.
- خُصِّص **الفصل الثاني** للحديث عن أبرز مجالات تطبيق تقانة اكتشاف المعرفة من قواعد البيانات، والتي انقسمت إلى ثلاثة أقسام ويرتبط القسم الأول بالمنشآت الربحية، ويتعلق القسم الثاني بالاكتشافات العلمية، ويتكون القسم الثالث من مجالات مختلفة لا تندرج تحت تصنيف محدد.
- يتضمن **الفصل الثالث** عرضاً للمراحل النظرية الستة لعملية اكتشاف المعرفة من قواعد البيانات، ويُشكل هذا الفصل الأساس النظري الذي سيتم الاستناد إليه في انجاز المراحل التطبيقية لاكتشاف المعرفة من قواعد البيانات المدروسة.

<u>الجزء الثاني</u> هو الجزء التطبيقي، ويتكون من الفصلين التاليين:

- يختص **الفصل الرابع** بالتعريف بالملامح العامة لمجال مخابر التحاليل الطبية، وتُعتبر المعلومات التي يتضمنها هذا الفصل جزءاً من الأساس المعرفي الضروري للتعرف على المجال المدروس، والذي يخص المرحلة الأولى من مراحل عملية اكتشاف المعرفة. ويتضمن هذا الفصل عرضاً لبعض المفاهيم الأساسية ولآلية تجميع وتخزين البيانات في مخبر التحاليل الطبية، ويهتم بالتعريف بنظم المعلومات في المنشآت الطبية. كما ويعرض الأساليب المتبعة في إجراء الدراسات العلمية المستندة إلى بيانات التحاليل المخبرية الطبية، وبعض الأمثلة عنها.
- يُعتبر **الفصلُ الخامس** الفصلَ الأهم والأكبر مقارنةً ببقية الفصول، وذلك لأنه يختص بعرض الحالة العملية الخاصة بالبحث. ويتناول هذا الفصل عرضاً لمراحل عملية اكتشاف المعرفة الكامنة في بيانات مخبر التحاليل الطبية المدروس. ويُقسم هذا الفصل إلى ستة مباحث يختص كل منها بعرض إحدى المراحل التطبيقية لاكتشاف المعرفة، حيث تُشكل مخرجاتُ كل مرحلة مدخلاتٍ للمرحلة أو المراحل التالية لها، كما تنشأ تغذية راجعة بين تلك المراحل تؤدي إلى تعديلات في المراحل السابقة. ونبين فيما يلي المهام التي يختص بعرضها كل من مباحث هذا الفصل:
  - يتناول **المبحث الأول** مرحلة استيعاب مجال مسألة اكتشاف المعرفة من بيانات مخبر التحاليل الطبية، وهي المرحلة التمهيدية التي يتم فيها الإلمام بمجال التطبيق، وصياغة مسائل اكتشاف المعرفة، والتعرف على المستفيدين من هذه العملية، وتحديد الأدوات البرمجية التي ستُستخدم في المراحل القادمة.
  - يختص **المبحث الثاني** بعرض مرحلة استيعاب بيانات مخبر التحاليل الطبية والبيانات المرتبطة بها، والتي يتم فيها الاطلاع على كل التفاصيل المتعلقة بفهم جميع مكونات قواعد البيانات المتاحة لهذا البحث، وهذا ما يساعد في التعرف على البيانات اللازمة لحل مسائل اكتشاف المعرفة من أصل جميع البيانات المتاحة.
  - يُظهر **المبحث الثالث** العملياتَ التي أجريت في مرحلة تهيئة بيانات مخبر التحاليل الطبية، ويتضمن انتقاء البيانات التي تفيد في انجاز مراحل اكتشاف المعرفة، وتنظيفها

وتنسيقها وتحويلها إلى بنية مستودع بيانات، وأخيراً إجراء بعض عمليات المعالجة التحليلية الآنية عليها.

- يُعبِّر **المبحث الرابع** عن لب عملية اكتشاف المعرفة لأنه يتضمن كافة التفاصيل المحيطة بمرحلة التنقيب في بيانات المرضى الشخصية والمرضية ونتائج التحاليل الطبية، والتي ستجري فيها محاولة إيجاد الحلول المناسبة لمسائل اكتشاف المعرفة من خلال بناء واستعراض نماذج التنقيب الملائمة.
- يَعرض **المبحث الخامس** مرحلةَ تقييم المعرفة المكتشفة من بيانات مخبر التحاليل الطبية، والتي يتم فيها تقييم ومقارنة قدرات نماذج التنقيب التي أُنشئت في المرحلة السابقة من حيث قدرتها على تقديم تنبؤات صحيحة.
- يَتناول **المبحث السادس** والأخير مرحلةَ استخدام المعرفة المكتشفة من بيانات مخبر التحاليل الطبية، ويتضمن عرضاً لبعض الاستخدامات المفيدة للمعرفة المكتشفة، مثل بناء تقارير النتائج، وإنشاء الاستعلامات عن توقعات نماذج التنقيب على أية بيانات جديدة، والقيام بعمليات التنبؤ المستقبلي، والتعرف على العوامل الأكثر تأثيراً على متغير ما.

وأخيراً ينتهي البحث **بخاتمة** تتضمن جملة من النتائج والاقتراحات المستخلصة من مختلف الفصول والمباحث.

دراسات سابقة

تبيَّن من خلال الاطلاع على الدراسات المحلية السابقة المتعلقة بموضوع اكتشاف المعرفة من قواعد البيانات، أن معظم تلك الدراسات قد أُنجزت في كليات الهندسة الكهربائية والالكترونية وكليات الهندسة المعلوماتية، ولذلك فإنها كانت تهتم بانجاز جوانب تقنية برمجية مثل تطوير بعض الخوارزميات، وتصميم برامج تطبيقية (أو واجهات ويب) تُمكِّن المستخدمين في المجالات المختلفة من الحصول على نتائج عمليات التنقيب على البيانات الجديدة لديهم، لاستخدامها في أغراض التنبؤ أو دعم القرار. وقد تم اختيار عرض لمحة مختصرة عن رسالة ماجستير في كلية الهندسة الكهربائية والالكترونية، ومشروعي تخرج لطلاب كلية الهندسة المعلوماتية.

تصميم برمجيات لتحليل الروابط بين الزبائن لأنظمة البيع بالتجزئة باستخدام تقنيات التنقيب عن البيانات

- رسالة ماجستير في الهندسة الكهربائية والالكترونية - قسم هندسة الحواسيب - جامعة حلب - 2009
- مقدمة من قبل م. حسنين علي طالب
- إشراف د. م. زياد سالم

هَدَف هذا البحث إلى دراسة خوارزميات التنقيب المناسبة لتحليل الروابط والعلاقات في فواتير الزبائن في أنظمة البيع بالتجزئة، وقد جرى تطوير خوارزمية تستفيد من خوارزمية قواعد الاقتران في كشف المواد شائعة الشراء معاً، وبَعدها تم تجميع قواعد الاقتران الناتجة معاً بهدف التمكن من معرفة مجموعات المواد المتشابهة سلوكياً. وقد قام الباحث بتصميم أداة برمجية للقيام بهذه المهمة، والتي يمكن أن تُقدَّم إلى بائعي التجزئة لمساعدتهم في تصريف بضائعهم، وبناء استراتيجيات علاقاتهم مع الزبائن. ولكن لم تتطرق هذه الدراسة إلى مخرجات التنقيب في البيانات المدروسة، ولم تُظهر أية معرفة مكتشفة.

مشروع التنقيب في المعطيات ومخازن المعطيات: تطبيق على أرصاد جوية

- مشروع تخرج في الهندسة المعلوماتية – جامعة دمشق – 2008
- مقدم من قبل ايفان حسن، جواد الخلخالي، عمر أبو عمار، فاضل حموي وهوازن إبراهيم.
- إشراف د. راكان رزوق

قام الطلاب في هذا المشروع ببناء مستودع للمعطيات يحتوي على المعطيات المتوفرة لدى الهيئة العامة للاستشعار عن بعد، ثم جرى تحليل وتنقيب هذه المعطيات بهدف إيجاد نموذج بياني يمثل تغيُّرات الطقس عبر الزمن، ويُساهم في تحسين مستوى التنبؤ الجوي. وقد تم بناء نظام للتنبؤ الجوي يعتمد على إحدى الخوارزميات التالية (أشجار القرار (Decision Trees) والتصنيف العنقودي (Clustering) ونايف (Naive) وقواعد الاقتران (Association Rules)). وقُدِّمت النتائج على شكل واجهات ويب للمستفيدين، لكي يتمكن هؤلاء من الحصول على جداول للتنبؤ بالأحوال الجوية كدرجات الحرارة ومعدلات الأمطار، وعلى مخططات بيانية لعرض تلك التنبؤات. ولكن لا يوجد في هذا البحث تقييم أداء لنموذج التنبؤ الأفضل، كما بإمكان المستخدم أن يطلب نتائج التنبؤ بحسب أي خوارزمية يختارها من القائمة المخصصة في البرنامج.

تصميم مستودع معطيات وأنظمة التنقيب في المعطيات لشركة سيريتل للاتصالات

- مشروع تخرج في الهندسة المعلوماتية- جامعة دمشق – 2004
- مُقدم من قبل سوزان الحموي، مها الكمشة ورأفت سمور
- اشراف د. راكان رزوق

تقوم فكرة هذا المشروع على أساس بناء نماذج للتنبؤ تساعد على تقييم سلوك الزبائن، والتنبؤ بالزبائن المرشحين للدخول ضمن فئة الديون المعدومة أو الانتقال إلى الشركة المنافسة. وقد قام الطلاب في هذا البحث بتصميم مستودع للبيانات بالاستفادة من نظام الفوترة الموجود في المؤسسة، وجرت معالجة البيانات والتنقيب فيها، ومن ثم تصميم واجهات للمستخدم تساعده على الحصول على نتائج التنقيب، وقد تم فيها استخدام كل من خوارزميتي أشجار القرار(Decision Trees) ونايف بايز (Naïve Bayes). ولكن لم تعرض الدراسة أية مخرجات معرفية من البيانات المدروسة.

يتقاطع البحث الحالي مع سابقيه في بعض الخطوط النظرية العريضة، ولكنه يختلف عنها في العديد من التفاصيل النظرية، وفي الناحية التطبيقية من حيث نوعية البيانات المدروسة (بيانات طبية) ومسائل اكتشاف المعرفة وطريقة معالجتها وأدوات معالجتها وأهداف البحث. كما وانصب اهتمام معظم الدراسات السابقة نحو انجاز أجزاء من مراحل عملية اكتشاف المعرفة وليس كل المراحل، ولم تُعرض في أغلب الأحيان أية مخرجات لعمليات التنقيب في البيانات المدروسة، ولم تهتم أغلب تلك الدراسات بتحسين جودة البيانات أو بالتوصل إلى معرفة مكتشفة من البيانات. في حين يسعى البحث الحالي – المُنجز ضمن اختصاص قواعد البيانات في قسم الإحصاء التطبيقي– إلى تنفيذ جميع مراحل عملية اكتشاف المعرفة على البيانات المتوفرة للدراسة، ويعمل على تركيز الاهتمام لاكتشاف المعارف المخبأة في البيانات وبتقييم المعرفة المكتشفة وبطرح بعض الاستخدامات لها، ولكن لن يتم تصميم واجهات ويب تفاعلية مع المستخدمين.

أما عن الدراسات ذات الصلة بالمجال التطبيقي لهذا البحث، فسيتم التطرق إليها في الفصل الرابع. وتستخدم بعضُ هذه الدراسات تقانةَ التنقيب في البيانات لدراسة نتائج التحاليل الطبية للمرضى.


الجزء النظري
الفصل الثالث
مراحل عملية اكتشاف المعرفة من قواعد البيانات

الفصل الأول

مفاهيم أساسية ومقارنات في مجال اكتشاف المعرفة من قواعد البيانات

الفصل الأول

مفاهيم أساسية ومقارنات في مجال اكتشاف المعرفة من قواعد البيانات

مقدمة

تقوم فكرة اكتشاف المعرفة على أساس استخراج حقائق مخفية تصف البيانات كالأرقام والرموز والصور والنصوص، وهي ليست بفكرة جديدة فالناس دائماً يقومون بذلك في جميع مجالات الحياة، ويعتمدون في ذلك إما على بعض الأساليب الإحصائية مثل نظرية بايز وتحليل الانحدار، أو بالاعتماد على الطريقة الذهنية كالطبيب عندما يحاول اكتشاف المرض الذي يعاني منه المريض، والمحقق الذي يحاول اكتشاف مرتكب الجريمة، وكأي شخص يحاول تحليل تصرفات الآخرين واكتشاف المعالم المخفية في شخصياتهم. وتتصف جميع تلك الطرق بأنها تبحث عن أشياء تقع في حدود الإدراك والمنطق والخبرة السابقة والعدد المحدود من المتغيرات، والجديد في تقانة اكتشاف المعرفة أنها تبحث عن أشياء قد لا تقع في هذه الحدود.

أدى انتشار وتزايد قوة تقنيات الحواسيب في العصر الحالي إلى توسيع قدرة قواعد ومستودعات البيانات على استيعاب كميات هائلة من البيانات. وقد ظهرت مع تزايد أحجام البيانات المتراكمة في الحواسيب[1] فكرة أنه لابد من أن يختفي خلف هذه الحجوم الهائلة العديدُ من المعلومات التي قد تحمل أهميةً معرفيةً أو إستراتيجية[2]. ولكن من جهة أخرى شكَّل الحجم الضخم للبيانات والتشعب الكبير في خاصيات البيانات عائقاً أمام الغوص في مفردات البيانات، واكتشاف المجهول فيها اعتماداً على العلوم والتقانات والأدوات المتوفرة...فأين الحل؟.

[1] وقد أصبح العديد منها يقاس بمضاعفات التيرا بايت (1Terabyte =1,000,000,000,000 byte)

[2] Adriaans, P. (1999), **Introduction to Data Mining and Knowledge Discovery**

لقد بدأ يظهر الحل الذي فتح الباب نحو طريقة تفكير جديدة عندما بُدأ العملُ على دمج أساليب التحليل اليدوي المباشر للبيانات مع أساليب المعالجة الآلية للبيانات، حيث تمت الاستعانة بعدد من حقول المعرفة كعلم الإحصاء، وبعض حقول علم الحاسوب مثل حقل تعليم الآلة (Machine Learning)، والعرض المرئي (Visualization)...الخ، ودُمجت المبادئ والأدوات المنتقاة من هذه الحقول معاً لتشكل حقل اكتشاف المعرفة من قواعد البيانات. والذي استطاع أن يحقق انجازات كبيرة في مجال تبسيط فهم البيانات[1] وفي التمكن من تحليل مجموعات الخاصيات التي تتصف بصعوبة التعرف على أنماطها لأن الروابط بينها مجهولة وأحجامها كبيرة. وإن تطبيق تقانة اكتشاف المعرفة لا يعني اكتشاف كافة المعلومات الموجودة في البيانات إنما جزء منها فقط ويقدر بنسبة 20% تقريباً[2] وهو الجزء الخاص بالمعرفة المخفية التي تعجز عنها الأساليب الأخرى.

ويختص الفصل الحالي بالتعريف بالمفاهيم الأساسية المرتبطة بمفهوم اكتشاف المعرفة من قواعد البيانات، وبإجراء بعض المقارنات بين عدد من المفاهيم في هذا المجال والتي تمتلك نقاط مشتركة أو نقاط اختلاف. ويهدف هذا الفصل إلى تقديم أفكار مختصرة عن المفاهيم التي سيرد ذكرها في الفصول القادمة من هذه الدراسة، وإلى إعطاء فكرة عن اختلاف أو تشابه المهام التي تستطيع أن تقوم بها الأساليب الآلية المختلفة التي تتعامل مع البيانات.

1•1 المفاهيم الأساسية

1•1•1 اكتشاف المعرفة من قواعد البيانات (Knowledge Discovery in Database (KDD))

هي إحدى حقول علم الحاسوب، وتُعبِّر عن عملية البحث الآلي في كمية هائلة من البيانات بهدف استنباط أنماط (Patterns) معرفية عن البيانات[3]. وتعَّرف بحسب [Fayyad, Piatesky-Shapiro & Smyth 1996] [4] بأنها عملية غير بسيطة في استنباط أنماط صحيحة وجديدة ومفيدة ومفهومة من البيانات. أي أنها عبارة عن طريقة للتعلم من الماضي (البيانات والمعلومات المخزنة والمتاحة)، بهدف الوصول إلى معرفة تساعد على اتخاذ قرارات أفضل في

[1] http://www.cs.ou.edu/academics/syllabi_08/spring09/CS_5083_McGovern

[2] بحسب ما أشارت إليه معظم المراجع

[3] http://en.wikipedia.org

[4] Fayyad, U; Piatetsky, G. & Smyth, P. (1996), **"From data mining to knowledge discovery in databases"**, P.41

المستقبل[1].

وسنتعرف فيما يلي على معاني المفردات الواردة في هذا التعريف.

البيانات (Data): هي عبارة عن الحقائق والأرقام والنصوص والصور، والأحداث.. التي يمكن أن تُعالَج وتتحول إلى معلومات أو معرفة.

النمط (Pattern): هو عبارة عن طريقة للتعبير عن فهم البيانات من خلال سلوكها أو صفاتها، أو هو مخرجات النموذج الذي يصف البيانات، ويكون قابلاً للتطبيق على مجموعة بيانات أخرى مشابهة للبيانات التي استُخرج منها النمط.

عملية (Process): يقصد بالعملية أن اكتشاف المعرفة من قواعد البيانات يتألف من عدة مراحل وهي بشكل عام تبدأ بوجود مسألة أو مشكلة مطروحة في مجال ما، ثم يتم استيعاب وتهيئة وتحليل البيانات المتاحة والمناسبة لهذه المشكلة، وتُطبق عليها بعد ذلك خوارزميات التنقيب في البيانات، ويجري البحث فيها عن الأنماط التي تصفها والتي تقدم حلاً للمشكلة المطروحة، وأخيراً تقييم المعرفة المستخلصة ووضعها في الاستخدام بالشكل المناسب. وهذه المراحل متداخلة ومتكررة بحسب المستجدات والاحتياجات التي تنشأ أثناء بقية المراحل، وسيتم التعرف بالتفصيل على هذه المراحل في الفصلين الثالث والخامس.

استنباط النمط (Extracting A Pattern): ويعني التوصل إلى النمط الذي يُعبِّر عن البيانات من خلال إيجاد بنية تُعبِّر عن البيانات، وتصف كل مجموعة جزئية من البيانات المدروسة.

غير بسيطة (Nontrivial): أي غير عادية ويقصد بها أن معظم عمليات البحث والاستدلال تحتاج إلى خوارزميات أو عمليات حسابية معقدة، وليست بسيطة كتلك التي تنتج عن الاستعلامات العادية التي تَحسب المجموع والعدد والقيمة المتوسطة.

صحيحة (Valid): تستند هذه العملية إلى بيانات حقيقية موجودة وتصف الواقع، وستكون بالتالي الأنماط المكتشفة منها معبِّرة عن الواقع، وقابلة للتطبيق على البيانات الجديدة المشابهة بدرجة احتمالية جيدة.

---

[1] Michael, J. A. & Berry, (2004), **Data Mining Techniques For Marketing, Sales, and Customer Relationship Management**, P.44

جديدة (Novel): أي مبتكرة ومختلفة عما تستطيع أن تقدمه التقانات الأخرى، وذلك أنها لا تستند غالباً إلى فرضيات مسبقة في البيانات، وإنما تعتمد على البحث عن الصفات المخفية في البيانات.

مفيدة (Useful): أي تستطيع أن تؤدي بعض المنفعة إلى المستخدم أو سير العمل أو الجهة المستفيدة.

مفهومة (Understandable): أي أن تكون مفهومة إما مباشرة أو بعد إجراء بعض المعالجات اللاحقة عليها، ويتم التركيز على تمثيل أو إظهار المعرفة المكتشفة بأسلوب بسيط وقابل للتفسير، مثل استخدام المدرجات التكرارية ومخططات الانتشار، وقواعد الاقتران بين البيانات.

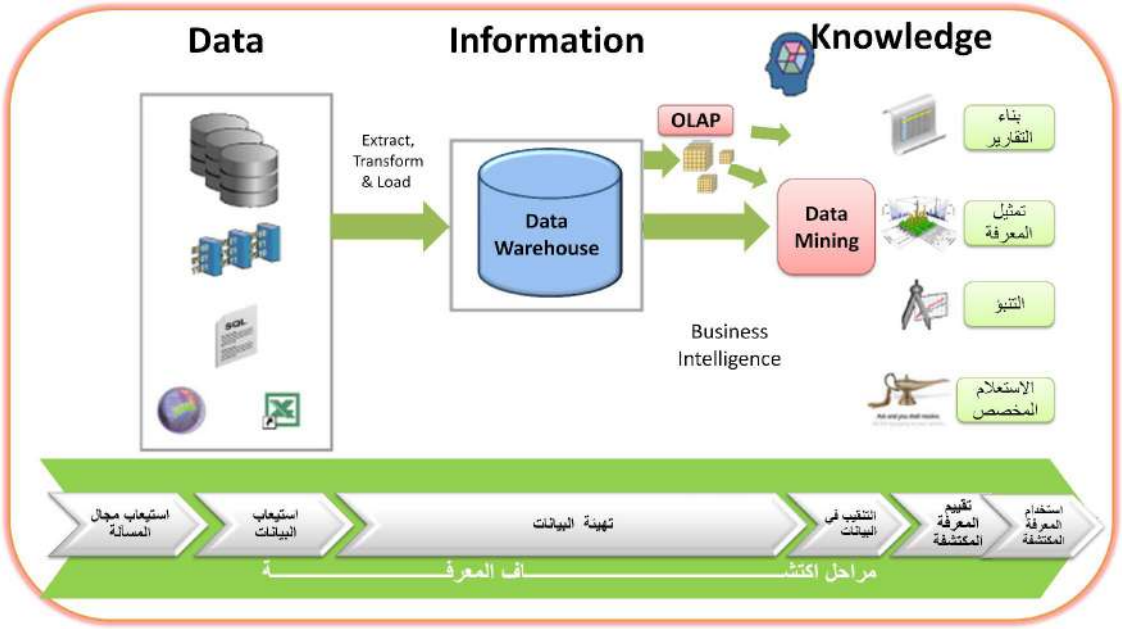
المعلومات (Information): هي البيانات التي تمت معالجتها وتحولت إلى علاقات ونماذج. وتقدم المعلومات الإجابة عن التساؤلات التي من نمط (من، ماذا، أين، متى).

المعرفة (Knowledge): هي المعلومات المسبقة التي تمت معالجتها بطرق خاصة، وتم التمكن من تحويلها إلى استنتاجات هامة تفيد في بناء أحكام وقرارات واقعية. وتكون المعرفة الناتجة إما توصيفية حول الأنماط الحالية للحالة المدروسة أو تنبؤية حول التوقعات المستقبلية، وتجيب المعرفة عن التساؤلات التي من نمط (كيف).

ويمكن تعريف عدد من الطرائق أو المقاييس لتقييم الأنماط المستنبطة مثل مقاييس الدقة (accuracy) التي تُستخدم لتحديد دقة تنبؤ نماذج التنقيب المكتشفة، ومقاييس المنفعة (utility) التي تُقيِّم كمية النقود الممكن تحقيقها وخسارتها بنتيجة التنبؤات الأفضل. وفي الوقت ذاته إن تقييم بعض الأمور الخاصة بالأنماط المكتشفة كالابتكار وبساطة الاستيعاب هي أمور شخصية تختلف من شخص لآخر، وتتعلق عادةً بالمسألة المدروسة، وبالخلفيات العلمية للمستفيدين.

تحتوي مراحل اكتشاف المعرفة عادةً على مجموعة من التقنيات بحيث تتلقى كل تقنية مدخلاتها من البيانات من التقنية أو المرحلة السابقة لها ثم تحولها بأسلوبها الخاص إلى مجموعة من المخرجات وتزود بها التقنية أو المرحلة التالية لها، انظر الشكل (1-1). وأهم هذه التقنيات مستودع البيانات، والمعالجة التحليلية الآنية والتنقيب في البيانات، ومن الممكن

بحسب المسألة المدروسة أن يتم الاستغناء عن إحداها، باستثناء تقنية التنقيب في البيانات فهي أهم مكون من مكونات اكتشاف المعرفة.



الشكل (1– 1): تدفق البيانات عبر التقنيات المختلفة في تقانة اكتشاف المعرفة

2•1•1 التنقيب في البيانات (Data Mining)

هو مرحلة جزئية من العملية الكلية لاكتشاف المعرفة من قواعد البيانات[1] وهي المرحلة التي يتم فيها التعرف الآلي العميق على البيانات المدروسة، وذلك لتحقيق نوعين من الأهداف هما تصنيف البيانات ضمن مجموعات متشابهة في قيم الصفات (أو الخاصيات)، والتنبؤ بقيم إحدى الصفات استناداً إلى قيم الصفات الأخرى.

تتعدد طرق التنقيب في البيانات وتختلف باختلاف الخوارزميات التي تُطَبَّق فيها، وتُعرف خوارزميات التنقيب[2] بأنها آلية لتوليد نماذج التنقيب، حيث تعمل الخوارزمية على تحليل مجموعة البيانات المدروسة، وتبحث فيها عن الأنماط والاتجاهات، وتتوصل بالتالي إلى مجموعة من النتائج (أنماط واتجاهات) التي تُخزن في نموذج التنقيب، وبعد ذلك ترفدُ عمليةُ التنقيب تقانةَ اكتشاف المعرفة بالنتائج ليُصار إلى اختبارها وتقييمها ومقارنتها واختيار الأنسب والأقدر واعتبارها على أنها معرفة مكتشفة قابلة للاستخدام بطرق متعددة.

[1] Fayyad, U.; Piatetsky & G. Smyth, P. (1996), **"From data mining to knowledge discovery in databases"**, P.39

[2] SQL Server Tutorials 2008

تعمل عمليات التنقيب على اكتشاف المعارف الجديدة وغير المألوفة في البيانات المدروسة، ولكن قد تُركز هذه العمليات على تعليل بعض الظواهر أو التحقق من صحة بعض النظريات التي يفكر بها الباحثون ومتخذو القرارات. كما يمكن أن تُطبَّق عمليات التنقيب أحياناً في المراحل الأولى لعملية اكتشاف المعرفة للمساعدة في اختيار البيانات وتنظيفها ودمجها. وتطبق كذلك بعض عمليات التنقيب لكشف الحالات التي تتصرف بطريقة خاطئة أو استثنائية أو مختلفة عن المألوف[1] أو التي لا تستجيب لنفس السلوك العام لبقية البيانات، مثل كشف أخطاء الصناعة وكشف الاحتيال في استخدام البطاقات الائتمانية، والتعرف على الحالات النادرة مثل الأمراض النادرة.

1•2•1•1 أنواع التنقيب في البيانات

تتجه عمليات التنقيب في البيانات لانجاز هدفين أساسيين كما سبق وذُكر، ويتحدد هدف التنقيب بحسب مسألة التنقيب المطلوب حلها، وبحسب طبيعة البيانات، ويوجد مقابل ذلك نوعان أساسيان من التنقيب هما:

1- **التنقيب الوصفي** (Descriptive Data Mining)، ويعتمد على محاولة فهم البيانات والبحث عن نقاط التشابه بينها، وتصنيفها ضمن مجموعات تحمل عناصرها صفات متشابهة. ويمكن إنجاز هذا التنقيب باستخدام عدد من خوارزميات التنقيب، مثل خوارزميات التجزئة العنقودية، وخوارزميات قواعد الاقتران، وخوارزميات التحليل التتابعي.

2- **التنقيب التنبؤي** (Predictive Data Mining)، ويعمل على إيجاد أفضل التنبؤات أو توقع ما سيحدث في المستقبل اعتماداً على البيانات الحالية، ويمكن استخدام عدد من خوارزميات التنقيب في انجاز هذا التنقيب، مثل خوارزميات أشجار القرار وخوارزميات الشبكات العصبونية وخوارزميات الانحدار.

ولا يوجد فواصل حادة بين هذين الهدفين حيث قد تحقق بعض الأنماط كلا الهدفين، ولكن يُقصد من التمييز بينهما أن يتم استيعاب الهدف الكلي لعملية اكتشاف المعرفة[2]. كما أن انتماء خوارزمية ما إلى إحدى نوعي التنقيب السابقين لا ينفي إمكانية استخدامها في النوع الآخر من

[1] MacLennan, J.; Tang, Z. Crivat, B. (2009), **Data Mining with Microsoft SQL Server 2008**, P.9
[2] Fayyad, U.; Piatetsky, G. & Smyth, P. (1996), **"From data mining to knowledge discovery in databases"**, P43

التنقيب، وذلك بحسب المسألة المطلوب حلها، وعلماً أن معظم الخوارزميات تتطلب أثناء بناء نماذج التنقيب التي تحتويها أن يتم تعريف متغير متنبأ به (Predicted Variable).

2•2•1•1 أنماط خوارزميات التنقيب في البيانات

يختص كل نمط من أنماط خوارزميات التنقيب بنوع معين من المهام، وتتنافس الشركات المنتجة للبرامج الحاسوبية للتنقيب في البيانات على تطوير خوارزميات تنقيب خاصة بها انطلاقاً من الخوارزميات الأساسية الشائعة. وتختلف لذلك برامج التنقيب عن بعضها في التعليمات البرمجية الداخلية لخوارزميات التنقيب المُستخدمة فيها، والتي تبقى في كثير من الأحيان سرية للحفاظ على الوضع التنافسي للشركة المنتجة. يرتبط اختيار خوارزمية التنقيب بسؤال التنقيب المطروح، وتستطيع عدة خوارزميات أن تجيب عن مسألة التنقيب المطروحة بطريقتها الخاصة، ولذلك يمكن تطبيق عدد من الخوارزميات ومقارنة أدائها واختيار الأفضل بينها، ونبين فيما يلي أبرز أنماط خوارزميات التنقيب في البيانات.

أ- خوارزميات التصنيف (Classification Algorithms)

تُعنى بتخصيص كل سجل للانتماء إلى إحدى المجموعات المحددة مسبقاً، ويمكن أن تتنبأ هذه الخوارزميات بمتغير واحد أو أكثر استناداً إلى المتغيرات (الخاصيات) الأخرى في قاعدة البيانات، مثل تصنيف درجات تخرج طلاب الجامعة (ممتاز، جيد جداً، جيد، وسط، رسوب..) بحسب العمر، والجنس، ومكان، الإقامة، والتحصيل العلمي للأبوين، وعمل الأب، وعمل الأم، والعلامات الفردية للمواد، ومصدر الشهادة الثانوية، والعلامات الفردية لمواد الشهادة الثانوية. وتُستخدم عدة خوارزميات تنقيب في هذا النمط، ويُذكر منها خوارزمية مايكروسوفت لأشجار القرار (Microsoft Decision Trees Algorithm) وخوارزمية مايكروسوفت للشبكات العصبونية (Microsoft Neural Networks Algorithm) وخوارزمية مايكروسوفت للاحتمالات نايف بايز (Microsoft Naïve Bayes Algorithm).

ب- خوارزميات الانحدار (Regression Algorithms)

تقوم بتحليل العلاقة بين قيم الخاصيات (attributes) في نفس السجل واكتشاف التابع الذي يصف البيانات، ويتنبأ بإحداها، ثم تقوم الخوارزميات بالتوليد الآلي للنماذج التي تستطيع أن

تتنبأ بقيم الخاصيات في السجلات المستقبلية. ويشبه عمل هذه الخوارزميات عمل خوارزميات التصنيف، باستثناء أنها بدلاً من البحث عن الأنماط التي تصف الطبقة التصنيفية (Class)، فإنها تهدف إلى إيجاد أنماط لتحديد القيمة العددية (numerical value)، ومثال ذلك التنبؤ بحجم طلب الزبائن على سلعة جديدة استناداً إلى نفقة الإعلان، وأماكن التوزيع، ونوع الإعلان (طرقي، إذاعي، مجلات..)، وزمن الإعلان. ومن هذه الخوارزميات خوارزمية مايكروسوفت للانحدار الخطي (Microsoft Linear Regression Algorithm)، وخوارزمية مايكروسوفت للانحدار المنطقي (Microsoft Logistic Regression Algorithm).

ج- خوارزميات التقطيع (Segmentation Algorithms)

تقوم بتقسيم مجموعة البيانات إلى تجمعات عنقودية جزئية حيث تتشابه صفات البيانات التي تنتمي إلى نفس التجمع، فمثلاً أدى تطبيق هذه الخوارزمية على خاصيات مكان الإقامة ولون السيارة وحجمها إلى التوصل إلى عدة فئات من الزبائن، وتتضمن الفئة الأولى (الدول الخليجية واللون الأسود والحجم الكبير)، وتتضمن الفئة الثانية (الدول الأوربية واللون الأزرق الفاتح والحجم الصغير)..الخ ويستفاد من هذه المعلومات في التحديد الأولي لرغبات الزبائن واختيار الدعايات الموجهة بحسب هذه الفئات، وتُعتبر خوارزمية مايكروسوفت للتجزئة العنقودية (Microsoft Clustering Algorithm) مثالاً عن هذه الخوارزميات.

د- خوارزميات الاقتران (Association Algorithms)

تعمل على اكتشاف القواعد التي تُزامن بين مجموعة من قيم خاصيات البيانات المدروسة مع قيم مجموعة أخرى من الخاصيات في قاعدة البيانات، ومن أبرز هذه الخوارزميات هي خوارزمية مايكروسوفت لقواعد الاقتران (Microsoft Association Algorithm). ويعتبر تحليل سلة التسوق التطبيق الأكثر شيوعاً لهذا النوع من الخوارزميات، وينتج عن هذه الخوارزمية قواعد تصف الترافق بين السلع المباعة في نفس الفاتورة، ويعبَّر عنها كما في القاعدة التي تقول { إن زبائن المتجر الذين يشترون أجبان ماركة (أ) وشاي ماركة (ب) يشترون عصير ماركة (ج) بدعم قدره (1500 فاتورة) وباحتمال قدره (0.70%) }.

ه- خوارزميات التحليل التتابعي (Sequence Analysis Algorithms)

تعمل على اكتشاف العلاقات السببية بين الأحداث العابرة المتتالية، فمثلاً إن حدوث أكثر من 3 هزات أرضية قوية أو زلازل مدمرة في أماكن متفرقة من العالم خلال ستة أشهر، يقود إلى توقع انفجار بركان باحتمال قدره 75%.

أو تقوم هذه الخوارزميات بتلخيص السلاسل المتعاقبة أو الأحداث المتكررة في البيانات، مثل تحليل وتوصيف مسارات تدفق الوثائق المترابطة عند استخدام صفحات الانترنت للتعرف على أنماط الاستخدام والمستخدمين للانترنت، وتُعتبر خوارزمية مايكروسوفت للتجزيء العنقودي التتابعي (Microsoft Sequence Clustering Algorithm) كمثال على هذه الخوارزميات.

وتُطبق عمليات التنقيب على أشكال مختلفة من البيانات وعلى أشكال مختلفة من مخازن البيانات كقواعد البيانات العلاقاتية ومستودعات البيانات وملفات الجداول والملفات النصية، وملفات الصور.. وتزداد قوة وجودة وديمومة عمليات التنقيب عندما يتم تحويل أي من بيانات المخازن السابقة إلى بنية مستودع بيانات.

3•1•1 ذكاء الأعمال (Business Intelligence (BI))

يقصد بهذا المفهوم جعل المعلومات متاحة في الشركة باستمرار، وتقديمها لمساعدة متخذي القرارات في اتخاذ قراراتهم بشكل أفضل وأسرع[1] من خلال استخدام التقنيات الحاسوبية التي تدعم كل من إنشاء التقارير، والمعالجة التحليلية الآنية، والتنقيب في البيانات، وإدارة أداء الأعمال، والتنقيب في النصوص، والتحليل التنبؤي..الخ. ويُستخدم هذا المفهوم في بعض البرامج الحاسوبية عند الرغبة في إنشاء مشاريع لأغراض المعالجة التحليلية الآنية أو التنقيب في البيانات مثل برامج (Microsoft SQL Server 2008).

4•1•1 قواعد البيانات العلاقاتية

هي عبارة عن مجموعة من الجداول ولكل منها اسم مميز وحيد، ويتكون كل جدول من مجموعة من الخاصيات (الحقول أو الأعمدة)، ومن مجموعة من السجلات أو الأسطر. ويتم

[1] Jacobson, R. (2006), **SQL Server 2005 Analysis Services**, P.3

تمثيل العلاقات بين الجداول باستخدام نموذج الكيان– العلاقة (Entity-Relationship Diagram (ERD))، ويمكن الوصول إلى أي من البيانات العلاقاتية من خلال الاستعلامات المبنية باستخدام إحدى لغات البرمجة أو من خلال واجهات بيانية سهلة على المستخدمين. كما يمكن الحصول في قواعد البيانات العلاقاتية على بعض المقارنات وعلى بعض القيم التلخيصية مثل المجموع والمتوسط وعدد القيم.. وتحمل قواعد البيانات بحد ذاتها فوائد عظيمة للقائمين عليها وتساعدهم في إعداد تقارير عن البيانات وعن أداء العمل خلال فترة زمنية معينة، ويمكن تطبيق التنقيب في البيانات على هذه القواعد للحصول على اتجاهات البيانات وأنماطها[1]، واكتشاف المعارف الكامنة فيها إما مباشرةً أو بعد تحويل بيانات هذه القواعد إلى مستودع بيانات.

5•1•1 مستودع البيانات (Data Warehouse)

تمتلك كل جهة سواء أكانت منظمة أو قسم من منظمة قاعدة بيانات خاصة بها لإدارة أعمالها. وقد برزت الحاجة لدى العاملين في الإدارات العليا مع توسع أعمالهم إلى ضرورة وجود البيانات الخاصة بموضوع معين كالزبائن مثلاً في وعاء تخزيني واحد بدلاً من تجزؤ بيانات الزبائن في عدة قواعد بيانات بحسب أماكن إنتاج هذه البيانات (كمراكز التوزيع المختلفة، وأقسام الدعاية والتسويق والمبيعات)، وذلك بهدف استخدام بيانات هذا الغرض في بناء القرارات الهامة وفق أسس حقيقية مستمدة من مسيرة عمل الشركة، ومن بياناتها التاريخية. ومن هنا جاءت فكرة مستودعات البيانات التي تختصر جميع قواعد البيانات المطلوبة وفق أهداف محددة في مكان تخزيني واحد هو مستودع البيانات، ويتم تغذيته دورياً بالبيانات الجديدة. ويُستخدم مستودع البيانات في كثير من الأحيان كمرحلة انتقالية في حياة البيانات أثناء استخدامها في عملية اكتشاف المعرفة حيث تنتقل البيانات من أماكن تخزينها الأصلية إلى مستودع البيانات، الذي يُقدِّم بدوره البيانات المناسبة إلى عمليات التحليل والتنقيب، راجع الشكل (1–1).

[1] Han, J. & KAmber, M. (2001), **Data Mining Concepts and Techniques**, P.12

يُعرَّف مستودع البيانات بحسب [W. H. Inmon, 1992][1] بأنه مجموعة من البيانات الموجهة بحسب الموضوع (subject-oriented)، والمدموجة أو المتكاملة (integrated)، والمتعلقة بالزمن (time-variant)، وغير المتقلبة (nonvolatile)، والتي تهدف إلى دعم عملية اتخاذ القرارات الهامة.

يستمد مستودع البيانات بياناته من عدة مصادر للبيانات وينظمها حول مواضيع أساسية تلخيصية كالزبون والمورِّد والسلعة خلال فترة زمنية معينة.

وقد تكون مصادر البيانات غير متجانسة بمعنى أن لكل منها نمطها ونظامها التخزيني المختلف، وتجري المحاولات لتوحيدها وتوفيقها معاً عندما يتم جمعها في مستودع البيانات فتظهر وكأنها مدموجة أو متكاملة. وتستطيع مستودعات البيانات أن تستوعب أنماطاً مختلفة من البيانات كالمبادلات، والأرقام والرموز، والصور والنصوص ورسائل البريد الالكتروني والعقود والتقارير[2].

تعتبر بيانات المستودعات بأنها غير متقلبة لأنها تُخزن في أماكن معزولة عن العمليات اليومية حيث لا يُستخدم المستودع في إدخال بيانات العمليات اليومية كالمبادلات أو الواقعات (Transactions)، وإنما يتم فقط تحميل المستودع بالبيانات في فترات متباعدة ليتم الوصول إليها من قبل أدوات برمجية خاصة بدراسة هذه البيانات.

ويعتبر الزمن عنصراً أساسياً في وصف البيانات حيث يتم إنشاء جدول خاص بالزمن يتضمن حقلاً خاصاً بالتاريخ الممتد خلال الفترة الزمنية التي يمثلها المستودع ويتضمن أيضاً مجموعة من الحقول التي تصف هذا التاريخ بحسب السنة وربع السنة والشهر واسم اليوم..، ويُستفاد من هذه الحقول في الاستعلام وتقسيم البيانات وفقها والتعرف على التغيرات التي تطرأ على البيانات بالترافق مع تغير خاصيات الزمن كتغير الأشهر أو الفصول أو الأيام. ويطلق على جدول الزمن في مستودع البيانات اسم البعد الزمني. ويراعى أثناء تصميم مستودع البيانات أن تتمكن الجداول فيه من تشكيل بنية تخيلية هي مكعب البيانات الذي سيرد تعريفه في الفقرة القادمة. ويوجد ثلاثة استخدامات لمستودعات البيانات[3] هي:

---

[1] The previous reference
[2] Inmon, B. (2010), **"Including Unstructured Text in the Data Warehouse"**
[3] Han, J. & KAmber, M. (2001), **Data Mining Concepts and Techniques**, P.94

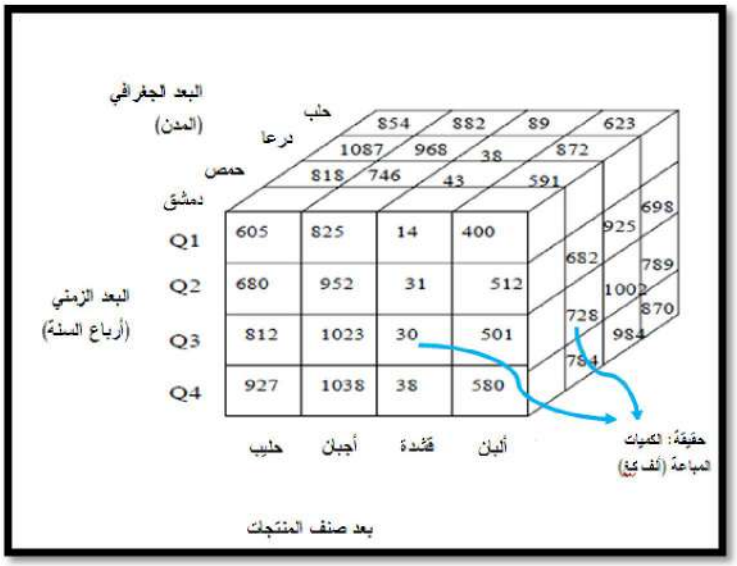
أ- معالجة البيانات، وتدعم إجراء الاستعلامات والتحليل الإحصائي وإنشاء التقارير باستخدام الجداول والرسوم البيانية.

ب- المعالجة التحليلية، وتدعم عمليات المعالجة التحليلية الآنية (On Line Analytical Processing) التي تتضمن اقتطاع شرائح أو أجزاء من البيانات، وعمل استعلامات تلخيصية أو تفصيلية.

ج- التنقيب في البيانات، وتدعم اكتشاف المعرفة من خلال إيجاد الأنماط الكامنة والارتباطات ونماذج التحليل البنيوي، وإيجاد تصنيفات وتنبؤات وأخيراً تمثيل المعرفة المكتشفة من خلال أدوات العرض المرئي (visualization tools).

6•1•1 مكعب البيانات (Data Cube)

هو عبارة عن طريقة أو مرأى لربط البيانات مع بعضها ضمن نموذج تخيلي متعدد الأبعاد للبيانات، ويتكون مكعب البيانات من صنفين من الجداول هما جداول الأبعاد وجداول الحقيقة الذين سيأتي ذكرهم تباعاً. فبينما يتألف الجدول من بعدين (الأسطر والأعمدة)، فإن مكعب البيانات يتألف من عدة أبعاد وكل بعد هو جدول، ويمكن تعريف أي نقطة من مكعب البيانات من خلال قيمتها بالنسبة إلى كل من أبعاد المكعب، انظر الشكل (1-2). ويُستخدم مكعب البيانات كأساس لترتيب تخزين البيانات واسترجاعها في مستودعات البيانات وفي تنفيذ عمليات المعالجة التحليلية الآنية.



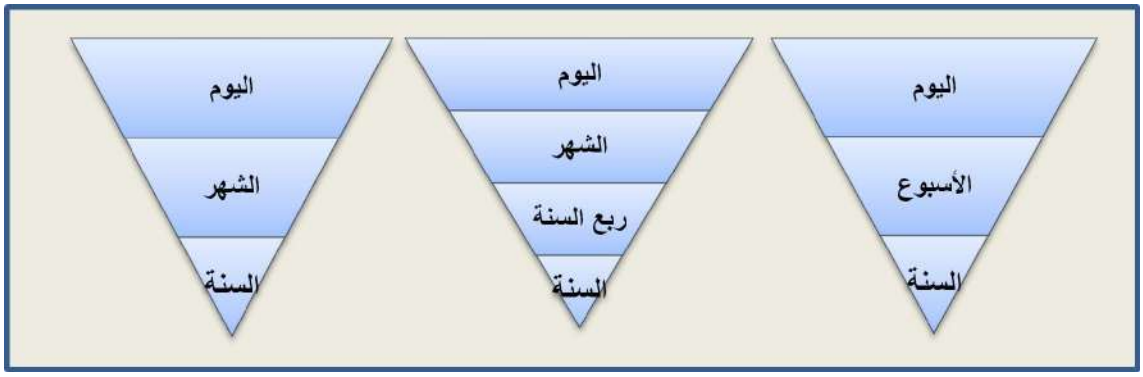
الشكل (1- 2): مكعب بيانات ثلاثي الأبعاد لكميات مبيعات (حقائق) شركة ألبان وأجبان بحسب البعد الزمني والجغرافي والصنف

أ- الأبعاد (dimensions)

هي خاصيات (Attributes) أو مرائي (Views) عن الخاصيات متعلقة بالأشياء التي ترغب المنظمة بحفظ سجلات عنها. فمثلاً قد ترغب المنظمة بإنشاء مستودع بيانات للمبيعات بغرض حفظ سجلات عن مبيعات متاجرها بحسب بعد الزمن والموقع الجغرافي وحجم المتجر والصنف المباع..الخ، وكل من هذه الأبعاد له جدول يرتبط به هو جدول البعد (dimension table). وتعتبر المهمة الأساسية للبعد هي في انتقاء (filtering) وتصنيف (grouping) وعنونة البيانات (labeling). وغالباً ما يتكون كل بعد من سلسة أو بنية هرمية (Hierarchy) أو أكثر من الخاصيات، فمثلاً يمكن تمثيل البعد الزمني من خلال البنى الهرمية التالية:

(اليوم⊃ الشهر⊃ السنة)، (اليوم⊃ الشهر⊃ ربع السنة⊃ السنة)، (اليوم⊃ الأسبوع⊃ السنة)،

انظر الشكل (1-3)، كما يمكن تعريف بنى هرمية زمنية أخرى تتعلق بتعريف السنة كالسنة الدراسية وأيام العمل. ويوجد بنى هرمية للموقع الجغرافي تبدأ مثلاً من اسم المدينة، ثم المحافظة، ثم المنطقة الجغرافية.



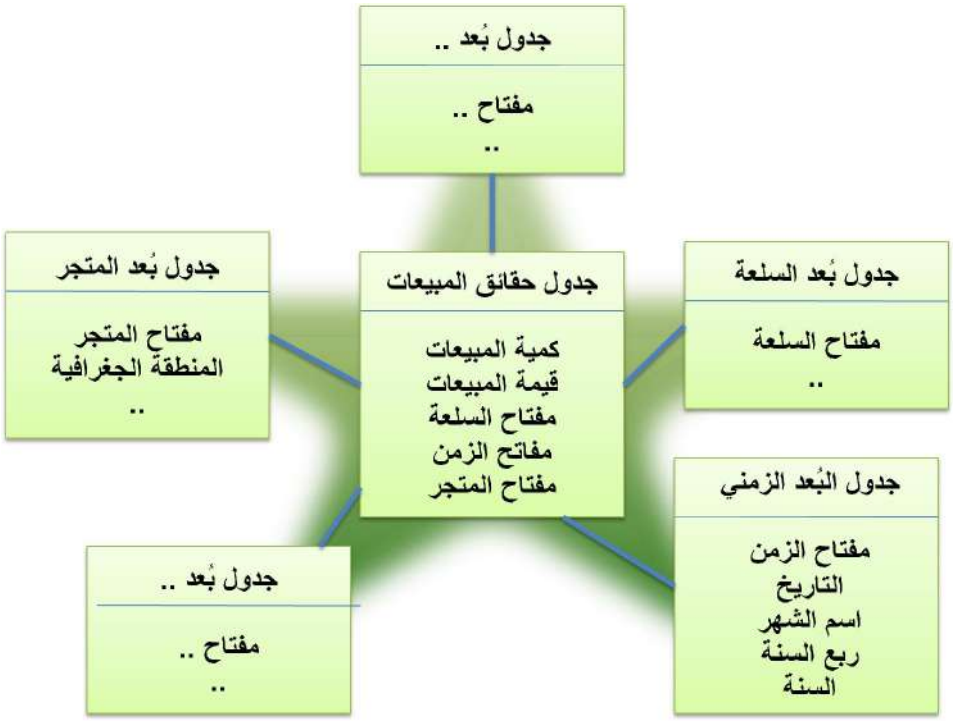
الشكل (1- 3): أمثلة عن بعض التراتبات الهرمية للبعد الزمني

ب- الحقائق (Facts) أو المقاييس (Measures)

هي مقاييس رقمية قابلة لإجراء عمليات التجميع عليها (كالمجموع والوسطي)، وتُعبِّر عن القيم التي ترغب المنظمة بإجراء استعلامات معقدة عنها، وذلك نسبةً إلى الأبعاد المرتبطة بها. وتخزن الحقائق في جداول خاصة بها تدعى جداول الحقيقة، والتي تحتوي أيضاً على مفاتيح الربط الثانوية بجداول الأبعاد. فمثلاً يساعد وجود بنية مكعب البيانات في تمكين المنظمة من التعرف على قيم وكميات المبيعات الشهرية، لأصناف معينة للمتاجر التي تقع في

منطقة جغرافية معينة... وتتمكن من ذلك من خلال الاستعلام المتعدد الأبعاد عن قيم وكميات المبيعات المخزنة في جدول الحقيقة الخاص بالمبيعات مع انتقاء أصناف معينة من بُعد السلعة، وانتقاء منطقة جغرافية معينة من البُعد المكاني وعرض البيانات وفق خاصية الأشهر من البُعد الزمني.

ويتم عادة تمثيل نموذج البيانات متعدد الأبعاد من خلال مخطوطة تحتوي على جدول حقيقة واحد على الأقل وعدد من جداول الأبعاد، وتختلف أشكال المخطوطات بحسب عدد جداول الحقيقة التي تتألف منها، وبحسب علاقتها المباشرة وغير المباشرة مع جداول الأبعاد. وأبسط شكل للمخطوطة هو المخطوطة النجمية (star schema) التي تحتوي جدول حقيقة واحد تربطه علاقات مباشرة مع الأبعاد، انظر الشكل (1-4)، أما إذا ارتبط جدول الحقيقة الوحيد بعلاقات مباشرة مع بعض الأبعاد التي ارتبطت بدورها مع أبعاد أخرى (ارتباط غير مباشر بين جدول الحقيقة وهذه الأبعاد) فتتشكل بالتالي مخطوطة رقاقة الثلج (snowflake schema)، وإذا تعددت جداول الحقيقة تشكلت مخطوطة أكثر تعقيداً تدعى مخطوطة مجموعة النجوم (constellation schema).



الشكل (1- 4): المخطوطة النجمية لمكعب بيانات المبيعات

7•1•1 المعالجة التحليلية الآنية (Online Analytical Processing (OLAP))

هي تقنية تحليلية تتألف من مجموعة من الوظائف التي تُستخدم في الاستعلام عن البيانات المتعددة الأبعاد كتلك المخزنة في مستودعات البيانات، وتعمل على إظهار معلومات عن البيانات من زوايا مختلفة. وتتمكن هذه التقنية من القيام بعملها استناداً إلى بنية مكعب البيانات الذي تتحدد فيه الحقائق والأبعاد وبناها الهرمية[1]، والذي يُمكِّنها أيضاً من اقتطاع شرائح أو مكعبات صغيرة أو إجراء استعلامات تفصيلية أو تلخيصية بحسب ما يراه المستخدم مناسباً لبناء استنتاجات مفيدة من البيانات المدروسة. وتعتبر المعالجة التحليلية الآنية أحياناً كإحدى الأدوات التي تُستخدم أثناء مراحل اكتشاف المعرفة (راجع الشكل (1-1))، حيث تساعد في التعرف على البيانات المتعددة الأبعاد وتستخدم المعلومات الناتجة عنها في دعم القرارات أو في التمهيد لدراسة البيانات بأساليب مختلفة كالأساليب الإحصائية أو أساليب التنقيب في البيانات.

وتعتمد هذه التقنية على الجهد البشري في توجيه عملياتها، فهي تقنية للاستعلام متعدد الأبعاد بحسب الطلبات التي يوجهها إليها المستخدم، وتقوم هذه التقنية بأربعة عمليات أساسية مبينة في الشكل (1-5)، وهي[2]:

أ- **طي للأعلى (Roll-up)**: وتفيد الحصول على نتائج تجميع البيانات عند مستوى تصنيفي أعلى في البنية الهرمية لأحد أبعاد مكعب البيانات، مثل عرض مجاميع كميات السلع اليومية المباعة وفق السنوات التابعة لها.

ب- **نثر للأسفل (drill-down)**: وتعني الحصول على نتائج تجميع البيانات بشكل تفصيلي عند مستوى تصنيفي أدنى في البنية الهرمية لأحد أبعاد مكعب البيانات، وهي عكس العملية السابقة، فمثلاً يمكن طلب عرض تفاصيل المبيعات السنوية وفق الأشهر التابعة لكل سنة.

ج- **تقطيع إلى شرائح (Slice)**: أي اختيار شريحة من المكعب تحتوي على أحد أبعاد مكعب البيانات، وبالتالي الحصول على جزء من المكعب الكلي، كأن تُعرض المبيعات المحققة في الربع الثالث فقط من السنوات المتوفرة.

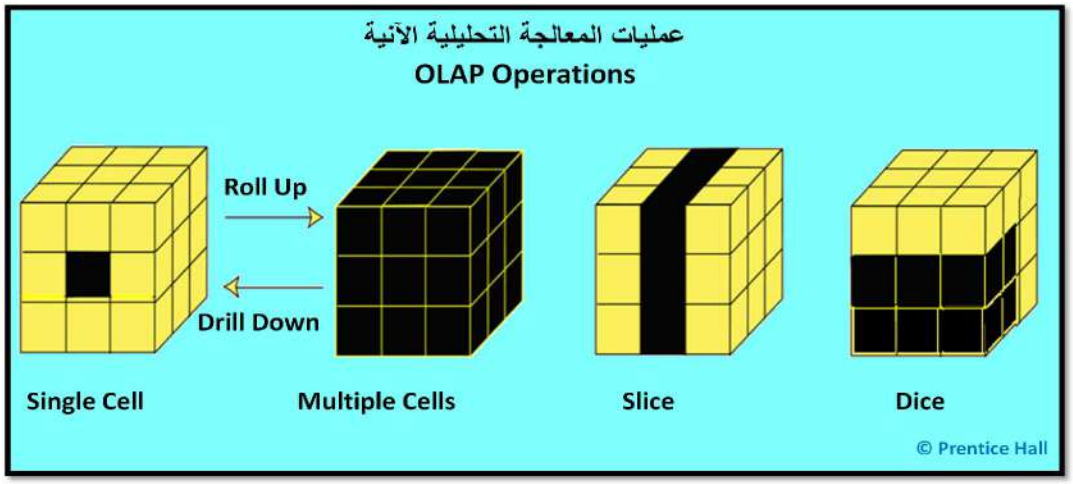
---

[1] Michael J. A. Berry, (2004), **Data Mining Techniques For Marketing, Sales, and Customer Relationship Management,** P.496

[2] Han,J. & KAmber,M. (2001), **Data Mining Concepts and Techniques**, P.59

د- **تقطيع إلى مكعبات أصغر (Dice)**: أي تعريف مكعب جزئي من خلال الاختيار ضمن بعدين أو أكثر من أبعاد مكعب البيانات، كأن يتم عرض المبيعات المحققة في المنطقة الجغرافية الجنوبية في الربع الثالث من السنوات.

وغالباً ما تندمج أكثر من عملية من العمليات السابقة في استعراض البيانات، وإضافةً إلى ذلك يمكن تدوير محاور استعراض مكعب البيانات فيصبح المحور الأفقي عمودياً والمحور العمودي أفقياً. وتتيح أدوات المعالجة التحليلية الآنية إضافة مقاييس جديدة كالقيم الوسطية، وبالإمكان أيضاً طلب تحويل قيم المقاييس إلى نسب مئوية أو طلب المجاميع واستعراض البيانات تنازلياً أو تصاعدياً، وانتقاء عرض أعلى القيم أو أدنى القيم..الخ من العمليات التي ستظهر أثناء التطبيق العملي في المبحث الثالث من الفصل الخامس.



الشكل (1- 5): العمليات الأساسية في المعالجة التحليلية الآنية

8•1•1 تعليم الآلة (Machine Learning)

هو أحد فروع الذكاء الاصطناعي الذي يهتم بتصميم وتطوير خوارزميات وتقنيات تسمح للحواسيب بامتلاك خاصية "التعلم". حيث يتم فيه التعرف على مجموعات البيانات الموجودة، وبالتالي استنتاج قواعد وأحكام عامة عنها، واستخدامها في الحكم على أية بيانات جديدة تشبه البيانات التي تمت دراستها. ويستخدم تعليم الآلة عادةً في تحسين أداء الأعمال، ويعتبر حقل تعليم الآلة من الحقول التي اعتمدت عليها تقانة اكتشاف المعرفة في نشأتها، وحيث أن تعليم الآلة يَستخدم بعض الخوارزميات في استنباط المعلومات من البيانات الخام في قواعد البيانات

فقد قدم جزءاً من القاعدة التقنية للتنقيب في البيانات[1]، ومن هذه الخوارزميات خوارزمية الشبكات العصبونية، وخوارزمية أشجار القرار.

9•1•1 العرض المرئي (Visualization)

يُعرَّف العرض المرئي بأنه استخدام لتقانة التمثيل البصري كأداة لفهم البيانات[2]، ويتم ذلك من خلال تحويل المعلومات إلى رسوم بيانية وأشكال توضيحية، وبالتالي فإن الهدف منه هو إظهار المعلومات بطريقة واضحة وكفوءة عبر الوسائل التصويرية[3]. وقد اعتمد حقل اكتشاف المعرفة على بعض أدوات العرض المرئي في تمثيل الأنماط المكتشفة من البيانات.

2•1 مقارنات في مجال اكتشاف المعرفة من قواعد البيانات

تتسع المقارنات كثيراً في حقل اكتشاف المعرفة، وسيتم التركيز في الفقرات القادمة على أبرز المقارنات التي تقع ضمن مجال اهتمام البحث الحالي.

1•2•1 مقارنة بين قاعدة البيانات (Database) ومستودع البيانات (Data Warehouse)

يتشابه مستودع البيانات مع قاعدة البيانات في أنه يعتبر هو أيضاً قاعدة بيانات مؤلفة من جداول وبيانات وأدلة ومفاتيح ومرائي وعلاقات[4]. ويكمن الفارق الأساسي بينهما في أن قاعدة البيانات قد صُمِّمت ونظِّمت بهدف التسجيل والتوثيق للعمليات الجارية الكترونياً، بينما صُمم ونُظِّم مستودع البيانات فقط للقراءة، وليستجيب للاستعلامات التحليلية التي تقدم معلومات حاسمة عن العمل. ويتم الاستعلام في قاعدة البيانات بوساطة معالجة المبادلات الآنية (On Line Transaction Processing)، بينما يتم الاستعلام في مستودعات البيانات من خلال المعالجة التحليلية الآنية (OLAP). كما يوجد فارق آخر بينهما، وهو استخدام نموذج كيان–علاقة في تصميم قاعدة البيانات، واستخدام نموذج البيانات (Data–Modeling) في تصميم مستودع البيانات[5].

---

[1] Witten, I. & Frank, E. (2005), **Data Mining: Practical Machine Learning Tools and Techniques**

[2] Schulz, H. (2007), **"Visual Graph Mining for Regenerative System"** from (Haber & McNabb (1990))

[3] http://en.wikipedia.org From (Vitaly Friedman (2008))

[4] http://it.toolbox.com/blogs/opensource-analytics/database-vs-data-warehouse-8286

[5] http://www.shvoong.com/internet-and-technologies/blog/1827096-database-vs-data-warehouse/

2•2•1 مقارنة بين معالجة المبادلات الآنية (OLTP) والمعالجة التحليلية الآنية (OLAP)

تُستخدم معالجة المبادلات الآنية في إجراء استعلامات بسيطة من قواعد البيانات، وغالباً ما تتعلق هذه الاستعلامات بالعمليات اليومية التي تُجرى على البيانات. أما المعالجة التحليلية الآنية فتُستخدم في إجراء استعلامات معقدة (مركبة، متعددة الأبعاد) من مستودعات البيانات، ويتم الحصول من خلالها على معلومات تاريخية تلخيصية عن البيانات.

ونورد فيما يلي أهم الفروقات بين هاتين المعالجتين.

الجدول (1– 1): مقارنة معالجة المبادلات الآنية مع المعالجة التحليلية الآنية

	OLTP	OLAP
البيانات	حالية، حديثة، تفصيلية، أحادية البعد، منفصلة عن بيانات الأنظمة الأخرى	تاريخية، تلخيصية، متعددة الأبعاد، مدموجة من عدة مصادر
الاستخدام	العمليات اليومية	العمليات التحليلية ودعم القرار
المستخدمين	مستخدمي قواعد البيانات	خبراء تحليل بيانات ومدراء
حجم البيانات	أحجام صغيرة أقل من 100 غيغا بايت	أحجام كبيرة تصل إلى مئات التيرا بايت

3•2•1 مقارنة بين لغة الاستعلام المهيكلة (Structural Query Language (SQL)) ومكعب البيانات (Data Cube)

أُنشئت لغة SQL من أجل العمل في أنظمة المبادلات، وليس من أجل العمل في التطبيقات ذات التقارير، وتتصف العمليات التي تقوم بها بالمحدودية. ولا يعود السبب في ذلك إلى ضعف في هذه اللغة بحد ذاتها، وإنما تكمن الأسباب في أن قاعدة البيانات لا تُقدم سوى معلومات بسيطة نسبياً عن ذاتها، وهذه المعلومات تتضمن هيكلية وكيفية تخزين هذه البيانات، بالإضافة إلى معنى هذه البيانات، وذلك ضمن ما يسمى بالبيانات المفسِّرة (metadata). وتتضمن قواعد البيانات العلاقاتية القليل من البيانات المفسِّرة، مما يتطلب من المستعلم أن يقدم الكثير من المعلومات عن قاعدة البيانات أثناء كتابة الاستعلام، مثل تعليمات ربط الجدول، وشروط الانتقاء والتصنيف.

ابتُكر تصميم مكعب البيانات خصيصاً لربط البيانات بوضوح وتعريف صفات وأدوار كل منها قبل بناء الاستعلامات التحليلية عنها، ويحتوي لذلك مكعب البيانات – الذي يُستخدم في المعالجة التحليلية الآنية – على الكثير من البيانات المفسِّرة. فعندما يُعرَّف المكعب لا يتم فقط تعريف ماهية المقاييس، وإنما تتحدد خصائصها وتفاصيل ظهورها أثناء استعراض المكعب، ويجري تحديد كيفية تكامل هذه البيانات وارتباطها معاً، ويتم التمييز بين جداول الحقيقة وجداول الأبعاد، وتُعرَّف أيضاً البنى الهرمية للخاصيات في الأبعاد.. ويساعد كل ذلك على تبسيط عمل طالب الاستعلام المتعدد. وبهذا يتبين أن الوصول إلى بنية مكعب البيانات هو أسهل بكثير من الوصول إلى قواعد البيانات العلاقاتية[1].

4•2•1 مقارنة بين التنقيب في البيانات (Data Mining) والمعالجة التحليلية الآنية للبيانات (OLAP)

تُعتبر المعالجة التحليلية الآنية أداة تجميع وتلخيص تساعد في تبسيط تحليل البيانات، بينما يمكِّن التنقيب في البيانات من الاكتشاف الآلي للأنماط الضمنية، وللمعرفة المفيدة المخبأة في الكميات الضخمة من البيانات.

وتأتي عادةً عملية المعالجة التحليلية الآنية للبيانات (OLAP) كخطوة مصاحبة أو سابقة للتنقيب في البيانات[2]، حيث أنها تستطيع أن تشتق معلومات تلخيصية من عدة مجموعات جزئية من مستودعات البيانات، وبالتالي يُمكِن أن تُستخدم هذه المعلومات لاحقاً في عمليات التنقيب في البيانات.

كما تقدم أنظمة المعالجة التحليلية الآنية توصيفاً عاماً للبيانات، وهي تهم غالباً المستخدمين الذين يحتاجون إلى تلخيص عن البيانات أو مقارنات بسيطة بين البيانات. بينما تقدم أنظمة التنقيب في البيانات معرفة أوسع بالبيانات من خلال كشف الارتباطات الكامنة فيها، وتجزئتها إلى مجموعات تصنيفية، واستخراج قواعد الاقتران، وتحليل السلاسل الزمنية وعمل التنبؤات المستقبلية.

---

[1] Michael, J. A. & Berry. (2004), **Data Mining Techniques For Marketing, Sales, and Customer Relationship Management**, P.498

[2] Han, J. & KAmber, M. (2001), **Data Mining Concepts and Techniques**, P.94

وتختلف أيضاً المعالجة التحليلية الآنية عن التنقيب في البيانات في أن الأولى تُستخدم في عمل توصيفات عامة عن بيانات تتضمن مقاييس رقمية ولها بنية مستودع بيانات، بينما يطبَّق التنقيب في البيانات على أنماط أخرى من البيانات كالبيانات النصية وصفحات الانترنت والتنقلات عبر الانترنت، والوسائط المتعددة.. وبهذا تكون التطبيقات العملية للتنقيب في البيانات[1] أوسع وأهم بكثير من تلك التي تخص المعالجة التحليلية الآنية.

5•2•1 مقارنة التنقيب في البيانات (Data Mining) والإحصاء (Statistics)

يتشابه علم الإحصاء مع التنقيب في البيانات في المنهجية، حيث أن كلاهما يهتمان بجمع البيانات وتحليلها، ويقال بأنه لا تكاد تخلو مرحلة من مراحل اكتشاف المعرفة من استخدام تقنيات الإحصاء، حيث تُستخدم التقنيات الإحصائية في التعرف على البيانات وفي المساعدة في تنظيف وإزالة التكرارات ورسم الانتشار..الخ في المراحل الأولى. كما تُستخدم التقنيات الإحصائية داخلياً في البرامج الحاسوبية التي تُعنى بالتنقيب وذلك لسحب مجموعات بيانات الاختبار التي ستُستخدم في عمليات التنقيب. ويتم أيضاً الاستعانة ببعض التقنيات الإحصائية مثل التحليل العنقودي، ونظرية بايز، وتحليل الانحدار وتحليل السلاسل الزمنية في بناء الخوارزميات التي تُستخدم في البحث عن أنماط البيانات.

وعلى الرغم من الاعتماد الكبير للتنقيب على الإحصاء، فإن المعرفة المكتشفة التي تستنبطها عمليات التنقيب قد تشكل نظريات جديدة تُبنى عليها دراسات إحصائية جديدة.

ويكمن الاختلاف الجوهري بين الإحصاء والتنقيب في البيانات في أنه يتم في الدراسات الإحصائية تصميم النموذج من عينة بيانات ممثلة لمجموعة من البيانات، وتُسحب هذه العينة بناءً على فرضيات مسبقة توضع في بداية الدراسة عن النتائج المتوقعة، ومن ثم يتم التأكد من أن عينة البيانات تحقق هذه الفرضيات أم لا. أما في التنقيب في البيانات فإن البيانات تكون موجودة مسبقاً –لأغراض أخرى– ومتراكمة في الأجهزة الحاسوبية عبر الزمن، ومن ثم يتم تطبيق التنقيب في البيانات عليها بهدف اكتشاف نماذج منها دون وجود فرضيات مسبقة، ولكن

[1] انظر الفصل الثاني من هذه الدراسة

يوجد مسائل أو قضايا يتم اقتراحها في بداية عملية اكتشاف المعرفة بهدف توجيه عمليات التنقيب. ونبين في الجدول (1–2) أبرز الفروقات بين الإحصاء و التنقيب في البيانات.

الجدول (1– 2): أبرز الفروقات بين التنقيب في البيانات والإحصاء التقليدي

	الإحصاء	التنقيب في البيانات
بنية المشكلة	بنية مرتبة وواضحة	بنية غير مرتبة أو مرتبة جزئياً
دور الاستدلال	يلعب الاستدلال دوراً واضحاً في أي تحليل إحصائي	لا يوجد استدلال واضح
هدف التحليل وجمع البيانات	يحدد الهدف ثم تجمع البيانات	غالباً ما تكون البيانات موجودة ثم يحدد الهدف
حجم مجموعة البيانات	مجموعة البيانات صغيرة ومتجانسة وتمثل حجم العينة	مجموعة البيانات كبيرة وغير متجانسة لأنها قد تكون مستمدة من عدة قواعد بيانات غير متجانسة
المنهجية	منهجية استنتاجية واستدلالية (Deductive)	منهجية استقرائية (Inductive)
نمط التحليل	تأكيدي لأنه يستند إلى الفرضيات	استكشافي لأنه يعتمد على اكتشاف المعرفة في ضوء مسألة التنقيب المقترحة
عدد المتغيرات	قليل ويُشتق من الاستبيان الذي يصمم لدراسة فرضيات البحث	كبير وضخم ومستمد من الخاصيات المتوفرة في قواعد البيانات
المخرجات	معروفة مسبقاً ويتم البحث عنها أو برهانها	جديدة وكامنة في البيانات

وبهذا يمكن القول بأن الإحصاء والتنقيب متداخلين ومتكاملين، ولكل منهما المجالات التي يختص ويبدع بها ولا يمكن الاستغناء عنها في العصر الحالي.

قدمت المفاهيم والمقارنات السابقة الخاصة بحقل اكتشاف المعرفة صورة أولية عن مجال اهتمام الدراسة الحالية، وتُشكل المعلومات المعروضة في هذا الفصل الأرضيةَ التي تستند إليها وتنطلق منها الفصول القادمة.

الفصل الثاني

أبرز مجالات تطبيق تقانة اكتشاف المعرفة من قواعد البيانات

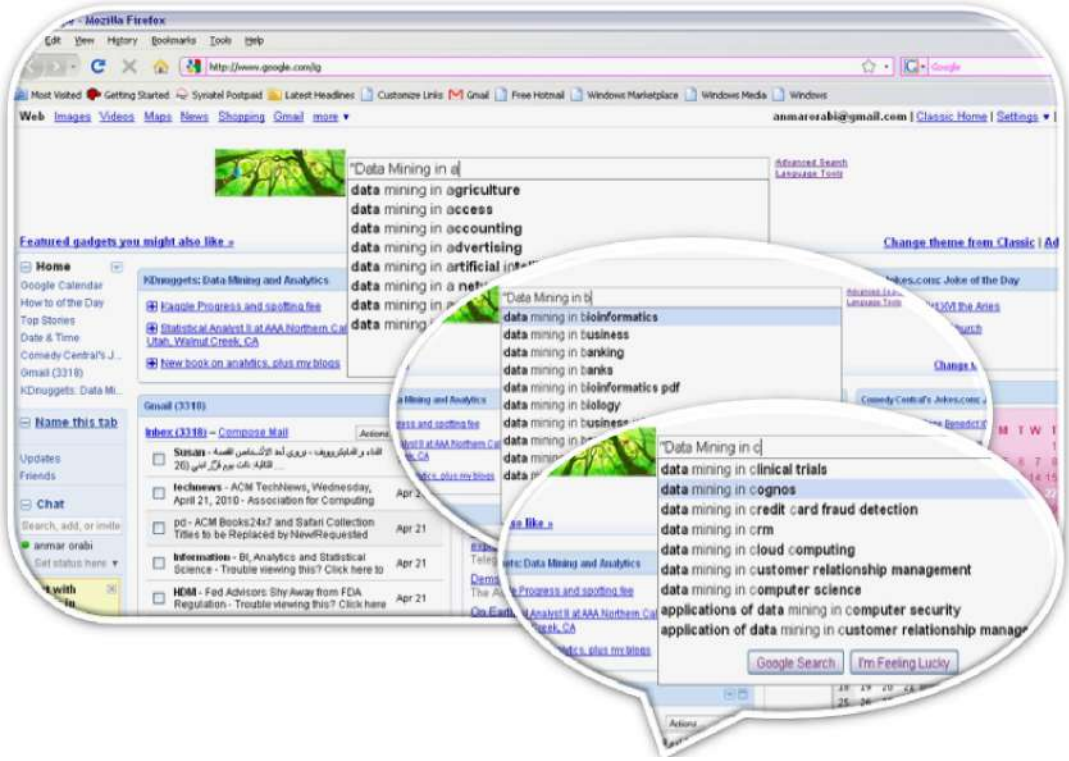
الفصل الثاني

أبرز مجالات تطبيق تقانة اكتشاف المعرفة من قواعد البيانات

مقدمة

أصبحت تقانة اكتشاف المعرفة من قواعد البيانات كالعصا السحرية، فمن يمتلكها يمتلك المعرفة التي تساعده في بناء القرارات المناسبة في الوقت المناسب. وتملك كل عملية اكتشاف معرفة أهدافاً ومهاماً تنقيبية مختلفة مرتبطة بالمجال الذي تُطبَّق فيه هذه التقانة، وبالمخرجات النهائية المطلوبة من هذه العملية. وقد تعرفنا في الفصل الأول على مهام خوارزميات التنقيب في البيانات والتي يمكن من خلالها التعرف على الحقائق والخفايا الكامنة في مختلف أشكال البيانات ومختلف مجالات الحياة.. وسنتحدث في هذا الفصل عن مجالات تطبيق "تقانة اكتشاف المعرفة من قواعد البيانات" من ناحية التنقيب في البيانات لأنها هي المرحلة الأساسية التي تتحدد فيها الأدوات المناسبة للمجال الذي تطبق فيه تقانة اكتشاف المعرفة.

حققت تقانة التنقيب في البيانات عالمياً نجاحات كبيرة منذ منتصف التسعينات تقريباً، وقد بات لا يخلو مجال من مجالات الحياة من اللمسات السحرية لهذه الأداة التي تتطور باستمرار لتصبح أكثر قدرة على استيعاب أنواع متعددة من البيانات كالأرقام، والرموز، والنصوص، والخرائط، والصور، والأصوات، والفيديو...الخ والدليل على تشعب تطبيقات هذه التقانة أنه عندما تتصفح الانترنت وتبحث ضمن محرك البحث عن تطبيقات التنقيب في البيانات وتكتب عبارة "Data Mining in" فإنك ما إن تتابع وتكتب أي حرف بعد هذه العبارة (A, B, C...) حتى يظهر العديد من مجالات تطبيق هذه التقانة، انظر الشكل (2-1).



الشكل (2- 1): مجالات تطبيق تقانة التنقيب في البيانات

وتُصنَّف المجالات التي يطبق فيها التنقيب في البيانات ضمن مجموعتين رئيستين، تتعلق المجموعة الأولى بمجال الأعمال، وهو المجال الذي غالباً ما يرتبط بالشركات والمنظمات الربحية، بينما تتعلق المجموعة الثانية بالمجالات البحثية والتي تدعم الاكتشافات العلمية، كما أن هنالك مجموعة ثالثة تتضمن المجالات التي لا يمكن تصنيفها ضمن المجموعتين السابقتين. وقد يصعب أحياناً تمييز انتماء مجال ما إلى إحدى هذه المجموعات فمثلاً يدور التنقيب في البيانات الطبية حول محورين منفصلين ومتداخلين بآن معاً، الأول يخدم صحة الإنسان وهدفه المباشر بحثي علمي، أما الثاني فيخدم عمليات الإدارة وهدفه المباشر مادي.

كما وتتشابه طرق وأساليب التنقيب في مجالات تطبيقه مع بعضها البعض مثل تلك المطبقة على مواقع التجارة الالكترونية سواء السلعية (سيارات) أو الخدمية (سياحة)، كما تتشابه طرق التنقيب في الصور سواء كانت طبية أو فضائية أو غيرها.. وهذا ما قد يولِّد بعض التكرار في جزء من الأفكار التي سنأتي على شرحها في الفقرات الخاصة بمجالات تطبيق تقانة التنقيب في البيانات، ولكن لكل فكرة خصوصيتها.

1•2 مجالات الأعمال (Business Fields)

يتوجه التنقيب في البيانات في مجال الأعمال نحو مساعدة أصحاب القرار على اكتشاف أهم المعلومات في قواعد بياناتهم، والاستفادة منها في تطوير أساليب أعمالهم المستقبلية. وقد ازدادت الحاجة في العصر الحالي لدى المنظمات الكبيرة نحو استخدام هذه التقانة بسبب اشتداد المنافسة، واتساع الأعمال، ورغبة المدراء في الحصول على معلومات تلخيصية سريعة وجيدة عن كافة فروع المنظمة الكبيرة التي يديرونها. كما أصبح من المرغوب به الحصول على أساليب جديدة في جذب الزبائن، وتقديم العروض المناسبة لهم، ونيل رضاهم، والحفاظ على ولائهم وانتمائهم المستمر للمنظمات، وذلك بما يحقق الهدف الأسمى لدى هذه المنظمات وهو الربح الذكي والمستمر. ونبين فيما يلي بعضاً من مجالات تطبيق تقانة التنقيب في عالم الأعمال.

1•1•2 الاتصالات (Telecommunications)

ينقسم التنقيب في الاتصالات إلى ثلاثة أقسام هي التنقيب في بيانات المكالمات، والتنقيب في بيانات الزبائن الشخصية، والتنقيب في بيانات عمل شبكة الاتصالات.

أ- التنقيب في بيانات المكالمات

يستند التنقيب في بيانات المكالمات إلى الكم الهائل من البيانات المتوفرة في قطاع الاتصالات والمتعلقة بتفاصيل (خاصيات) المكالمات، حيث يتم تجميع هذه البيانات بغرض التنقيب بحسب الرقم المتصِل على النحو التالي: (تجميع بحسب أوقات الاتصال، وسطي مدة المكالمة، وسطي تكلفة المكالمة، وسطي عدد المكالمات بالنسبة للزمن، نسب الاتصالات بدون رد من المتصَل به....). ويهدف التنقيب في بيانات المكالمات إلى بناء نماذج عن تصنيفات أرقام المتصلين وفقاً لنوع خطوطهم وكيفية استخدامهم لخدمة الاتصال ولقيم فواتيرهم.. وبالتالي يمكن استخدام المعرفة المكتشفة مثلاً في تطوير أنظمة الفوترة الحالية، وفي التعرف على مدى تأثير أنظمة الفوترة المقترحة على الأرباح المتوقعة من كل صنف من أصناف أرقام الاتصال.. وهذا ما يساعد في اتخاذ قرارات أجود وفي تحقيق فوائد تنافسية لشركة الاتصالات في تقديم خدمات مناسبة أكثر لزبائنها، وفي الوقت نفسه ضمان تحقيقها لكم أكبر من الأرباح.

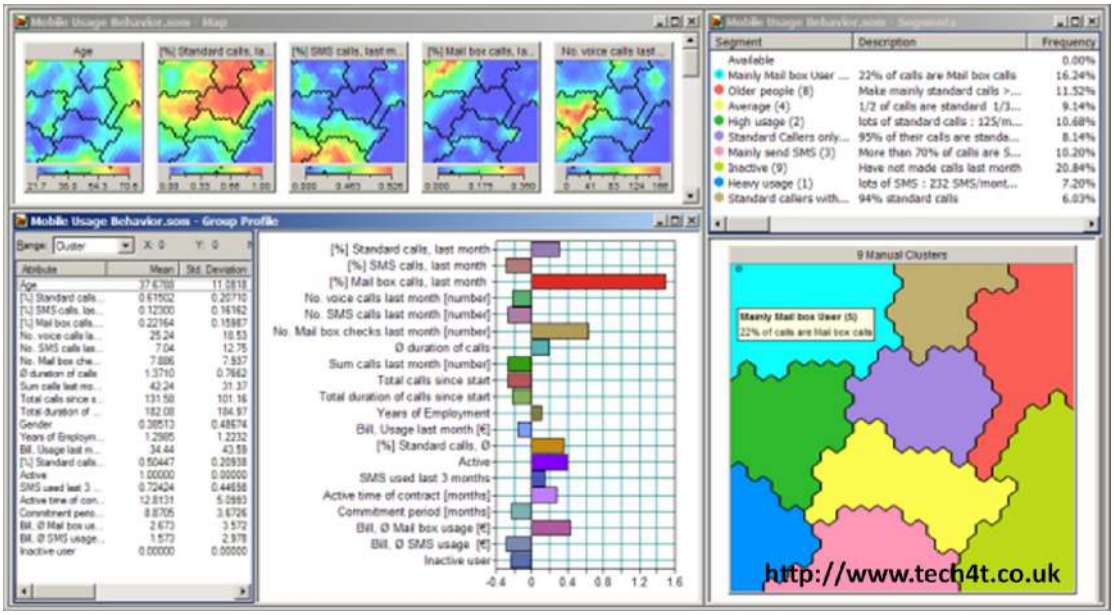
ب- التنقيب في البيانات الشخصية

يُعتبر قطاع الاتصالات من أكثر قطاعات العمل تجميعاً للبيانات الشخصية للزبائن كالعمر، والجنس، والمهنة، ومكان الإقامة..الخ، وتُستخدم هذه البيانات في عمليات التنقيب بالترافق مع بيانات المكالمات، حيث يتم الكشف عن أنماط الزبائن وفقاً لبياناتهم الشخصية وسلوكهم الاتصالي وكيف ومتى وأين يسدِّدون الفواتير. وتَستخدم شركة الاتصالات هذه المعرفة المكتشفة في عدة نواحٍ مثل تقديم عروض خدمية متناسبة مع أصناف المستخدمين المكتشفة، وتوجيه إعلانات غير عشوائية بل متناسبة مع كل صنف، وأيضاً تحديد سقف شهري للفواتير لبعض أصناف الزبائن المتوقع أن يتأخروا بتسديد ذممهم أو ألا يسددوها.

ويُعتبر من أهم تطبيقات التنقيب في بيانات المكالمات والبيانات الشخصية في حقل الاتصالات هو التقاط المحتالين الذين يقومون مثلاً بإجراء اتصالات دولية بكميات كبيرة، ومن ثم يمتنعوا عن تسديد فواتيرهم، حيث تقوم شركة الاتصالات باكتشاف أنماط هؤلاء الزبائن من البيانات التاريخية الموجودة لديها، وبالتالي تقوم بإضافة تعليمات برمجية تفحص الاتصالات والمتصلين. وعندما تتوقع أن يكون متصل ما هو من فئة المحتالين، تعطي تعليمات آلية مباشرة لإطلاق إنذار بتوقع حدوث احتيال، وبالتالي يمكن أن يتم قطع هذا الاتصال أو قطع الخط في حال كانت معطيات سلوك الاحتيال كثيرة. كما تساعد آليات التنقيب في البيانات في التعرف على الأرقام التي يتكرر الاتصال بها في حالات الاحتيال، وبالتالي تكون هذه طريقة ذكية في التنبؤ بالمحتالين المستقبلين بمجرد اتصالهم بأحد هذه الأرقام[1]. وإن اتخاذ القرار بوجود احتيال واتخاذ التدابير المناسبة هو ليس بالأمر البسيط طالما أن الأمر هو احتمالي، وقد ينطبق أيضاً على الأبرياء الذين قد يتضررون خطأً، مما يتسبب في استيائهم من الشركة. وتبعاً لذلك يتم دائماً تشديد وتعقيد تصميم نظام كشف الاحتيال لكي يكون دقيقاً بما فيه الكفاية في تقديراته، كما ويتم دائماً مقارنة العوائد مع الخسائر المترتبة على أي إجراء ستتخذه الشركة بحق زبائنها. وهنا نشير إلى أن شركات الاتصالات التي لا تستخدم تقانة التنقيب في البيانات تكون أكثر قسوة على زبائنها من الشركات التي تستخدم التنقيب، وذلك لأنها تعامل زبائنها جميعاً معاملة موحدة. فعلى سبيل المثال قد توجه الشركة إنذاراً بقطع الخط خلال 24

[1] Santangelo, J. & Rogers, P. & Others. (2007), **"Using Data Mining Tools to Discover Novel Clinical Laboratory Test Batteries"**

ساعة لأحد المشتركين ما لم يسرع بدفع فاتورته التي بلغت في أحد الأشهر ضـــعف المبلـــغ الوسطي لاستهلاكه الشهري المقدر لدى الشركة عن الفترة السابقة، في حين أن هذا المشترك لا يتأخر عادةً في التسديد ويمتلك أيضاً صفات ينتفي معها انتماءه إلى فئـــة المتخلفـــين عـــن الدفع، وهذا ما يتسبب في الإساءة إلى هذا الشخص دون مبرر منطقي لذلك.



الشكل (2– 2): نتائج تحليل سلوك استعمال الهاتف الخلوي باستخدام التجزئة العنقودية والتصنيف

ج– التنقيب في بيانات عمل شبكة الاتصالات

تُستخدم تقانة التنقيب في بيانات عمل شبكة الاتصالات في توقع مشاكل خطوط هذه الشـــبكة، ويتم بالتالي استثمار المعرفة المكتشفة هنا في تصميم أنظمة اختبار ذكية لعمل الشـــبكة تتنبـــأ بالمشاكل قبل حدوثها.

تمتلك شبكة الاتصال العادية نظام إنذار للإعلام عن أية أخطاء أو مشاكل تتعرض لها الخطوط أو الشبكة، حيث يرسل هذا النظام رسائل أو تقارير عن حالة الشبكة في حال تعرضها لأي خلل. ومن مساوئ هذا النظام أنه لا يساعد في التنبؤ آلياً بالمشاكل، وبالتالي لا يساعد في اتخاذ القرار المناسب في الوقت المناسب. ومن هنا جاءت فكرة دراسة بيانات الإنذارات والتقارير، وتطبيق تقانة التنقيب في البيانات عليها، حيث تقوم تقانة التنقيب بتوصيف أنماط الإنذارات التي تحصل بحسب أنواع الأخطاء والمشاكل، ومن ثم يقوم محلل إنذار تعاقبي بالعمل على تتبع تسلسل وتكرار الإنذارات والمشاكل، ويتمكن بالتالي من

استخراج قواعد عنها، لاستخدامها فيما بعد في التنبؤ بحدوث المشاكل الكبيرة قبل وقوعها وبالتالي اتخاذ التدابير المناسبة في الوقت المناسب[1]. فمثلاً إذا توصَّل محلل الإنذار التعاقبي إلى القاعدة التالية:

{إذا كان الإنذار من نوع إنذار اتصال ثم حدث انقطاع اتصال خلال 5 ثوانٍ، فإنه باحتمال قدره 70% سيحصل إنذار من نمط معدل فشل عالٍ خلال 60 ثانية}[2]

فإن هذه القاعدة تُبرمَج آلياً لتعمل في المستقبل على إطلاق منبه يعمل بمجرد حدوث التعاقب المعروض في الشق الأول من الشرط، ويحذر من احتمال حدوث فشل عالٍ خلال 60 ثانية، وهذا ما قد يساعد أصحاب القرار على إصلاح الخلل، أو تطبيق حل بديل قبل أن تصل الشبكة إلى مرحلة الفشل.

2•1•2 التنقيب في استخدام شبكة الانترنت (Web Usage Mining (WUM))

تعمل بعض مواقع الانترنت على التنقيب في بيانات عمليات الولوج إلى صفحات الانترنت، بهدف استنباط القواعد الكامنة المختبأة فيها، حيث تحمل عمليات الولوج هذه بين طياتها أساليب تفكير ورغبات كل من متصفحي هذه المواقع (تطبيقاً لمبدأ قل لي ماذا تتصفح أقل لك من أنت). ويساعد تتبع تنقلات المتصفحين عبر الارتباطات (Links) إلى المواقع الأخرى[3] في اكتشاف أنماط هؤلاء، وفي التعرف على مدى قدرة وكفاءة هذا الموقع على تلبية رغبات كل نمط من أنماط المتصفحين. وتُستخدم المعرفة المكتشفة من هذا التنقيب في تطوير تصميم مواقع الويب بما يحسِّن تصفح الانترنت، ويقلص من زيارة الواجهات الخاطئة، ويحقق رضى الزبائن في تلبية رغباتهم بسهولة[4]. كما وتستخدم المعرفة المكتشفة في تصميم مواقع ويب تكيفية تتغير من مستخدم لآخر بحسب النمط الذي يتم تصنيفه ضمنه مما يساعد في تحسين ملاحة (تصفح) المستخدمين من خلال الإحضار المسبق والتقاط ما يُتوقع أن يكون مرغوباً به من كل صنف من أصناف الزبائن. ويكثر تطبيق تقانة التنقيب أثناء تطوير استخدام مواقع

[1] Fayyad, U.; Piatetsky, G. & Smyth, P. (1996), **"From data mining to knowledge discovery in databases"**

[2] Santangelo, J. & Rogers, P. & Others (2007), **" Using Data Mining Tools to Discover Novel Clinical Laboratory Test Batteries "**

[3] Wang, H.; Wang, Ch. & Zeng, Z. (2006), **Design and Implementation of a Web Usage Mining Model Based On Fpgrowth and Prefixspan**, P.72

[4] The previous reference, P.84

التجارة الالكترونية وفي وضع إعلانات في الملفات الأكثر زيارة، أو وضع إعلانات متناسبة مع أنماط الزائرين. وتعتمد الطرق النموذجية في تنقيب استخدام شبكة الانترنت على أساس قواعد الاقتران (association rule) وتنقيب الأنماط التتابعية (sequential pattern mining) باستخدام بنية الشجرة وسلاسل ماركوف لتمثيل أنماط الملاحة (التصفح)، كما يتم استخدام نوعين من الخوارزميات الخاصة لرفع كفاءة التنقيب عن قواعد الارتباط وعن الأنماط التسلسلية وهما (FPgrowth and PrefixSpan)[1].

3•1•2 التنقيب في بيانات الأسواق المركزية

اختارت بعض إدارات الأسواق المركزية (Supermarkets) الكبيرة أن تقتحم عالم المعرفة الكامن في بياناتها و أن تستفيد من تقانة اكتشاف المعرفة في تحسين مبيعاتها وطرق عرضها وترويجها. وقد وَجدت هذه الأسواق أن بياناتها الهائلة التي تمتلكها تحتاج إلى الربط مع الأشخاص الذين قاموا بعمليات الشراء المتتابعة عبر الزمن. وللحصول على البيانات الشخصية للزبائن يقوم السوق المركزي بالاقتراح على زبائنه الدائمين أن يستفيدوا من الحسومات التي تقدم لحاملي بطاقة الولاء أو الحسم الخاصة به (Loyalty Card) والتي يحصل عليها الزبون مقابل تقديم بعض بياناته الخاصة (مكان الإقامة، الحالة الاجتماعية، عدد الأولاد، مستوى الدخل، مستوى التعليم، المهنة..) التي يحتاجها السوق



المركزي في التعرف على زبائنه، والتي تبقى سرية لدى السوق المركزي لاستخدامها في عمليات التنقيب فقط. وهكذا تضاف بيانات تراكمية باستمرار عن الزبائن الذين يتابعون مشترياتهم عن طريق هذه البطاقة، وتتزايد أيضاً معرفة السوق المركزي بهؤلاء الزبائن عند كل عملية شراء منه.

وتطبق عمليات التنقيب في بيانات الأسواق المركزية بعدة طرق أهمها تحليل سلة التسوق (Market Basket Analysis)، والتي تعرف بأنها استخدام تقانة التنقيب في البيانات للبحث عن قواعد الاقتران أو الترافق (association rules) التي تُفسِّر عادات شراء الزبائن من خلال

[1] The previous reference , P.73

إيجاد الترافقات بين المواد التي يضيفها الزبون إلى عربة تسوقه أو إلى بطاقة مشترياته[1]. ويعبر عن قاعدة الترافق كما سبق وورد في الفصل الأول في الفقرة **(1•1•2•2– د)**.

وتبرز أهمية طريقة تحليل سلة التسوق لأنها تقوم بتحليل جميع السلع المباعة خلال فترة زمنية معينة، وبأنها غالباً ما تكشف عن علاقات ترافقية غير بديهية بين السلع المباعة معاً، وهذا ما يميزها عن الطرق التقليدية في التسويق، والتي تعتمد في ترويج المنتجات على العلاقات البديهية المحددة مسبقاً بين السلع المترافق بيعها معاً مثل (قلم الرصاص والممحاة). ويستفاد من تحليل سلة التسوق في اختيار مواد الحزم التسويقية بالاعتماد على مجموعات المواد المستنبطة من أنماط شراء الزبائن[2]، مما يؤدي إلى التقليص من الإنفاق الترويجي غير الهادف، وإلى الرفع بالوقت نفسه من فائدة الحملات الترويجية[3]. وتطبيقاً على ذلك نفترض مثلاً أنه تم التوصل من خلال تحليل سلة التسوق إلى الاستنتاجات التالية:

80% من الزبائن الذين يشترون المشروب الغازي ماركة "BB" يشترون رقائق البطاطا بالخل ماركة "CC"

و75% من الذين يشترون المشروب الغازي ماركة "DD و" يشترون رقائق البطاطا المملحة ماركة "EE"

وهنا تستفيد إدارة السوق المركزي من هذه الاستنتاجات، وتبيعها إلى موردي المشروبات الغازية أو رقائق البطاطا لكي تتوافق عروضهم الترويجية مع أذواق الزبائن. كما تستطيع إدارة السوق المركزي أيضاً أن تحقق بعض المكاسب المادية، وذلك من خلال مساعدة المنتج في التعرف والوصول إلى الزبائن المناسبين بطريقة غير عشوائية لمنحهم قسائم شراء تشجيعية لمنتجاته بقيم معينة. وهنا أيضاً تعمل القسائم الموزعة بدورها كمدخلات جديدة إلى قواعد بيانات مناقلات المنتج لتساعده في التعرف على رد فعل الزبائن المستهدفين، وفي التعرف على نقاط الضعف والقوة في حملته الترويجية، وبالتالي بناء نماذج تنبؤية مستقبلية كفوءة[4].

كما يمكن تطبيق آليات التجزئة العنقودية أو التصنيف على بيانات السوق المركزي، بهدف التوصل إلى تقسيم الزبائن إلى مجموعات بحسب الماركات التي يشترونها من السلع المختلفة فمثلاً يشتري زبائن المجموعة الأولى الماركات التالية (ماركة شين1 للشاي، وماركة باء1

[1] http://www.micsymposium.org/mics_2004/Vaughn.pdf
[2] Han,J. & KAmber,M. (2001), **Data Mining Concepts and Techniques**, P.95
[3] The previous reference , P.95
[4] Michael, J. A. Berry**,** (2004)**, Data Mining Techniques For Marketing, Sales, and Customer Relationship Management**, P.16

للبن، وماركة راء1 للرز، وماركة ميم1 لمسحوق الغسيل، وماركة غين1 للمشروبات الغازية، وماركة سين1 لمعجون الأسنان...)، وهكذا بالنسبة لبقية المجموعات. وتحمل كل مجموعة عدداً من الصفات التي تميز منتجاتها وزبائنها عن بقية المجموعات، ويساعد التعرف على هذه المجموعات في الوصول الكفء إلى أي منها. فمثلاً قد يرغب مدير السوق المركزي في تقديم ماركة منتج جديدة إلى زبائنه، وهنا يحاول البحث عن مجموعات الزبائن المناسبة لتقبُّل هذا المنتج بنجاح من حيث (الذوق الاستهلاكي، السعر، حجم زبائن المجموعة، سلوكهم الشرائي..)، وبالتالي يمكن أن يقدم هدايا مجانية تجريبية لمجموعات الزبائن اللاتي تم اقتراحها على أنها ستقتنع بهذا المنتج الجديد.. وبعد ذلك يتم التأكد من نجاح هذا التنبؤ من خلال مراقبة مشتريات أولئك الزبائن المستهدفين وفيما إذا دخل المنتج الجديد فعلاً إلى سلة استهلاكهم بعد انتهاء الحملة الترويجية أم لا؟.

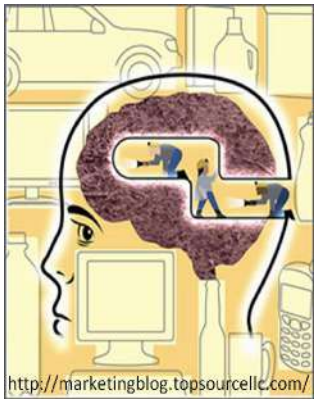


الشكل (2- 3): التنقيب في بيانات الأسواق المركزية

4•1•2 استخدام التنقيب في البيانات في فهم الزبائن وتوقع مجالات اهتمامهم

تقوم بعض المواقع الخدمية عبر الانترنت باستخدام تقانة التنقيب في البيانات في تصنيف بيانات زبائنها السابقين من حيث معلوماتهم الشخصية واهتماماتهم التي تظهر أثناء تصفحهم للموقع. ثم تَستخدم أنظمة تستفيد من هذه التصنيفات في فهم الزائرين الجدد واقتراح أشياء تتناسب مع كل صنف من أصناف الزبائن. فمثلاً عندما يدخل زائر ما إلى موقع موسيقي على

الانترنت –يطبق هذه التقانة – فإنه يطرح عليه [1]عدة أسئلة عن نمط حياته والموسيقى التي يحبها وأوقات سماعها.. وعن بعض صفاته الشخصية (عمره، جنسه، جنسيته، مستوى تعليمه..) وبالتالي يتمكن هذا الموقع من اقتراح مجموعة من المقطوعات الموسيقية التي من المتوقع أن تقع ضمن اهتمامات هذا المتصفح[2]. كما أنه من الممكن أن يقترح عليه أشياء مختلفة للشراء أو للتصفح في مجالات مختلفة تماماً عن الموسيقى، ولكنها تتناسب مع نمط شخصيته التي تم التعرف على ملامحها من خلال الأسئلة التي طرحت عليه.. مثل الألعاب الالكترونية، والأدوات الرياضية وبعض المفروشات وبعض المقاهي والمطاعم وكل ذلك بما يتناسب مع نمط هذا المتصفح. فمثلاً لو توصل هذا الموقع إلى المجموعة التصنيفية التالية:



{يميل الشباب الذكور ممن تتراوح أعمارهم بين 18–25 ويسكنون في المدن الكبيرة ويدرسون في الكليات الهندسية إلى سماع الموسيقى الهادئة، وإلى اقتناء كراسي مريحة أثناء دراستهم وإلى ارتياد المطاعم التي تتصف ببساطة الديكور وفيها عازف ناي }

فإنه يمكن أن يقدِّم عروضاً ترويجية لهؤلاء المتصفحين فقط، كأن يحصلوا مجاناً على مجموعة من 12 قرص موسيقى هادئة لدى شرائهم أحد الكراسي المريحة المعروضة في هذا الموقع.

5•1•2 استخدام التنقيب في البيانات في الإعلانات الموجهة لمستخدمي البريد الالكتروني

تعتبر شركة غوغل (Google) أول موقع الكتروني يستخدم ويعترف باستخدامه لتقانة التنقيب في البيانات في صناعة إعلانات موجهة لمستخدمي البريد الالكتروني لديه[3]، حيث يعتمد هذا الموقع على أسلوب ذكي جداً فهو يقوم بالاحتفاظ ببيانات البحث لمستخدميه لمدة لا تقلُّ عن 18 شهراً[4]، كما أنه يقدم أحجاماً هائلة للبريد الالكتروني، وكان يذكِّر مستخدميه في السنوات الأولى من عمر هذا البريد الالكتروني بعدم حاجتهم لحذف محتويات بريدهم بسبب كبر مساحات التخزين المعطاة لهم، مما يمكِّن هذا المحرك من التنقيب في عبارات البحث، وفي

[1] http://marketingblog.topsourcellc.com/page/2
[2] Fayyad, U.; Piatetsky, G. & Smyth, P. (1996), **"From data mining to knowledge discovery in databases"**
[3] http://www.bnikhaled.com/vb/t3464.html
[4] http://informatics.suite101.com/article.cfm/text_analytics_in_data_mining_software

نصوص البريد الالكتروني لدى مستخدميه، ويتمكن بالتالي من التعرف على العادات الشرائية، والاهتمامات الشخصية، وكل المعلومات التي تساعده في توقع الاحتياجات الفردية لكل مستخدم، وبالتالي يعرض له إعلانات موجهة وغير عشوائية وتتناسب مع اهتماماته وفقاً لفئة المستخدمين التي صُنِّفَ بأنه ينتمي لها، وهي طريقة ذكية في إدارة الإعلانات بحيث تُفيد المستخدم وتُفيد الشركة المعلنة.

6•1•2 استخدام تقانة التنقيب في صناعة السياحة

تَحمل صناعة السياحة بين طياتها فرصاً استثماريةً كبيرةً، ولا يزال العديد من هذه الفرص مجهولاً وغير واضح المعالم. ويعود ذلك إلى ضعف توصيف هذه الصناعة –على المستوى الوطني أو مستوى الشركات السياحية– من حيث مساهمة وأهمية كل من العوامل المباشرة وغير المباشرة المرتبطة بها. وهذا ما يتحقق باستخدام أدوات التنقيب في البيانات، التي تساعد في الكشف عن أنماط السياح، وفي فهم احتياجاتهم، وتلبيتها بسرعة، وكفاءة مناسبين.

وقد جرت محاولات لتطبيق تقانة التنقيب على بيانات السياح في سلسلة فنادق عالمية، حيث اعتمدت على بيانات حقيقية عن السياح، وتم التوصل إلى تصنيف السياح اعتماداً على بياناتهم الشخصية (الجنس، العمر، الجنسية، الحالة الاجتماعية، الوضع المادي، الحالة التعليمية..) وعلى سلوكهم السياحي (مدة الإقامة، الأماكن التي زاروها، الإنفاق داخل الفندق..)، وتمكَّنت بالتالي إدارة هذه الفنادق من معرفة السياح الأكثر ربحيةً، والتعرف على أساليب تلبية رغبات كل نمط من أنماط السياح.

كما استفادت بعض المواقع الالكترونية التي تقدم خدمات سياحية من إمكانية تطبيق التنقيب على بيانات استخدام شبكة الانترنت التي سبق ذكرها، حيث عملت هذه المواقع على الاستفادة من بيانات تنقلات متصفحي الانترنت عبر المواقع التي تقدم خدمات سياحية، وتمكنت من التعرف على العلاقات الارتباطية بين المواقع التي تتناسب مع أنماط المتصفحين، كما تم التعرف على رغبات كل نمط من أنماط السياح المتصفحين.. وجرى بناءً على ذلك تطوير تصميم بعض مواقع الخدمات السياحية بحيث تَظهر وتستجيب بطرق مختلفة متناسبة مع توقعات الموقع عن رغبات كل نمط من أنماط الزبائن. فمثلاً وبصورة مبسطة عندما يتصفح الانترنت شخص من أوربا، ويطلب معلومات عن السياحة في سورية، فالموقع يقدم له مباشرةً

صوراً عن الآثار والمتاحف القديمة، ويقترح عليه الإقامة في إحدى الفنادق الصغيرة ذات الطابع التراثي وفي حارات شعبية، ويقدم له فرصة طباعة خرائط سياحية للمدن السورية، وأيضاً للمطاعم والمأكولات العربية التي يُتوقع أن تتناسب مع ذوقه... بينما إذا تصفح الانترنت شخص من إيران وطلب السياحة في سورية فالموقع مباشرة يُظهر له صوراً عن الأماكن الدينية التي يُتوقع أن تقع ضمن اهتماماته، ويقترح عليه الإقامة أيضاً في فنادق قريبة من الأماكن الدينية المقترحة، ويمكن أن يقترح عليه الاشتراك بمجموعة من البرامج السياحية المناسبة مع انتمائه الديني المتوقع.. وهكذا. وقد يكون هذا المثال بسيطاً وبديهياً ويمكن التوصل إلى نتائجه بدون استخدام هذه التقانة.. ولكنه في الحقيقة يشكل جزءاً صغيراً من الاكتشافات التي تستطيع أن تقدمها تقانة التنقيب في الكم الهائل من البيانات والخاصيات للسياح المتصفحين لشبكة الانترنت من كل أنحاء العالم والراغبين بالسياحة في أي منطقة من العالم.

7•1•2 استخدام التنقيب في البيانات في شركات التأمين

تحتاج شركات التأمين إلى إنشاء نماذج تنبؤية تتناسب مع موضوع التأمين ومع البيانات الشخصية للزبون. وتُستخدم أدوات التنقيب بكفاءة للقيام بهذه المهمة نظراً للعدد الكبير من المتغيرات والبيانات التي يمكن دراستها. فعلى سبيل المثال تفيد هذه التقانة في توقع احتمال تعرض السيارات خلال فترة من الزمن إلى الحوادث، وذلك استناداً إلى مجموعة من المتغيرات المتعلقة بالصفات الشخصية للزبائن وبالموصفات الفنية للسيارة وتاريخ الحوادث ونوعياتها التي تَعرَّض لها الزبائن القدامى للشركة. كما تفيد هذه التقانة في حالات التأمينات الصحية حيث تساعد في اقتراح المبالغ التأمينية المتناسبة مع نتائج التنقيب في بيانات السجلات الصحية، وذلك لكل نمط من أنماط الزبائن.

8•1•2 استخدام التنقيب في البيانات في التعليم العالي

يعتبر التنبؤ بمسار الطلاب والمتخرجين من كبرى التحديات التي تواجه القائمين على إدارة التعليم العالي، حيث ترغب بعض المنشآت التعليمية بالتعرف على الطلاب الذين قد ينتسبون إلى برامج فصلية جزئية، وتعريف أنماط معينة من الطلاب الذين قد ينتقلون لمتابعة تعليمهم في مؤسسة تعليمية أخرى، وتحديد الطلاب الذين يُتوقع أن لا يوفوا بالتزاماتهم المالية، والتنبؤ

بالطلاب الأكثر ربحية[1] (المتوقع أن يتابعوا في الدراسات العليا)، والتعرف أكثر على الميول الدراسية للطلاب من خلال علاماتهم في المواد الدراسية المترابطة، وبالتالي توقع أعداد الطلاب في فروع الاختصاصات المختلفة في السنوات اللاحقة وفي الدراسات العليا، ومن ثم تقدير احتياجاتهم إلى الكوادر التدريسية المناسبة[2].

ويُستفاد من هذه التقانة أيضاً في التعرف على الفجوات التعليمية، وفي معرفة من هم الطلاب الذين قد يحتاجون إلى مساعدة تعليمية في النجاح والتخرج؟، وبالتالي مساعدتهم في وقت مبكر، وحتى قبل أن يعي هؤلاء أنهم في خطر وبحاجة للمساعدة. كما يمكن تطبيق هذه التقانة على إدارة الحياة الطلابية حيث يتم دراسة تأثير نسبة حضور الطالب، وتأثير الصحبة الدراسية، والوضع العائلي، ومهنة الأب.. على الأداء الدراسي[3]. وبذلك تتمكن الإدارة التعليمية المتطورة من تسيير أمور الطلاب بطريقة حكيمة، ومن رفع قدرتها على اتخاذ التدابير المناسبة بوقت مناسب وقبل وقوع المشاكل.

2•2 مجالات الأبحاث العلمية

يمكن أن يُستخدم التنقيب في البيانات في الأوساط العلمية بكثرة، ويستفاد منه في الكيمياء والفيزياء، والطب، وعلوم الأحياء، والفلك، والجيولوجيا والجغرافيا..الخ. وتمتلك أدوات التنقيب القدرة على مساعدة الأساليب التقليدية المستخدمة في كل من تلك العلوم على إنجاز الجوانب التي تصعب عليها، ودون أن تكون بديلة لها، وقد تم اختيار استعراض مجالات تطبيق هذه التقانة في البيانات الطبية، والأبحاث الجينية، ومضمون المؤلفات والكتابات الالكترونية، وفي علم الفلك.

1•2•2 التنقيب في البيانات الطبية

يُستخدم التنقيب في البيانات الطبية في اتجاهين الأول هو دعم المعرفة الطبية، والثاني هو دعم القرارات الإدارية. ويُعتبر الاتجاه الأول أوسع وأهم من الاتجاه الثاني، كما أن دعم القرارات الإدارية قد يَنتُج في كثير من الأحيان من المعارف الطبية المكتشفة. وبناءً على ذلك فإن

[1] Luan, J. (2004), **"Data Mining Applications in Higher Education"**
[2] Bresfelean, A. & Vasile P. (2008), **Data Mining Applications in Higher Education and Academic Intelligence Management**
[3] Ranjan, J. & Ranjan, R. (2010), **"Application Of Data Mining Techniques In Higher Education In India"**

التنقيب في البيانات الطبية هو أقرب إلى مجالات الأبحاث العلمية منه إلى مجالات الأعمـــال، وهذا ما جعل التنقيب في البيانات الطبية يُدرج ضمن مجالات الأبحاث العلمية.



الشكل (2– 4): التنقيب في البيانات الطبية

أسفر تطبيق التجارب الطبية الواسعة وتطبيق تقانة المعلومات الصحية على بيانات المرضى عن جبال من البيانات الطبية الالكترونية، و تعتبر هذه البيانات موارد إستراتيجيةً للمنشآت الطبية إذا عرفت كيف تلخصها وتستفيد منها. وتكمن أهمية البيانات الطبية القديمة في أنها تتضمن توثيقاً للأساليب الطبية التشخيصية والعلاجية التي طبقت في علاج المرضى، وكانت تعتبر في حينها أساليب صحيحة.. ولكن هل هي فعلاً كذلك؟. وما هو الفرق بين الأساليب الطبية المختلفة التي تم تطبيقها على المرضى فيما مضى؟. وأي الأساليب أفضل؟. وهل يختلف نجاح الأساليب باختلاف أنماط المرضى؟.... أسئلة كثيرة من هذا النوع يمكن طرحها.. ولكن كيف للمهتمين بهذه الأمور أن يعرفوا الحقيقة؟ **الحقيقة كامنة في البيانات**.. ونظراً لكبر حجم هذه البيانات ولكثرة المتغيرات فهي على درجة من التعقيد يصعب معها

دراستها وتحليلها بالطرق التقليدية. ومن هنا برزت أهمية استخدام أدوات التنقيب في البيانات لدراسة هذه البيانات والكشف عن أنماط المعرفة الكامنة فيها. ولكن يكمن نجاح التنقيب في البيانات هنا في تعريف صحيح للمسائل الطبية المدروسة، وفي التفسير والاستخدام المناسبين للمعرفة المكتشفة التي تحتاج إلى الخبرة الإنسانية في تفسيرها، وليس إلى القدرة الآلية.

1•1•2•2 المحاور الأساسية للتنقيب في البيانات الطبية

يدور التنقيب في البيانات الطبية حول ثلاثة محاور أساسية كما هو آت:

أ- التشخيص

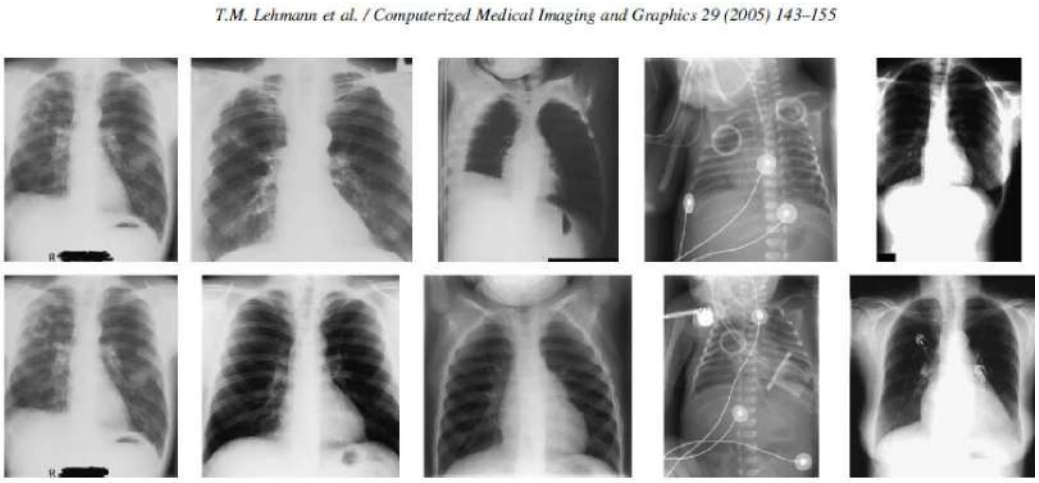
يُستخدم التنقيب في البيانات في تقييم أساليب ووسائل التشخيص المطبقة على المرضى، كما تُستخدم تقانة التنقيب كأداة مساعدة في التشخيص، وذلك من خلال التحليل الآلي للإشارات الطبية (تخطيط القلب، تخطيط الدماغ، التخطيط الكهربائي للعضلة)، ولمكونات الصور الطبية (Images (mammograms, ultrasound, X-ray, CT, and MRI)، وتحويلها إلى رموز قابلة للقراءة الآلية. ومن أمثلة آليات التنقيب الأولية المستخدمة في ذلك هي آلية التصنيف الآلي للصور الشعاعية، والتي تقوم بتحليل مكونات الصورة الجديدة الملتقطة، وتقارنها مع مثيلاتها المخزنة مسبقاً في الأنظمة الحاسوبية المرتبطة بأجهزة التصوير، ومن ثم تقوم بانتقاء الصور المخزنة الأقرب لها، وتستفيد من بياناتها المخزنة معها في قراءة الصورة الجديدة، انظر الشكل (2-5). كما تساعد آلية التصنيف في أرشفة جميع الصور الملتقطة وتصنيفها ضمن مجموعات تحمل عناصرها صفات متشابهة.

ب- العلاج

تساعد آليات التنقيب في البيانات الطبية في اختيار أساليب العلاج الأكفأ والأنسب لأنواع الأمراض ولفئات المرضى، وذلك من خلال تقييم أساليب العلاج المتبعة سابقاً أو في التجارب الطبية، ومدى تأثيرها على المرضى بكل الجوانب بما فيها الجوانب المتعلقة بالتأثيرات الجانبية وبتتابع التعرض لأمراض أخرى، ودراسة الاختلاف بين المرضى في تقبل أساليب العلاج. فمثلاً تم تطبيق التنقيب في البيانات على بيانات لمرضى سرطان الثدي في إحدى الدراسات[1]، وتبين أنه لدى دراسة تأثير عقار طبي معين في علاج هذا المرض أن المرضى

[1] http://informationweek.com/story/showArticle.jhtml?articleID=165702390

الذين كانت نسب الكولسترول مرتفعة لديهم لم يظهروا استجابة جيدة للعلاج بهـذا العقـار... وتعتبر هذه النتيجة اكتشاف لعلاقة بين متغيرات ليس بينها علاقة منطقية، ولم يكن بالإمكـان التعرف عليها واختبارها ما لم تكشف عنها أدوات التنقيب في البيانات. وهذا ما يسـتدعي أن يتم إجراء دراسات متممة لهذه الدراسة ولكن من الجوانب الطبية بهدف التقصي عن الحقـائق الطبية المرتبطة بهذا الموضوع.



الشكل (2- 5): مثال عن استدعاء الصور الشعاعية المخزنة (السطر الثاني) الأقرب إلى الصور الملتقطة (السطر الأول)[1]

ج- التنبؤ

يُستخدم التنقيب في البيانات الطبية للتنبؤ بالمخرجات المستقبلية استناداً إلى التجـارب السـابقة والظروف الحالية، وهنا توجد فوائد كبيرة من دراسة الظواهر الطبية، وتوقع تفشـي بعـض الإصابات أو الأوبئة، وفي الكشف المبكر عن بعض الأمراض كمرض السرطان، وفي وضع تقديرات صحيحة عن خطر إجراء عمل جراحي قلبي لمريض معين، وفي تحليل مدة البقـاء لمرضى معينين كمرضى الايدز..

2•1•2•2 فوائد تطبيق تقانة التنقيب في البيانات الطبية

يوجد بشكل عام عدة جهات تستفيد من تطبيق التنقيب في البيانات في المجال الطبي، ونورد فيما يلي أبرز الفوائد والمستفيدين من هذه التقانة.

[1] Lehmann, Th. , Guld, M & Others (2005), **Automatic categorization of medical images for content-based retrieval and data mining**, P.115

أ- تفيد هذه التقانة في دعم القرار الطبي أي في مساعدة الأطباء على تطوير أساليب التشخيص الطبي، وفي تقييم أساليب العلاج المتبعة، وفي تطوير الأدوية الطبية.

ب- يساعد التنقيبُ في البيانات الأطباءَ السريريين والمدراءَ في إيجاد أنماط (patterns) جديدة قيِّمة وعلاقات ترافقية كامنة في البيانات، حيث يقوم الخبراء الأخصائيون باكتشاف الأنماط وتفسيرها واستخدامها في الحالات التي يُتوقع أن تقود إلى تحسينات في صحة المرضى، وفي الاستخدام الرشيد للموارد المادية والبشرية في المراكز الصحية[1].

ج- تقديم خدمات أفضل للمرضى مبنية على معارف حقيقية.

د- تقوم هذه التقانة بتزويد إدارات المستشفيات بالمعارف التي تحتاجها عند اتخاذ قرارات إدارية أو قرارات العناية الصحية، فمثلاً تحتاج المستشفى إلى تقدير احتياجاتها المستقبلية من الأَسِرَّة والأدوات والمواد والكادر الطبي في كل قسم، وخلال أي وحدة زمن، ويفيد ذلك في ضمان تحقيق تحسينات ملائمة في استخدام موارد المشفى مع ضمان صحة المرضى[2].

ه- تساعد هذه التقانة إدارة المشفى في رقابة وتقييم أداء الأطباء لديها من خلال دراسة مجمل المتغيرات المتعلقة بأساليب عملهم، وتتبع مدى نجاحهم في اختيار أساليب الاستقصاء الصحيحة، وأساليب العلاج المناسبة، مع التقليل من الفحوصات غير المفيدة، والتي تشكل هدراً لموارد المشفى، وترفع من تكاليف العلاج دون فائدة حقيقية، مثل طلب بعض التحاليل الطبية أو الصور الشعاعية دون وجود أسباب منطقية لها. ويزداد الحرص على هذه الأمور في ظل ضغوطات شركات الضمان الصحي على تقليص التكاليف الطبية.

و- تساعد هذه التقانة في تقديم المعرفة لشركات الضمان الصحي لكي تستخدمها في تقييم مدفوعاتها ومقبوضاتها التأمينية من كل نمط من أنماط المرضى، وفي التقاط حالات الاحتيال والإساءة والادعاءات الكاذبة.

ز- تفيد نتائج التنقيب في البيانات في توجيه بعض الأبحاث الطبية وفي دراسة نتائج البعض الآخر.

---

[1] Baylis, P. (2009), **"Better Healthcare with Data Mining"**

[2] The previous reference

لقيت الرعاية الصحية اهتماماً كبيراً من قبل الحكومة الأمريكية الجديدة منذ نيسان 2009[1]، حيث جرى التأكيد على ضرورة إيجاد الأساليب الكفيلة بتخفيض تكلفة الرعاية الصحية وتحسين نوعيتها. وذلك أن هنالك تزايداً كبيراً في تكاليف الرعاية الصحية وأن جزءاً من هذه التكاليف غير حقيقي بسبب الهدر واستخدام أساليب العلاج الباهظة، أو بسبب الاحتيال من قبل أي طرف من أطراف الرعاية الصحية، أو بسبب سوء استخدام التأمين الصحي[2].

وقد اكتشف مؤخراً باحثون[3] من جامعة (University of Louisville) في الولايات المتحدة الأمريكية في أيلول 2010 طرقاً أفضل لتقديم الرعاية الصحية استجابةً إلى مناداة الحكومة الأمريكية، وقد استخدموا البرنامج الحاسوبي SAS لتحليل وتنقيب البيانات الكمية المهيكلة، والمعلومات النصية غير المهيكلة من السجلات الطبية واستحقاقات الرعاية الصحية. وتشكل المعرفة التي تم الحصول عليها من التنقيب في البيانات والنصوص أساساً يمكن الاستناد إليه في المساعدة في مفاوضات الادعاءات بين مقدمي الخدمات الصحية وشركات التأمين الصحي. واتجهت بعض الأبحاث المنجزة في هذا الصدد نحو استخدام التحليل التنبؤي في الكشف عن العلاقات بين المعالجة والنتائج، وقد تم التوصل إلى تعريف عدد من خطط العلاج المثالية التي تُحسِّن بآن معاً نوعية العناية، وتُخفِّض التكاليف، وتضمن رضى واستحسان المستهلكين، فمثلاً أظهرت بعض النتائج أن استخدام بعض الأدوية باهظة الثمن قد يبدو مكلفاً للوهلة الأولى، ولكنه تبين في الحقيقة وعلى المدى البعيد أنه أقل كلفة من العلاجات البديلة.

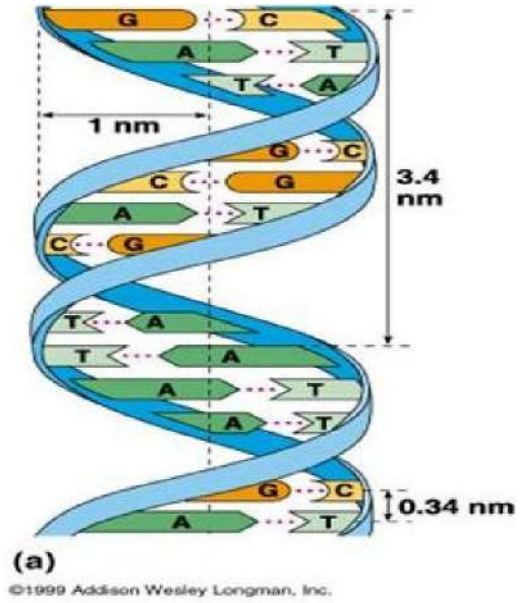
2•2•2 استخدام التنقيب في البيانات في الأبحاث الجينية[4]

تَستخدِم الخلايا الحية الحمضَ النووي (DNA) لتخزين معلوماتها الوراثية، حيث يحتوي الحمض النووي (DNA) على المورثات (الجينات) المسؤولة عن الصفات التي تميز كل إنسان (كائن حي) عن غيره من البشر (الكائنات الحية). ويتكون جزيء الحمض النووي (DNA) من شريطين ملتفين حول بعض بشكل لولبي. ويتألف كل حبل من أربعة أنواع من العناصر وهي (Adenine (A), Thymine (T), Cytosine (C), and Guanine (G)) حيث ترتبط

[1] http://www.america.gov/....
[2] http://www.healthdatamanagement.com/news/health_reform_obama_database_claims_fraud....
[3] http://www.sas.com/offices/europe/norway/nyheter/university-of-louisville-analytics.html
[4] [Maglothin, R. (2002), **"Data Mining in DNA: Using the SUBDUE Knowledge Discovery System to Find Potential Gene Regulatory Sequences"**] & [Shah, Sh. & Kusiak, A. (2004), **"Data mining and genetic algorithm based gene/SNP selection"**]

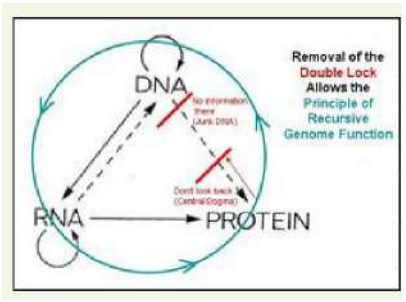
أعداد مختلفة من هذه العناصر وبترتيبات مختلفة مشكِّلةً السلسلة الخطية التي تتحدد وفقها المعلومات الوراثية الخاصة بالخلية والتي تختلف من كائن حي لآخر، وقد استطاع علماء الوراثة أن يفككوا جزيئات الحمض النووي (DNA) للتعرف على هذه السلسة، انظر الشكل (2–6). وبالتالي أصبح بالإمكان أن تخزن هذه السلسة في الحواسيب كسلسلة حرفية كما في النموذج التالي:

TTCAG CCGATATCCTGGTCAGAT TCTCT
AAGTCGGCTATAGGACCAGTCTAAGAGA



الشكل (2– 6): طريقة كتابة الشيفرة الوراثية في الحمض النووي

وعلى اعتبار أن الجينات تتكون من عدد كبير من البيانات، وأن معظم وظائف السلاسل والجينات مجهولة، فقد ظهرت الحاجة إلى استخدام خوارزميات كفوءة لتعريف الأنماط الهامة لأجزاء السلاسل، ولتعريف وتصنيف الجينات المحتملة، واستنتاج العلاقات بين الجينات، وأيضاً توقع تركيب البروتين الناتج عن الأوامر المخبأة في سلسلة الحمض النووي (DNA).



وقد وجد العلماء أن التوصل إلى تلك الأمور يحتاج إلى رفع كفاءة البحث وتوجيهه واعتباره اكتشاف للمعرفة المخبأة في البيانات الجينية، أي أنه يحتاج إلى تطبيق لتقنيات التنقيب في

البيانات للقيام بهذه المهمات الصعبة. فعلى سبيل المثال تقوم بعض مراكز الأبحاث[1] بمحاولة تمييز العلامات الجينية عند مرضى التهاب الكبد الفيروسي B (HBV) المؤهبة لسرطان الكبد عند البعض منهم مستقبلاً، وذلك بمقارنة السلاسل الجينية بين فئتي مرضى التهاب كبد B الذين لديهم سرطان كبد والذين ليس لديهم. حيث تجري المحاولات لإيجاد العلامات الجينية وربطها مع المعلومات السريرية والشخصية للمرضى، وبالتالي توقع حدوث سرطان الكبد عند فئات من هؤلاء المرضى، وتوقع مدى استجابة كل فئة جزئية منهم للعلاج. ولتحقيق ذلك يتم تطبيق عدد من طرق التنقيب في البيانات منها التجزئة العنقودية والتصنيف وأشجار القرار.

وبشكل عام يوجد سعي عالمي بقيادة الولايات المتحدة الأمريكية نحو الكشف عن تعاقب الحمض النووي الكامل للبشر (The complete DNA sequence for humans) ولبعض أنواع الأحياء الأخرى، وذلك لاستخدامها في الكشف عن آليات الأمراض، وتحسين فعالية الأدوية وتتبع النمو والتطور في الكائنات الحية، واكتشاف العلاقات التطورية..الخ. وتعتبر تطبيقات التنقيب في البيانات الجينية على درجة عالية من الأهمية.

3•2•2 التنقيب في مضمون المؤلفات والكتابة الالكترونية

تتزايد أعداد الكتب الرقمية في الوقت الحالي بصورة متسارعة، ويزداد معها الوقت والجهد اللازمين لقراءتها والتعرف عليها ومقارنتها مع غيرها.. ومن هنا نشأت الحاجة إلى إيجاد طريقة لتحويل النصوص الالكترونية الحرة غير المهيكلة (Non Structured) إلى نصوص يمكن تحليلها آلياً مما يُمكِّن من التعامل مع الكم الهائل من الكتابات النصية، واستنباط المادة المعرفية الموجودة بين مفرداتها بأسلوب مبسط. ولتحقيق ذلك تم تطوير آلية التنقيب في النصوص (Text Mining)، وهي عبارة عن مجموعة من معالجات الحوسبة الذكية مع توظيف خوارزميات معالجة اللغة الطبيعية (Natural Language Processing (NLP))[2] بهدف اكتشاف العلاقات القائمة بين المفردات الموجودة في النصوص، وبالتالي توليد عددٍ كافٍ من الأنماط

---

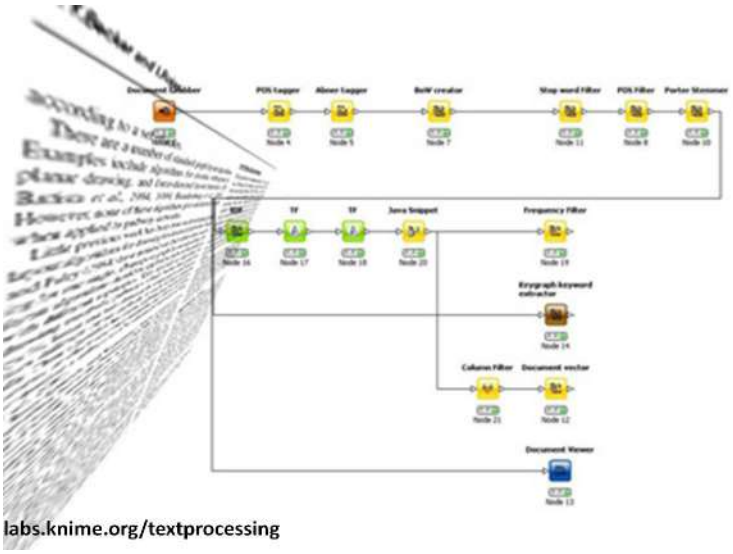
[1] Leung, K. & Eddie, Y.(2004), **Data Mining on DNA Sequences of Hepatitis B Virus by Nonlinear Integrals**

[2] هي إحدى أوجه الذكاء الاصطناعي حيث يقوم الحاسوب بترجمة اللغة البشرية المدخلة من خلال مقابلة الرموز المدخلة مع المعلومات المخزنة في الحاسوب بغية فهم المعنى العام المقصود من النص.

المعرفية السائدة في النص.

وتتحقق هذه العملية من خلال معالجة أولية للنص يتم فيها استخراج الكلمات والمفاهيم الأساسية وتمييز البنى المختلفة للألفاظ (فعل، فاعل، حرف..)، وتجري بعد ذلك محاولة إيجاد العلاقات بين المفاهيم، فتنتج اصطلاحاتٌ ومعانٍ تمكِّن من تمثيل النص في قواعد ربط وتصنيف، ثم تُعرض للمستخدم بطريقة سهلة الفهم وضمن نماذج تصنيفية مناسبة.

وتطبق هذه التقانة في العديد العديد من ميادين الحياة أهمها تلخيص مضمون الرسائل الالكترونية التي تأتي إلى شركة ما، والكشف في المراجع البحثية في موضوع معين عن نَهْجَ المؤلف، والنسق المفاهيمي الذي انتهجه عند معالجة المسألة المطروحة أمامه، وعن مضمون مؤلفاته ومقارنتها بغيرها من المؤلفات.. وهنا نلفت الانتباه إلى أن العرب تمكنوا من استثمار هذه التقانة في قراءة المؤلفات العربية التراثية القديمة[1].



الشكل (2– 7): طريقة فهم النص باستخدام تقانة التنقيب في النصوص

ويعتبر التنقيب في النصوص من مجالات التطبيق الجديدة والمتنامية في التطور والاستخدام، وخاصة مع انتشار الانترنيت، بحيث أصبح من المرغوب فيه إيجاد حلول للوصول الذكي والهادف إلى المعرفة في الكم الهائل من النصوص الموجودة على شبكة الانترنيت، وهو ما تقوم به عادةً محركات البحث النشيطة مثل (Google) الذي يعتبر أقوى محرك بحث على

[1] الرزو، حسن مظفر، (2008)، **تطبيقات التنقيب المعلوماتي على موارد المعرفة الإسلامية**

الانترنت، حيث يقوم باستخدام هذه التقانة في تلبية طلبات البحث لمستخدميه، فمثلاً عندما يطلب المستخدم الاستعلام عن عبارة محددة مكتوبة ضمن علامتي الاقتباس "..." ولا تكون هذه العبارة موجودة تماماً في صفحات الانترنت، فإن محرك البحث يزيل علامتي الاقتباس في بحثه ويحاول استنتاج معنى تقريبي للجملة المطلوبة باستخدام خوارزميات (NLP) وبالتالي جلب جميع العناوين التي يَعتقد أنها قريبة من قصد المستخدم، كما أنه قد أصبح لدى هذا المحرك مثلاً القدرة على تصحيح الأخطاء الطباعية للمستخدم عندما يخطيء (أو لا يصيب) في كتابة الكلمات التي سيستعلم عنها، فيسأله المحرك "Did you mean?"، ويقترح عليه بعض العبارات التي يتوقع أنها تناسب عبارة طلب الاستعلام.

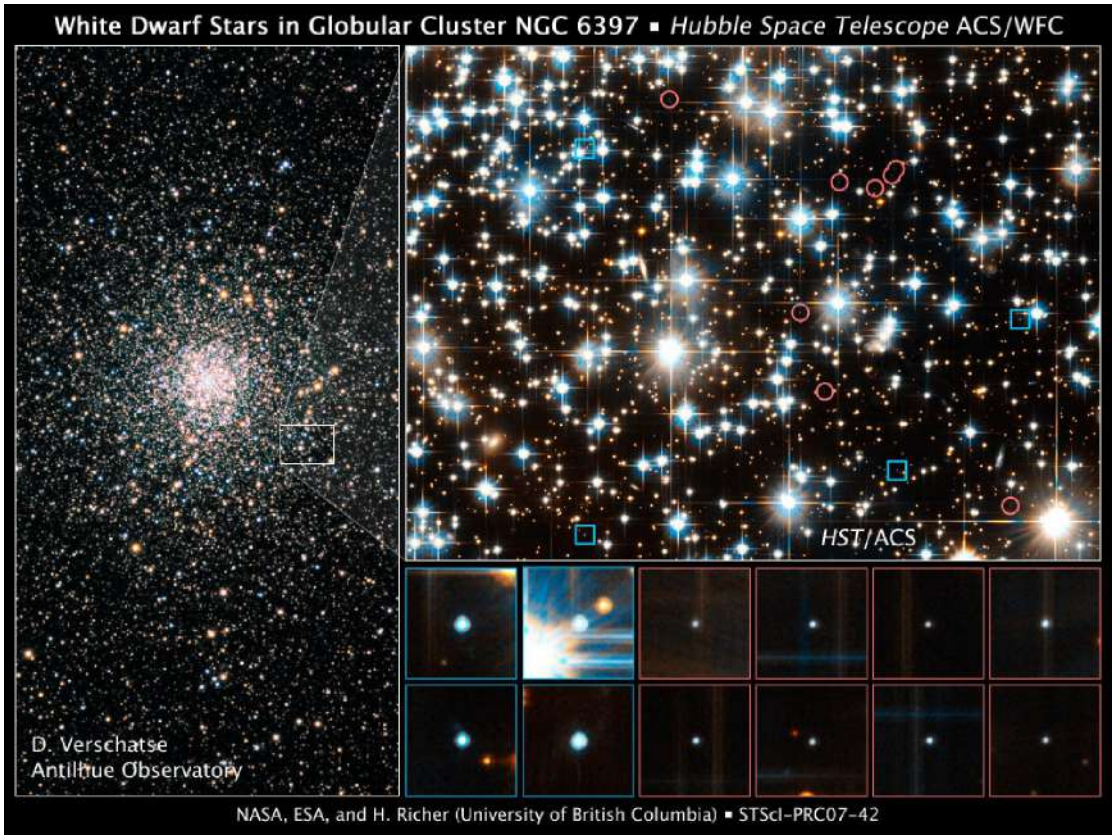
وقد تمكن محرك البحث غوغل حديثاً من تقديم خدمة تحويل لغة مواقع الانترنت التي يجلبها إلى اللغة التي يختارها المستخدم في صفحة غوغل الخاصة به، ولم تعد اللغة العربية عائقاً على الفهم من قبل محرك البحث هذا، فهو يستطيع أن يتعامل معها بمرونة ويفهم بنية الكلمات العربية، ويخفف من الالتباس الذي قد يقع بسبب الهمزات والمذكر والمؤنث والمفرد والجمع..الخ، ويقدم بالتالي نتائج متميزة للاستعلامات التي لا يجد لها مقابل حرفي في مواقع الانترنت.

4•2•2 استخدام التنقيب في البيانات في علم الفلك

تحمل الصور الملتقطة من المسوحات الفضائية بين طياتها كميات ضخمة من البيانات عن العالم الفضائي المحيط بنا –خصوصاً مع تطور أجهزة وطرق التصوير، ومع التزايد المضطرد في أعداد الصور الملتقطة– والتي أصبح تحليلها والتعرف عليها بحاجة إلى طرق شبه آلية تحققها تقانة التنقيب في البيانات مع تقانات أخرى أهمها تقانة تعليم الآلة. وقد تم من خلال هذه التقانات تطوير القدرة الآلية على التمييز واكتشاف الفروق بين المجموعات والأجسام السماوية (مجرات، نجوم، كواكب أطياف، أشعة غاما..) بناءً على الكثير من الأسس، منها قوى جاذبية النجوم للكواكب، وطول الأمواج الضوئية، وعمق الصورة، وشكلها التصميمي..الخ، وتستخدم بالتالي هذه النتائج في فهرسة وتصنيف أجسام السماء، والتعرف على مكونات أية صورة فضائية جديدة، وتقدير أعداد كل نوع من الأجسام السماوية فيها.
كما وتساعد هذه التقانات في التوصل إلى العلاقات بين الظواهر الفضائية، وتوقع الأحداث

المستقبلية بناءً على التحليل السلوكي والتتابعي للظواهر الفضائية المصورة والمتراكمة في الأجهزة الحاسوبية عبر الزمن[1].

ويعتبر علم الفلك واحداً من التطبيقات الأولى القديمة لتكنولوجيا تنقيب البيانات، حيث قدَّمت هذه التقانة العديد من الاكتشافات إلى علم الفلك، منها أنها كشفت عن أن الكون ممتد إلى ما لا نهاية، وأنه يوجد فيه العديد من المجرات، وأن مجرة درب التبانة التي ننتمي إليها ليست الوحيدة كما كان الاعتقاد سائداً من قبل، وأنه يوجد في مجرتنا والمجرات الأخرى بُنى مشابهة لنظامنا الشمسي، وتحتوي على كواكب أكثر مما تتضمنه المجموعة الشمسية بعشرات أو مئات المرات، وأن فيها كواكب أكبر بأربعة أضعاف من المشتري أكبر كواكب المجموعة الشمسية[2].. وبالتالي استطاعت هذه التقانة أن تثبت بأنها أكثر قدرة من الأداء البشري ومن الطرق الحسابية التقليدية على تصنيف حتى الأجسام السماوية الباهتة. ويبين الشكل (2-8) التالي صور عامة للأجرام السماوية مع صور جزئية تفصيلية.



الشكل (2- 8): إلى اليسار صورة فضائية ملتقطة، وإلى اليمين التعرف على مكونات جزء من هذه الصورة[3]

[1] http://www.slideshare.net/dataminingcontent/data-mining-the-sky-3191852
[2] Ciriani, T.; Fasano, G. & Haydock, M. (2003), **Operations Research in Space and Air**, P.143-145
[3] http://www.scientificblogging.com/news releases/ngc_6397_shows_the_natal_kick_of_white_dwarf_stars

3•2 مجالات أخرى للتنقيب في البيانات

تتشعب المجالات التي تُطبَّق فيها تقانة التنقيب في البيانات، وقد تم العثور على مجالات مختلفة لا يمكن تصنيفها ضمن مجموعة محددة. وسنتطرق إلى مجالين من الاستخدامات الجديدة وهما توقع جماعات الإرهاب والتنقيب في الصور المتوفرة على شبكة الانترنت.

1•3•2 استخدام التنقيب في مراقبة الأشخاص عبر الانترنت وفي توقع جماعات الإرهاب

تُشكِّلُ شبكة الانترنت العالمية أكبر مصدر للخطر المستقبلي على الأفراد، وذلك لأن أفراد الشعوب في العالم أصبحت تنتمي إلى مجتمع جديد هو المجتمع الالكتروني، وقد أدى توسع هذا المجتمع الالكتروني إلى نشوء أحجام هائلة ومتنوعة جداً من البيانات الالكترونية المتاحة للمعالجات الآلية بشتى أنواعها. وأصبحت بعض الجهات تستخدم تقانة التنقيب في مراقبة تحركات المتصفحين للانترنت من حيث المواقع التي تتم زيارتها، والأشخاص الذين يُتبادل معهم الآراء والرسائل، والمنتديات والمواقع التي ينتمون إليها ومضمون الوثائق التي يتبادلونها..الخ. وتتكون لديها بنتيجة ذلك مجموعاتٍ تصنيفية للأشخاص على الانترنت.

وتعتبر وكالات الاستخبار العالمية من أبرز هذه الجهات حيث تقوم بالتوسع باستخدام أسلوب التنقيب في البيانات في جمع وتحليل المعلومات الاستخبارية، لاستخدامها في ما يسمى مكافحة الإرهاب[1]، حيث تُطبق هذه الوكالات برامج تنقيب في البيانات لتتبع تصفح المواقع ولتحليل المشاعر والأفكار في الوثائق والمستندات غير المنظمة التي ينشرها الأشخاص على الانترنت – أي دمج نتائج التنقيب في استخدام الانترنت مع نتائج التنقيب في النصوص– وتستخلص منها الآراء واتجاهات التفكير وارتباطات الأشخاص الذين يتبادلون هذه المستندات، وتتمكن بالتالي هذه الوكالات من استخلاص وتكوين ما يسمى صورة نمطية عن الأشخاص على الانترنت، ويتم على أساسها وضع صفات معينة للشخصيات المتوقع أن تكون إرهابية، وتَستخدم الوكالات هذه المعلومات في التعرف على الأشخاص الذين تنطبق عليهم صفات معينة وتعمل على مراقبتهم، وحرمانهم أحياناً من بعض الحقوق المدنية كالسفر جواً، والمعاملات المصرفية، والعمل في بعض الأماكن الحساسة أمنياً.

---

[1] شاينين، مارتين، (شباط 2009)، **"تعزيز وحماية جميع حقوق الإنسان، المدنية والسياسية والاقتصادية والاجتماعية والثقافية، بما في ذلك الحق في التنمية"**

ويؤدي التوسع في الاعتماد على نتائج التنقيب في البيانات إلى المساس بحقوق الإنسان، ويمكن أن يتسبب في ظلم العديد من الناس الأبرياء، ولاسيما أن البيانات قد تعاني من بعض العيوب مما يؤدي إلى أخطاء في البيانات المسجلة وأخطاء في تفسيرها.

ومن هذا المنطلق فإن مجلس حقوق الإنسان في الأمم المتحدة أصبح يطالب هذه الوكالات بعدم تجاوز الحدود في التصنيف العرقي الذي يؤدي إلى التمييز العرقي بين الناس، كما يطالب بعدم استخدام الصور النمطية في حرمان بعض الأشخاص من حقوقهم بسبب فقط امتلاكهم لصفات مشبوهة، بناءً على التصنيفات الآلية التي تقدمها لهم طرق التنقيب في البيانات. كما يطالب هذا المجلس باعتماد تشريع يتضمن المبادئ الأساسية لحماية البيانات الشخصية الالكترونية، ووضَع لذلك جملة من المبادئ التوجيهية، وطالب الدول بالتقيد بها، وأهمها عدم الحصول على البيانات الشخصية ومعالجتها بأساليب غير نزيهة، وعدم جمع البيانات الخاصة بالأصل العرقي والآراء السياسية والمعتقدات الدينية والانتماء إلى الجمعيات والنقابات[1]..الخ.

نلاحظ أنه يوجد خطورة كبيرة في تداول البيانات والآراء على الانترنت، إذ لا يوجد حتى الآن قيود أو اتفاقيات بين الدول على وضع عتبات تُنظم صلاحيات وحدود جمع وتفسير البيانات الشخصية والمستندات الالكترونية من قبل وكالات الاستخبار العالمية، بل تشكِّل هذه الوكالات خطراً كبيراً على حياة الناس الخاصة، وخصوصاً أن العديد من هذه الوكالات مرتبطة مع بعضها وتتقاسم البيانات دون أن تخضع لأي نوع من أنواع الرقابة، ودون أن يتوفر أية ضمانات لحقوق الإنسان.

2•3•2 استخدام التنقيب في الصور (Image Mining) في محرك البحث Google

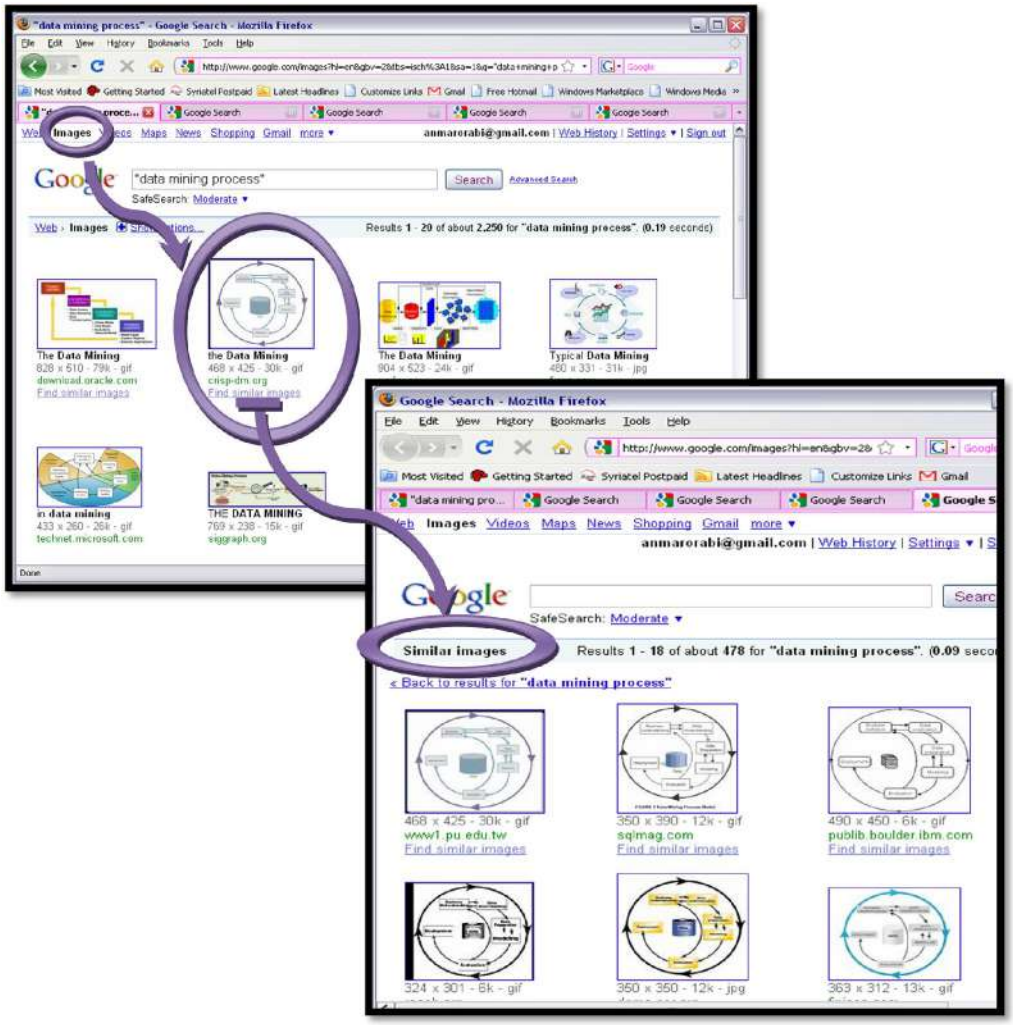
يُعتبر محرك البحث في الانترنت اليوم وخصوصاً Google[2] أهم أداة يمتلكها الإنسان، شريطة أن يتحلى هذا الإنسان بروح البحث والفضول والمبادرة لاكتشاف المزايا الضخمة التي تتطور وتتزايد باستمرار.. وتأتي أهمية هذا المحرك من كونه يعمل على تطوير ذاته وفي مختلف

---

[1] المرجع السابق

[2] كثُر الحديث عن محرك البحث غوغل (Google) في هذا الفصل، ولا يُقصد من ذلك الدعاية لهذا المحرك، وإنما دعت الصدفة إلى وروده كمثال عن العديد من مجالات التنقيب، ويعود ذلك إلى أنه أشهر محرك بحث، وأنه يقوم بدور رائد في تقديم أنسب الخدمات والاستعلامات إلى المستخدمين، والأهم من ذلك هو أنه يستخدم تقانة التنقيب في البيانات بكفاءة واضحة.

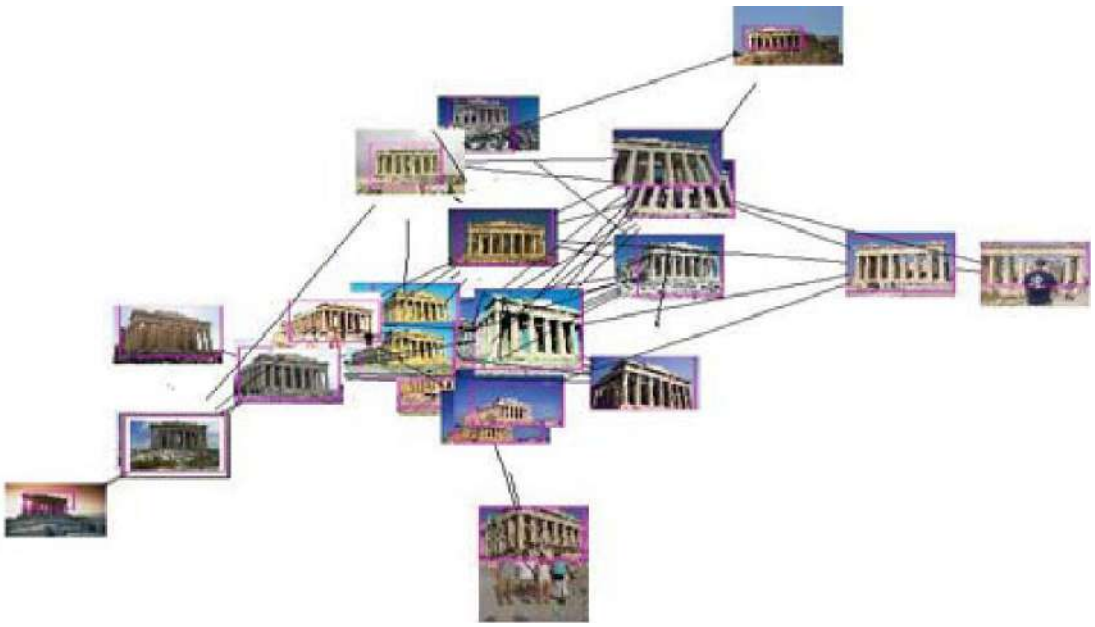
الاتجاهات باستخدام التقنيات المتطورة الممكنة ومنها تقنيات التنقيب في الصور. فلقد أصـــبح بالإمكان اليوم أن تبحث عن الصور بحسب أسمائها ومن ثم تبحث عن صور شبيهة بالصورة المطلوبة، انظر الشكل (2–9).



الشكل (2– 9): مثال عن استخدام محرك البحث Google لتقانة التنقيب في الصور لايجاد صور مشابهة

يعتمد التنقيب في الصور على أساس تمييز معالم الصورة، والبحث عن صور تمتلك نسبة عالية من هذه المعالم، باستخدام قاعدة الجار الأقرب (nearest neighbor)، فمثلاً تتكون صورة مراحل التنقيب في البيانات في الشكل (2–9) من حلقة دائرية فيها أربع إشارات سهمية عند كل ربع منها، كما فيها ستة مستطيلات واسطوانة في المنتصف.. ولدى الضغط على عبارة (Find similar images) أسفل هذه الصورة يقومُ محرك البحث بإحضار جميع الصور التي تحملُ معالم مشابهة للصورة المطلوبة.

وتعتمد طرق التنقيب هنا على التجزئة العنقودية البصرية (visual clustering) للصور[1]، حيث يتم تصنيف الصور وفقاً للمعالم الرئيسة التي تتكون منها، انظر الشكل (2–10).



الشكل (2– 10): مثال على التجزئة العنقودية للصور[2]

يختلف التنقيب في الصور عبر الانترنت عن التنقيب في الصور الطبية أو الصور الفلكية التي تحدثنا عنها سابقاً في أن الصور الموجودة على الانترنت هائلة ومتشعبة كثيراً، ولا تنتمي إلى مجال واحد فقط، وبالتالي لا يمكن تعريف قائمة محددة لمعاني المعالم التي تتكون منها كل الصور. وعلى العكس من ذلك فعندما تكون الصور محصورة في مجال محدد مثل علم الفلك فإنه يصبح بالإمكان تعريف قائمة مسبقة بمعاني المعالم التي تتكون منها الصور، فمثلاً يمكن التمييز بدقة بين الأجرام السماوية حسب شكلها ولونها وحجمها وسلوكها وإضاءتها.. وبالتالي التعرف على مكونات أية صورة جديدة ملتقطة. وتعتبر فكرة إيجاد مقاييس عالمية لقراءة أية صورة من التحديات التي يطمح الباحثون في الوصول إليها، مع العلم أنه تم تحقيق نجاحات جزئية في المجلات المتعلقة بالصور الجغرافية و الصور السياحية.

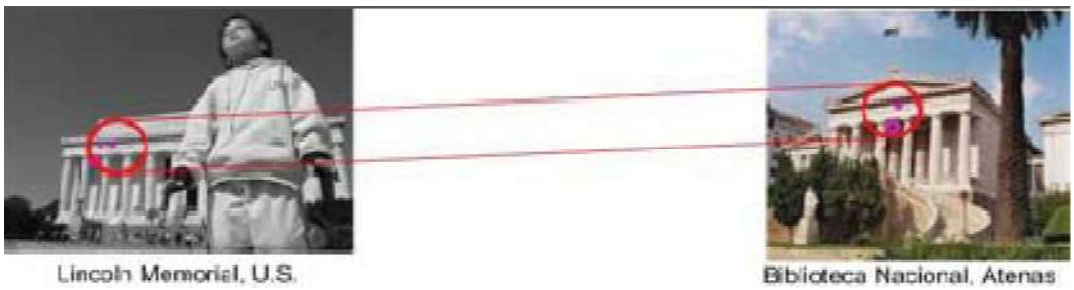
وقد يخطئ التنقيب في الصور ويُحضر صوراً غير مناسبة تماماً للمطلوب، فمثلاً تم اعتبار الصورتين في الشكل (2–11) على أنهما متشابهتين وذلك لأن تحليل الصورتين انطلق من

[1] Zheng, Y. , Zhao, M. & Others, (2008), **Tour the World: building a web-scale landmark recognition engine**

[2] The previous reference

وجود مجسم لعمود شاقولي وفي أعلاهم منحني للأعلى لونه أغمق وخلفه سماء وبناء له أعمدة، ولكن تبين أن المجسم الشاقولي في الأولى هو شجرة وفي الثانية إنسان..

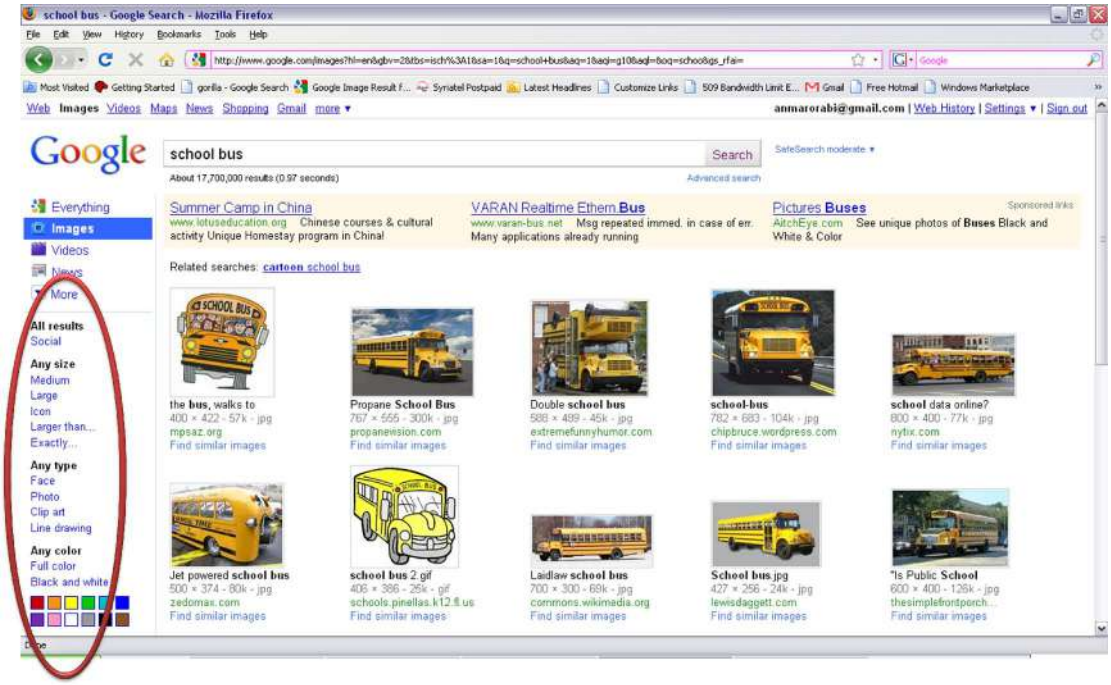
ويتضح من ذلك أن التنقيب في الصور يقترب من الدقة ولكنه قد لا يصيب الهدف في بعض الأحيان، ولذلك فهو إحدى مجالات التنقيب التي لا تزال تواجه تحديات، وتحتاج إلى ابتكار آليات تمييز أفضل للصور.



الشكل (2– 11): مثال عن محدودية قدرات التنقيب في الصور[1]

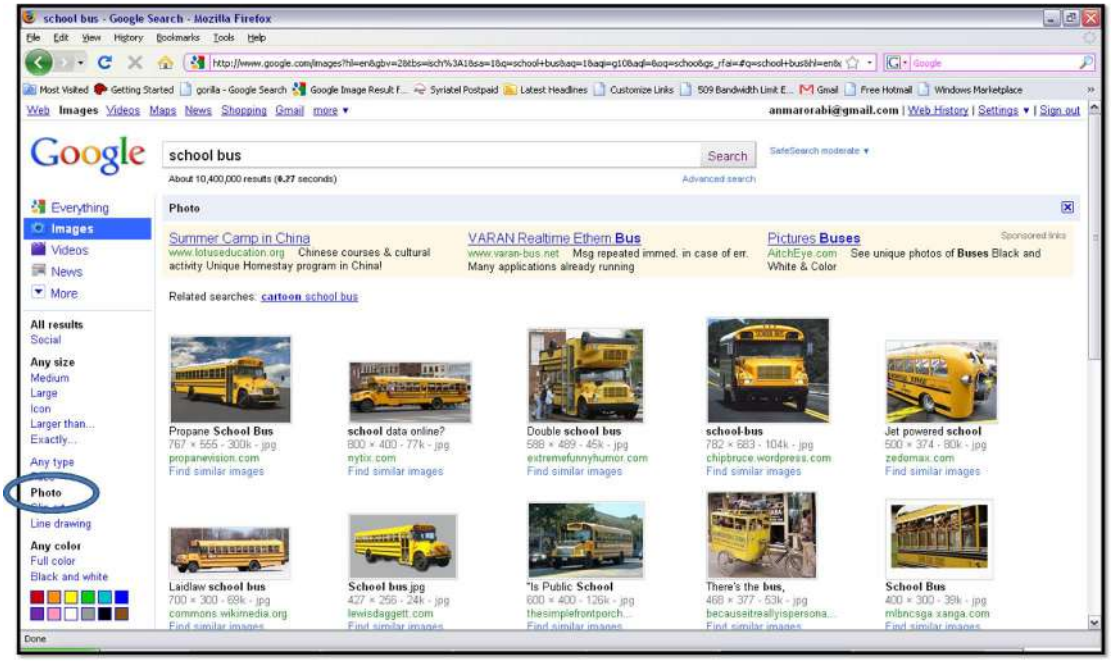
ومن الملفت للانتباه أن التطورات التي تحصل في عالم التكنولوجيا سريعة جداً لدرجة أنه بعد أيام قليلة من أخذ لقطة من الانترنت المعروضة في الصورة (2–9)، تمت إضافة مزايا جديدة إلى واجهة البحث في الصور في Google – في شهر أيار 2010، بحيث أصبح بالإمكان إدراج اختيارات فرعية تصنيفية لعملية البحث الأساسية وانتقائها من القائمة الموجودة إلى اليسار في نافذة التصفح المتعلقة بالصور. ويظهر الشكل (2–12) مثلاً توضيحياً عن هذه الإمكانيات الجديدة، فلدى البحث عن صور لباص مدرسة، ظهرت مجموعة من الصور في نافذة التصفح، وظهرت إلى اليسار قائمة من التصنيفات الفرعية للصور بحسب النوع والحجم واللون..الخ من التفاصيل التي تسهل عملية البحث، ويلاحظ أن جميع الصور الأولية تحمل اللون الأصفر وذلك لأنه اللون التقليدي والأكثر شيوعاً عالمياً كلون لباص المدرسة.

[1] The previous reference



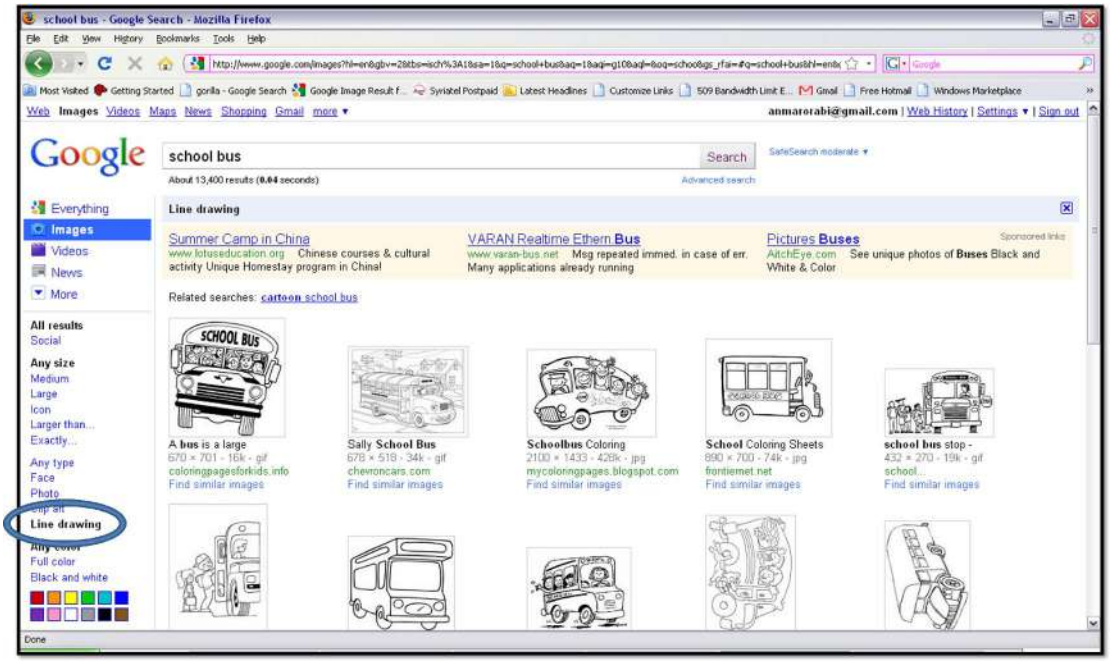
الشكل (–): النتيجة الأولى للبحث عن صور لباص مدرسة

ثم تم الطلب من محرك البحث أن يعطي فقط الصور الفوتوغرافية لباص المدرسة، فظهرت النتيجة المبينة في الشكل (–).



الشكل (2– 13): صور فوتوغرافية لباص مدرسة

ثم جرى طلب رسومات غير ملونة فظهرت النتيجة كما هو مبين في الشكل (2–14). وبالإمكان اختيار أن يكون لون الباص أحمر، أو أي لون آخر.



الشكل (2- 14): رسومات غير ملونة لباص مدرسة

وطبعاً يأتي دور التنقيب في الصور في المرحلة التي تلي مرحلة البحث السابقة، أي عندما يُطلب البحث عن صور مشابهة للصور المعروضة (Find similar images). أي أن محرك البحث يقدم مجموعة من الأدوات التي تساعد في تخصيص البحث في الصور، وبالتالي تسهيل عملية البحث ضمن ملايين الصور المتوفرة.

قدم الفصل الحالي عرضاً لأبرز تطبيقات تقانة اكتشاف المعرفة في حقول مختلفة من الحياة، وقد تقاطعت في كثير من الأحيان الأفكار المعروضة عن آليات التنقيب المتبعة في تلك التطبيقات، وذلك لأن العنصر المشترك فيها هو تقانة التنقيب التي تحمل أنماطاً محددة من المهام، فهي تنقسم كما سبق وورد في الفصل الأول إلى نوعين أحدهما وصفي ويبحث عن الصفات المتشابهة بين العناصر ويضعهم في مجموعات، والآخر تنبؤي ويعمل على توقع الأحداث المستقبلية من خلال فهمه للأحداث الماضية. وقد اهتم هذا الفصل بتسليط الضوء على الجندي المجهول الذي حقق انجازات عالمية كبيرة قد نكون سمعنا عنها ولكننا لم نعرف من صنعها..

الفصل الثالث

مراحل عملية اكتشاف المعرفة من قواعد البيانات

الفصل الثالث

مراحل عملية اكتشاف المعرفة من قواعد البيانات

مقدمة

قام خبراء عملية اكتشاف المعرفة من قواعد البيانات باقتراح وتطوير عدد من طرق اكتشاف المعرفة بما يتناسب مع طبيعة المسائل المدروسة وطبيعة البيانات والأدوات التقنية المستخدمة، وبشكل عام تتكون عملية اكتشاف المعرفة من قواعد البيانات من عدة مراحل متتابعة، وتستند كل مرحلة إلى نجاح المرحلة السابقة حيث تعتبر نتائج كل مرحلة على أنها مدخلات للمرحلة التالية لها. وتختلف طرق اكتشاف المعرفة عن بعضها في عدد المراحل وفي النقاط التي تركز عليها كل مرحلة، وتشترك طرق اكتشاف المعرفة في أنها تبدأ بالمدخلات وتنتهي بالمخرجات حيث [1]تتكون المدخلات من بيانات بأشكال مختلفة كالبيانات الرقمية، والبيانات النصية المخزنة في قواعد البيانات أو الملفات المنبسطة (flat files)، والصور والفيديو، والبيانات شبه المهيكلة كملفات (XML, HTML..)، أما المخرجات فهي عبارة عن معرفة مكتشفة ويتم عرضها بأشكال مختلفة كقواعد الاقتران، ومجموعات البيانات، والشبكات التبعية، والاتجاهات التنبؤية، وجداول المقارنات، والتمثيلات البيانية..الخ.

تم وضع مراحل عملية اكتشاف المعرفة في منتصف التسعينات تقريباً بشكل بحثي أكاديمي (Academic Research Models) من قبل الأستاذ الدكتور أسامة فياض[2]، وكانت هذه العملية تتضمن تسعة مراحل منها أربعة مراحل لتجهيز البيانات، وثلاثة مراحل لإجراء التنقيب في البيانات (اختيار مهام التنقيب، اختيار خوارزميات التنقيب، التنقيب في البيانات)، ومرحلتين لتقييم وتفسير النماذج المكتشفة. واتصفت مراحل عملية اكتشاف المعرفة هذه بأنها تقدم

[1] Cios, K.J.; Pedrycz, W. & Swiniarski, R. W. (2007), **Data Mining: A Knowledge Discovery Approach**

[2] (**Usama Fayyad**) من المؤسسين والمطورين لتقانة اكتشاف المعرفة من قواعد البيانات، حيث كان يعمل باحثاً في شركة ميكروسوفت.

توصيفاً تقنياً للعمليات التي تُجرى على البيانات ولكنها تفتقد إلى التوصيفات المتعلقة بمجالات الأعمال، ولذلك تم اعتبارها حجر الزاوية الذي أسست عليه المراحل التطبيقية لاكتشاف المعرفة فيما بعد[1].

وقد حصلت جملة من التطورات في أدوات اكتشاف المعرفة عندما تزايد تطبيقها في المجالات الصناعية والتجارية، وقد أدت هذه التطورات إلى اختصار الجهد اليدوي وتجنيب المستخدمين من الخوض في التفاصيل البرمجية لمهام وخوارزميات التنقيب، حيث تتحدد مهام وخوارزميات التنقيب بمجرد اختيار الأداة البرمجية الجاهزة الخاصة بالتنقيب، وهذا ما أدى إلى اندماج المراحل الثلاثة لإجراء التنقيب في البيانات في مرحلة واحدة، وتم بذلك اعتماد ستة مراحل أساسية لاكتشاف المعرفة –منذ أواخر التسعينيات– ودعيت بالعملية القياسية للتنقيب في البيانات في الصناعة (Cross-Industry Standard Process for Data Mining) وتختصر بالأحرف (CRISP-DM).

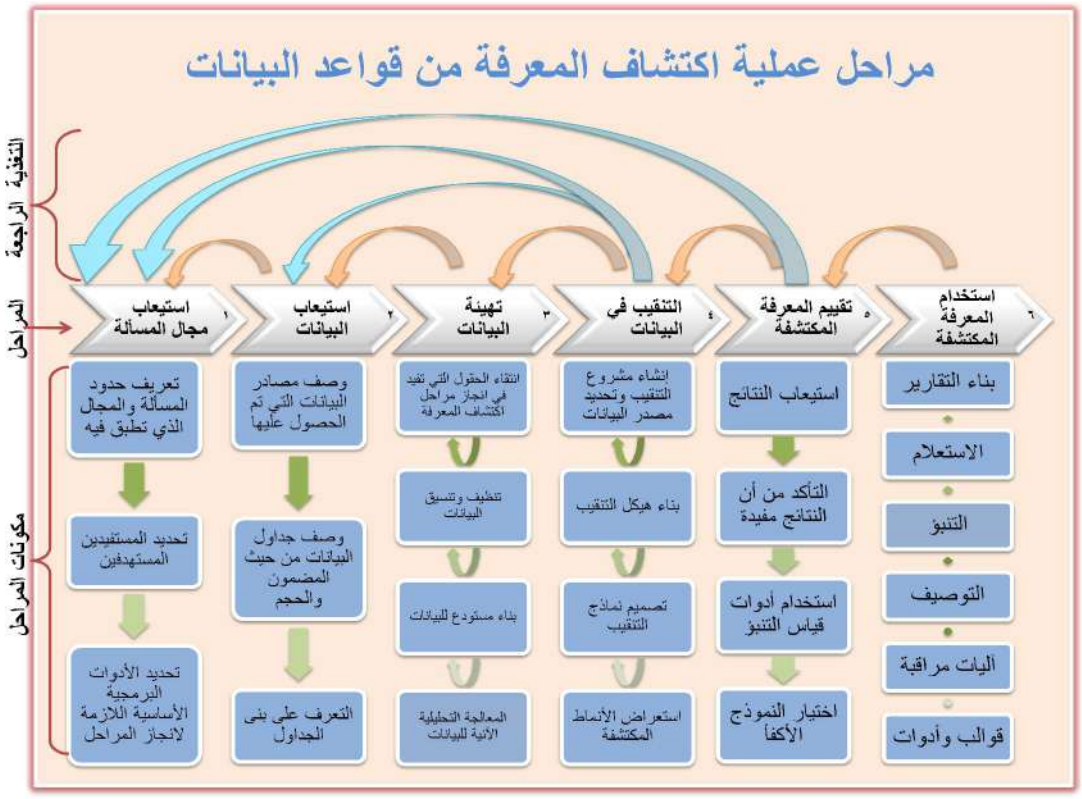
ثم جرت عدة محاولات لتطوير ودمج طريقتي اكتشاف المعرفة الأكاديمية والصناعية ضمن طريقة واحدة هجينة تتضمن مزيجاً من الطريقتين في الفترة ما بين عامي 2000–2004، وكان من أبرز هذه الطرق الهجينة طريقة المراحل الستة لاكتشاف المعرفة (Six-Step KDP) والتي طورها الخبير سيوس (Cios) وزملاؤه[2]، وقد اتصفت هذه الطريقة بالمرونة في التنقل بين مراحل التنقيب وبأنها تدعم التغذية الراجعة في جميع المراحل، وبأنها الطريقة المتوافقة مع استخدام البرمجيات الحديثة الخاصة بانجاز مراحل اكتشاف المعرفة.

وقد تم في الدراسة الحالية إتباع طريقة المراحل الستة لاكتشاف المعرفة من قواعد بيانات التحاليل الطبية، بسبب ملاءمتها للمسألة المدروسة إضافةً إلى تمتعها بالمزايا المذكورة أعلاه، ويختص الفصل الحالي من الدراسة بعرض المراحل النظرية لعملية اكتشاف المعرفة، في حين يتناول الفصل الخامس من الدراسة التطبيق العملي لهذه المراحل على بيانات مخبر التحاليل الطبية التابع لإحدى المشافي الحكومية في دمشق.

---

[1] The previous reference, P.12
[2] The previous reference, P.14-16

تبدأ مراحل عملية اكتشاف المعرفة بالتعرف على حقل (مجال) المسألة المدروسة من خلال الاحتكاك مع العاملين في هذا الحقل وجمع المعطيات الأساسية من المصادر المختلفة والتي إما ستُستخدم في حل مسألة التنقيب أو أنها ستُفيد في تكوين المعرفة الأولية، ثم تأتي مرحلة التعرف على البيانات التي تم الحصول عليها في المرحلة السابقة، من حيث معانيها والعيوب التي تشوبها. وتأتي بعد ذلك مرحلة تحضير هذا البيانات، وتتضمن التخلص من عيوبها وتحويلها إلى أشكال جديدة تتناسب مع دخولها إلى عالم التحليل والتنقيب. ثم يحين دور مرحلة التنقيب في البيانات التي تَستخدم جزءاً من البيانات المتاحة لها في بناء نماذج التنقيب المختلفة، وتَعرض نتائج هذه النماذج بأنماط عرض متنوعة تختلف باختلاف خوارزمية التنقيب المطبقة. ويتم بعد ذلك استخدام الجزء المتبقي من البيانات في المرحلة التالية التي يتم فيها اختبار قدرة النماذج المكتشفة على التنبؤ بمخرجات هذا الجزء من البيانات. وأخيراً تأتي المرحلة التي ترتدي فيها البيانات حلتها الجديدة وتظهر على شكل حقائق يمكن استخدامها بطرق مختلفة في الحقل المدروس الذي خرجت منه كبيانات وعادت إليه كحقائق... تلك كانت الصورة المصغرة عن مراحل عملية اكتشاف المعرفة، انظر الشكل (3-1)، وسنعرض الأسس النظرية لهذه المراحل في الفقرات التالية.



الشكل (3- 1): مراحل عملية اكتشاف المعرفة من قواعد البيانات

1•3 مرحلة استيعاب مجال المسألة (Understanding of the problem domain)

تُعتبر مرحلة التعريف بالمسألة المطروحة على أنها المرحلة التأسيسية للمراحل التالية، وتُشكل هذه المرحلة مفتاح النجاح الأول لعملية التنقيب في البيانات، فالتعريف الصحيح للمسألة المطروحة هو العامل الأول في التوصل إلى نتائج جيدة لعملية اكتشاف المعرفة. ويمكن تجزئة هذه المرحلة إلى أربعة محاور أساسية هي:

1•1•3 تعريف حدود المسألة والمجال الذي تطبق فيه

تعتبر هذه الخطوة على أنها الخطوة الأولية التي تتضمن التعامل مع الخبراء ومصادر المعلومات في الحقل المدروس، بهدف تعريف المسألة المدروسة وتحديد أهدافها، والتعرف على الحلول الحالية لمثل هذه المسألة من خلال الاطلاع على آلية جمع البيانات وعلى الدراسات العلمية في هذا الحقل، والتعرف على المصطلحات التخصصية فيه. ويمكن في بعض الحالات أن يتم تقسيم المسألة إلى مسائل جزئية قابلة للحل باستخدام البيانات المتاحة[1].
ويتم عادةً في هذه المرحلة طرحُ بعض التساؤلات عن التصورات المبدئية لنوع المعرفة التي من المرغوب التوصل إليها، ثم تَجري صياغة هذه التساؤلات بطريقة مسائل التنقيب في البيانات، وهذا ما يساعد في انتقاء البيانات المناسبة التي تختص بها المرحلة الثانية من اكتشاف المعرفة. ومن الأمثلة على مسائل التنقيب (اكتشاف المعرفة) الممكن طرحها في الحقل الطبي ما يلي[2]:

أ- تقسيم المرضى بدقة إلى مجموعات ذات أنماط صحية متماثلة.
ب- التعريف السريع لأكثر الزبائن (المرضى) ربحية والأسباب الكامنة وراء ذلك.
ج- فهم أو التعرف على الزبائن الذين يتوجهون إلى المنافسين (attrition, churn analysis)
د- الاستعداد للطلب الطبي على الموارد من كوادر بشرية ومواد وأدوات وأجهزة وأثاث..
ه- التنبؤ بالتصرفات المستقبلية للزبائن (المرضى) بالاعتماد على تاريخهم وخصائصهم.
و- توقع التشخيص الطبي الأكثر احتمالاً.
ز- التنبؤ بكلف المعالجة للمجموعات المكتشفة من المرضى.

[1] Mayatt, G. J. (2007), **Making Sense of Data A Practical Guide to Exploratory and Data Mining Data Analysis**
[2] Rogers, G. & Joyner, E. (1997), **Mining Your Data For Health Care Quality Improvement**

ح- توقع طول الإقامة في المشفى (دراسة العوامل المساهمة المؤثرة على طول مدة إقامة المرضى في المشفى أثناء العلاج، وهي المكون الأهم في رفع تكاليف العلاج في المشفى) وذلك بهدف تعريف الأنماط المؤثرة على مدة الإقامة في المشفى، والتي يمكن أن تساعد في تخفيض تكاليف العلاج بما فيه مصلحة الجهات التي ستدفع التكاليف مثل شركات التأمين أو الحكومة.

ط- التعرف على الأساليب الطبية الأفضل، مثل "ما هو أسلوب الجراحة الأفضل بالنسبة لاستئصال المرارة ومن هو الجراح الأفضل في المشفى خلال السنوات الخمسة الأخيرة، من حيث نجاح النتائج، وقصر مدة الإقامة في المشفى، واختصار التكاليف، ووقت انجاز العملية....؟."

وحالما تتم صياغة مسألة (مسائل) التنقيب المطلوبة فإنه يتحدد نمط التنقيب الذي سيحل هذه المسألة، أي هل من المطلوب وصف وتعميق المعرفة بالبيانات من خلال إيجاد حقائق وعلاقات وأنماط كامنة أو مخفية في البيانات المدروسة، أم أن المطلوب هو إنشاء نموذج تنبؤ (prediction model) قادر على إجراء التنبؤات المستقبلية؟ أم النمطين؟

3•1•2 تحديد المستفيدين المستهدفين من عملية اكتشاف المعرفة

تهدف عملية اكتشاف المعرفة إلى خدمة جهة معينة أو عدة جهات، وإن التعرف على الجهات المستفيدة يساعد على توجيه مسائل التنقيب نحو تلبية احتياجات هؤلاء. وقد تعمل الجهات المستفيدة من البيانات في نفس مجال المسألة المدروسة مثل الأطباء الذين يرغبون بالتعرف على نجاعة أساليب العلاج بحسب خاصيات المرضى، كما يمكن أن تكون جهة خارجية ترغب بالحصول على بعض المعارف، مثال شركات التأمين التي ترغب في التعرف على النفقات الصحية بحسب خاصيات المرضى وبيانات سجلاتهم الطبية.

3•1•3 تحديد الأدوات البرمجية الأساسية اللازمة لانجاز مراحل اكتشاف المعرفة

يتوفر في الأسواق العالمية عدد من الأدوات البرمجية التي تساعد في انجاز مراحل اكتشاف المعرفة، وتختلف الأدوات باختلاف المهام المطلوب انجازها في كل مرحلة، فمنها من يهتم بإدارة البيانات، وبعضها يهتم بعرض مخرجات المراحل، وأخرى تختص بانجاز عمليات

التنقيب، والتي تتنوع بدورها وتختلف عن بعضها في عدة نواحٍ أهمها خوارزميات التنقيب المستخدمة في كل منها. ومن أشهر أدوات التنقيب وأكثرها انتشاراً وتطبيقاً هي برنامج (SAS) وبرنامج (Microsoft SQL Server 2008) وبرنامج (Clementine). انظر الشكل (3–2).



الشكل (3– 2): بعض الأسماء التجارية للأدوات البرمجية المستخدمة في عمليات التنقيب في البيانات

وواقعياً تساهم عدة عوامل في اختيار برامج (أدوات) التنقيب المناسبة، وأهمها:

أ– أن يكون البرنامج متناسباً مع طبيعة البيانات المدروسة فالبرامج التي تنقب في جداول البيانات تختلف عن تلك التي تنقب في الصور وأفلام الفيديو.

ب– قدرة البرنامج على حل مسائل التنقيب المطروحة في المرحلة الأولى.

ج– مدى توفر البرنامج والقدرة على اقتنائه، حيث لكل برنامج سعره الخاص به والذي قد يقع خارج القدرة المادية للمنظمة أو الجهة البحثية.

د– مدى سهولة البرنامج وإمكانية التدرب على استخدامه.

2•3 مرحلة استيعاب البيانات (Understanding of the Data)

تتحدد في هذه المرحلة البياناتُ التي تحتاجها عملية اكتشاف المعرفة، ويتم عادةً انتقاء مجموعات البيانات بالاستعانة بالخلفية المعرفية عن الحقل المدروس، وبما يتناسب مع احتياجات مسألة التنقيب المطروحة، ويمكن التعرف على البيانات بإجراء بعض التحليلات

الإحصائية البسيطة التي تفيد في التعرف على طبيعة البيانات. كما يتحدد هنا فيما إذا كان هنالك أية احتياجات إضافية من البيانات من مصادر أخرى غير المصدر المتاح. وفي حال وجود ضَعف في ملاءمة البيانات المتاحة مع المشكلة المطروحة، فإنه إما أن يتم البحث عن بيانات جديدة أو يتم إعادة صياغة مسألة التنقيب المطروحة في المرحلة السابقة بما يتناسب مع البيانات المتوفرة. [1]وتتكون هذه المرحلة من خطوات جزئية يتم فيها التعرف على البيانات من حيث مصادر الحصول عليها، وحجمها وصفات الحقول التي تتكون منها، والتحقق من جودتها والتعرف العام على المشاكل التي تعاني منها.

أ- وصف مصادر البيانات التي تم الحصول عليها

تتضمن هذه الخطوة توصيفاً عاماً لمصادر الحصول على البيانات اللازمة للدراسة، حيث يتم التعرف على نوع البيانات (Data types). التي قد تكون على شكل قواعد بيانات علاقاتية، أو ملفات جداول، أو ملفات نصية، أو ملفات الوسائط المتعددة..الخ.

كما يتم التعرف على هذه المصادر من حيث أماكن الحصول عليها وتقنيات تخزين البيانات (Data Storage techniques) أو البرامج التي ولَّدتها، وحجم البيانات فيها، ومضمونها العام. وتعتمد الدراسة الحالية على بيانات من مصدر قواعد بيانات علاقاتية، ولذلك فإنه سيتم التركيز في المضمون النظري لمراحل وخطوات اكتشاف المعرفة اللاحقة على الأسس النظرية التي تتناسب مع هذا النمط من البيانات.

ب- وصف جداول البيانات من حيث المضمون والحجم

يتم في هذه الخطوة التعرف على الجداول المكونة لقواعد البيانات المدروسة والمضمون العام لكل منها، والعلاقات التي تربط بينها، وعدد السجلات التي يتكون منها كل جدول. وهي خطوة تمهيدية للخطوة التالية، وتفيد في تكوين فكرة عامة عن جداول البيانات.

ج- التعرف على بنى الجداول في قواعد البيانات

تُعتبر هذه الخطوة الأوسع والأطول بين خطوات مرحلة استيعاب البيانات، حيث يجري فيها الخوض في تفاصيل مضمون الجداول والحقول (الخاصيات) المكونة لها. كما يتم في هذه الخطوة تصفح قيم الحقول والتعرف على نوعها (رقمية، نصية، تاريخ)، ومضمونها (معناها)،

[1] Chapman, P.; Khabaza, Th.; Shearer, C. (2000), **CRISP-DM: Step-by-step data mining guide**

وإظهار بعض العيوب العامة التي تعاني منها مثل القيم المفقودة والمشاكل الناجمة عن أنماط الحقول. وتُشكل البيانات الناتجة عن هذه المرحلة مدخلات للمرحلة التالية التي يتم فيها تجاهل البيانات غير المتناسبة مع مسائل التنقيب المطروحة، وانتقاء المفيد منها فقط.

3•3 مرحلة تهيئة البيانات (Data Preparation)

تختص هذه المرحلة بانتقاء الجداول والحقول المفيدة في عملية اكتشاف المعرفة، وبتجهيز البيانات للاستخدام في عمليات التنقيب من خلال تنظيفها وتنسيقها وتحويلها إلى حقول وجداول جديدة تكون أكثر انسجاماً مع حل مسائل التنقيب المقترحة في المرحلة الأولى.

ويتم العمل بعد ذلك على جمع وتوحيد الجداول الناتجة من جميع قواعد البيانات المدروسة في وعاء واحد هو قاعدة بيانات جديدة أو مستودع بيانات جديد، وتُطبّق بعدها أدوات المعالجة التحليلية الآنية على البيانات الجديدة التي بُنيت بشكل ينسجم مع بنية مكعب البيانات، وينتج عن ذلك مجموعة من المعلومات التي قد تفيد في التعرف على البيانات وبناء التقارير عنها، أو قد تفيد في إجراء عمليات التنقيب. وتعتبر هذه المرحلة من أطول المراحل في اكتشاف المعرفة حيث أن العمل فيها يستغرق أكثر من 70% من إجمالي العمل. وتنقسم هذه المرحلة إلى أربعة مراحل أو خطوات فرعية كما يلي:

1•3•3 انتقاء الحقول التي تفيد في انجاز مراحل اكتشاف المعرفة

يجري في هذه الخطوة انتقاء الجداول والحقول من قواعد البيانات المتوفرة، والتي تستطيع أن تقوم بمهام متنوعة في خدمة كل من مراحل اكتشاف المعرفة.

وتنقسم البيانات إلى ثلاثة أقسام بحسب المهام التي يمكن أن تؤديها على النحو التالي:

- البيانات التي تُستخدم مباشرةً في عمليات التحليل والتنقيب.
- البيانات التي يتم استخدامها في عمليات التحليل والتنقيب بعد أن يتم تحويلها إلى أشكال جديدة.
- البيانات التي تُستخدم في تنظيف البيانات الأخرى، أو التحقق من صحتها.

ويتم حذف بعض الخاصيات (الحقول) من البيانات عندما تكون غير مفيدة في استنباط المعرفة (كأسماء الموردين للبضائع)، أو عندما تكون هذه الخاصيات على علاقة ارتباطية قوية وواضحة مع خاصيات أخرى ويمكن بالتالي الاستغناء عنها.

2•3•3 تنظيف وتنسيق البيانات وإنشاء بيانات تركيب جديدة

1•2•3•3 تنظيف وتنسيق البيانات (Data Cleaning and Formatting)

يرتبط الحصول على نتائج صحيحة من البيانات بمدى جودة وصحة البيانات الموضوعة قيد التنقيب. وتفتقد البيانات الموجودة في أي نظام حاسوبي – واقعياً وبدرجات مختلفة – إلى الكمال والدقة، وذلك لأن هذه البيانات هي ثمرة العمل الإنساني الذي يتأثر بالظروف المختلفة المحيطة بجو العمل. فمثلاً قد يُقصِّر مبرمج النظام في وضع القيود والضوابط اللازمة للتأكد من صحة واتساق البيانات مع بعضها، أو قد لا يصرح الزبون أو المريض بكافة البيانات الشخصية التي يتضمنها ويطلبها النظام الحاسوبي، أو قد يدلي ببعض البيانات الخاطئة المتوهة، أو قد يخطئ الموظف في إدخال البيانات إلى النظام الحاسوبي..الخ. وكل من ذلك يؤدي إما إلى وجود قيم مفقودة في البيانات، أو إلى بيانات خاطئة ومشوشة (noisy) وغير منسجمة مع غيرها من البيانات.

وتبعاً لذلك فإن استعمال أية بيانات في عملية اكتشاف المعرفة يتطلب أولاً إيجاد حلول للعيوب التي تشوبها. وتُسمى هذه العملية بتنظيف البيانات، وهي إحدى المراحل الأساسية والضرورية في عملية اكتشاف المعرفة والتي تهدف إلى التوصل إلى بيانات ذات كفاءة وجودة مناسبة للحصول على نتائج معرفية حقيقية ومفيدة مما يزيد من فعالية نماذج التنقيب المكتشفة.

وتتم عمليات تصحيح أخطاء البيانات بطرق يدوية وآلية معاً، ويمكن الاستعانة ببرامج إدارة قواعد البيانات في تصفح وتنظيف البيانات، وذلك بإتباع عددٍ من الأساليب، منها:

أ- إجراء الاستعلامات الموجهة لاختيار مجموعات من البيانات ومقارنة السجلات فيها مثل اختيار بيانات المرضى التي تحتوي كلمة "احتشاء" في حقل سبب القبول في المشفى.

ب- ترتيب البيانات تصاعدياً أو تنازلياً.

ج- طلب الاستعلامات التلخيصية للحصول على الحسابات الإحصائية البسيطة كالوسط الحسابي والقيمة الدنيا والقيمة العظمى.

كما وتدعم بعض البرامج الالكترونية المختصة بتنفيذ عمليات اكتشاف المعرفة التعرُّفَ والتخلص من عيوب البيانات مثل برامج (Clementine) وبرامج (SQL Server 2008) وبرنامج (SAS) وبرنامج اكسل. ويرتبط اختيار البرنامج المناسب للتنظيف بمكان وجود البيانات، وباحتياجات عمليات التنظيف.

1•1•2•3•3 التعامل مع القيم المفقودة (Missing Values)

يوجد طريقتان أساسيتان للتغلب على مشكلة القيم غير الموجودة، الأولى بالاستغناء عن السجلات التي تحتوي على القيم المفقودة، والثانية بتعبئة القيم المفقودة وتُنفَذ بعدة طرق، منها[1]:

أ- التعبئة اليدوية للقيم المفقودة، وهي طريقة فعالة نسبياً ولكنها تحتاج إلى وقت طويل ويصعب تطبيقها عند التعامل مع كم هائل من البيانات.

ب- استخدام الوسط الحسابي (أو الوسيط أو المنوال) لخاصية ما في تعبئة القيم المفقودة، مثل استبدال القيم المفقودة في حقل الدخل بالدخل المتوسط لزبائن الشركة.

ج- الاستبدال بقيمة ثابتة عامة مثل (Unknown, NA[2], ∞) لجميع القيم المفقودة.، ولكن لا يُنصح بهذه الطريقة لأنها قد تجعل النظام يعتقد أن هذه البيانات تمتلك نفس المعنى..

د- استخدام الوسط الحسابي للمجموعة التصنيفية الجزئية التي تنتمي لها كل مجموعة من السجلات، مثل استخدام (الدخل الوسطي لفئة المهندسين العاملين في المدينة) في تعبئة قيم الدخل المفقودة عند هذه الفئة، وهنا يمكن أن يُستعان بتقانة التنقيب في البيانات كأداة مساعدة في التعرف على المجموعات.

ه- الاعتماد على القيم الأكثر احتمالاً في تعبئة القيم المفقودة. وتستخدم لذلك توابع الانحدار، وطرق الاحتمال، وأشجار القرار. وهنا يتم الاستناد إلى قيم الخاصيات الأخرى التي تؤثر على قيم الخاصية التي تحتوي على قيم مفقودة، وذلك بناءً على العلاقات بين الخاصيات التي تم تقديرها من بيانات سابقة.

و- ترك القيم المفقودة فارغة دون تعبئة، ويُنظر إلى هذه البيانات في نماذج التنقيب على أنها فارغة.

قد تؤدي طرق تعبئة القيم المفقودة إلى حدوث تحيُّز في البيانات، ولذلك يُفضل التروي عند اتخاذ قرار تعبئة القيم المفقودة وعند انتقاء الطريقة المناسبة وعند تعبئة القيم، كما يجب أن تقدِّم تعبئة القيم المفقودة الفوائد لعملية اكتشاف المعرفة.

---

[1] [Cios, K.J.; Pedrycz, W.& Swiniarski, R. W. (2007), **Data Mining: A Knowledge Discovery Approach**, P. 41] & [Han,J. & KAmber,M. (2001), **Data Mining Concepts and Techniques**, P. 67-69]

[2] اختصار لعبارة (Not Available)

2•1•2•3•3 التعامل مع البيانات المشوشة (Noisy Data)

تُعرف البيانات المشوشة بأنها قيم الخاصيات ذات الأخطاء العشوائية، أو الأخطاء في قياس قيم الخاصية[1]، وقد يؤدي الإبقاء على هذه البيانات – عندما تكون نسبتها إلى بقية القيم كبيرة– دون تصحيح إلى الإضرار بعملية اكتشاف المعرفة. وتقل القيم المشوشة عادةً في نظم قواعد البيانات الجيدة التي تتواجد فيها القيود المناسبة والفعالة أثناء إدخال القيم (البيانات) إلى النظام، فمثلاً لا تسمح أنظمة قواعد البيانات المُحْكمة بتسجيل تاريخ ميلاد للزبون أكبر من التاريخ الحاضر، ولا تسمح بتسجيل تاريخ قديم بحيث يجعل العمر يزيد عن الحدود الطبيعية. ويمكن التعرف على القيم المشوهة وتصحيحها بعدة طرق بحسب طبيعة البيانات والمشاكل التي تعاني منها، ونبين ما يلي بعضاً من هذه الطرق:

أ- طريقة الفحص الآلي–اليدوي، ويمكن استخدام هذه الطريقة في الكشف عن عيوب البيانات بمساعدة البرامج الحاسوبية واختيار قيم مناسبة للقيم المشوشة، فمثلاً إذا كانت قيم خاصية ما تقدر بالأرقام الحقيقية التي تقع في المجال [102.5–521.3]، وتبين من خلال ترتيب قيم هذه الخاصية تنازلياً وجود بعض القيم التي تزيد عن الألف مثل (3526، 4864، 1506) فإنه يُتَوقع غالباً أن هنالك نقص في تدوين الفاصلة، وبالتالي يمكن أن تُستبدل تلك القيم بالقيم التالية (352.6، 486.4، 150.6).

ب- طريقة الوضع في الصندوق (Binning)، ويتم في هذه الطريقة ترتيب قيم الخاصية تسلسلياً، ثم يتم تقسيم قيم هذه الخاصية إلى مجموعات تحتوي كل منها على عدد ثابت من العناصر، ثم تُستبدل عناصر كل مجموعة بقيمة ما مثل الوسط الحسابي أو الوسيط.

ج- طريقة التجزئة العنقودية (Clustering)، وهي طريقة لتقسيم البيانات إلى مجموعات تحتوي على عناصر متشابهة الخاصيات، حيث تَعتبر هذه الطريقة أن القيم المشوهة هي القيم النائية أو المتطرفة (Outliers) التي لم تتمكن من الانتماء إلى أية مجموعة.

د- استخدام معادلات الانحدار، حيث يجري البحث عن أفضل خط انحدار يمثل العلاقة بين متغيرين أو أكثر، ويمكن بالتالي وضع القيم المقدرة في تابع الانحدار بدلاً من القيم المشوهة.

---

[1] [Cios, K.J.; Pedrycz, W.& Swiniarski, R. W. (2007), **Data Mining: A Knowledge Discovery Approach**, P. 41] و [Han,J. & KAmber,M. (2001), **Data Mining Concepts and Techniques**, P. 67-69]

ولكن بشكل عام من الأفضل أخذ الحذر في اختيار طريقة التعامل مع القيم المفقودة والمشوهة وخصوصاً في البيانات الحساسة (مثل البيانات الطبية) التي قد يؤدي تغييرها إلى أخطاء في النتائج، ومن الممكن أن تبنى عليها قرارات مُسيئة لصحة الإنسان. ومن المحبذ أيضاً في المجالات التي تُطَبَّق تقانة التنقيب فيها حديثاً أن يتم الاعتماد على الفحص الشخصي للبيانات، وأن يتم التعامل مع القيم المفقودة والقيم المشوهة بأسلوب أقرب إلى اليدوي، ريثما يجري التعرف أكثر على الأخطاء المرتكبة في هذا المجال، وبالتالي يتم التمكن من إنشاء طرق آلية لتصحيح مثل هذه الأخطاء في المستقبل إذا تطلب الأمر.

3•1•2•3•3 تنسيق البيانات (Data Formatting)

يُقصد بتنسيق البيانات إجراء عمليات التحويل أو التحسين على الجداول أو الحقول أو القيم دون أن يؤثر ذلك على معاني البيانات، وبما يتناسب مع متطلبات نماذج التنقيب. ويجرى تنسيق البيانات بالعمل اليدوي وليس الآلي. فمثلاً قد يأخذ أحد الحقول النمط النصي بينما يتضمن بيانات رقمية، بسبب تقصير في عملية البرمجة لقاعدة البيانات الأصلية. ويتطلب إدراج هذه البيانات في نماذج التنقيب أن تأخذ قيماً متناسبة مع أنماطها، أي يجب تغيير نمط الحقل إلى رقمي، ولكن طبعاً بعد تنظيف القيم التي لابد وأن تذخر بأخطاء الإدخال لأن نمط النص يتقبل جميع أنماط البيانات. كما تظهر هذه المشكلة عند دمج قواعد بيانات من مصادر مختلفة معاً وتكون تسميات وأنماط الجداول والخاصيات مختلفة باختلاف المصدر، وعندها يجب توحيدها قبل ضم البيانات معاً. كما تظهر مشكلة تناسق البيانات عندما يختلف التعبير عن القيم باختلاف المصدر أو حتى في نفس المصدر، فمثلاً يُرمز للذكور في حقل الجنس بإحدى الرموز (M, Male, 1)، وبالتالي يجب توحيدها لأن الإبقاء على كل هذه الرموز يسبب إشكالية كبيرة في التعرف عليها أثناء عمليات التحليل والتنقيب.

2•2•3•3 إنشاء بيانات تركيب جديدة (New Constructed data)

تحتاج عادةً عمليات اكتشاف المعرفة إلى خاصيات بيانات جديدة غير متوفرة في قواعد البيانات الأصلية حيث لا توجد حاجة لها في أنظمة العمل التي تحتويها. ولكن يمكن الحصول عليها من خلال اشتقاقها من البيانات الأصلية، وتدعى هذه العملية بإنشاء بيانات تركيب

جديدة، كأن يتم إنشاء خاصية جديدة مركبة من خاصيتين موجودتين مثل إنشاء خاصية المساحة من حاصل جداء قيم خاصية الطول بقيم خاصية العرض، أو كإيجاد خاصية جديدة للزبائن تحمل بياناتها إحدى قيمتين، وتدل الأولى (Yes) على شراء الزبون لسلعة ما خلال السنة الأخيرة، وتدل الثانية (No) على عدم الشراء. كما يمكن تحويل عمود ما إلى عدة أعمدة لكي تسهل المقارنة، فمثلاً يمكن تقسيم الحقل الذي يحتوي اسم الجريدة التي يشتريها الزبون من بين أربعة أنواع جرائد إلى أربعة حقول، ويختص كل منها باسم جريدة معينة، وتكون قيم البيانات في كل منها تساوي (1) في حال كان الزبون يشتري هذه الجريدة، وتكون قيمته (0) إذا لم يكن يشتري هذه الجريدة.

وتعتبر عملية إنشاء بيانات التركيب الجديدة من العمليات المستمرة طوال مراحل عمليات اكتشاف المعرفة، حيث كثيراً ما تظهر الحاجة لها في أثناء بناء نماذج التنقيب، ويمكن الحصول على هذه البيانات من خلال بناء مرأى للبيانات في قاعدة أو مستودع البيانات باستخدام لغة (SQL) ويتضمن البيانات الجديدة المشتقة.

3•3•3 بناء مستودع بيانات جديد (New Data Warehouse)

تظهر الحاجة إلى بناء مستودع للبيانات من أجل دمج قواعد البيانات من مصادر مختلفة في وعاء تخزيني واحد، ومن أجل تنظيم البيانات حول مواضيع أساسية تلخيصية بعيدة عن تفصيلات العمليات اليومية كالمناقلات (المبادلات، المعاملات)، وأيضاً للحصول على بيانات مخزنة بطريقة تسمح بإنشاء جداول الحقيقة وجداول الأبعاد وخصوصاً البعد الزمني، وأخيراً للتوصل إلى بيانات ثابتة لا تتغير وذلك لاستخدامها في عمليات المعالجة التحليلية الآنية والتنقيب. ويتم البدء ببناء مستودع البيانات بعد انتهاء العمليات الأولية للتنظيف والتنسيق وإعادة بناء الحقول والجداول الضرورية في قاعدة البيانات.

ويمكن الحصول على جداول مستودع البيانات من خلال المراحل التالية:

أ- تجهيز البيانات المراد نقلها من قاعدة البيانات الأساسية

تحتاج بعض البيانات قبل نقلها لمستودع البيانات إلى تحويلها إلى أشكال مختلفة، ويتم لذلك بناء استعلامات مناسبة في قاعدة البيانات الأساسية المدروسة بشكل يتلاءم مع هدف استخراج

المعلومات والمعارف منها. وتقوم هذه الاستعلامات على أساس دمج أو تجميع البيانات من عدة جداول أو سجلات. وكمثال على ذلك يمكن دمج جدول بيانات المتاجر مع جدول أرباح ومبيعات المتاجر وجدول المواقع الجغرافية للمتاجر في جدول واحد. كما يمكن أن تحتوي البيانات المدموجة على تلخيص للبيانات من عدة سجلات أو جداول. فمثلاً يمكن تحويل جدول مشتريات الزبائن –الذي يُخصص فيه سجل لكل عملية شراء للزبون– إلى جدول تجميعي لكل بيانات الزبائن، ويُخصص فيه سجل واحد لكل زبون، ويتضمن السجل حقول بيانات الزبون ويضاف إليها حقل تجميعي لعدد عمليات الشراء، وحقل تلخيصي لقيمة المشتريات الوسطية، وحقل لنسبة المبيعات التي دفعت عبر بطاقة الائتمان، وحقل لنسبة المواد التي تم شراؤها في العروض الترويجية..الخ. ويمكن أن تتحقق هذه الجداول بناءً على الاحتياجات المتقدمة التي تظهر أثناء عمليات المعالجة التحليلية الآنية أو التنقيب، أي أنها قد تنتج من خلال بناء مرأى لها في مستودع البيانات الجديد وليس من قاعدة البيانات الأساسية.

ب- إنشاء بنية مستودع البيانات الجديد

يتشابه الهيكل العام لمستودع البيانات مع الهيكل العام لقاعدة البيانات، ويختلف عنه في عدة نواحٍ أهمها محتويات الجداول وأهدافها. ويتم تأسيس البنية التصميمية لهذا المستودع باستخدام تعليمات إنشاء قاعدة بيانات على مخدم الحاسوب المستخدم في عملية البناء وهو المخدم المحلي (Localhost) وذلك عبر البرنامج الحاسوبي المناسب.

ج- سحب البيانات المراد نقلها من قاعدة البيانات إلى مستودع البيانات

يعتمد عادة سحب البيانات من مصدر البيانات (قاعدة البيانات الموجودة في أي نظام لقواعد البيانات) إلى وجهة البيانات (مستودع البيانات الموجود على مخدم الحاسوب) على الاستعانة ببرنامج وسيط يرتبط مع كل منهما، ويساعد على انتقاء البيانات المراد نقلها، وإجراء بعض التعديلات المناسبة عليها مثل تغيير نمط الحقل (رقمي، نصي..)، ثم يعمل على نقلها. ويتم نقل المحتويات من جدول أو استعلام واحد في كل مرة.

تتعدد غالباً مصادر قواعد البيانات التي يَستمد منها المستودع بياناته، وقد تختلف أنظمة قواعد البيانات في تلك المصادر عن بعضها مثل (Oracle, Access, SQL Server)، مما يقود غالباً

إلى عدم تجانس البيانات المراد نقلها. وتساعد هذه الخطوة في التغلب على هذه المشكلة، وفي التوصل إلى مستودع بيانات متجانس.

د- تطوير مستودع البيانات

يُصبح مستودع البيانات بعد الانتهاء من العمليات السابقة ممتلئاً بالجداول وجاهزاً ومستقلاً بنفسه شكلاً ومضموناً عن البيانات السابقة التي أنتجته. ويمكن أن يتولد عنه أي مرأى بيانات تحتاجه عمليات المعالجة التحليلية أو عمليات التنقيب. كما أنه قد تظهر الحاجة إلى كتابة بعض الدوال لبناء مرائي جديدة للبيانات..الخ. أي أنه يتعرض للعديد من عمليات التطوير اللاحقة، والتي تفرضها الخصوصية التي تتمتع بها عملية اكتشاف المعرفة بحسب المجال الذي تُطبق فيه وبحسب البيانات التي تشملها. وتستمر عمليات التطوير حتى آخر مرحلة من مراحل عملية اكتشاف المعرفة، وهذا ما سيتم التعرف عليه في الجزء التطبيقي من الدراسة.

4•3•3 المعالجة التحليلية الآنية للبيانات (OLAP)

يمكن أن تُطبَّق المعالجة التحليلية الآنية كجزء من عمليات اكتشاف المعرفة، أو بشكل مستقل عنها، وهذا يعود إلى الدور الذي تستطيع أن تؤديه في خدمة اكتشاف المعرفة. وتُعتبر المعالجة التحليلية الآنية في هذه الدراسة جزءاً من تقانة اكتشاف المعرفة في البيانات، وتعمل كخطوة تمهيدية أو تأسيسية لعمليات التنقيب في البيانات، لأنها تساعد في التعرف بدقة على اتجاهات البيانات، وتلخيصها.

تُطبَّق عمليات المعالجة التحليلية الآنية أحياناً على الجداول الموجودة في مستودع البيانات، حيث يمتلك مستودع البيانات بنية تصميمية لجداوله تساعد على تلبية احتياجات هذه العمليات. وتقوم فكرة المعالجة التحليلية الآنية على أساس إجراء الاستعلام المتعدد الأبعاد من مكعب البيانات الذي يتم تأسيسه لهذا الغرض. وتتألف المعالجة التحليلية الآنية من عدة مراحل **تبدأ** بإنشاء مشروع خاص بالمعالجة التحليلية الآنية من خلال الاستعانة ببرنامج حاسوبي مختص، وتتشكل بالتالي البنية العامة لهذا المشروع، وتصنَّف المعالجة التحليلية ضمن مجموعة مشاريع خدمات التحليل التابعة لمشاريع ذكاء الأعمال (راجع فقرة 3•1•1). **ثم** يَجري تعريف مصدر ومرأى البيانات الذي تتزود منه عملية المعالجة التحليلية الآنية ويتمثل هنا بمستودع

البيانات الناتج عن الخطوة السابقة. **ويلي** ذلك تعريف أبعاد مكعب البيانات وجداول الحقيقة. **ومن** ثم يتم تطوير المكعب الناتج بما يتناسب مع احتياجات المعالجة التحليلية الآنية، مثل تعريف التراتبات الهرمية لبعد الزمن (السنة، نصف السنة، ربع السنة، الشهر، اليوم)، وإضافة مقاييس جديدة إلى المكعب بحسب الحاجة..الخ. **وتأتي** بعد ذلك خطوة حصاد النتائج، وهي عبارة عن استعراض مكعب البيانات من زوايا مختلفة، وتلخيص مضمونه بالطريقة التي يجدها المستخدم مناسبة، وذلك بتطبيق العمليات الأساسية في المعالجة التحليلية مثل الطي للأعلى والنثر للأسفل، والتقطيع إلى شرائح (راجع فقرة 1•1•7). وينتج عن هذه المعالجة عددٌ كبير من المخرجات أو المعلومات، ويمكن تمثيلها بوساطة الجداول والأشكال البيانية المحورية، كما يمكن تلخيص المعلومات الناتجة وترتيبها بطرق مختلفة وانتقاء بعضاً منها، وهذا ما سيتم التعرف عليه بالتفصيل في الجانب التطبيقي من الدراسة.

4•3 مرحلة التنقيب في البيانات (Data Mining)

تستند عمليات التنقيب في البيانات إلى مخرجات جميع مراحل اكتشاف المعرفة السابقة، ويتم أولاً في هذه المرحلة مقاربة أهداف اكتشاف المعرفة الموجودة في المرحلة الأولى مع طرق وخوارزميات التنقيب المناسبة[1]، وتحويل مسائل التنقيب الموجودة في تلك المرحلة إلى مشاريع تنقيب يختص كل منها بتنفيذ مهام معينة. وتتزود هذه المرحلة دائماً من مرحلة استيعاب البيانات بالمعارف اللازمة عن معاني البيانات وعلاقاتها وبنيتها الأساسية قبل أن تطرأ عليها التعديلات.. كما تتعامل مرحلة التنقيب مباشرة مع مخرجات مرحلة تهيئة البيانات حيث تعتمد على مستودع البيانات كمصدر أساسي لبيانات التنقيب، وتستفيد من مخرجات الاستعلامات متعددة الأبعاد (OLAP) في توجيه بعض عمليات التنقيب.

تتألف مرحلة التنقيب في البيانات من الناحية النظرية من أربعة أجزاء منفصلة ومتداخلة، وهي إنشاء مشروع التنقيب، وبناء هيكل التنقيب، وتصميم نماذج التنقيب، واستعراض الأنماط المكتشفة.

[1] Fayyad, U.; Piatetsky, G. & Smyth, P. (1996), **"From data mining to knowledge discovery in databases"**

1•4•3 إنشاء مشروع التنقيب وتحديد مصدر البيانات

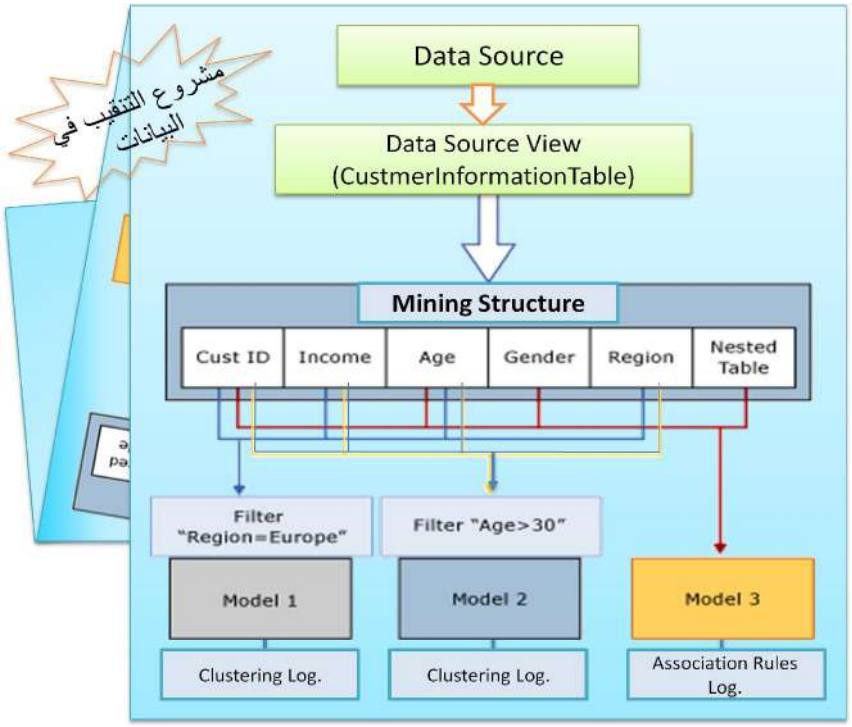
تتألف عملية اكتشاف المعرفة من مشروع تنقيب واحد على الأقل، ويختص مشروع التنقيب بدراسة موضوع معين في البيانات، ويتعلق عادةً بحل مسألة على الأقل من مسائل التنقيب التي طُرحت في المرحلة الأولى. وتتحدد إجراءات إنشاء المشروع بحسب البرنامج الحاسوبي الذي سيُستخدم في التنقيب. ويصنَّف التنقيب في البيانات ضمن مجموعة مشاريع خدمات التحليل التابعة لمشاريع ذكاء الأعمال التي تحتضن أيضاً المعالجة التحليلية الآنية. ويقدم مشروعُ التنقيب القالبَ العام الذي يساعد في انجاز عمليات التنقيب في البيانات، ويتكون هذا المشروع من مجموعة من المكونات، تبدأ بتحديد الاتصال المناسب بمصدر البيانات اللازمة للتنقيب (Data Source) وهو غالباً مستودعُ البيانات الناتج عن المرحلة السابقة، ثم يتم أخذ مرأى من هذا المصدر (Data Source View) عن البيانات المطلوبة للتنقيب سواء أكانت من الجداول أو المرائي التي يتضمنها.

2•4•3 بناء هيكل التنقيب

يحتوي هيكل التنقيب على البيانات (الجداول والخاصيات) التي ستُستخدم في بناء نماذج التنقيب، وهذه البيانات هي من البيانات التي عُرِّفت في مرأى مصدر البيانات في الخطوة السابقة، وتتحدد هنا بعض التفاصيل المرتبطة بطبيعة الجداول والحقول المنتقاة للتنقيب فيها. وتعتبر الاختيارات التي تحصل في هذه الخطوة اختيارات أولية يمكن تغييرها بحسب الحاجة في الخطوات اللاحقة. ومن المهم معرفته هو أن هيكل التنقيب الواحد يسمح بأن يُبنى عليه العديد من نماذج التنقيب التي قد تختلف عن بعضها في واحد على الأقل من مكوناتها أو خواصها، انظر الشكل (3–3)، ومن أبرز هذه الاختلافات:

- اختلاف الخاصيات (الحقول) التي يتم انتقاؤها في كل نموذج من بين الخاصيات المتاحة في هيكل التنقيب.
- اختلاف خوارزميات التنقيب المستخدمة في كل نموذج.
- اختلاف شروط انتقاء (filter) جزء من البيانات المتاحة في الهيكل مثل انتقاء فئة عمرية، أو منطقة سكنية.

- اختلاف مهام حقول نموذج التنقيب، فقد يُعتبر حقل ما من حقول الإدخال في أحد النماذج وفي الآخر يتم تجاهله.



الشكل (3– 3): مكونات مشروع التنقيب في البيانات، ويتكون بشكل أساسي من مصدر البيانات ومرأى مصدر البيانات وهيكل التنقيب ونماذج التنقيب

تتألف عملية بناء هيكل التنقيب من سلسلة خطوات، وتبدأ بإنشاء هيكل عام للتنقيب في البيانات دون التخصص بنموذج تنقيب معين، ثم يجري انتقاء الجداول من مرأى مصدر البيانات، بالإضافة إلى الانتقاء الأولي للحقول التي ستُستخدم في بناء نماذج التنقيب، حيث يمكن لاحقاً إجراء أي تعديل على هذه الحقول (كالحذف، والإضافة، وتغيير الخواص..) بحسب الضرورات التي تظهر أثناء تطبيق نماذج التنقيب المختلفة. كما وتقدر في إحدى هذه الخطوات نسبة بيانات كل من الاختبار (Testing Data) وبيانات التجريب (Training Data)، حيث يُنفَّذُ التنقيب في البيانات في أغلب الأحيان على جزء من البيانات وهي بيانات التجريب، وتُختَبر النتائج على جزء آخر هو بيانات الاختبار. وتبعاً لذلك فمن الممكن أن لا يكون التنقيب في البيانات فعالاً ولا يمكنه اكتشاف الأنماط الموجودة في كامل البيانات في حال كانت بياناتُ التجريبِ لا تمثِّل كاملَ البيانات الضخمة التي سُحبت منها. أي أن عمليات التنقيب في البيانات

ليست طريقة لبرهان عبثي (fool proof) عن أية بيانات، ولكنها طريقة مفيدة فقط عندما تكون بيانات التجريب ممثلة للبيانات بما فيه الكفاية[1]، ويجب بالتالي اختيار نسبة بيانات التجريب بعناية. وهكذا تكون الطلبات الأساسية في بناء الهيكل قد أصبحت جاهزة للخضوع إلى إجراءات التدقيق والبناء والتفعيل من قبل البرنامج الحاسوبي الذي يحتويها.

3•4•3 تصميم نماذج التنقيب

يتكون نموذج التنقيب بشكل أساسي من بيانات ومن خوارزمية تنقيب، حيث يستمد بياناته من هيكل التنقيب ثم يحللها باستخدام إحدى خوارزميات التنقيب. ويتم بناء نموذج التنقيب من خلال الخطوات الأربعة التالية:

1•3•4•3 اختيار حقول البيانات ومهامها في النموذج

يَستخدمُ نموذج التنقيب بعض الحقول المتوفرة في هيكل التنقيب، وتتحدد استخدامات هذه الحقول بحسب مسألة التنقيب المطروحة، ويُستخدم كل حقل لتأدية واحدة من المهام التالية:

أ- التنبؤ فقط (Predict Only)، وتُعطى هذه المهمة للحقول التي تتطلب مسألة التنقيب أن يتم التنبؤ بها. وتحتاج معظم الخوارزميات إلى تسمية حقل واحد على الأقل للتنبؤ به.

ب- الإدخال (Input)، وتُعطى هذه المهمة لجميع الحقول التي يُتوقع أن تساهم في التأثير على قيم المتغير المتنبأ به.

ج- التجاهل (Ignore)، حيث تأخذ بعض الحقول صفة التجاهل عندما لا يكون لها دور مناسب كمدخلات أو كمتغير متنبأ به في النموذج. ومن الممكن أن يحمل حقل ما مهمة الإدخال في نموذج ومهمة التجاهل في نموذج آخر، وهذا يعود إلى تقديرات الشخص الذي يصمم النماذج ويختبرها ويقارن النتائج ويبحث عن معرفة ما ترتبط بمهام الحقول. كما وتُعطى صفة التجاهل أحياناً لبعض الحقول التي ستُستخدم فقط لانتقاء قيم معينة أو فئات معينة من البيانات لكي تُطبَّق عليها عمليات النقيب، فمثلاً لدراسة العوامل المؤثرة على طول مدة الإقامة في المشفى عند مرضى شعبة الجراحة العامة، يتم انتقاء هذه

[1] http://en.wikipedia.org

الشعبة بواسطة الإجراء الخاص بالانتقاء، وإعطاء صفة التجاهل لحقل الشعبة الطبية لأنه لم يعد يحتوي سوى قيمة واحدة هي الشعبة المختارة، ولن يكون له أي فائدة إدخال أو تنبؤ في هذا النموذج.

د- التنبؤ (Predict)، تتعدد أحياناً الحقول المتنبأ بها في نموذج التنقيب، وقد يكون لبعض هذه الحقول القدرة على التأثير كمدخلات على قيم الحقول الأخرى المتنبأ بها، ولذلك تأتي مهمة التنبؤ هذه لتحل هذه المشكلة، وتَمنح المتغير الذي يحملها نوعين من المهام وهما كمتغير متنبأ به، وكمتغير إدخال للحقول الأخرى المتنبأ بها (عندما لا يكون هو الحقل المتنبأ به)، وستتضح هذه الفكرة لاحقاً أثناء التطبيق العملي.

2•3•4•3 تخصيص نموذج التنقيب بمجموعة جزئية من البيانات

تحتاج بعض عمليات التنقيب إلى التعرف على نتائج نماذج التنقيب عند فئة معينة من البيانات، مثل مستوى دخل محدد أو فئة عمرية معينة. وتزود بعض برمجيات التنقيب المستخدمَ بالقدرة على انتقاء المجموعات الجزئية من البيانات من داخل النموذج[1]، وذلك من خلال إجراء الانتقاء (set model filter) الملحق بالنموذج، راجع المثال المطروح في الفقرة (**3•4•3•1**-ج). وهي من المزايا الهامة جداً وكثيراً ما تظهر لها الحاجة أثناء تنفيذ عمليات التنقيب.

3•3•4•3 اختيار خوارزمية التنقيب

تتوفر في البرامج الحاسوبية التي تُعنى بالتنقيب في البيانات الإمكانيةُ لاستخدام عدد من خوارزميات التنقيب، ولكل برنامج طريقته الخاصة في انجاز العمليات الداخلية للخوارزميات، وطريقته الخاصة في عرض النتائج. ونبين فيما يلي لمحة مختصرة عن أهم الخوارزميات الشائعة الاستخدام في التنقيب في قواعد البيانات، والتي استُخدمت في الجانب التطبيقي من هذا البحث، وسيتم التعرف في المبحث الرابع من الفصل الخامس على التفاصيل الدقيقة لمخرجات هذه الخوارزميات في أداة التنقيب المستخدمة.

[1] وهي من المزايا الجديدة في نسخة 2008 لبرامج (SQL Server)، حيث كان من المفترض أن يعود المستخدم في كل مرة إلى مرحلة تهيئة البيانات وينشئ مرأى للمجموعة الجزئية التي يرغب بها، ثم يستدعيه في مرأى مصدر البيانات ويصمم له هيكل جديد ونموذج جديد.

أ- خوارزمية أشجار القرار (Decision Trees Algorithm)

تعتبر خوارزمية أشجار القرار من أشهر خوارزميات التنقيب في البيانات بسبب أدائها السريع على بيانات الاختبار، ودقتها العالية، وسهولة استيعاب الأنماط الناتجة عنها. وتُستخدم هذه الخوارزمية للقيام بعدة مهام أهمها التصنيف. وتطبَّق هذه الخوارزمية على المتغيرات المستمرة (Continues) والمتقطعة (Discrete)، [1]وتقوم الفكرة الأساسية فيها على أساس تقسيم البيانات بشكل متكرر إلى مجموعات جزئية، وتقيَّم كل خاصية إدخال لتحديد كيف تتمكن من تقسيم البيانات بشكل نظيف عبر تقسيمات (أو قيم) الخاصية المتنبأ بها، ويجري تكرار تقييم كل الخاصيات عند كل مجموعة جزئية، وعندما تكتمل هذه العملية التكرارية تكون شجرة القرار قد تشكَّلت.

تقوم خوارزمية أشجار القرار ببناء شجرة قرار مستقلة خاصة بكل حقل من الحقول المتنبأ بها، أي يُنتج النموذج المبني على هذه الخوارزمية أشجارَ قرار بعدد المتغيرات المتنبأ بها.

ب- خوارزمية التجزئة العنقودية (Clustering Algorithm)

تعمل هذه الخوارزمية على إيجاد التصنيفات الكامنة وغير الواضحة في البيانات[2] دون وجود معلومات مسبقة عن الأصناف، وذلك على عكس خوارزمية أشجار القرار التي تصنف البيانات بحسب الخاصيات الموجودة[3]، ويعتبر هذا النوع من التصنيف مفيد لاستعراض واستكشاف المعطيات، وكشف الحالات الشاذة[4]، وأخيراً امتلاك القدرة على إيجاد أو توقع انتماء الحالات الجديدة إلى المجموعات التي تم تعريفها.

تَستخدم خوارزمية التجزئة العنقودية تقنيات التكرار لتصنيف الحالات (cases) الموجودة ضمن بيانات الاختبار، حيث تعمل أولاً على تعريف العلاقات (relationships) الموجودة في هذه البيانات، ثم تقوم بالتالي بتوليد سلسلة من العقد (clusters) استناداً إلى هذه العلاقات، وبعد ذلك تعمل على اختبار قدرات العقد على تمثيل البيانات التابعة لها، وتعاود تشكيل عقد جديدة

---

[1] Cios, K.J.; Pedrycz, W. & Swiniarski, R. W. (2007), **Data Mining: A Knowledge Discovery Approach**, P. 236

[2] The previous reference , P. 292

[3] www.acm.org

[4] http://msdn.microsoft.com/en-us/library/ms175595.aspx

أفضل تمثيلاً للبيانات من السابقة. وهكذا تستمر محاولات التطوير مرات عدة إلى أن يتم التوصل إلى أفضل مجموعات ممكنة، وتظهر نتائج نموذج التنقيب المستند إلى هذه الخوارزمية على شكل أنماط قابلة للتصفح. وإن تصنيف البيانات ضمن عقد أو فئات لا يعني أن كل عقدة تختلف بكامل محتوياتها عن العقد الأخرى، وإنما هنالك تقارب أو تشابه بين بعض العقد، وهذا ما سيظهر بشكل أوضح أثناء التطبيق العملي في هذه الدراسة.

ج- خوارزمية نايف بايز (Naive Bayes Algorithm)

تُعتبر هذه الخوارزمية إحدى خوارزميات التصنيف (Classification) التي تُستخدم في النماذج التنبؤية. وتقوم هذه الخوارزمية بحساب الاحتمال الشرطي بين حقول المدخلات والحقول التنبؤية، حيث تفترض أن حقول المدخلات مُستقلة عن الحقول التنبؤية. وسميت بهذا الاسم (خوارزمية بايز البسيطة) لأنها تستخدم نظرية بايز في الاحتمالات، ولكنها تفترض عدم تبعية حقول المخرجات إلى حقول المدخلات. وتبعاً لذلك فهذه الخوارزمية أقل قوة من الخوارزميات الأخرى، وتُستخدم غالباً في استكشاف البيانات، والتعرف في كل قيمة من قيم الحقل المتنبأ به على احتمالات حدوث القيم في كل حقل من حقول المدخلات.

د- خوارزمية قواعد الاقتران (Association Algorithm)

تتبع هذه الخوارزمية إلى خوارزميات الاقتران، ويُستفاد منها في الحصول على أداة توصية مفيدة وخصوصاً في مجال التسويق. وتعمل هذه الخوارزمية على اجتياز وفحص بيانات الاختبار بقصد العثور على العناصر التي تظهر معاً، وينتج عنها قواعد اقتران يتألف كل منها من طرفين، الأيسر يتضمن الشرط والأيمن يحتوي على النتيجة. وتَستخدم هذه الخوارزمية محددين (Parameters) هما الدعم (Support) والثقة (Confidence) لوصف مجموعات العناصر والقواعد التي أنتجتها. ويعبِّر الدعم عن تكرار ظهور جميع عناصر القاعدة معاً في بيانات الاختبار، بينما تُمثل الثقة – وتسمى الاحتمال (Probability)– احتمال تحقق القاعدة، وتحسب بقسمة تكرار ظهور جميع العناصر الواردة في طرفي القاعدة على تكرار العناصر في طرف الشرط، وكلما ازدادت قيمة هذين المحددين ازدادت أهمية القاعدة المكتشفة. ويعمل نموذج التنقيب على إظهار أهم قواعد الاقتران المكتشفة وليس الكل، وذلك لأن هذه الخوارزمية تشترط أن تحقق القواعد المكتشفة الحد الأدنى المقبول لكل من هذين المحددين،

مما يتحكم بعدد القواعد التي سيظهرها النموذج، فكلما ارتفعا قلت القواعد التي ستظهر. ونبين فيما يلي مثالاً توضيحياً عن قواعد الاقتران التي أتنتجها هذه الخوارزمية:

If customer buys potatoes (A) and pickles (B) then→ customer buys sausage (Z)
(Support = 40%, Confidence = 80%)
Support: How often do potatoes, pickles and sausage occur together (A,B,Z)
Confidence (A,B → Z) = Support (A,B,Z) ÷ Support(A,B)

تم بيع البطاطا والمخلل معاً في 100 عملية بيع من أصل 200 عملية بيع في البيانات المدروسة ، وتم بيع السجق معهما في 80 عملية من هذه العمليات.

ويعتبر مفهوم تحليل سلة التسوق (Market Basket Analysis) حالة خاصة من خوارزمية قواعد الاقتران (Association Rules) ، ولكنه يختلف عنها من حيث أنها تُنتج قواعد اقتران بغض النظر عن الأسعار، أما هو فيَبْني قراراته استناداً إلى الأسعار ويَهدف إلى اقتراح المزيج الأكثر ربحية.

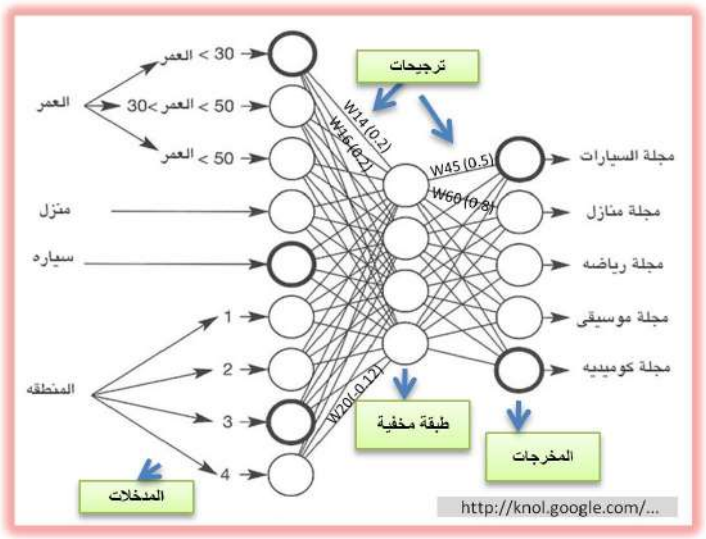
ه- خوارزمية الشبكات العصبونية[1] (Neural Network Algorithm)

يحاكي أسلوب هذه الخوارزمية طريقة عمل العقل البشري، فعندما يصادف العقل البشري مشكلة ما، فإنه يقوم أولاً بتحليلها ثم يعطي ترجيحات معينة للحقائق المحيطة بها بحسب إدراكه لها، وتوضع بعد ذلك هذه الحقائق في مسارات عصبية لتعمل على غربلتها استناداً إلى الأنماط المعروفة سابقاً بالنسبة لها والتي تكونت بالخبرة. وتقود هذه العمليات إما إلى اقتراح العقل لحلول ممكنة للمشكلة المطروحة، أو يضيف الحقائق الناتجة إلى الحقائق القديمة وتعاد غربلتها في المسارات العصبية ثانيةً وهكذا[2]..

تقوم خوارزمية الشبكات العصبونية ببناء نماذج تصنيف (Classification) وانحدار (Regression) من خلال بناء شبكة مكونة من ثلاثة طبقات على الأقل من الخلايا العصبونية (Neurons)، وهي طبقة المدخلات وطبقة مخفية (Hidden) واحدة على الأقل وطبقة مخرجات. ويتم في طبقة المدخلات تعريف جميع قيم خاصيات الادخال في نموذج التنقيب، واحتمالات كل منها. أما الطبقة الخفية فهي تستقبل المدخلات من الخلايا العصبونية للمدخلات

[1] تُعتبر الشبكات العصبونية إحدى أنواع نُظُم الذكاء الصنعي التي تُعنى بتمكين الحواسيب من محاكاة مظاهر من الذكاء الإنساني.
[2] Cios, K.J.; Pedrycz, W. & Swiniarski, R. W. (2007), **Data Mining: A Knowledge Discovery Approach**, P. 371

وتقوم بتحديد الترجيحات المناسبة لهذه المدخلات، ثم تزود هذه الطبقة طبقة المخرجات بالقيم المناسبة. وكلما كانت قيم الترجيحات أكبر كلما دل على كبر أهمية قيمة المدخلات. وفي حالات خاصة قد تكون قيم الترجيحات سلبية مما يدل على ضعف تأثير هذه المدخلات ويمكن بالتالي تجاهلها. وتمثل طبقة المخرجات قيم الخاصيات المتنبأ بها لنموذج التنقيب والمقابلة لكل قيمة من قيم خاصيات المدخلات، وتختلف طريقة عرض نتائج التنقيب أو الأنماط المكتشفة باختلاف برنامج التنقيب المستخدم، ويوضح الشكل (3-4) آلية عمل هذه الخوارزمية.



الشكل (3- 4): مثال عن آلية عمل خوارزمية الشبكات العصبونية

و- خوارزمية الانحدار المنطقي (Logistic Regression Algorithm)

تعمل خوارزمية الانحدار المنطقي[1] بطريقة مبسَّطة تشبه طريقة عمل خوارزمية الشبكات العصبونية ولكنها تقوم بقياس تأثير كل من المدخلات على المخرجات وتُعطي ترجيحات للمدخلات مباشرة دون وجود طبقة مخفية. وبشكل عام تُعتبر دقة خوارزمية الانحدار المنطقي أقل من دقة خوارزمية الشبكات العصبونية[2]، ولكن لكل حالة وضعها الخاص، ويتحدد النموذج الأفضل بالاستعانة بأدوات أو مقاييس الدقة التي تُستخدم في المرحلة الخامسة من مراحل اكتشاف المعرفة.

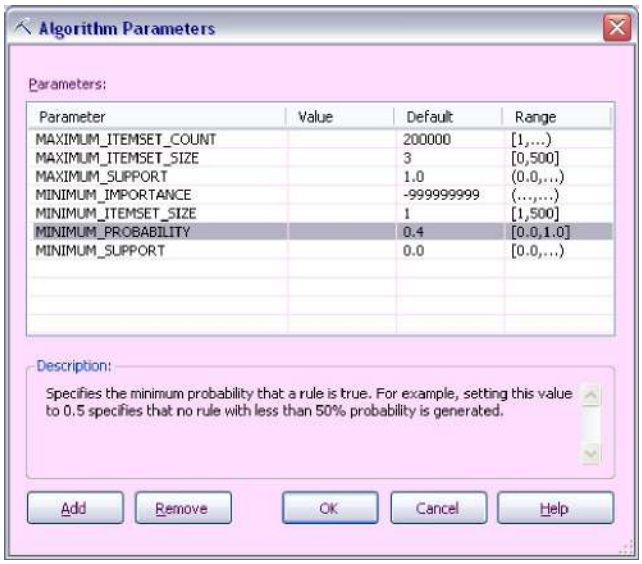
[1] يُعتبر الانحدار المنطقي تقنية إحصائية شائعة تُستخدم لتحديد تأثير العوامل المتعددة على المخرجات الثنائية (Yes-No).

[2] The previous reference, P. 396

جرى التركيز في هذه الفقرة على النقاط الأساسية في كل خوارزمية من خوارزميات التنقيب، وعُرضت في كل منها التفاصيل التي تميزها والتي تَصبُّ في مجالات اهتمام الدراسة الحالية، كما تم تجنب عرض الرسوم البيانية للخوارزميات باستثناء خوارزمية الشبكات العصبونية، لأن تلك الرسوم تسبب الكثير من الإرباك في ربط الأساس النظري مع الأساس التطبيقي، حيث أن مخرجات الخوارزميات في نماذج التنقيب تختلف كثيراً من الناحية التطبيقية عن الرسوم البيانية الموجودة في المراجع النظرية التي تتحدث عن خوارزميات التنقيب في البيانات.

4•3•4•3 تعديل محددات خوارزمية النموذج (Parameters) في حالات الضرورة

يمكن التحكم بعتبات أي من خوارزميات التنقيب وانقساماتها وبعض ظروفها، إذا احتاج الأمر لتحسين نتائج التنقيب. وتوجد هذه العناصر في قائمة المحددات التابعة لكل خوارزمية والتي يمكن استدعاؤها من داخل نموذج التنقيب الذي يحتوي على هذه الخوارزمية. وتختلف المحددات باختلاف الخوارزمية، وتحتوي كل قائمة على قيم افتراضية تتناسب تقريباً مع الكثير من عمليات التنقيب، وقلما يحتاج المستخدم إلى تغييرها. ويبين الشكل (3–5) قائمة محددات خوارزمية قواعد الاقتران، وتتضمن العتبات العليا والصغرى التي تتحكم بعدد قواعد الاقتران التي سيُظهرها النموذج، فمثلاً حُددت القيمة الافتراضية للاحتمال الأدنى المقبول للقواعد بقيمة 0.40، ويمكن تعديل هذه القيمة إلى 0.80 وذلك للتقليل من عدد القواعد الناتجة، وهي الحالة التي يرغب فيها المستخدم بالحصول على القواعد العالية الثقة فقط وبتجاهل البقية.



الشكل (3– 5): قائمة محددات خوارزمية قواعد الاقتران

4•4•3 استعراض مخرجات نماذج التنقيب

تتيح برامج التنقيب في البيانات إمكانية استعراض مخرجات نماذج التنقيب مباشرة في إحدى نوافذ البرنامج، وتختلف طريقة العرض باختلاف الخوارزمية، وباختلاف برنامج التنقيب المستخدم، ويترافق غالباً استعراض نتائج النموذج مع عملية تصميمه. وتبعاً لذلك يتم غالباً دمج خطوتي التصميم والاستعراض معاً أثناء التعرف على نماذج التنقيب المدروسة. كما ويمكن الاستعانة ببرامج حاسوبية إضافية لعرض مخرجات نماذج التنقيب بشكل غير مباشر مثل برامج بناء التقارير وبرنامج اكسل وبرنامج (Microsoft Visio).

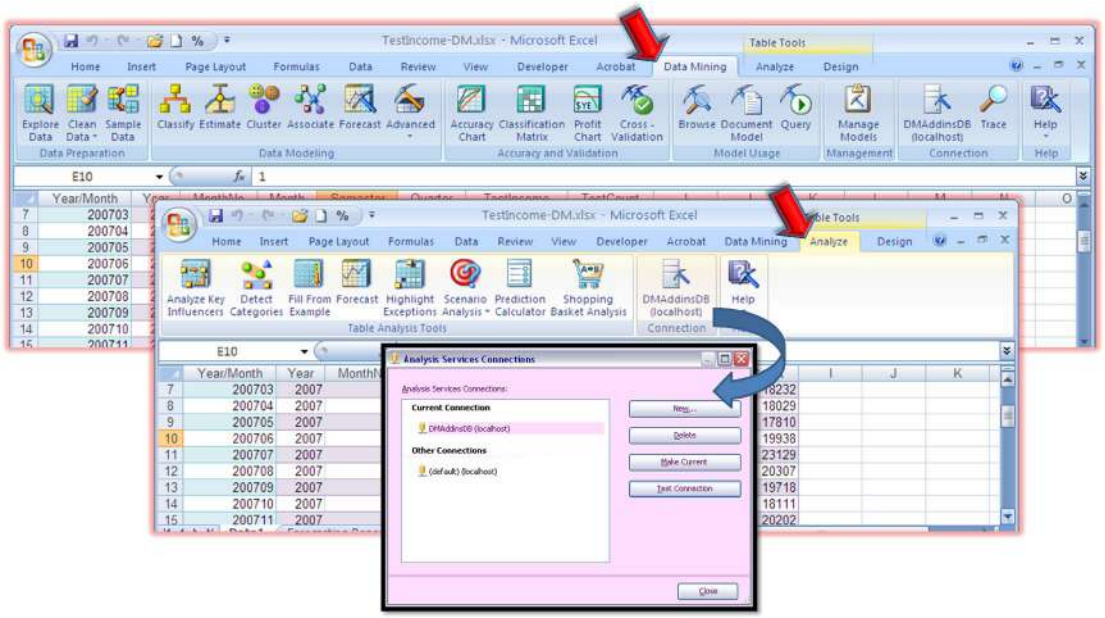
5•4•3 استخدام اكسل في عمليات التنقيب في البيانات[1]

قامت شركة مايكروسوفت بتزويد برنامج اكسل ابتداءً من نسخة 2007 بالقدرة على تنفيذ بعض عمليات التنقيب في البيانات، من خلال الوظائف المضافة[2] للتنقيب في البيانات عبر مخدم لغة الاستفسار المهيكلة (SQL Server 2008 Data Mining Add-ins). ويمكن للمستخدم أن يحصل على هذه الوظائف من تلك الشركة، ويضيفها إلى برنامج اكسل ، ويُفعِّل الرابط بين اكسل ومخدم (Server SQL) الذي يُشترط أن يكون متواجداً على حاسوب المستخدم أو على مخدم الشبكة المرتبط بها. ويؤدي تفعيل هذه الوظائف إلى انضمام قائمتين جديدتين للتنقيب إلى مجموعة قوائم برنامج اكسل الأساسية . وبهذا يمكن القيام بالتنقيب باستخدام واجهات اكسل المرتبطة داخلياً مع تقنيات التنقيب في (SQL Server)، انظر الشكل (3–6). وتتصف عمليات التنقيب في اكسل بالسهولة والوضوح في بناء النماذج، كما تُعرض النتائج بطرق جذابة وصديقة للمستخدم ومرنة في التنقل بين مفرداتها، ويمكن الاستفادة مباشرةً من هذه النتائج بتحليلها ودراستها إحصائياً وتحويلها بمرونة إلى رسوم بيانية ، وباستخدامها في عمل التنبؤات وبناء التقارير..الخ. ولكن يتطلب استخدام اكسل في التنقيب أن يكون المستخدم ملماً بأسس التنقيب ومراحله، وإلا سيكون من الصعب التعرف على المدخلات وآلية العمل واستيعاب النتائج.

---

[1] Microsoft SQL Server (2008) Data Mining Add-Ins for Office 2007 Help

[2] تُعرف الوظائف المضافة (add-ins) بأنها عبارة عن وظائف تكميلية تضيف أوامر أو سمات معينة إلى برامج موجودة، وتساعد هذه الوظائف على توسيع إمكانيات هذه البرامج لإجراء عمليات إضافية إلى جانب عملياتها التقليدية المعتادة.



الشكل (3- 6): قائمتي التنقيب في البيانات المضافتين إلى برنامج اكسل (Excel)،
ونافذة الربط مع مخدم (SQL Server) على مخدم الحاسوب المحلي

5•3 مرحلة تقييم المعرفة المكتشفة (Evaluation of the Knowledge Discovered)

تأتي مرحلة تقييم المعرفة المكتشفة بعد مرحلة الحصول على نتائج التنقيب، وكثيراً ما تكثر التغذية الراجعة بين هاتين المرحلتين من أجل تطوير النماذج وتحسين المعرفة المكتشفة. وهي المرحلة الأكثر تشويقاً حيث تظهر فيها خلاصة الأعمال والجهود السابقة جميعاً. وتتكون مرحلة تقييم المعرفة المكتشفة من عدة محاور كما يلي:

أ- استيعاب النتائج، والتحقق من أن نتائج التنقيب استطاعت أن تجيب على مسائل التنقيب المطروحة في المرحلة الأولى، ويمكن الاستعانة بالخبراء في المجال المدروس للمساعدة في تفسير النتائج.

ب- التحقق من أن المعارف المكتشفة جديدة ومبتكرة ومفيدة، ويمكن أن تُعتبر المعرفة المكتشفة مفيدة وهامة حتى لو كانت غير جديدة، وذلك عند التطبيق الحديث والأول لتقانة اكتشاف المعرفة على البيانات المدروسة، حيث يكون من الأفضل أن يتم التعرف على نجاح عملية التنقيب من خلال قدرتها على برهان العلاقات والحقائق المعروفة، وهذا ما يزيد الثقة بها وبقدرتها على تقديم معارف جديدة.

ج- استخدام أدوات قياس دقة التنبؤ للنماذج المكتشفة، واختيار النموذج الأكفأ. وتقاس دقة التنبؤ عادةً من خلال تطبيق الأنماط المكتشفة على مجموعة بيانات الاختبار التي

احتُجزت سابقاً لهذه الغاية عند بناء هياكل التنقيب، فينتج قيمة متنبأ بها مقابلة لكل سجل من سجلات بيانات الاختبار، ثم تقوم أدوات قياس الدقة بمقارنة القيم الحقيقية مع القيم المتنبأ بها وتعرض نتائج التقييم بأشكال مختلفة . ويتمكن المستخدم بالتالي من قراءة النتائج ومقارنتها والتعرف على النموذج الأكفأ في التنبؤ، ومن أمثلة هذه الأدوات مخطط الارتقاء (Lift Chart)، ومصفوفة التصنيف (Classification Matrix) المعتمدتين في برامج (SQL Server).

د- استخدام أدوات قياس المنفعة المادية المتحققة أو التكاليف المترتبة على اختيار كل نموذج تنقيب، ويرتبط استخدام هذه الأدوات عادة بمجالات التسويق، ومن هذه الأدوات مخطط المنفعة (Profit Chart).

ه- الأخذ بعين الاعتبار طبيعة المسألة المدروسة عند اختيار النموذج الأكفأ من حيث دقة التنبؤ.

و- تطوير النماذج الموافق عليها من ناحية دقة التنبؤ، حيث يجرى تفحصها مرة ثانية للتعرف على الإجراءات الممكن اتخاذها لتحسين النتائج أكثر وأكثر[1].

6•3 مرحلة استخدام المعرفة المكتشفة (Use of the Discovered Knowledge)

تبقى المعارف المكتشفة عاطلة عن العمل ما لم تجد مكاناً مناسباً لاستثمارها، وترتبط استخدامات المعرفة المكتشفة بالمسألة المدروسة والأهداف المرغوبة. ويستفاد من المعرفة المكتشفة في نفس المجال أو في مجالات أخرى. وقد تأخذ شكلاً واحداً في الاستخدام أو عدة أشكال كما هو آت:

أ- كتابة تقارير (Reports) عن نتائج عمليات التنقيب.

ب- بناء استعلامات جديدة من نماذج التنقيب.

ج- توصيف البيانات المدروسة بطرق مختلفة بحسب المسألة المدروسة.

د- اقتراحات عمل (Action suggestions) [2] على أساس النتائج.

ه- إنشاء قوالب أو أدوات للاستعمال في سيناريوهات التحليل المماثلة.

---

[1] Cios, K.J.; Pedrycz, W. & Swiniarski, R. W. (2007), **Data Mining: A Knowledge Discovery Approach**, P.16

[2] people.dsv.su.se/~benkt/2I1224/oh2.ppt

و- التنبؤ بالاتجاهات المستقبلية.

ز- تزويد أنظمة عمل قواعد البيانات الأساسية بآليات مراقبة (Monitoring mechanisms) تُدرك المتغيرات التي شملها التنقيب ، وتصدر تبليغات عندما تلاحظ ظهور ظروف معينة مشابهة في البيانات[1]. مثل التحذير من فشل شبكات الاتصال، والتحذير من المحتالين على البطاقات الائتمانية.

7•3 التغذية الراجعة (Feed Back)

تنشأ الحاجة إلى التغذية الراجعة عندما تتوقف مرحلة ما عند حد معين، وتحتاج فيه إلى التزود بأشياء جديدة من المراحل السابقة، أو إلى إجراء تعديلات عليها ..، ونورد فيما يلي بعضاً من أسباب التغذية الراجعة في مراحل اكتشاف المعرفة[2]:

أ- ينتج عن مرحلة فهم البيانات الحاجة أحياناً إلى معرفة إضافية عن المجال المدروس، من أجل استيعاب البيانات بصورة أفضل.

ب- ينجم عن مرحلة إعداد البيانات الحاجة إلى فهم البيانات بصورة أوسع، من أجل المساعدة في اختيار خوارزميات إعداد البيانات.

ج- تظهر عند التنقيب في البيانات الحاجة للعودة إلى المرحلة الأولى، حيث قد تحدث نتائج تنقيب غير مرضية، ولا يمكن تحسينها، مما يستدعي تغيير أهداف اكتشاف المعرفة الموضوعة في المرحلة الأولى.

د- قد يحتاج التنقيب في البيانات للعودة إلى مرحلة فهم البيانات، حيث يؤدي التعريف الضعيف للبيانات إلى اختيار طرق تنقيب خاطئة، وبالتالي التوصل إلى نتائج فاشلة، وهنا لابد من العودة إلى خطوة تعريف البيانات.

ه- تحتاج طرق التنقيب في البيانات إلى متطلبات خاصة في البيانات قد لا تكون معروفة أثناء إعداد البيانات، لذا قد يحتاج الأمر لاحقاً إلى إضافة إعدادات جديدة إلى البيانات.

و- قد تتم العودة من مرحلة تقييم المعرفة المكتشفة إلى المرحلة الأولى، وذلك بسبب عدم فعالية المعرفة المكتشفة الناتج عن سوء فهم مجال التطبيق، والتصميم الخاطئ لمشكلة

---

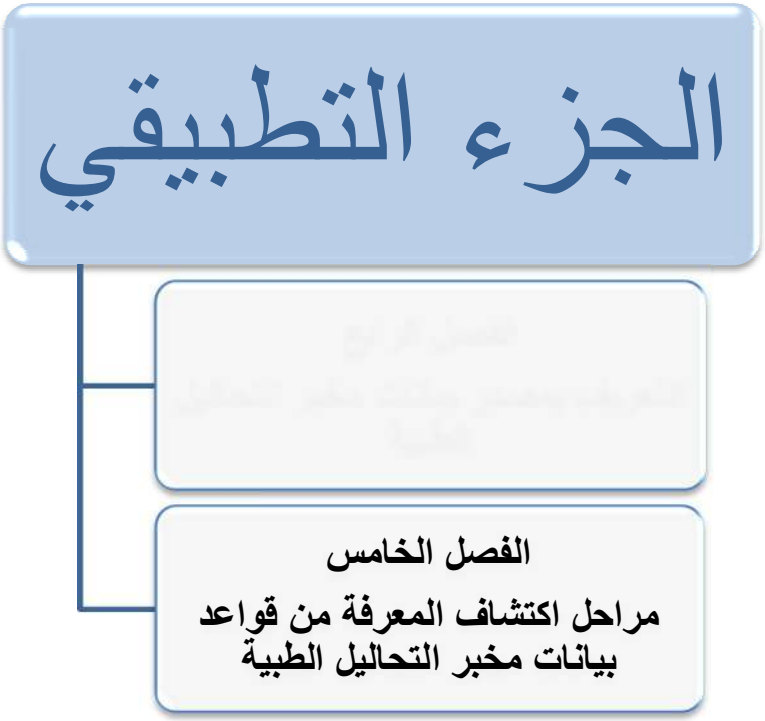
[1] The previous reference

[2] Cios, K.J.; Pedrycz, W. & Swiniarski, R. W. (2007), **Data Mining: A Knowledge Discovery Approach**, P.16

التنقيب (القيود، الاحتياجات والأهداف)، وفي هذه الحالة يجب إعادة صياغة مضمون مسألة اكتشاف المعرفة.

ز- تظهر الحاجة للعودة من مرحلة تقييم المعرفة المكتشفة إلى مرحلة التنقيب من أجل تطوير النماذج المكتشفة. كما يحدث أحياناً أن تكون المعرفة المكتشفة ليست مفيدة أو ليست هامة أو ليست بالجديدة، وفي هذه الحالة يمكن الاستعانة بأداة (برنامج حاسوبي) أخرى للتنقيب في البيانات المدروسة وتكرر فقط مرحلتي التنقيب والتقييم للتأكد من جودة النتائج المكتشفة كإجراء أولي وسريع قبل الاضطرار إلى إعادة المراحل السابقة جميعاً.

تتسع الأسس النظرية لعملية اكتشاف المعرفة كثيراً، وقد جرت المحاولة في هذا الفصل لإعطاء فكرة أولية عن طريقة العمل التي سيتم إتباعها في الفصل الخامس من هذه الدراسة، وذلك دون الخوض – قدر الإمكان– في التفاصيل التي تسبب التكرار في طرح الأفكار بين هذين الفصلين.


الجزء التطبيقي
الفصل الخامس
مراحل اكتشاف المعرفة من قواعد بيانات مخبر التحاليل الطبية

الفصل الرابع

التعريف بمصدر بيانات
مخبر التحاليل الطبية

الفصل الرابع

التعريف بمصدر بيانات مخبر التحاليل الطبية

مقدمة

يُعتبر مخبر التحاليل الطبية من وجهة نظر البحث الحالي على أنه نظام يقوم بتوليد البيانات. ولا بد للتعرف على هذا المصدر بحسب اهتمام البحث الحالي من الإلمام ببعض المفاهيم المرتبطة بنظام المخبر، والتعرف على آلية إنتاج البيانات المتعلقة بالتحاليل الطبية، والاطلاع على الدراسات التقليدية والحديثة التي تَستخدم البيانات الناتجة عن مخابر التحاليل الطبية. ويحتوي هذا الفصل على الكثير من المفردات النظرية التي كان من الممكن أن تُدرج ضمن الجزء النظري من هذا البحث، ولكنها في الحقيقة شكَّلت الأساس المعرفي الذي يخص المرحلة الأولى من عملية اكتشاف المعرفة من قواعد بيانات مخابر التحاليل الطبية، مما جعلها تنتمي إلى الجزء التطبيقي.

1•4 مفاهيم أساسية

1•1•4 مخبر التحاليل الطبية

يُعرَّف مخبر التحاليل الطبية بأنه مكان مؤلف من مجموعة من الأدوات والمواد والأجهزة المختلفة التي تُستخدم في إجراء التحاليل أو الفحوصات المخبرية المختلفة التي يحددها الطبيب المعالج للمريض، وتُجرى التحاليل الطبية على عينات من أحد سوائل جسم الشخص المريض ومفرزاته كالدم والبول والسائل الدماغي الشوكي واللعاب والقشع أو أجزاء من جسمه كالأنسجة. وتقدَّم نتائج التحاليل الطبية عادةً إلى الطبيب المعالج لتساعده في تشخيص المشاكل المرضية التي يعاني منها المريض، وفي مراقبة استجابته لبعض العلاجات، وفي التأكد من تحسن حالته الصحية. ويشرف عادةً على العمل في مخبر التحاليل الطبية إما طبيب متخصص بالطب المخبري بعد تخرجه الجامعي أو صيدلي متخصص بالتشخيص المخبري بعد تخرجه الجامعي.

وقد تتخصص المخابر الطبية ببعض أنواع التحاليل الطبية كما في المخابر الصـــغيرة، أو قـــد تشتمل على كل أنواع التحاليل كما في المخابر الكبيرة ومخابر المستشفيات الكبيـــرة. ويبـــين الشكل (4–1) مجموعة من الصور لأدوات وأجهزة مخابر التحاليل الطبية.



الشكل (4- 1): صور لبعض أدوات وأجهزة مخبر التحاليل الطبية

2•1•4 بيانات مخبر التحاليل الطبية

يُقصد ببيانات مخابر التحاليل الطبية تلك البيانات التي تُعبِّر عن التحاليل الطبية التي أجريـــت في المخبر من حيث أسماء التحاليل وقيمها الطبيعية المرتبطة بالعمر والجـــنس، وتصـــنيفاتها بحسب مجموعات التحاليل أو الأمراض، ونتائج إجراء هذه التحاليل الطبيـــة علـــى العينـــات المأخوذة من المرضى، والبيانات الشخصية والأسباب المرضية الداعيـــة لإجـــراء التحاليـــل، وتكاليف التحاليل المترتبة على المرضى...الخ

3•1•4 السجل الطبي الالكتروني (Electronic Medical Record (EMR))

هو توثيق الكتروني منظم للتاريخ الطبي للمرضى في قواعد بيانات، ويشتمل على البيانات والمعلومات المتعلقة بصحة الأفراد في مركز واحد للرعاية الصحية، ويدار من قبل فريق

مختص في هذا المركز[1]. ويتضمن السجل الطبي بيانات شخصية عن المرضى (اسم، عمر، جنس، مكان الإقامة، الحالة الاجتماعية..)، وبيانات طبية عن حالة المرضى والخدمات التي تُقدم لهم (الأمراض، العلاج، التحاليل الطبية، الأدوية..). ويستفاد من السجلات الطبية الالكترونية في تبادل المعلومات الطبية بين أقسام مركز الرعاية الصحية الكترونياً وليس ورقياً، وفي توثيق ورقابة الخدمات التي تُقدَّم للمرضى، وفي تتبع الأوضاع الصحية للمرضى باستمرار. ويساعد في تمكين بعض الجهات المختصة من الاطلاع على صحة المرضى في هذا المركز، كما يساهم أيضاً في حماية المصلحة القانونية للمريض ولمقدمي الخدمات الطبية. وبالإضافة إلى ذلك فإن السجل الطبي يعمل كوثيقة لتعليم طلاب الطب كيفية التعامل مع المرضى ومراحل علاجهم، ويُستفاد أيضاً من البيانات الموجودة في هذه السجلات في تتبع حالات المرضى عند إجراء بعض الأبحاث الطبية[2].

4•1•4 نظام المعلومات الطبية في المشفى (Hospital Information System (HIS))

هو نظام معلومات شامل ومتكامل يُستخدم في إدارة المعلومات الإدارية والطبية في المستشفيات، ويهدف إلى تمكين المشفى والعاملين فيها من تغطية الجوانب الطبية والمالية والإدارية للمشفى ومن تقديم أفضل وأسرع رعاية للمرضى.

ويتكون نظام معلومات المشفى من عدة أنظمة فرعية، منها نظام السجلات الطبية ونظام الأشعة ونظام المخابر الطبية ونظام الصيدلية ونظام العيادات الخارجية ونظام العاملين ونظام التموين الطبي..، حيث ترتبط الأنظمة الفرعية معاً وتتكامل في العمل والمعلومات من خلال تدفق وتبادل البيانات آلياً فيما بينها.

وتتشارك الأنظمة الفرعية بقاعدة البيانات التي تحتوي على بيانات السجل الطبي الالكتروني، ويتمكن الأطباء في هذه الأنظمة من الاطلاع على السجل الطبي لمرضاهم بمنتهى الدقة والسرعة والمحافظة على سرية المعلومات دون عناء انتظار وصول السجل الطبي من قسم السجلات الطبية، وأيضا التعامل مع سجلات المرضى الالكترونية بإعطاء أوامر إجراء الفحوصات (المخبرية والإشعاعية) اللازمة دون استخدام النماذج الورقية، ويترتب بذلك على المريض فقط مراجعة القسم المعني سواء قسم المخبر أو الأشعة وإعطاء رقم ملفه الطبي ليتم

[1] http://healthit.hhs.gov/ (By The Office Of The National Coordinator (ONC))
[2] http://en.wikipedia.org

إجراء التحليل الطبي أو التصوير الإشعاعي اللازم له. ويُسهِّل هذا النظام من صرف الأدوية من الصيدلية حيث يتيح للطبيب معرفة الأدوية المتوفرة والموجودة في الصيدلية، وفي حالة عدم تواجد الدواء المطلوب تَظهر الأدوية البديلة ليختار منها الطبيب ما يناسب المريض، مما يعني تجنيب المريض الانتظار ومعاودة الذهاب والإياب بين الطبيب و الصيدلية [1].

ويحقق هذا النظامُ المتكاملُ العديدَ من المزايا، أهمها أنه يرفع من أداء إدارة المشفى، ويوفر من الجهود المبذولة في التوثيق الالكتروني للبيانات في كل قسم، ويقلص من الوثائق الورقية اللازمة لتبادل البيانات، كما أنه يضمن المتابعة الالكترونية الدقيقة لحالة المرضى والخدمات الطبية التي تقدم لهم خلال كل مراحل تواجدهم في المشفى. ويعتمد نجاح النظام على كفاءته في تخزين واستدعاء مختلف المعلومات المتعلقة بأنشطة الرعاية الصحية في الوقت والمكان المناسبين، وتتوفر في بعض الأنظمة المتطورة أجهزةُ طبية تتصل بالنظام المعلوماتي وترفده بالنتائج مباشرةً وآلياً.

وعلى الرغم من الأهمية الكبيرة المرتبطة بوجود نظام معلومات المشفى فإن الأهمية تكبر وتتوسع عندما يطبق هذا النظام على المستوى الوطني في جميع المشافي، حيث ترتبط كل أنظمة معلومات المشافي معاً ضمن نظام المعلومات الصحي، وبالتالي تتشارك بالبيانات الطبية الخاصة بالمريض كلُ الأطراف المعنية بذلك وليس جهة واحدة.

5•1•4 سجل الصحة الالكتروني (Electronic Health Record (EHR))

هو سجل طبي موحد[2] يحتوي على مجموعة منظمة من المعلومات الصحية الالكترونية حول أفراد السكان منذ ولادتهم، وهو قابل للوصول والمشاركة عبر كافة المراكز الصحية المختلفة من خلال تضمينه في أنظمة المعلومات ذات الارتباط الشبكي. وتتضمن مثل هذه السجلات بيانات بشكل تفصيلي أو تلخيصي عن المرضى وتتعلق بالخصائص السكانية، والتاريخ الطبي، والتداويات والحساسيات، وحالة اللقاحات، ونتائج التحاليل الطبية، والصور الشعاعية، والمعلومات المالية المتعلقة بالفواتير الطبية[3]، وتدار سجلات الصحة الالكترونية مركزياً وتشاركياً بين كل المنظمات المشمولة وذلك من قبل فرق مختصة بالصحة والرعاية، ويمكن

[1] http://www.alriyadh.com/2007/05/01/article246118.html
[2] http://healthit.hhs.gov/ (By The Office Of The National Coordinator (ONC))
[3] http://en.wikipedia.org

أن يستمد سجل الصحة الالكتروني البيانات السكانية من السجل المدني للسكان، وبالتالي فإن سجل الصحة الالكتروني أكثر شموليةً من السجل الطبي، وذلك لأنه يتضمن البيانات الصحية لكافة السكان منذ الولادة وفي كل المراكز الصحية والطبية التي زودتهم بالخدمات، كما أن ارتباطه المباشر بالبيانات السكانية يوفر المعرفة الكاملة عن الأوضاع الصحية لعائلة المريض وأقربائه.

6•1•4 تقانة المعلومات الصحية (Health Information Technology (HIT))

هي التقانة التي تحتضن سجلات الصحة الالكترونية وهي المسؤولة عن الإدارة الشاملة للمعلومات الصحية وتبادلها الآمن بين الأفراد ومقدمي الخدمات الطبية والجهات الحكومية، وشركات التأمين[1]. وينظر إلى هذه التقانة على أنها الأداة الواعدة في تحسين النوعية العامة والأمان والكفاءة لأنظمة تبادل البيانات الصحية. حيث أنها تساهم في تحسين نوعية الرعاية الصحية، وتمنع الأخطاء الطبية، وتخفف من تكاليف الرعاية الصحية، وتدعم القرار الطبي، وتحسِّن من كفاءة الإدارات، وتخفف من العمل الورقي، وتساهم في تقييم أداء الرعاية الصحية على المستوى الوطني.

وتفيد هذه التقانة في تزويد الأبحاث الطبية والصحية الشاملة للسكان بكم هائل من البيانات لدراستها والتنقيب فيها واستخراج المعلومات والمعارف الهامة الكامنة. وتحمل هذه التقانة بالتالي فوائد طبية عظيمة للسكان حيث أنها تساهم في تقديم المعلومات الكفيلة بالكشف المبكر عن تفشي الأمراض المعدية في مختلف أنحاء البلاد، أو حتى تمركز بعض الأمراض في أماكن معينة، مما يساعد في اتخاذ التدابير الوقائية أو التخفيفية من الإصابات. وتسمح هذه التقانة بالوصول إلى كافة السكان واستهداف فئات منهم وحثهم على إجراء فحوصات وتحاليل معينة بصورة دورية أو في أعمار معينة للتأكد من سلامتهم. فمثلاً تساهم عدة عوامل في نشوء اعتلال الكلية المزمن مثل مرض السكري غير المراقب، وارتفاع الضغط الشرياني، والتأثيرات الجانبية لبعض الأدوية، وارتفاع العمر عن السبعين عاماً، وتأثير العوامل الوراثية..الخ. وتبعاً لذلك فإن الدول المتقدمة تتعرف إلكترونياً على السكان المعرضين لاعتلال الكلية المزمن وتحثهم على الإجراء الدوري للتحاليل المخبرية المناسبة[2].

---

[1] المرجع السابق

[2] د. بشور، نهيد ، 2010، **"اعتلال الكلية المزمن"**

أما في سورية فيوجد برنامج لتحديث القطاع الصحي وبناء نظام معلومات ممول من الاتحاد الأوربي، حيث بُدأ منذ بداية عام 2009 بالعمل على أتمتة ثلاثة مستشفيات حكومية هي مشفى دمشق الوطني ومشفى درعا الوطني ومشفى الحفة في مدينة اللاذقية[1]، وذلك بهدف الوصول إلى ملف طبي الكتروني وأتمتة كافة الأقسام والشعب الطبية والإدارية في كل من هذه المشافي. وقد أُطلق هذا المشروع إلى حيز التطبيق بتاريخ 2010/12/8،[2] وستعمل أنظمة المعلومات المصممة بصورة مستقلة في كل من هذه المشافي في المراحل الأولى للتطبيق، وسيتم العمل فيما بعد على ربطها معاً، ويمكن التوسع مستقبلاً بتعميم هذه التجربة على المشافي الحكومية الأخرى في سورية.

2•4 آلية تجميع وتخزين البيانات في مخبر التحاليل الطبية

تعتمد مخابر التحاليل الطبية في الوقت الحالي على استخدام برامج حاسوبية مختلفة لإدارة أعمالها، وتصنف هذه البرامج الحاسوبية بحسب المهام التي تقوم بها إلى الأنماط التالية:

أ- البرامج الالكترونية البسيطة

هي البرامج الشائعة التي تلبي فقط احتياجات طباعة نتائج التحاليل والفواتير للمرضى، فمثلاً تقوم بعض المخابر بطباعة نتائج التحاليل على برنامج معالجة النصوص (Word)، وهي طريقة قديمة، ويقتصر استخدامها في الوقت الحالي على المخابر الصغيرة جداً في الأماكن الشعبية، وهي لا تقدم أية بيانات الكترونية متراكمة أو قابلة للدراسة.

ب- البرامج الالكترونية الخاصة بمخابر التحاليل الطبية

هي البرامج التي تعمل على تخزين بيانات ونتائج إجراء التحاليل ضمن قواعد بيانات قادرة على تلبية احتياجات توثيق وطباعة نتائج العمل داخل المخبر، مثل طباعة نتائج التحاليل الحالية والسابقة للمرضى، ومتابعة الأمور المحاسبية للمخبر، ومتابعة مخزون المواد والأدوات المخبرية، وإجراء بعض الإحصاءات البسيطة عن التحاليل الطبية مثل مقارنة أعداد أو نسب التحاليل بحسب نوعها، أو بحسب أصناف المرضى. وتعتمد على الجهد

[1] جريدة الثورة 2010/3/26 (http://thawra.alwehda.gov.sy)

[2] http://damascushospital.org/?page=518&lang=Ar

البشري كلّيةً في إدخال البيانات إلى الحاسوب، ولا تمتلك القدرة على تبادل البيانات مع أجهزة التحاليل الطبية العاملة داخل المخبر. ويوجد في الأسواق عدد كبير من البرامج المخصصة للمخابر الطبية، وصُمِّمت بطرق مختلفة بحسب وجهات نظر القائمين عليها، ولم تصمم وفق معايير ثابتة وموحدة، وقد صُمِّم معظمها دون مراعاة فكرة الربط المستقبلي لها مع السجلات الطبية للمرضى. وتعاني هذه البرامج من عدم وجود أسماء أو أرقام موحدة للتحاليل الطبية من قبل جهة ما[1] مثل هيئة مخابر التحاليل الطبية السورية، مما يجعلها تختلف من مخبر لآخر وتزداد بالتالي صعوبة ربط برامج مخابر التحاليل مع بعضها البعض في المستقبل.

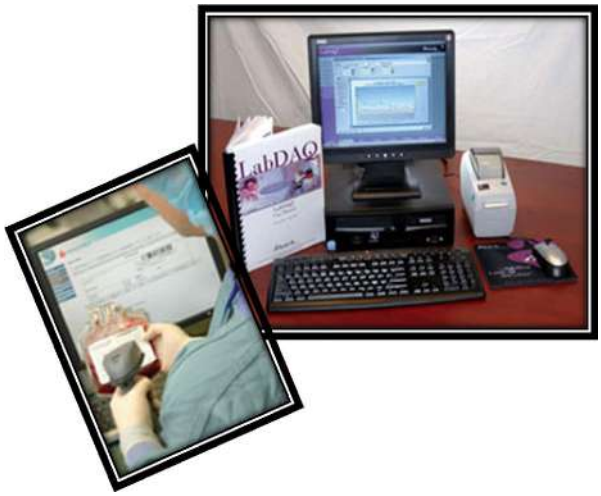
ج- نظام معلومات مخبر التحاليل الطبية (Laboratory Information System (LIS))

وهو نظام برمجي يؤتمت عمل مخبر التحاليل الطبية ابتداءً من استقبال طلبات التحاليل الطبية ووصولاً إلى تحصيل النتائج وطباعتها، وتخزن النتائج ضمن قواعد بيانات. ويتم تطبيق هذه النظم في المخابر الكبيرة وخصوصاً المخابر التابعة للمستشفيات. وتعمل البرامج الحاسوبية في هذه النظم على أساس المراسلات الآلية للبيانات بين أجزاء شبكة الحواسيب الموجودة في الأقسام المختلفة للمخبر كإدارة المخبر وقسم الاستقبال وقسم سحب العينات وقسم إجراء التحاليل وقسم تحصيل النتائج وطباعتها.

كما ويتم تبادل البيانات بين البرامج الحاسوبية وأجهزة إجراء التحاليل الطبية إما يدوياً في بعض أنواع التحاليل أو آلياً في حال استخدام المخبر لأجهزة متطورة قابلة للربط مع هذه البرامج بحيث تستقبل هذه الأجهزة أوامر إجراء التحاليل من قسم الاستقبال وتلبيها بعد أن تصلها عينات التحاليل الموافقة لهذه الأوامر، حيث توضع لصاقات على عبوات العينات عند تلقيها من المرضى مطبوع عليها رِمَازٌ قُضْباني (barcode) لتعريف هويّة هذه العيّنة والبيانات الضرورية لإجراء التحليل الطبي المناسب لها. وبعد أن تنتهي الأجهزة من إجراء التحاليل تقوم بتحصيل النتائج ومن ثم ترسلها آلياً عبر نظام المعلومات إلى قسم الاستقبال لطباعتها وتسليمها إلى الجهة الطالبة للتحليل الطبي (انظر الشكل (4-2))، وهي طريقة متطورة في العمل المخبري ولها فوائد كثيرة[2] أهمها رفع كفاءة ودقة العمل والمعلومات، وتوفير بيانات

[1] إلياس، محمود، 2008، **"أتمتة مخابر التحاليل الطبية في سورية بين الواقع والمأمول"**
[2] المرجع السابق

إحصائية متعددة عن تحاليل المرضى بالإضافة إلى تقديم بيانات تاريخيـــة عـــن أداء أجهـــزة التحاليل الطبية المرتبطة بنظام معلومات المخبر.



الشكل (4- 2): استخدام الرماز القضباني للتعرف على العينات في مخبر التحاليل الطبية

ويعتبر نظام معلومات مخبر التحاليل الطبية – في المشافي المتطورة – أحدَ الأنظمة الفرعيـــة التي يتكون منها نظامُ معلومات المشفى (طبياً وإدارياً ومحاسبياً). فمثلاً يـــؤدي ربـــط نظـــام معلومات المخبر مع نظام معلومات الصيدلية إلى تحقيق فوائد مباشرة إلى المرضى وأهمهـــا الحيلولة دون إساءة فهم الوصفات الطبية المتعلقة بالمضادات الحيوية المناسبة لنتائج التحاليـــل الحيوية[1].

3•4 الأبحاث العلمية المستندة إلى بيانات التحاليل المخبرية الطبية

يُعتبر مخبر التحاليل الطبية في العصر الحالي مصـــدراً غنيـــاً بالبيانـــات القابلـــة للدراســـة والاكتشاف، وخصوصاً الذي يعتمد في بياناته على نظم معلومات المخابر، والذي يرتبط مـــع السجلات الطبية أو الصحية للمرضى.

تتطور العلوم الطبية المخبرية وتزداد أهميتها يوماً بعد يوم، وقد تم الإقرار حـــديثاً بـــأن مـــا يقارب 70% من القرارات الطبية تتأثر بنتائج القيم المخبرية[2]، وهذا يعني أن الطب المخبـــري هو اختصاص طبي يقع في لب الرعاية الصحية، ويرافق مختلف مراحل الفحوصات الطبيـــة، وحتى أنه أصبح يُستخدم بشكل متزايد في الكشف المبكر عن بعض الأمراض. ويعود الفضـــل

[1] Bossuyt, X. & Verweire, K. 2007, **Laboratory Medicine: Challenges and Opportunities**.
[2] Beastall, GH. (2008), **The modernization of pathology and laboratory medicine in the UK: networking into the future**.

في ذلك إلى تزايد مساهمة الباحثين المخبريين في تطوير اكتشافاتهم واختراعاتهم البحثية التي ترفد الطب بكل جديد في عالم الفحوص المخبرية.

تقع الأبحاث المخبرية ضمن الحقل الطبي، وتتضمن الكثير من المصطلحات والأساليب والأفكار الموجهة بالدرجة الأولى لأصحاب الاختصاصات الطبية. ولا تهدف الدراسة الحالية إلى الخوض في التفاصيل الطبية لهذه الأبحاث لأنها تقع خارج اهتمام واختصاص البحث الحالي، وإنما سيتم الاطلاع على بعض النقاط الأساسية في هذه الأبحاث مثل مصادر بيانات هذه الأبحاث والأساليب المتبعة في انجازها وكيفية تطورها بالترافق مع التطورات التقنية.
وتُقسم الأبحاث المخبرية من وجهة نظر الدراسة الحالية إلى قسمين، هما الأبحاث المخبرية التقليدية والأبحاث المخبرية المستندة إلى التطورات التقنية في مجال مخابر التحاليل الطبية.

1•3•4 الأبحاث المخبرية التقليدية

تُجرى البحوث الطبية المخبرية التقليدية عادةً من قبل الباحثين من أخصائيي الطب المخبري أو التشخيص المخبري، وتندرج هذه البحوث ضمن محوريين رئيسين. ويجري العمل في المحور الأول على دراسة طرق وأساليب الحصول على نتائج التحاليل الطبية حيث قد تختلف نتائج التحاليل باختلاف الكواشف المستخدمة أو اختلاف طرق حساب النتائج أو اختلاف أجهزة الاختبار..الخ. ولذلك تقوم بعض الدراسات المخبرية على اختبار الطرق والأساليب المختلفة والمقارنة بينها وترجيح الأفضل.
ويقوم الباحثون في المحور الثاني بدراسة دور الفحوصات المخبرية في الكشف عن الأمراض، واكتشاف أنواع جديدة من التحاليل الطبية ودراسة علاقة قيم نتائج التحاليل المخبرية مع صفات المرضى وحالتهم المرضية..الخ.

1•1•3•4 مراحل انجاز الأبحاث المخبرية الطبية التقليدية

يحمل البحث الطبي المخبري بشكل عام أهدافاً أو فرضيات واضحة تحدد مسبقاً قبل الشروع به، ويتم إجراء هذا البحث وفق المراحل الأساسية التالية:

أ- اختيار موضوع أو مشكلة للبحث.

ب- سحب عينة إحصائية من المرضى تتوافق مع موضوع البحث.

ج- تعبئة استبيانات تصمم خصيصاً بما يخدم أهداف البحث، وتتناول البيانات الشخصية للمرضى، والأعراض السريرية لهم، ونتائج التحاليل الطبية التي تتوافق مع مشكلة البحث.

د- دراسة حالات هؤلاء المرضى وإجراء الفحوصات اللازمة لهم كالفحوصات السريرية والتحاليل الطبية.

ه- الاستمرار بمراقبة حالات المرضى خلال فترة الدراسة إذا تطلب الأمر ذلك، مع الاستمرار بتعبئة الاستبيانات الخاصة بمراقبة حالة هؤلاء.

و- إجراء التحليل الإحصائي للبيانات الناتجة عن استبيانات البحث بهدف التوصل إلى نتائج تلخيصية إحصائية وإجراء المقارنات والتحقق من بعض العلاقات الارتباطية مثل العلاقة الارتباطية بين نتائج التحاليل والأعراض السريرية، واختبار فرضيات البحث إن وجدت..الخ.

ز- عرض نتائج البحث الطبي المخبري التي تم التوصل إليها.

2•1•3•4 أمثلة عن الأبحاث المخبرية التقليدية

جرى الاطلاع العام على العديد من الأبحاث المخبرية التقليدية المنجزة في رسائل الماجستير في التشخيص المخبري (كلية الصيدلة) وفي الطب المخبري (كلية الطب البشري)، بهدف التعرف على المواضيع التي تتطرق لها، والأساليب التي تُستخدم في انجازها.

ويعرض الجدول (4-1) بعض الأمثلة عن تلك الأبحاث والتي استَخدَمت الطرق الإحصائية التقليدية في الكشف عن بعض العلاقات بين التحاليل المخبرية والأمراض، والتي يُعتقد أنها يمكن أن تُنجَزَ يوماً ما بالاستفادة من البيانات المتراكمة في قواعد بيانات السجلات الطبية (وليس فقط قواعد بيانات مخابر التحاليل الطبية)، كما أنها يمكن أن تستفيد مستقبلاً من نتائج الأبحاث التي تعتمد على تقانة اكتشاف المعرفة في الحقل الطبي كتلك التي سيتم التطرق إليها في الفصل الخامس من هذه الدراسة.

الجدول (4- 1): عناوين بعض رسائل الدراسات العليا في التشخيص المخبري والطب المخبري في جامعة دمشق

الاسم	عنوان رسالة الماجستير	السنة	الاشراف	الاختصاص
ابراهيم الحاج خلف	تحري بعض عوامل الخطورة القلبية عند مرضى القصور الكلوي المزمن	2003	د. وفيقة زرزور	التشخيص المخبري
محمد صادق العلواني	مساهمة مخبرية لتشخيص الإصابة البشرية بداء الليستيرية	2004	د .ميخائيل عبيد	التشخيص المخبري
المثنى المطر	دور بعض المعالم المناعية في تقييم ومراقبة مرضى الربو	2005	د.خليل القوتلي	التشخيص المخبري
نسرين مرتيني	تقييم دور ال CTP في كشف النقائل العظمية من سرطان الثدي	2006	د.سحر الفاهوم	التشخيص المخبري
أسيل الحافظ	دراسة مستويات البروتين البلازمي المرافق للحمل عند مرضى الخناق الصدري الثابت والمزمن	2009	د.إميل شاهين و د.فايزة القبيلي	التشخيص المخبري
عيسى سعيد	دراسة مستويات البيتيد المشابه للفوكاكون - 1 لدى مرضى السكر النمط الثاني	2010	د.ميخائيل عبيد	التشخيص المخبري
د. هشام عبد الصمد	واصمات كيميائية حيوية مفيدة في تحديد التقلب العظمي وتقييم ICTP كواصمة ارتشافية لدى مرضى الفشل الكلوي.	2005	أ. د. سحر الفاهوم	الطب المخبري
د. رزان سمسمية هاشم	الأهمية التشخيصية لعيار الخمائر القلبية في احتشاء العضلة القلبية	2005	أ. د. محمد محجوب جيرودي	الطب المخبري
د. ندال كريّم	دراسة العلاقة بين مستويات الهوموسيستئين والتبدلات المخبرية عند الحوامل المصابات بما قبل الارتعاج	2010	أ.د. نزار الكشكة	الطب المخبري
د. وسام مكة	دراسة العلاقة بين ارتفاع الهوموسيستئين في الدم والأمراض القلبية الوعائية	2010	أ. د. عماد أبو عسلي	الطب المخبري

2•3•4 الأبحاث المخبرية المستندة إلى التطورات التقنية في مجال مخابر التحاليل الطبية

تتغير الأساليب المستخدمة في أية أبحاث علمية باستمرار، وتتصف الاتجاهات الحالية والقادمة في العلم والعمل المخبري بوجود عدة تغيرات في التقنيات المستخدمة في مخابر التحاليل الطبية، والتي أثرت بشكل غير مباشر على المنهجيات المستخدمة في الأبحاث الطبية. ويحتاج انجاز بعض هذه الأبحاث إلى المساعدة من قبل أخصائيين تقنيين إلى جانب الأخصائيين الطبيين. وسنعرض فيما يلي لتأثير أهم التغيرات التقنية في مجال مخابر التحاليل الطبية على طبيعة ومنهجية الأبحاث المستندة إلى بيانات التحاليل الطبية.

1•2•3•4 الأبحاث المخبرية المرافقة لتطور الأجهزة والتقنيات المخبرية

ساعدت التطورات الحاصلة في الأجهزة والتقنيات المخبرية على زيادة الدقة والسرعة والجودة في تقديم نتائج التحاليل المخبرية، كما ساعدت على تسهيل انجاز وتقييم الأعمال البحثية في مجال التحاليل المخبرية أو أي بحث طبي آخر يحتاج إلى نتائج متتابعة للتحاليل المخبرية، كما مكَّنت من التوسع في اكتشاف اختبارات جديدة لاستخدامها في الكشف عن بعض الأمراض.

وكمثال على ذلك وَرَدَ في دراسة نُشرت في كانون الثاني عام 2010 في مجلة التشخيص المخبري السورية بعنوان " اللعاب كسائل حيوي للتطبيقات التشخيصية للأمراض " أن اللعاب يُستخدم بشكل تقليدي في تشخيص عدد من الأمراض الجهازية لكونه يحتوي على نفس مكونات المصل، أما حالياً فقد أصبح اللعاب يُستخدم في تشخيص الأمراض المعدية، والخبيثة، والاضطرابات الوراثية، وأمراض المناعة الذاتية، والاضطرابات الصماوية، وتقييم مستويات الأدوية العلاجية، وبشكل خاص مراقبة تأثير أدوية الإدمان... ويعد تشخيص الأمراض عن طريق تحليل اللعاب ذا قيمة فعالة عند المرضى وخصوصاً الأطفال لأن جمع هذا السائل أسهل بالمقارنة مع سحب عينات الدم من هؤلاء المرضى، كما أنه يُمكِّن من التحري عن بعض المشاكل الصحية لعدد ضخم من الأفراد بسرعة وبتكلفة معقولة.

وفي دراسةٍ أخرى نُشرت في مجلة التشخيص المخبري السورية في نيسان عام 2010 بعنوان " استخدام اختبارات الرقاقات الحيوية النانوية المعتمدة على اللعاب في تشخيص

احتشاء عضلة القلب الحاد في وحدة العناية " تم توضيح إمكانية استخدام اللعاب في تشـــخيص احتشاء عضلة القلب الحاد في وقت مناسب يسمح باتخاذ التدابير العلاجية المناســـبة، وهـــي طريقة سريعة النتائج وأسرع من طريقة قياس الواصمات الحيوية في المصل[1].

وبهذا نجد أن عدداً من البحوث الطبية المخبرية تتجه نحو اكتشاف طـــرق وتحاليـــل جديـــدة للتحري عن الإصابة بالأمراض.

2•2•3•4 الأبحاث المخبرية المرافقة لتراكم البيانات في الحواسيب

أدى تزايد الاعتماد على أدوات تكنولوجيا المعلومات في إدارة العمليات والبيانات داخل مخابر التحاليل الطبية إلى تزايد هائل وتراكم لبيانات نتائج التحاليل المخبرية في الأجهزة الحاسوبية، مما ولَّد فكرة دراسة هذه البيانات إحصائياً للتعرف على التغيرات التي تطـــرأ عليهـــا عبـــر الزمن.

ومن الأمثلة المحلية على ذلك تلك الدراسة التي أجرتها الدكتورة عبير الكفري والدكتور محمد طاهر إسماعيل الأخصائيين في الطب المخبري في جامعة دمشق بعنوان " التراجع الهائل للطفيليات المعوية البشرية وبخاصة الديدان في دمشق وريفها– دراسة تشخيصية في الفترة بين عامي 1998–2007 "[2]. وقد استخدَم الباحثان بيانات فعلية لنتائج تحليل 4059 عينة براز من مخبرهم الخاص على مدى عشرة أعوام. وهدفَ البحث بشكل أساسي إلى دراسة آليات التعرف على أنواع الطفيليات، وتمييزها عن غيرها من الأشياء كحبات غبار الطلع. كما هدفَ ذلك البحث أيضاً إلى تحديد نسب انتشار الطفيليات المعوية الممرضة وغير الممرضة في البيانات المدروسة، وإلى التعرف على أصناف الطفيليات الموجودة، وتغيُّر نسب انتشارها عبر الزمن. وتبين لهما –من جملة النتائج الإحصائية– أن هناك انخفاضاً ملحوظاً في تواجد الطفيليات المعوية من 221 مريضاً في عام 1998 إلى 48 مريضاً في عام 2007، ويعود السبب الأساسي في ذلك إلى التأثير الايجابي لمعالجة مياه الصرف الصحي في محطة عدرا للصرف الصحي التي بدأت بالعمل منذ عام 1998.

[1] د. الساري، اسماعيل، 2010، **"استخدام اختبارات الرقاقات الحيوية النانوية المعتمدة على اللعاب في تشخيص احتشاء عضل القلب الحاد في وحدة العناية: دراسة الجدوى"**

[2] د. الكفري، عبير & أ.د. اسماعيل، محمد طاهر، 2008، **"التراجع الهائل للطفيليات المعوية البشرية وبخاصة الديدان في دمشق وريفها- دراسة تشخيصية في الفترة بين عامي 1998-2007"**

كما أصبح بإمكان الباحثين الطبيين الاستفادة من فكرة تراكم البيانات في تقدير مدى نشاط وفعالية الآفة المرضية، وفي تتبع ومراقبة طويلة الأمد –قد تمتد عشرات السنوات– للأشخاص الذين خضعوا لتجربة ما في فترة من فترات حياتهم مثل الدراسة التي قام بها مجموعة من الباحثين منهم سامية مورا ونادر رفاعي في عام 2009 بعنوان " مقارنة تراكيز كولسترول LDL المحسوبة رياضياً وفق معادلة فريدوالد Friedewald؛ مع المُقاسة مباشرة وارتباطها بحوادث القلب الوعائية CVD في 27331 امرأة[1] ". وقد خضعت لهذه الدراسة 27331 امرأة سلمية ظاهرياً من مناطق مختلفة في أميركا في عام 1995 ممن تزيد أعمارهن عن 45 عاماً، حيث أجريت لهن عدة فحوصات مخبرية تتعلق بالكولسترول الصيامي وغير الصيامي المقاسين بطريقة مباشرة أو بطريقة حسابية وفق معادلة فريدوالد، وجرى تتبع نتائج هذه الفحوصات المخبرية لأفراد العينة عبر السنوات بهدف كشف ارتباطها مع أمراض القلب الوعائية (CVD) الحاصلة عند بعض النساء في العينة على مدى 11 عاماً (لغاية عام 2006). وقد أجريت التحليلات الإحصائية للبيانات باستخدام البرنامج الإحصائي Stata 8.2 ، وتبين من جملة النتائج التي توصلت إليها الدراسة أنه يوجد ارتباط بين قيم الكولسترول وبين أمراض القلب الوعائية، واتضح أن كلتا طريقتي القياس المباشرة والمحسوبة رياضياً متقاربتين في النتائج مع انخفاض بسيط في قيم الطريقة المباشرة والتي تؤدي –في حال اعتمادها– إلى تصنيف خاطئ لبعض النساء ضمن فئة خطورة أقل من الحقيقة.

والمهم في هذه الدراسة من وجهة نظر اهتمامات البحث الحالي هو أن نبيِّن دور التقانات الحديثة في توسيع القدرة على إجراء البحوث الطبية طويلة الأمد ودراسة البيانات الطبية المتراكمة لسنوات عديدة.

3•2•3•4 الأبحاث المخبرية المرافقة لتقانة اكتشاف المعرفة

اتجهت حديثاً مجموعة من الأبحاث المخبرية في البلدان المتقدمة نحو الاستفادة من البيانات المتراكمة في قواعد البيانات الطبية للكشف عن وجود معارف طبية جديدة كامنة فيها باستخدام

[1] مورا، سامية، نادر رفاعي، وآخرون، ترجمة د. جالا قاسم العتمة، "مقارنة تراكيز كوليستيرول الـ LDL المحسوبة رياضياً وفق معادلة فريدوالد Friedewald؛ مع المُقاسة مباشرة وارتباطها بحوادث القلب الوعائية CVD في 27331 امرأة"

تقانة اكتشاف المعرفة. وتبرز أهمية هذه التقانة في دراسة العلاقات الترافقية والتقسيمات الممكنة للبيانات بوجود عدد كبير من المتغيرات وعدد كبير من البيانات المتراكمة زمنياً. فقد أجرى الباحثان المختصان بتقنية المعلومات الطبية (**Shoji Hirano and Shusaku Tsumoto**) عام 2007 دراسةً[1] بعنوان " Temporal Data Mining in Hospital Information Systems Analysis of Clinical Courses of Chronic Hepatitis "، تم فيها استخدام تقانة التنقيب في البيانات الزمنية لدراسة حالة مرضى التهاب الكبد المزمن بالاعتماد على بيانات نظم معلومات المشافي، وشملت الدراسة على بيانات طويلة الأمد لـــ 488 مريض بين عامي 1982–2003 مصابين بالتهاب الكبد المزمن في اليابان من أحد النمطين B أو C. وقد استندت الدراسة إلى فكرة أن عدد صفيحات الدم (Platlates) يختلف بحسب مراحل الالتهاب حيث يقل عدد هذه الصفيحات كلما تفاقم الالتهاب. وتم في هذه الدراسة تطبيق التجزئة العنقودية على بيانات السلاسل الزمنية لهؤلاء المرضى وفقاً لتعداد صفيحات الدم لديهم خلال حياتهم المرضية. وجرى البحث عن تصنيفات عنقودية للمرضى بالاستناد إلى عدة خاصيات مثل عدد السنوات التي وصل خلالها المرضى (من النمطين B وC) إلى المرحلة الرابعة وهي مرحلة التهاب النسيج الليفي للكبد، والسنوات التي قضاها المرضى بين المراحل، وعلاقتهم بأنواع الفيروسات وبمرحلة التهاب النسيج الليفي وبنوع العلاج...الخ. وقد أظهرت النتائج أنه يمكن تصنيف مراحل تعداد الصفائح ضمن عدة مجموعات عنقودية تحمل عناصر كل منها صفات متشابهة، وتم التوصل إلى وصف للحالات التي يتفاقم فيها التهاب النسيج الليفي للكبد بشكل سريع.

نلاحظ أن هذه الدراسة المعتمدة على تقانة التنقيب في البيانات قد تمت بوجود مسألة تنقيب ولكن بدون وجود فرضيات تمهيدية، وتفيد هذه الدراسة في تكوين فكرة سريعة عن مراحل تطور المرض المدروس وفقاً لتتبع نتائج تحاليل المرض، وتفيد أيضاً في دراسة استجابة أصناف المرضى لأنواع العلاج المختلفة، كما أنها قد تساعد في التوصل إلى فرضيات طبية جديدة[2] لاختبارها في الأبحاث الطبية المستقبلية.. وبالطبع هناك المزيد من الفوائد التي كانت

[1] Hirano, Sh. & Tsumoto, Sh. 2007, **"Temporal Data Mining in Hospital Information Systems Analysis of Clinical Courses of Chronic Hepatitis"**

[2] Kazmierczak, S. & Leen, T, 2007, **"Reduction of multi-dimensional laboratory data to a two-dimensional plot a novel technique for the identification of laboratory error"**

ستُجنى من هذه الدراسة، فيما لو دُرست بيانات إضافية عن المرضى، مثل البيانات الشخصية للمرضى وظروف معيشتهم ومدى تأثير المشروبات الكحولية على المرض...الخ من المتغيرات التي لم تكن متوفرة لهذين الباحثين في قواعد البيانات القديمة التي أتيحت لهما دراستها.

عرضنا فيما سبق للأساليب المتبعة في البحوث المخبرية وللتطورات الحاصلة فيها. وبشكل عام وفي جميع المجالات البحثية الطبية وغير الطبية، يبقى لكل نوع من أنواع البحوث دوره وأهميته التي لا يمكن الاستغناء عنها مهما تطورت الأساليب. وتسعى الاتجاهات الحديثة في البحوث نحو تكامل الأساليب أحياناً أو اندماجها أحياناً أخرى، ونحو تعاون الباحثين من الاختصاصات المختلفة، وذلك بما يحقق جودة وكفاءة وسرعة التوصل إلى النتائج المفيدة.

ابتعد الفصل الرابع قليلاً عن مسار عملية اكتشاف المعرفة الذي استمر خلال الفصول الثلاثة الأولى من هذه الدراسة، ولكنه جاء خصيصاً ليربط بين تلك الفصول والفصل الخامس من خلال التعرف على الملامح العامة لعالم التحاليل المخبرية الذي أنتج البيانات التي ستُستخدم في الحالة التطبيقية في هذه الدراسة، ومن خلال الاطلاع على اتجاهات الأبحاث الطبية المخبرية التي تساعد في التعرف على موقع الدراسة الحالية بالنسبة إلى تلك البحوث. وتتشابه الدراسة الحالية مع بعض تلك الأبحاث الطبية الحديثة في أنها تستخدم أدوات اكتشاف المعرفة من قواعد البيانات، ولكنها تختلف عنها في مسائل اكتشاف المعرفة المقترحة، حيث تركز تلك الدراسات الطبية على مناقشة مسائل طبية واحدة على الأغلب، وتُدرس بالتالي جميع الخاصيات المرتبطة بها، وتكون بعض هذه الخاصيات عبارة عن نتائج التحاليل الطبية ذات الصلة بالمسألة المدروسة وليس جميع التحاليل الطبية. أما في الدراسة الحالية فلا يوجد مسألة طبية محددة، إنما يوجد عدد من المسائل المدروسة، وتشارك جميع التحاليل الطبية في مناقشتها.

الفصل الخامس

مراحل اكتشاف المعرفة من قواعد بيانات مخبر التحاليل الطبية



الفصل الخامس

مراحل اكتشاف المعرفة من قواعد بيانات مخبر التحاليل الطبية

تمهيد

يختص الفصل الخامس من هذه الدراسة بعرض التطبيق العملي لمراحل عملية اكتشاف المعرفة على بيانات مخبر التحاليل الطبية والبيانات المرتبطة بها. وقد تم اختيار الحالة العملية من الواقع المحلي وليس من المصادر المتاحة عبر الانترنت، وذلك لأنه من الأفضل والواجب علينا كباحثين أن نساهم في دراسة تطوير أنظمتنا وقواعد بياناتنا المحلية في الدرجة الأولى، وأن نبين أهمية أن تكون البيانات المتراكمة في قواعد البيانات المحلية جيدة وكثيرة المتغيرات، وقادرة على تبادل البيانات مع قواعد البيانات الأخرى. فالمشاكل التي تعاني منها بياناتنا متشابهة في معظم أنظمتنا المحلية، ولكنها تختلف عن تلك التي تعاني منها حالياً البيانات في المجتمعات المتطورة. وربما تتشابه مشاكل بياناتنا الحالية مع جزء بسيط من المشاكل التي عانت منها البيانات العالمية منذ أكثر من 15 عاماً. وتظهر معظم هذه المشاكل بسبب قلة وجود القيود المنطقية على عمليات إدخال البيانات إلى الحواسيب مما يؤدي إلى إدخال بيانات خاطئة وغير منسجمة مع بقية البيانات، وإلى عدم اكتشاف الأخطاء أثناء إدخالها. وتتراكم بالتالي الأخطاء كثيراً ويصعب تصحيحها بعد مضي الزمن عليها، فالبيانات التي لا تُسجَّل بشكل صحيح في حينها قد تضيع وتضيع معها المعرفة المخبأة فيها.

وربما تكمن الثغرة الأساسية في أنظمة البيانات المحلية في عدم شعور معظم المعنيين بأمر جمع البيانات بالأهمية المستقبلية – وليس الآنية – لتراكم بيانات صحيحة وقليلة العيوب، ويترتب على ذلك الدخول في حلقة مفرغة من ضعف في البيانات من حيث الكم والكيف، وبالتالي تتزايد صعوبة إدخال أية تطويرات على الأنظمة الضعيفة الموجودة أو حتى الاستفادة منها طالما أن الأساسات ضعيفة وهشة وخالية من النظرة المستقبلية.. ولكن ما الحل؟؟.. الحل يحتاج إلى اختراق جريء وسريع ومتواصل لهذه الحلقة المفرغة من أجل كسرها أو تحويلها

إلى عجلة تقود عملية التطوير بدلاً من أن تعرقلها. ومن هذا المنطق يحمل هذا البحث رغبة التعرف على مدى قدرة إحدى قواعد بيانات التحاليل المخبرية المحلية المتوفرة على تقديم مواد أولية كفوءة ومناسبة لعمليات التنقيب في البيانات. وسنبين فيما يلي المراحل التطبيقية التي تم إتباعها لانجاز هذه المهمة استناداً إلى المراحل النظرية التي عُرضت في الفصل الثالث من هذه الدراسة.

وقد جَرت أثناء انجاز الجانب التطبيقي من هذه الدراسة مواجهةُ العديدِ من الصعوبات والعراقيل والمشاكل، مثل تنظيف البيانات، وربط الجداول، ونقل البيانات، وبناء مكعب البيانات واستعراضه.. والتي أوصلت العمل في بعض الأحيان إلى طرق مسدودة مما ولَّد ضرورة التوقف والتخلي عن عدد كبير من الأعمال التي تم انجازها ومن ثم البدء من جديد باستخدام طرق جديدة أو أفكار جديدة.. وقد اخترنا أن لا نخوض في تفاصيل هذه الأمور أثناء عرضنا للجانب التطبيقي من الدراسة إلا أحياناً، وإنما سنعرض غالباً الانجازات المفيدة أو النهائية التي أدت إلى النتائج السليمة، وذلك لأن الخوض في تفاصيل الصعوبات والحلول يؤدي إلى تشتت البحث وإطالته ويصرف القارئ عن التركيز على المسارات الأساسية للبحث.

يتوزع هذا الفصل على ستة مباحث، ويتناول كل منها إحدى مراحل اكتشاف المعرفة من قواعد بيانات مخبر التحاليل الطبية المتاحة للدراسة، وذلك على النحو التالي:

المبحث الأول	**المرحلة الأولى**	**استيعاب مجال مسألة اكتشاف المعرفة من بيانات مخبر التحاليل الطبية**
المبحث الثاني	**المرحلة الثانية**	**استيعاب بيانات مخبر التحاليل الطبية والبيانات المرتبطة بها**
المبحث الثالث	**المرحلة الثالثة**	**تهيئة بيانات مخبر التحاليل الطبية**
المبحث الرابع	**المرحلة الرابعة**	**التنقيب في بيانات مخبر التحاليل الطبية**
المبحث الخامس	**المرحلة الخامسة**	**تقييم المعرفة المكتشفة في بيانات مخبر التحاليل الطبية**
المبحث السادس	**المرحلة السادسة**	**استخدام المعرفة المكتشفة من بيانات مخبر التحاليل الطبية**

الفصل الخامس

المبحث الأول

مرحلة استيعاب مجال مسألة اكتشاف المعرفة من بيانات مخبر التحاليل الطبية

مقدمة

اختص الفصل الثالث بالحديث عن الأسس النظرية لمرحلة استيعاب مجال مسألة التنقيب المدروسة، وسنقوم فيما يلي باستعراض خلاصة ما تم التوصل إليه أثناء التطبيق العملي لهذه المرحلة وفقاً للمحاور الثلاثة الأساسية المذكورة في القسم النظري.

1•1•5 تعريف حدود المسألة والمجال الذي تطبق فيه

تندرج الدراسة الحالية ضمن دراسات اكتشاف المعرفة في الحقل الطبي، ولكنها ليست دراسة طبية بحتة إنما تهتم بتعزيز دور قواعد البيانات الطبية في التوصل إلى فوائد للطب ولإدارة المراكز الطبية وللإحصاء الطبي ولقطاع التأمين الصحي وللمنظمات التي تهتم بالتعرف على الصحة العامة للسكان.. وذلك باستخدام تقانة اكتشاف المعرفة.

وتتجه هذه الدراسة نحو إبراز أهمية نتائج التحاليل الطبية ودورها ليس فقط في تشخيص الأمراض ومتابعة الحالة الصحية للمرضى وإنما أيضاً في إجراء الدراسات الكبيرة والشاملة على البيانات الالكترونية المتراكمة عبر الزمن، حيث من الممكن أن تقدم هذه البيانات فوائد كثيرة كالتعرف على بعض العلاقات الترافقية بين نتائج التحاليل الطبية مع بعضها، ومع البيانات الشخصية للمرضى، ومع الأمراض..الخ.

وللإحاطة بمجال تطبيق المشكلة المدروسة جرى الاطلاع على واقع العمل المخبري الذي يولد البيانات التي ستُدرس، وجرى التخاطب مع الأخصائيين المَخبَريّين للتعرف على بعض

الجوانب التي تهم البحث، كما وتم تركيز بعض الاهتمام للاطلاع على المفاهيم الضرورية في هذا المجال وآليات التعامل مع البيانات في العمل المخبري وعلى بعض الدراسات العلمية العالمية والمحلية التي تستخدم بيانات مخابر التحاليل الطبية، وقد اختص الفصل الرابع من هذه الدراسة بالحديث عن جزءٍ من تلك الأمور التي تقع ضمن اهتمام مرحلة التعريف بمجال التطبيق، وسنتحدث فيما سيأتي عن الجزء الآخر من هذا المجال والذي يتعلق أولاً بالاطلاع على مضمون البرنامج الحاسوبي المولِّد للبيانات الالكترونية لمخبر التحاليل الطبية المدروس، ثم سيتم التعرف على مسائل اكتشاف المعرفة التي تستند إلى هذه البيانات المتاحة.

1•1•1•5 مجال تطبيق تقانة اكتشاف المعرفة في دراسة بيانات مخبر التحاليل الطبية

بدأ الاهتمام في مرحلة التحضير لهذه الدراسة بالبيانات الطبية المتراكمة في الأجهزة الحاسوبية للمستشفيات المحلية، وقد وقع الاختيار للانطلاق بمثل هذا النوع من الدراسات من البيانات المتوفرة في قسم التحاليل المخبرية وذلك للسببين الأساسيين التاليين:

أ- تبين أن قسم التحاليل الطبية المخبرية هو الأول والوحيد تقريباً من الأقسام الطبية في المستشفيات المحلية التي طُبِّقت فيها الأتمتة وتراكمت فيها البيانات.

ب- تعتبر التحاليل الطبية -الصحيحة طبعاً- مصدراً دقيقاً لتقييم حالة المرضى لأنها تعتمد على القياس وليس على وجهات النظر الطبية الشخصية.

وقد جرت المحاولات - في فترة التحضير للدراسة وضمن الظروف المتاحة وصعوبة الحصول على بيانات - للعثور على مخبرٍ يتوفر فيه الحد الأدنى من الشروط الضرورية لإجراء الجانب العملي من هذه الدراسة، وقد انطبقت هذه الشروط على مخبر التحاليل الطبية التابع لأحد المستشفيات الحكومية في دمشق، وهي:

1- أن يستخدم هذا المخبر برنامجاً حاسوبياً لإدارة وتخزين بيانات التحاليل الطبية في قاعدة بيانات كبيرة ولفترة زمنية لا تقل عن سنتين، وقد وُجد أن المخبر المدروس يستخدم برنامج للمخابر يدعى "الوافي للمخابر والتحاليل" منذ عام 2006.

2- أن تتوفر بيانات الكترونية أخرى عن المرضى كالبيانات الشخصية والمشاكل الصحية في أقسام أخرى من المشفى، وأن تكون قابلة للربط بطريقة ما مع بيانات التحاليل الطبية. وقد

وُجد في المشفى المدروس أنظمة قواعد بيانات في قسم قبول المرضى وقسم السجلات الطبية، ولكنها مصممة بصورة معزولة وغير مرتبطة مع غيرها وهدفها الأساسي التخزين الالكتروني الإضافي لبيانات المرضى إلى جانب الوثائق الورقية (أضابير المرضى).

3- أن يتم الحصول على موافقة الجهات الإدارية للمشفى في الحصول على البيانات الالكترونية المتوفرة لديها عن المرضى، والتي تم الحصول عليها بصعوبة وشريطة ألا يُدرج اسم مصدر البيانات المدروسة في هذا البحث.

1•1•1•1•5 واقع الأتمتة في مخبر التحاليل الطبية المدروس

يَستخدم مخبر التحاليل الطبية المدروس برنامجَ الوافي للمخابر والتحاليل بشكل مستقل في عمله عن أية جهة أخرى في المشفى، وقد قام مصمم البرنامج ببنائه باستخدام قواعد البيانات (Access 2003)، ونبين فيما يلي لمحة عامة عن بيئة عمل هذا البرنامج في مخبر التحاليل الطبية المدروس:

أ- يُستخدم البرنامج بالتشارك بين عدة حواسيب موجودة في الأقسام المختلفة في المخبر.

ب- يقوم المسؤول عن هذا البرنامج بحذف البيانات القديمة مع بداية كل عام بسبب صعوبة التعاطي مع الكميات الهائلة من البيانات باستخدام هذا البرنامج، وباستخدام الأجهزة الحاسوبية البطيئة نسبياً في المخبر المدروس، أي أن البيانات المتوفرة على البرنامج الحاسوبي لا تتجاوز مدة السنة، ولا يمكن استرجاع أية معلومات عن المرضى وتحاليلهم الطبية القديمة في برنامج الوافي إلا ضمن حدود البيانات المتوفرة.

ج- لا يرتبط البرنامج آلياً مع أجهزة التحليل الطبية في المخبر، بل يتم إدخال جميع النتائج يدوياً إلى واجهات البرنامج، ويمكن للبرنامج أن يجري العمليات الحسابية على النتائج التي تحتاج إلى حساب قبل نشرها مثل حساب النسب في تحليل الكريات البيض.

د- البرنامج غير مصمم بما يتناسب مع رفد السجلات الطبية الإلكترونية إن وجدت في المستقبل.

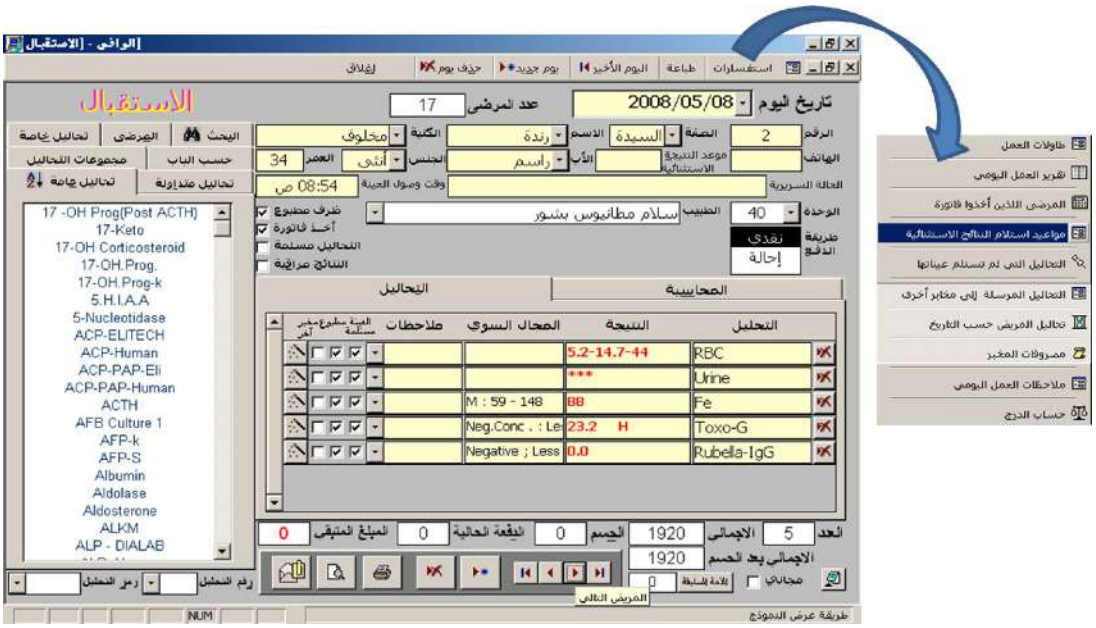
ه- لا يرتبط البرنامج مع أقسام المشفى الأخرى.

2•1•1•1•5 واجهات البرنامج الحاسوبي المستخدم في المخبر المدروس

يتكون برنامج الوافي للمخابر والتحاليل من عدد من الواجهات، ويختص كل منها بتنفيذ بعض المهام والإجراءات، وسنبين فيما يلي أهم مكونات برنامج الوافي والمهام التي يؤديها:

أ- واجهة الاستقبال

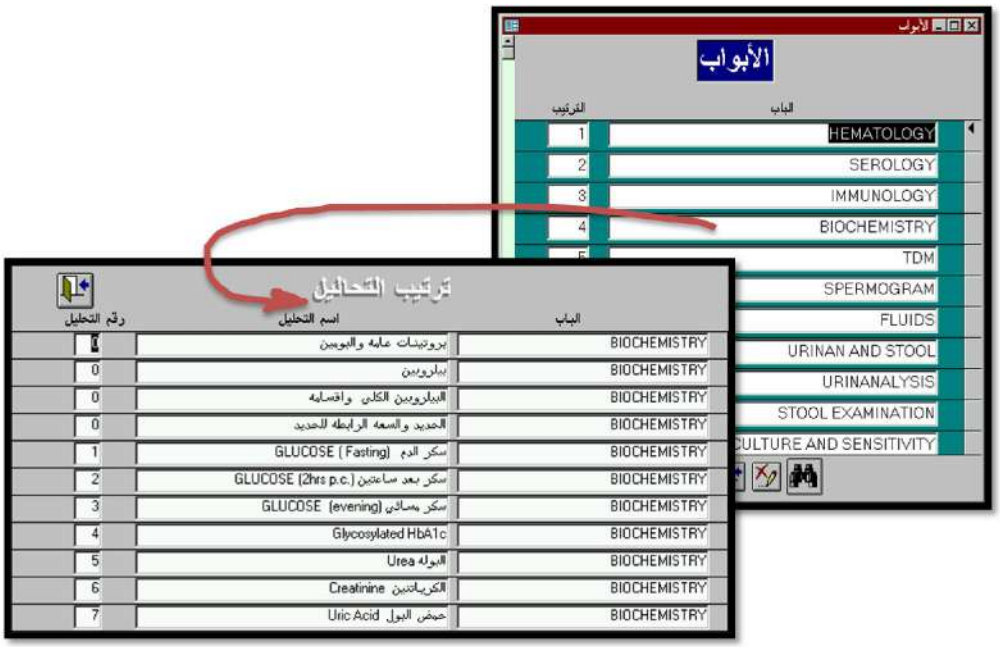
وهي الواجهة التي تُستعمل باستمرار في إدخال واستعراض كافة الأمور المتعلقة بالتحاليل الطبية للمرضى، وتتضمن بعض البيانات الشخصية كالاسم والعمر والجنس، والحالة السريرية للمريض، واسم الطبيب الطالب للتحاليل، وقائمة لتسجيل التحاليل الطبية المطلوبة ونتائجها، كما تتضمن كافة المعلومات المحاسبية المتعلقة بتكاليف التحاليل الطبية المترتبة على المريض، ويمكن إجراء عدد من الاستعلامات كما هو مبين في الشكل (5-1-1)



الشكل (5-1- 1): واجهة الاستقبال في برنامج الوافي للمخابر والتحاليل[1]

وتحتوي هذه الواجهة أيضاً على بعض القوائم الإضافية في الجهة اليسرى والتي تساعد في تسهيل انتقاء التحاليل الطبية المطلوبة، فمثلاً تتضمن قائمة التحاليل العامة كافة التحاليل مرتبة أبجدياً، أما قائمة أبرز التحاليل المتداولة فتتضمن أكثر التحاليل طلباً، ويمكن تعديلها بحسب وجهات نظر كل مخبر، وتتضمن قائمة التصنيف بحسب الباب أسماء التحاليل مرتبة تبعاً للأبواب التي تتبع لها، انظر الشكل (5-1-2)

[1] ملاحظة: أخذت صور واجهات برنامج الوافي من ملف دليل استخدام هذا البرنامج المرفق مع البرنامج.



الشكل (– –): الأبواب التي تصنف ضمنها التحاليل الطبية

ب- واجهة الإعدادات

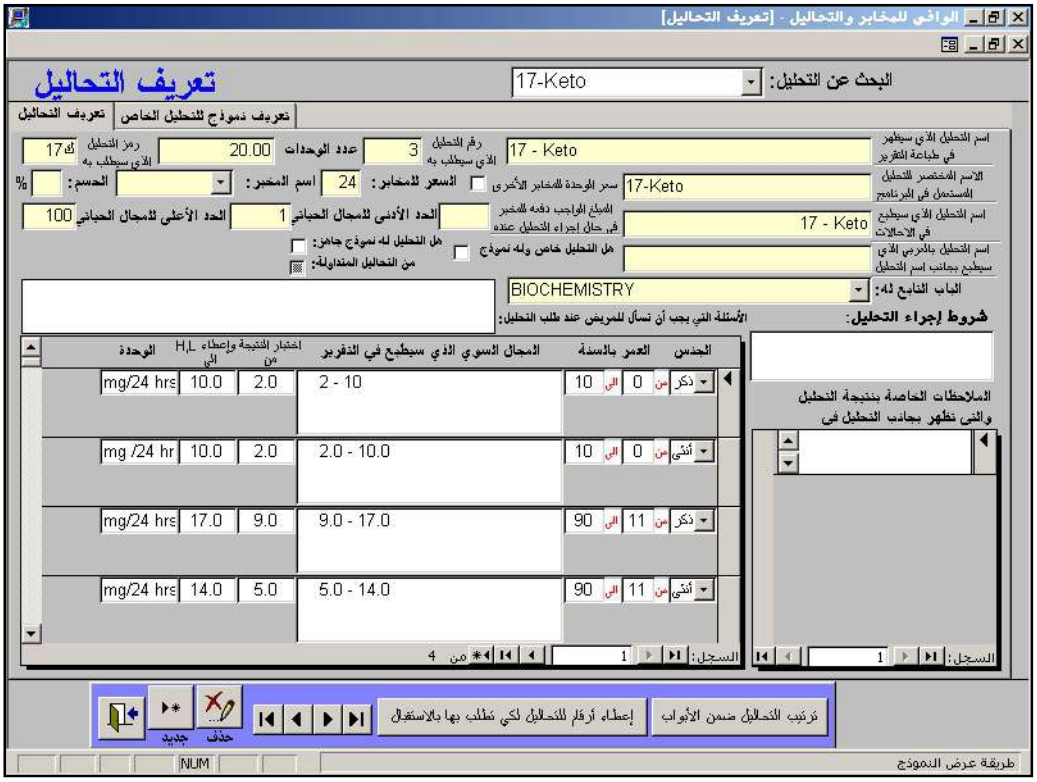
وتشتمل على القوائم التي تسمح بإضفاء الطابع الشخصي للمخبر وبتعريف العاملين والأجهـزة في المخبر وبتغيير في أسماء التحاليل الطبية الموجودة وفي المجموعات التي تنتمـي إليهـا.. انظر الشكل (– –).



الشكل (5-1- 3): واجهة الإعدادات في برنامج الوافي للمخابر والتحاليل

ج- واجهة تعريف التحاليل

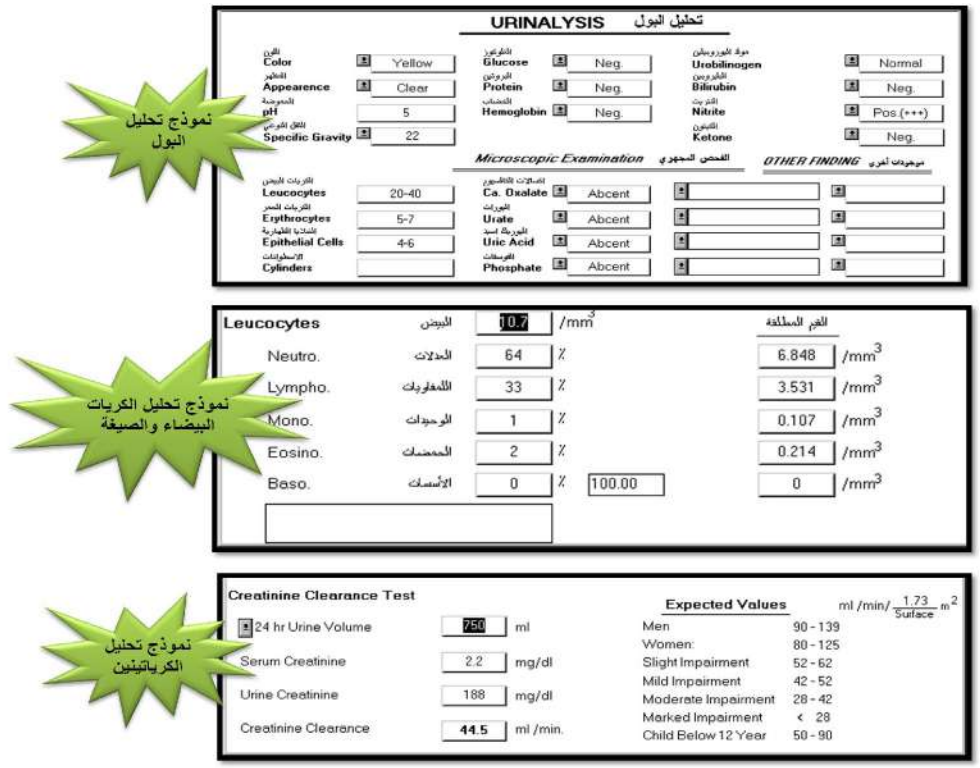
تفيد هذه الواجهة في إدخال المعلومات الضرورية عن التحاليل الطبية، مثل تسميات التحاليل بطرق مختلفة، وشروط إجراء التحاليل، وأسعار التحاليل التي تقاس بعدد الوحدات لكل تحليل مضروبة بسعر الوحدة لهذا التحليل. كما يتم تعريف الحد الأعلى والأدنى للمجال الحياتي لكل تحليل والذي يعبر عن أعلى وأدنى قيمة يمكن أن تصل إليها نتيجة التحليل.. انظر الشكل (5-1-4).



الشكل (5-1- 4): واجهة تعريف التحاليل في برنامج الوافي للمخابر والتحاليل

وتتضمن هذه الواجهة تعريف نماذج ظهور نتائج التحاليل والتي يُدَوِّن فيها المستخدم نتائج التحاليل، ومن ثم تُستخدم في طباعة نتائج التحاليل التي تقدم ورقياً إلى الزبون. وتتوفر في البرنامج نماذج جاهزة لطباعة نتائج معظم التحاليل المخبرية الموجودة، كما يتيح البرنامج الإمكانية للمستخدم لتصميم نماذج خاصة مختلفة لبعض التحاليل بمساعدة هذا البرنامج وذلك من خلال قائمة تعريف نموذج للتحليل الخاص الموجودة في الجهة اليسرى العليا في الشكل السابق. ومن الأمثلة على النماذج الجاهزة المتوفرة في البرنامج نموذج تحليل البول ونموذج

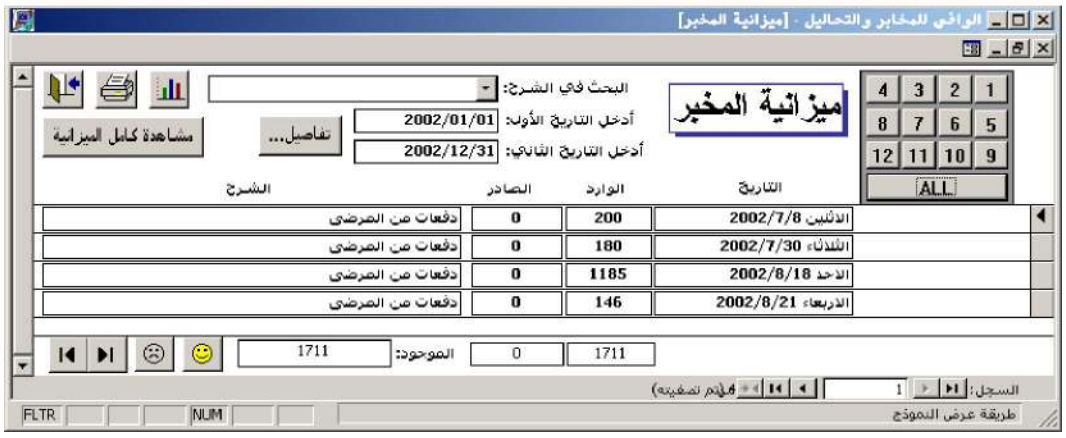
تحاليل الكريات البيض والصيغة ونموذج تحليل الكرياتينين[2]. انظر الشكل (5-1-5).



الشكل (- -): بعض نماذج طباعة نتائج التحاليل الطبية (تحليل البول، تحليل تعداد الكريات البيضاء والصيغة، وتحليل الكرياتينين) في برنامج الوافي للمخابر والتحاليل

د- واجهة ميزانية المخبر

وتُستخدم هذه الواجهة في الإطلاع على جميع واردات ومدفوعات المخبر التي تترحل تلقائياً من القوائم المحاسبية المختلفة في البرنامج إلى قائمة الميزانية. كما يتيح هذا البرنامج الحصولَ على مخطط بياني عن دخل المخبر في فترة معينة. انظر الشكل (- -).



الشكل (5-1- 6): واجهة ميزانية المخبر في برنامج الوافي للمخابر والتحاليل

[2] الكرياتينين هو مؤشر على سلامة وظائف الكلية، ويرتفع في حالات الفشل الكلوي الحاد والمزمن وحالات الانسداد البولي

ه- واجهة الموردين والمواد

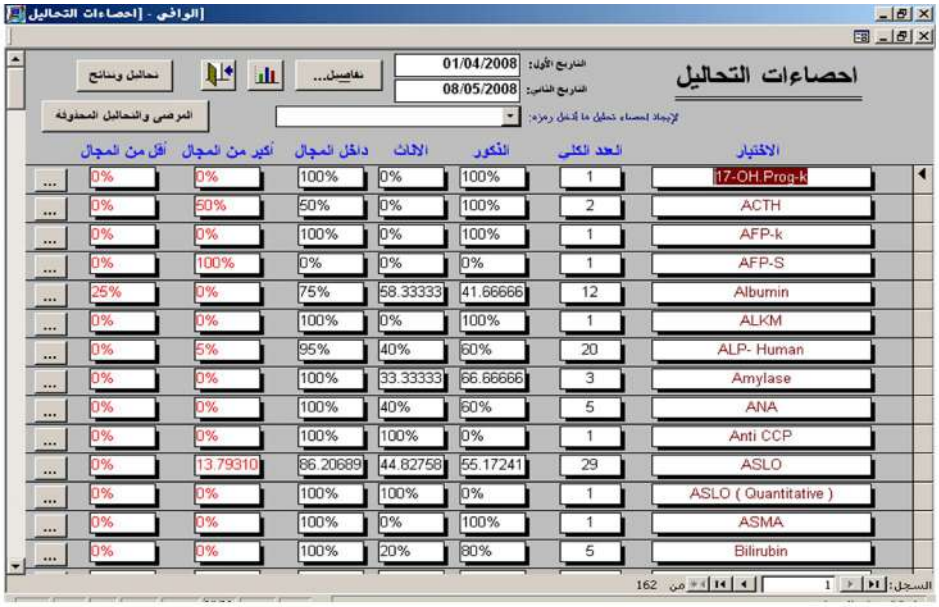
ويُستفاد من هذه الواجهة في أرشفة بيانات الموردين والمواد التي يشتريها المخبر من كل منهم، وتكاليف هذه المواد، وحسابات الموردين، كما يمكن أن يقوم البرنامج بالتنبيه التلقائي إلى قرب انتهاء كمية المواد التي يستعملها المخبر، وإلى قرب انتهاء فعالية المواد. انظر الشكل (5-1-7).



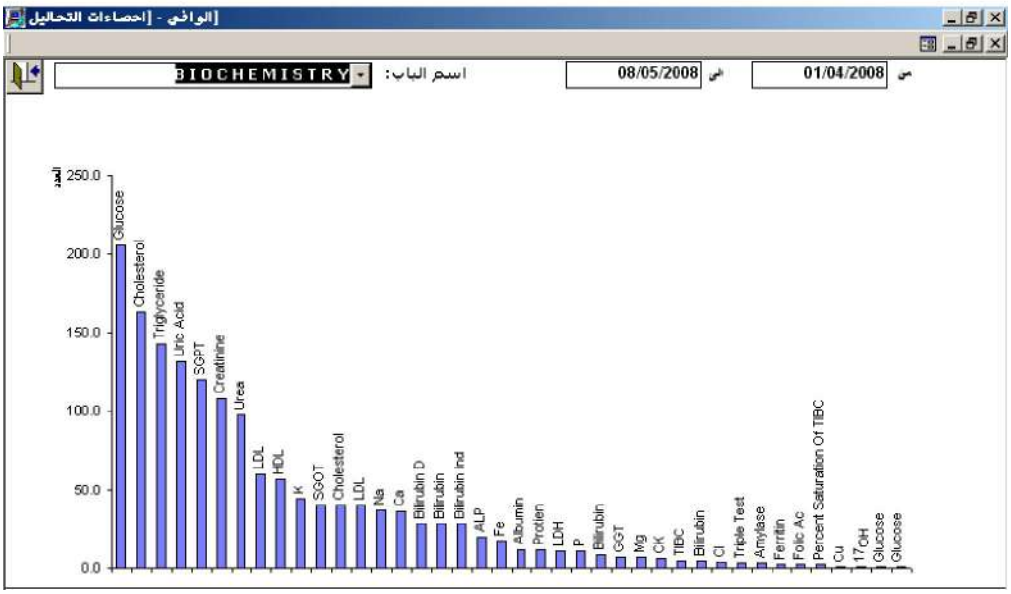
الشكل (5-1- 7): واجهة الموردين والمواد في برنامج الوافي للمخابر والتحاليل

و- واجهة إحصاءات التحاليل

وتفيد هذه الواجهة في معرفة جميع أنواع التحاليل التي أجراها المخبر وعددها الكلي بالإضافة إلى عدد المرضى الذكور والإناث الذين طلبوها وكذلك نسبة النتائج التي وقعت خارج المجال السوي أو داخله لبعض التحاليل.. وذلك كله خلال الفترة الواقعة بين تاريخين زمنيين يختارهما المستخدم. وتُعرض نتائج الإحصاءات إما في جدول كما هو مبين في الشكل (5-1-8)، أو في مخطط بياني كما في الشكل (5-1-9)، والذي يبين الترتيب التنازلي لعدد كل من التحاليل التابعة لباب التحاليل الكيميائية، والتي أجريت في المخبر خلال الفترة الممتدة من تاريخ 2008/4/1 ولغاية 2008/5/8.



الشكل (– –): واجهة طلب احصاءات التحاليل في برنامج الوافي للمخابر والتحاليل



الشكل (5–1– 9): واجهة احصاءات التحاليل في برنامج الوافي للمخابر والتحاليل

تعرفنا فيما سبق على أهم الأعمال التي يقوم بها برنامج الوافي للمخابر والتحاليل الطبية، وتبين أنه يساعد في إدارة المخبر محاسبياً بالإضافة إلى عمله الأساسي في تنظيم خدمة إجراء التحاليل الطبية وتقديم نتائجها إلى طالبيها. ولكن من الناحية العملية تبين أن المخبر المدروس لم يستخدم هذا البرنامج بكامل إمكانياته، بل اقتصر على بعض العمليات المحاسبية البسيطة المرتبطة بخدمة إجراء التحاليل الطبية وتحصيل تكاليفها من الزبائن. وسنتعرف في المبحث القادم وبالتفصيل على البيانات المستمدة من قاعدة بيانات هذا البرنامج في مخبر التحاليل الطبية المدروس والتي ستُستخدم في انجاز الجانب التطبيقي.

2•1•1•5 صياغة مسائل اكتشاف المعرفة من بيانات مخبر التحاليل الطبية المتاح

تتمحور المسألة الأساسية لاكتشافِ المعرفةِ في البحث الحالي حول التوصل إلى حقائق وأنماط مفيدة ومبتكرة كامنة في البيانات المخزنة في مخابر التحاليل الطبية والبيانات الشخصية والمرضية المرتبطة بها. وقد أُعِدَّت في البداية صياغة أولية عامة لبعض قضايا التنقيب المرغوب مناقشة حلها باستخدام هذه التقانة بالاسترشاد بالقضايا الطبية العالمية التي تستعين بهذه التقانة، وبما يتناسب مع الاطلاع الأولي على مضمون البيانات المتاحة للدراسة. وبعد الخوض أكثر في تفاصيل القضايا تبين أن الحالة ليست بهذه البساطة ومن الصعب تحويل العديد من هذه القضايا الأولية إلى مسائل تنقيب واضحة من البداية، وذلك لأنها المرة الأولى التي تخضع فيها هذه البيانات لأدوات اكتشاف المعرفة، ولأن البيانات المدروسة تنفرد بصفات خاصة نابعة من قلة عدد المتغيرات ومن كثرة الأخطاء مما يجعلها بعيدة نوعاً ما وغير صالحة تماماً لمناقشة القضايا الطبية العالمية التي تُستخدم تقانة اكتشاف المعرفة في حلها. ولذلك يمكن القول أن مسائل التنقيب هنا قد بنيت وطوِّرت شيئاً فشيئاً، وتبلور الكثير منها أثناء عمليات التغذية الراجعة الآتية من كل مراحل اكتشاف المعرفة. ونبين فيما يلي الشكل النهائي لمسائل التنقيب في بيانات مخابر التحاليل الطبية التي ستتم مناقشتها في هذا الفصل.

أ- اكتشاف أسباب الوفاة الأكثر تأثراً بالخاصيات المختلفة للمرضى من خلال دراسة خصائص المرضى المرافقة لكل سبب من أسباب الوفاة.

ب- التنبؤ بحالة خروج المرضى من المشفى سواء الشفاء أو الوفاة، من خلال دراسة بيانات الشخصية والمرضية.

ج- اكتشاف علاقات التأثير والاقتران بين نتائج التحاليل الطبية.

د- اكتشاف علاقات جديدة بين نتائج التحاليل الطبية والأمراض والصفات الشخصية للمرضى.

ه- البحث عن سجلات مرضى المشفى والتحاليل الطبية التي تحتوي على صفات مشتركة ووضعهم في فئات تحمل هذه الصفات.

و- هل يوجد علاقات بين أسباب القبول في المشفى والخصائص الكامنة في بيانات المرضى، أي دراسة تأثير قيم المتغيرات المختلفة (البيانات الشخصية والمرضية ونتائج التحاليل الطبية للمرضى) في كل نوع من أنواع أسباب القبول في المشفى.

ز- هل الدخل الوسطي الشهري للمخبر يختلف بحسب أشهر السنة أو بين الصيف والشتاء..؟ وكيف يمكن التنبؤ بالإيراد الشهري لمخبر التحاليل الطبية لعام 2009 باستخدام أدوات هذه التقانة؟

ح- التعرف على تسلسل (تعاقب) الأسباب المؤدية إلى تكرار القبول في المشفى خلال عامي 2007-2008.

ط- تصميم أداة للتنبؤ بالأشخاص المعرضين للوفاة باستخدام الصفات الشخصية والأسباب المرضية ونتائج التحاليل الطبية.

ي- اكتشاف المتغيرات الطبية الأكثر تأثيراً على تغير نتائج أحد التحاليل الطبية.

ك- استخدام تقانة التنقيب في البيانات في البحث عن الأسباب الكامنة وراء تحول المرضى إلى مشفى آخر.

يستدعي حل مسائل التنقيب السابقة أن يتم تطبيق عدة أنماط من التحليل على البيانات فمنها يحتاج إلى التنقيب الوصفي ومنها يحتاج إلى التنقيب التنبؤي وهذا ما سيظهر لاحقاً أثناء حل هذه المسائل، وسيتم العمل على حل بعض المسائل السابقة في مرحلة التنقيب في البيانات والبعض الآخر في مرحلة استخدام المعرفة المكتشفة.

2•1•5 تحديد المستفيدين المستهدفين

يستفيد من تطبيق هذه التقانة على البيانات المدروسة عددٌ من الجهات هي:

أ- الباحثون في المجال الطبي، حيث من الممكن أن تزود نتائج التنقيب في البيانات هؤلاء الباحثين بفرضيات وبعلاقات جديدة لكي يتعمقوا في دراستها من النواحي الطبية التخصصية.

ب- العاملون في مجال صحة السكان، كمنظمات الصحة والتي تهتم بدراسة المؤشرات الإحصائية عن الصحة العامة للسكان، وعن انتشار الظواهر المرضية زمنياً ومكانياً.

ج- مدراء المشافي ومدراء المخابر، حيث تساعدهم في وضع الخطط واتخاذ القرارات الإدارية المتعلقة بتنظيم وتنسيق الموارد المادية والبشرية مع احتياجات المشفى. فمثلاً قد تكثر حالات الأمراض الصدرية التي تحتاج إلى قبول في المشفى خلال أشهر معينة من العام، وإن التعرف على هذه الأزمنة بوساطة أساليب التنقيب يعني التنبؤ بها مما يساعد على تحضير كافة الاحتياجات اللازمة في الأوقات المناسبة.

د- الأطباء أثناء عملهم، حيث تساعد هذه الدراسة الأطباء في التعرف مثلاً على احتمالات تعرض المرضى المصابين بحالة معينة إلى الإصابة بأمراض أخرى، أو التنبؤ بحالة خروج المرضى من المشفى بناءً على خاصيات المرضى المختلفة.

ه- الجهات التي تغطي النفقات الطبية والتي تهتم بتقليص النفقات الطبية غير الضرورية، مثل شركات التأمين والمرضى.

3•1•5 تحديد الأدوات البرمجية التي ستُستخدَم في جميع المراحل

اعتمد انجاز الجانب التطبيقي من الدراسة الحالية على عددٍ من البرامج الحاسوبية، انظر الشكل (5-1-10)، وقد استُخدم كل منها لأداء مهام معينة في كل مرحلة من مراحل اكتشاف المعرفة. ونبين فيما يلي لمحة مختصرة عن البرامج الحاسوبية المستخدمة والمهام التي اعتُمد عليهم في انجازها.

أ- استُخدم برنامج أكسس **(Access 2007)** في احتواء البيانات الأولية التي تم الحصول عليها من مصادرها والتي بنيت وخزنت أصلاً في قواعد بيانات أكسس. وقد تم الاستمرار في التعامل مع هذا البرنامج أثناء مرحلة استيعاب البيانات وجزء من مرحلة تهيئة البيانات.

ب- استُعين ببرنامج **(Excel 2007)** في العديد من المراحل، وكان له استخدامات متنوعة منها الشائعة كالعمليات على الجدول والرسوم البيانية، ومنها التخصصية والجديدة كأداة لعرض البيانات الناتجة عن عمليات المعالجة التحليلية الآنية، وكواجهة لإجراء عمليات التنقيب في البيانات بالربط الداخلي مع مخدم **(SQL Server)**.

ج- استُخدم برنامج **SPSS** في مرحلة استيعاب البيانات، ولكن لم يتم التطرق إلى استخداماته خلال هذه الدراسة لأن فوائده انحصرت في دعم المعرفة الذاتية بالبيانات، وفي إجراء

بعض الأبحاث الإحصائية عن البيانات المتوفرة والتي قبلت للنشر في مجلة جامعة دمشق للعلوم الاقتصادية[3].

د- اعتُمد على مجموعة برامج (Microsoft SQL Server 2008) في تنفيذ العديد من المهام المتعلقة بعملية اكتشاف المعرفة، فقد استُخدم منها برنامج (Management Studio) في بناء مستودع البيانات، وتطويره وبناء الاستعلامات، واستُخدم برنامج (Visual Studio) في نقل البيانات من أكسس إلى (SQL Server) أثناء بناء مستودع البيانات، وفي بناء مكعب البيانات وانجاز عمليات المعالجة التحليلية الآنية، وأخيراً في تنفيذ المراحل الثلاثة الأخيرة من عملية اكتشاف المعرفة.



الشكل (5-1- 10): البرامج الحاسوبية المستخدمة في انجاز مراحل اكتشاف المعرفة من قواعد بيانات مخبر التحاليل الطبية

جرى الحديث في الفصل الرابع عن بعض الجوانب المتعلقة بمخبر التحاليل الطبية، وتابع المبحث الأول من الفصل الخامس مهمة التعريف بمصدر الحصول على البيانات المدروسة، بالإضافة إلى صياغة مسائل التنقيب وتحديد المستفيدين من نتائج عملية اكتشاف المعرفة، كما تم في هذا المبحث تحديد الأدوات البرمجية الضرورية لانجاز كل مرحلة من مراحل الجانب التطبيقي للبحث. ويعتبر هذا المبحث مبحثاً تمهيدياً لتحديد المسارات التي ستسير عليها بقية مراحل اكتشاف المعرفة.

[3] بحث (1): "**التحليل الإحصائي لبيانات المرضى المقبولين في مشفى في دمشق خلال عام 2007-2008 الذين أجري لهم تحليل السكر الصيامي في الدم**" وقد قُبل للنشر بتاريخ 2010/6/20. بحث (2): "**الأمراض العشرة الأكثر شيوعاً في دخول الأطفال الرضع إلى مشفى في دمشق خلال عامي 2007-2008**" وقد قُبل بتاريخ 2009/11/17.

الفصل الخامس

المبحث الثاني

مرحلة استيعاب بيانات مخبر التحاليل الطبية والبيانات المرتبطة بها

مقدمة

ذخرت مرحلة استيعاب بيانات مخبر التحاليل الطبية بالمحتويات والأعمال المتشعبة التي هدفت من جهة إلى التعرف على أدق التفاصيل المتعلقة بالبيانات المتاحة للدراسة كطبيعة البيانات وبنيتها والمصطلحات التخصصية، ومن جهة أخرى إلى بناء صورة عامة متكاملة عنها. وقد كان من الضروري انتقاء جزء فقط من الأفكار والمعارف عن البيانات لعرضها هنا، لأن التوسع في العرض يؤدي إلى ضياع التركيز على النقاط الأساسية التي تربطها مع المراحل القادمة. كما كانت صياغة التعابير اللغوية المعبرة عما تم انجازه في مرحلة التعرف على البيانات أصعب من المهام الحقيقية التي أُنجزت في هذه المرحلة، وذلك لأن بعض المعارف المتكونة اتصفت بالبساطة والبديهية لدرجة أنه كان من الصعب تحويلها إلى لغة أكاديمية وتقديمها بما يتناسب مع منهجية اكتشاف المعرفة. وسيتطرق المبحث الحالي إلى أهم النقاط المرتبطة بالتعرف على البيانات المتحققة خلال هذه المرحلة ووفقاً للخطوات النظرية الواردة في الفصل الثالث.

1•2•5 وصف مصادر البيانات التي تم الحصول عليها

يُعتبر مخبر التحاليل الطبية المدروس المصدر الأول للبيانات المراد دراستها، وقد تمت الاستعانة بمصادر إضافية للبيانات هما قاعدة بيانات مكتب القبول وقاعدة بيانات السجلات الطبية في المشفى الذي يتبع له المخبر، وذلك لأن البيانات الشخصية والمرضية هي من البيانات البديهية الواجب تواجدها أو ربطها مع بيانات مخبر التحاليل الطبية، ولكن لم توجه

العناية الكافية لتضمين هذه البيانات بدقة أثناء قيام العاملين في المخبر المدروس بالتوثيق الالكتروني لخدمات التحاليل. وتعتبر هذه القواعد قواعد بيانات علاقاتية، وسنتحدث فيما سيأتي عن هذه المصادر.

1•1•2•5 قواعد بيانات مخبر التحاليل الطبية

يَستخدمُ مخبرُ التحاليل الطبية المدروس برنامجَ الوافي للمخابر والتحاليل منذ عام 2006، كما سبق ذكره في المبحث السابق. ويقوم هذا المخبر بإجراء التحاليل الطبية لصنفين من المرضى هما المرضى المقبولين في المشفى ويُدْعَون المرضى الداخليين، والمرضى الآخرين غير المقبولين في المشفى ويُدْعَون المرضى الخارجيين ولكل من هذين الصنفين برنامج الوافي المستقل الخاص به. وعلى اعتبار أنه يصعب الاحتفاظ ببيانات المخبر لأكثر من سنة ضمن هذا البرنامج، فإن المسؤول عنه يحتفظ بنسخ احتياطية عن بيانات كل سنة في نهايتها ولكل من قاعدة بيانات المرضى الداخليين وقاعدة بيانات المرضى الخارجيين، ومن ثم يحذف تلك البيانات من البرنامج ليبدأ بتخزين بيانات جديدة مع بداية كل عام. ونتيجةً لذلك فقد توفرت لهذه الدراسة ستة قواعد بيانات من هذا المخبر كما هو مبين في الجدول (5-2-1).

الجدول (5-2- 1): قواعد بيانات مخبر التحاليل الطبية المتاحة للدراسة وأحجام كل منها

	اسم قاعدة البيانات	الحجم (ميغا بايت)	الحجم السنوي (ميغا بايت)	التزايد السنوي
2006	قاعدة بيانات المرضى الداخليين لعام 2006	358	750	-
	قاعدة بيانات المرضى الخارجيين لعام 2006	392		
2007	قاعدة بيانات المرضى الداخليين لعام 2007	596	1076	%43.5
	قاعدة بيانات المرضى الخارجيين لعام 2007	480		
2008	قاعدة بيانات المرضى الداخليين لعام 2008	806	1572	%46.1
	قاعدة بيانات المرضى الخارجيين لعام 2008	766		
		3398	3398	%44.8

ويتبين من الجدول السابق أن كمية البيانات قد تزايدت بنسبة وسطية تقدر بـــ 44.8% سنوياً خلال الأعوام من 2006 وحتى 2008، أي تضاعف تقريباً حجم القاعدة خلال سنتين، ولكن

لا نستطيع التوقع بحجم التزايد المستقبلي لهذه القواعد لأن ذلك يتعلق بالاحتياجات الصحية وبالطاقة الاستيعابية لهذا المخبر.

كانت تلك قواعد البيانات التي أتيحت لهذه الدراسة، وتم في البداية الاهتمام بالاحتفاظ بها جميعا للاستفادة منها، ولكن تقرر – بعد التعمق أكثر والتنقل التبادلي بين المراحل المختلفة لعملية اكتشاف المعرفة وخصوصاً المراحل الثلاثة الأولى– اختيار فقط قاعدتي بيانات المرضى الداخليين لكل من عامي 2007–2008 والاستغناء عن بقية القواعد الأخرى، وذلك تبعاً للأسباب التالية:

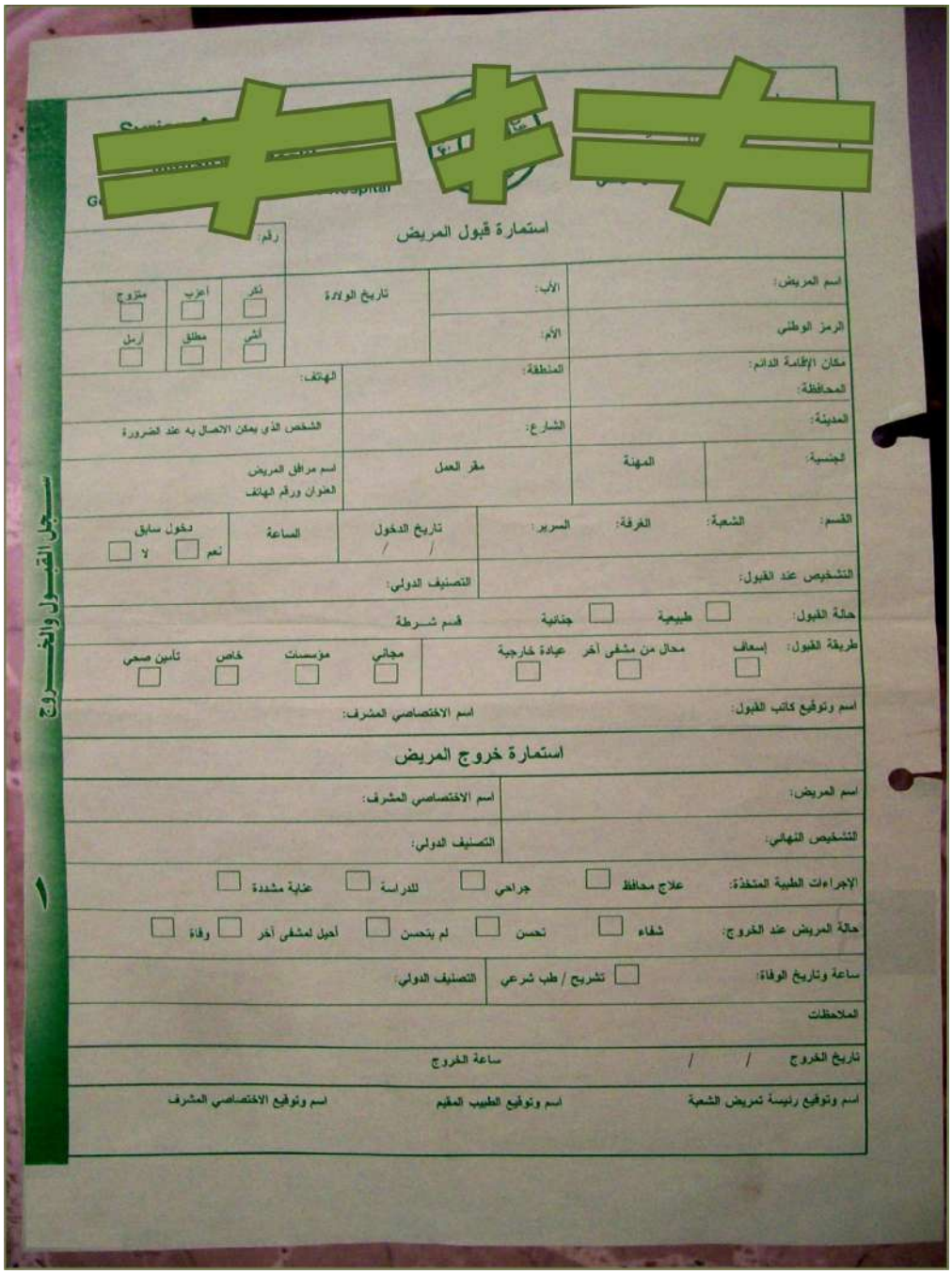
أ– تقرر الاستغناء عن قواعد بيانات المرضى الخارجيين الثلاثة، لعدم وجود أية بيانات شخصية أو بيانات عن الأسباب المرضية التي استدعت إجراء التحاليل للمرضى الخارجيين، سواء داخل قواعد بياناتهم أو في قواعد بيانات أخرى، مما يعني قلة عدد الخاصيات المتوفرة المفيدة في مناقشة قضايا البحث الأساسية، إضافةً إلى أن هذه القواعد تحتاج إلى تنظيف وإعادة تنظيم مما يتطلب المزيد من الوقت والجهد دون وجود جدوى حقيقية من ذلك.

ب– تم التخلي عن بيانات المرضى الداخليين لعام 2006 لسببين، الأول لأنها ضعيفة وناقصة حيث كان تطبيق البرنامج في بداياته الأولى، والسبب الآخر هو عدم وجود بيانات شخصية وطبية أخرى عن هؤلاء المرضى لأن الاستخدام الفعلي لقواعد بيانات مكتب القبول بدأ في عام 2007.

2•1•2•5 قاعدة بيانات مكتب القبول في المشفى

يجري غالباً الفحص الأولي للمرضى في المشفى إما عن طريق قسم الإسعاف أو عن طريق قسم العيادات التابعة للمشفى، ومن ثم يتم تحويل المرضى الذين تتوفر فيهم الأسباب المرضية التي تستدعي العلاج في المشفى إلى مكتب القبول لتسجيل بياناتهم ضمن وثيقة ورقية تدعى (استمارة قبول المريض)، وتتضمن هذه الاستمارة الورقية بيانات شخصية للمرضى مثل الاسم وتاريخ الميلاد ومكان الإقامة، والمهنة، كما تتضمن بيانات الحالة المرضية مثل تاريخ الدخول للمشفى، والتشخيص عند القبول، وطريقة القبول.. انظر الشكل (5–2–1).

ومن المفترض أن يأخذ كل مريض يدخل إلى المشفى رقماً خاصاً به هو رقم اضبارة المريض، وهو الرقم الذي يجب أن يُستخدم للتعرف على هذا المريض عند كل عملية قبول في المشفى، ولكن في الحقيقة لا يوجد اهتمام من قبل المشفى في التأكيد على هذا التوثيق مما سمح بوجود الكثير من التكرار، حيث من الممكن أن يأخذ مريض ما رقم جديد عند كل قبول في المشفى.


سجل القبول والخروج

استمارة قبول المريض

رقم:

اسم المريض:
الأب:
تاريخ الولادة
ذكر
أعزب
متزوج
الرمز الوطني
الأم:
أنثى
مطلق
أرمل
مكان الإقامة الدائم:
المحافظة:
المنطقة:
الهاتف:
المدينة:
الشارع:
الشخص الذي يمكن الاتصال به عند الضرورة
الجنسية:
المهنة
مقر العمل
اسم مرافق المريض
العنوان ورقم الهاتف
القسم:
الشعبة:
الغرفة:
السرير:
تاريخ الدخول
/ /
الساعة
دخول سابق
نعم
لا
التشخيص عند القبول:
التصنيف الدولي:
حالة القبول:
طبيعية
جنائية
قسم شرطة
طريقة القبول:
إسعاف
محال من مشفى آخر
عيادة خارجية
مجاني
مؤسسات
خاص
تأمين صحي
اسم وتوقيع كاتب القبول:
اسم الاختصاصي المشرف:

استمارة خروج المريض

اسم المريض:
اسم الاختصاصي المشرف:
التشخيص النهائي:
التصنيف الدولي:
الإجراءات الطبية المتخذة:
علاج محافظ
جراحي
للدراسة
عناية مشددة
حالة المريض عند الخروج:
شفاء
تحسن
لم يتحسن
أحيل لمشفى آخر
وفاة
ساعة وتاريخ الوفاة:
تشريح / طب شرعي
التصنيف الدولي:
الملاحظات
تاريخ الخروج / /
ساعة الخروج
اسم وتوقيع رئيسة تمريض الشعبة
اسم وتوقيع الطبيب المقيم
اسم وتوقيع الاختصاصي المشرف

الشكل (5-2- 1): استمارة قبول المريض، واستمارة خروج المريض

كما ويقوم مكتب القبول إضافةً إلى ذلك بالتوثيق الالكتروني لبيانات المرضى المقبولين في المشفى من خلال قاعدة بيانات بسيطة أنشأها أحد العاملين في هذا القسم بواسطة برنامج أكسس (Access)، وتشتمل هذه القاعدة على معظم البيانات الموجودة في استمارة قبول المريض الورقية، انظر الشكل (5-2-2). وقد ساهمت مشاكل نطق أو توثيق حروف اللغة العربية في تكرار أو ازدواجية إدخال البيانات إلى الحاسوب، فمثلاً قد يأخذ اسم ديمة عند الادخال إلى الحاسوب الأشكال التالية (ديمة، ديمه، ديما)، مما يَحول دون تمكين برنامج التوثيق الالكتروني من تمييز المرضى الجدد عن المرضى الذين يتكرر قبولهم في المشفى.

وقد بدأ العاملون في مكتب القبول بتطبيق هذا البرنامج في بداية عام 2005، ولفترة محدودة لا تتعدى بضعة أشهر، ثم عاودوا استخدامه مع بداية عام 2007. ويقدر حجم قاعدة بيانات مكتب القبول 496 ميغا بايت في نهاية عام 2008. ويساعد هذا البرنامج على طلب بعض الإحصاءات البسيطة عن البيانات المتوفرة فيه كما هو مبين في الشكل (5-2-3).

ويساعد توفر بعض البيانات عن المرضى ما قبل عام 2007 على حل مسائل في التنقيب في البيانات تتعلق بتتبع التاريخ المرضي أو الأسباب المرضية للمرضى الذين يتكرر دخولهم إلى المشفى. ولذلك سيتم الاحتفاظ في هذه الدراسة بجميع السجلات المتوفرة عن المرضى منذ عام 2005.



الشكل (5-2- 2): النسخة الالكترونية لاستمارة قبول المرضى في المشفى والموجودة في برنامج قاعدة بيانات مكتب القبول



الشكل (5-2- 3): واجهة التقارير الاحصائية في برنامج مكتب القبول

3•1•2•5 قاعدة بيانات السجلات الطبية في المشفى

يعمل قسم السجلات الطبية في المشفى على تتبع مصير المرضى بعد قبولهم في المشفى، ويعمل هذا القسم على التوثيق الورقي لبيانات المرضى وفق نموذج استمارة خروج المريض المرافقة لاستمارة قبول المريض، والمبينة في الشكل (5-2-1).

كما يقوم العاملون في هذا القسم بالتوثيق الالكتروني لبيانات استمارتي القبول والخروج في قاعدة بيانات بسيطة مبنية بواسطة برنامج أكسس من قبل أحد العاملين في هذا القسم.

وقد بدأ العاملون في قسم السجلات الطبية بتطبيق هذا البرنامج بشكل متقطع في عام 2006، وبشكل متواصل مع بداية عام 2007، وقد تم الحصول على نسخة عن البيانات الموجودة في قاعدة بياناتهم على شكل استعلام مخزن في جدول اكسل، وقد بلغ حجمه 14 ميغا بايت تقريباً، ويتضمن بيانات المرضى حتى نهاية الشهر الخامس من عام 2008 فقط، ويقدر حجم هذه القاعدة بـــ 350 ميغا بايت تقريباً. وقد رفض مكتب السجلات الطبية – الذي سمح سابقاً بالاستفادة من البيانات الالكترونية لديه– أن يتم الحصول على بقية بيانات عام 2008 أو أية بيانات مستقبلية.

تتمتع الأدوات البرمجية للتنقيب في البيانات بالقدرة على التعامل والتنقيب في قواعد البيانات الضخمة والتي تقاس أحجامها بين عدد من الغيغابايت (1 gigabyte= 10^9 byte) إلى عدد من

التيرابايت (1terabyte=10^{12} byte)، وهي من القدرات الأساسية التي تميز عمليات التنقيب وتميز أدواتها البرمجية عن الأدوات المختصة بالتحليل الإحصائي البحت. وقد قُدِّرَ الحجم الكلي لقواعد البيانات التي ستُطبَّق تقنيات التنقيب عليها بـــ (2.248 غيغا بايت) أي (350+496+806+596)، وهي من الأحجام الصغيرة نسبياً لعمليات التنقيب، ولكنها تعتبر مناسبة للحالة التي ستُدرس، وذلك لأن هذه البيانات لم يَسبق وأن طُبِّق عليها أي نوع من أنواع الدراسات سواء الإحصائية أو اكتشاف المعرفة، كما وتبين من خلال التجربة الحالية في تطبيق تقانة اكتشاف المعرفة على بيانات مخبر التحاليل الطبية المدروس أنها بحاجة إلى الكثير من الوقت والجهد خصوصاً أثناء عمليات التنظيف وعمليات ربط هذه القواعد مع قواعد بيانات مكتب القبول والسجلات الطبية، والتي ستزداد صعوبةً في حال كانت مثل هذه القواعد أكبر من ذلك. أما في المستقبل فيمكن الاستفادة من هذه الدراسة في تطبيق تقانة اكتشاف المعرفة على قاعدة بيانات ضخمة ومبنية وفق أسس سليمة، وتكون مشتركة بين عدد كبير من المخابر الطبية التابعة لنظام المعلومات الصحية الذي تسعى وزارة الصحة السورية إلى بنائه في المشافي المحلية مستقبلاً.

2•2•5 وصف جداول البيانات من حيث المضمون والحجم

تختص هذه الفقرة بالتعرف على الجداول التي يُعتقد بأنها ضرورية وتستطيع أن تؤدي خدمة ما لعملية اكتشاف المعرفة.

1•2•2•5 قواعد بيانات مخبر التحاليل الطبية

تتألف هذه القواعد من قاعدتي بيانات مخبر التحاليل الطبية لكل من عامي 2007 و2008، حيث تتكون كل منهما من 43 جدولاً أساسياً موزعين على النحو التالي:

أ- 38 جدولاً لتخزين نتائج التحاليل المتعددة النتائج (التي تتكون من عدة نتائج فرعية). مثل جدول (WBC) والخاص بنتائج تحليل الكريات البيض ويتكون من سبعة نتائج فرعية وتخزن في سبعة أعمدة. ويعاد تخزين بيانات بعض هذه الجداول في الجدول (PatTest) لهدف الطباعة. بينما يبقى البعض الآخر مستقلاً ولا يعاد تخزين نتائجه مثل الاختبارت المتعلقة بدراسة الأحياء المجهرية والفطريات (Microbiology & Mycology).

ب- جدول القيم الطبيعية للتحاليل الطبية (NormalRanges)، ويتألف من 773 سجل في عام 2007، ومن 793 سجل في عام 2008، حيث تمت إضافة تحاليل جديدة إلى المخبر وإلى قاعدة البيانات.

ج- جدول البيانات الشخصية للمرضى (Patient) ويتألف من 35398 سجل في عام 2007 ومن 32455 سجل في عام 2008.

د- جدول كل التحاليل الطبية التي أجريت للمرضى (PatTest) ويتضمن نتائج كافة التحاليل الطبية وحتى التي تتألف من عدة نتائج، ولكنه لا يحتوي على نتائج بعض التحاليل التي لها نماذج عرض خاصة كالمخططات البيانية. ويحتوي الجدول في عام 2007 على 325394 سجل، وفي عام 2008 على 259582 سجل.

ه- جدول أسماء التحاليل وبعض الخصائص عنها (TestInfo)، ويتألف من 361 تحليل طبي أو سجل في عام 2007، ومن 367 تحليل طبي في عام 2008.

و- جدول أسماء الأبواب أو الأنواع التي تندرج ضمنها التحاليل الطبية (Doors)، ويتكون من18 سجل أو باب في عام 2007 ومن 21 سجل في عام 2008، أي تمت إضافة بابين جديدين وهما في السجلين الأخيرين من الجدول المعروض في الشكل (5-2-4).

وتختلف أحياناً تسميات وتقسيمات التحاليل الطبية من مرجع لآخر ومن برنامج إدارة مخابر لآخر، وهي من الصعوبات التي تعيق توحيد قواعد البيانات الموجودة حالياً في قاعدة بيانات واحدة. وبشكل عام يمكن أن نبين بعض الأمثلة عن أنواع التحاليل الطبية وما يتفرع عن كل منها كما يلي[4]:

1- التحاليل الكيميائية الحيوية (Biochemistry)، ويتفرع عنها عدد من التحاليل مثل تحاليل السكر الصيامي والكرياتينين، وحمض البول والكوليسترول، ، والشحوم الثلاثية، والبروتينات وتتضمن (ألبومين، بيلروبين، الكالسيوم، المغنيزيوم، البوتاسيوم، الحديد، الصوديوم..)..

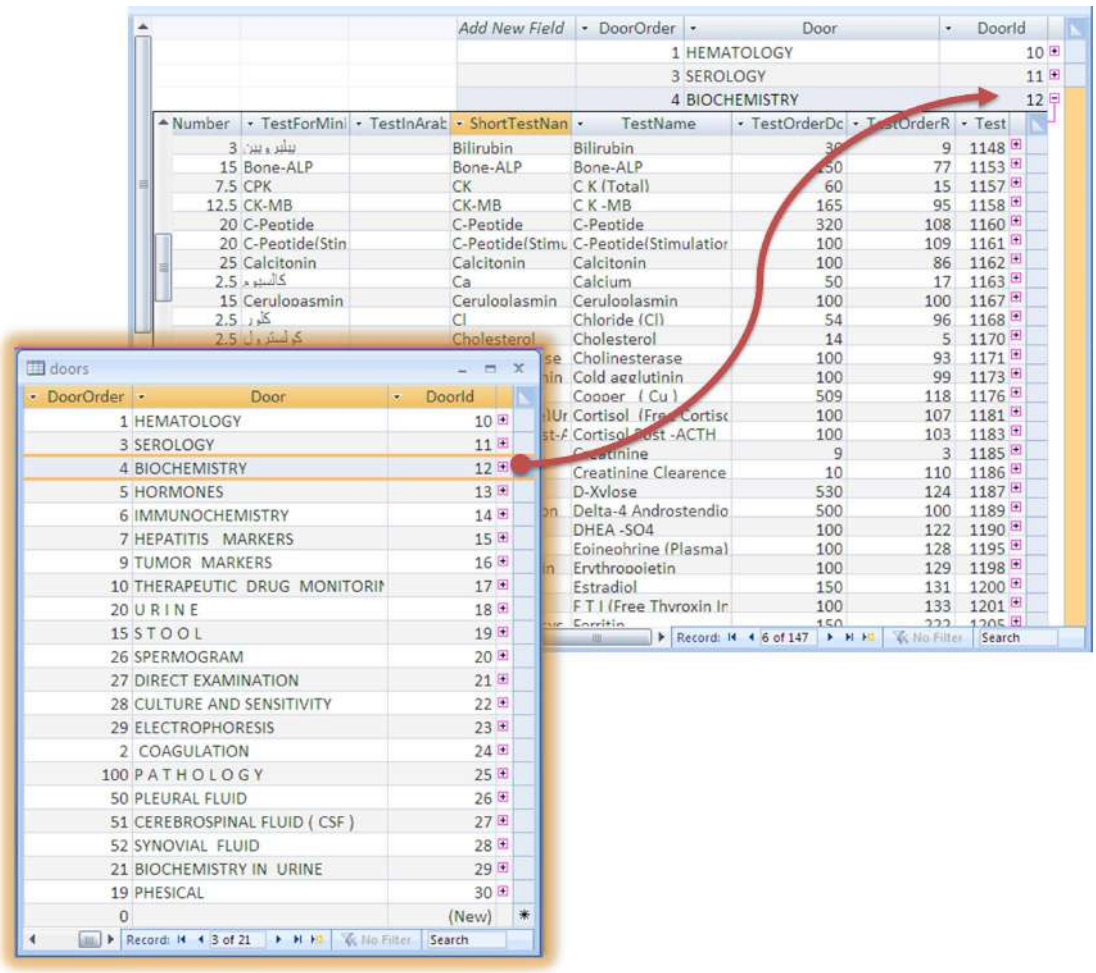
2- التحاليل الدموية أو علم الدم (Hematology)، وهي التحاليل المتعلقة بحالة خلايا الدم وهي من التحاليل الأساسية وتشكل نقطة البداية في التعرف على حالة المريض

[4] حاولنا الاختصار والتحدث بشكل عام دون التوسع في التفاصيل الطبية البحتة، التي قد تضيع نقاط التركيز في هذا البحث.

الصحية، وتتضمن هذه التحاليل: تحاليل كريات الدم البيضاء (WBC) وكريات الدم الحمراء (RBC)، والخضاب (Hemoglobin)، والهيماتوكريت (Hematocrit) والصفائح (Platelets) وسرعة التثفل (ESR)، وزمن النزف (Bleeding Time) وزمن التجلط (Clotting Time)، والخلايا المنجلية (Sickle Test)..

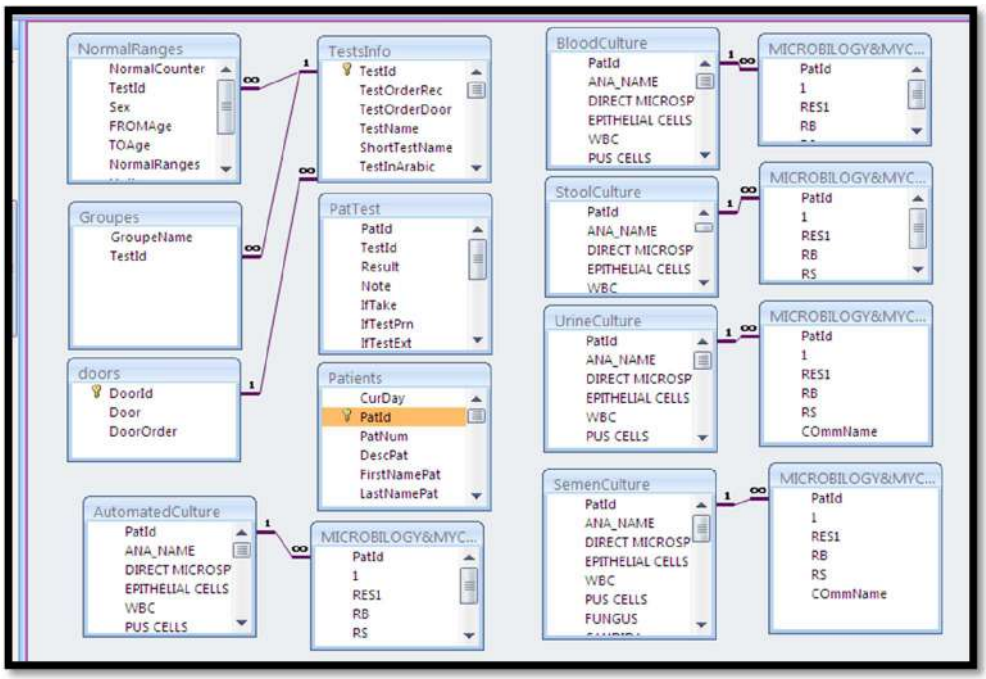
3- تحاليل الهرمونات (Hormones)، وهي التحاليل المتعلقة بالهرمونات التي تفرزها الغدد الصماء في الجسم مثل هرمونات الغدة الدرقية، وهرمونات الغدة النخامية، وهرمونات البنكرياس، هرمونات الكلية..

4- تحاليل علامات التهاب الكبد (Hepatitis Markers)، وتتضمن عدد من التحاليل الطبية لكل نوع من أنواع التهاب الكبد (A,B,C).



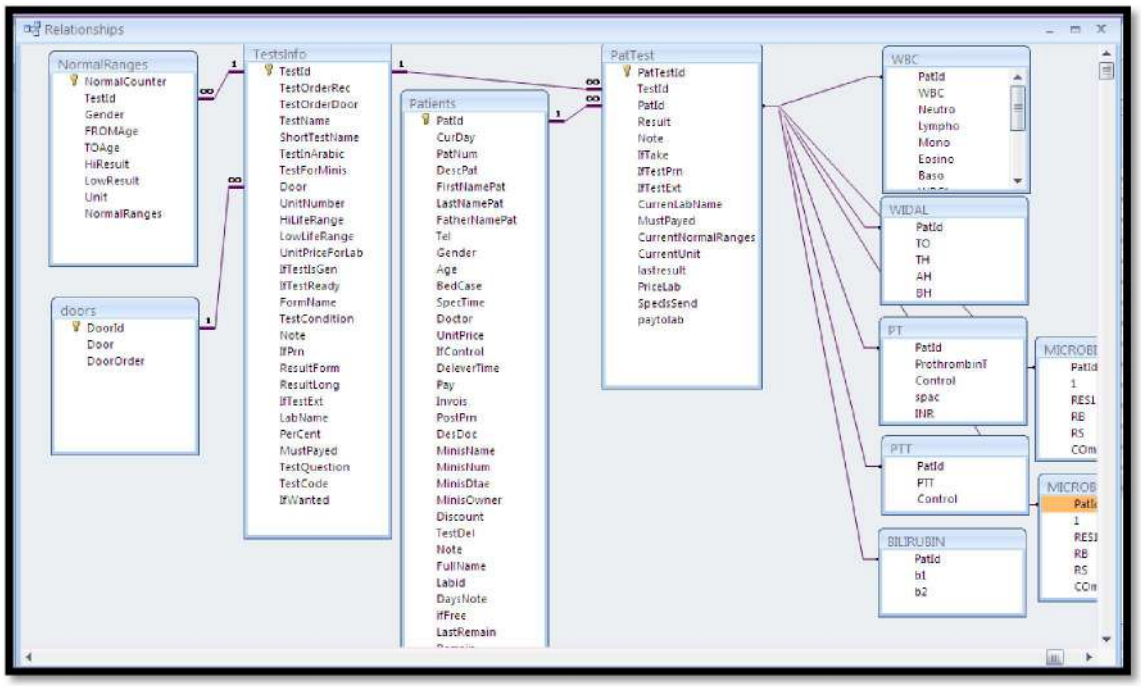
الشكل (5-2- 4): جدول أبواب التحاليل الطبية المخبرية (Doors)، وخلفه الجدول الفرعي لأسماء التحاليل المنتمية إلى باب التحاليل الكيمائية الحيوية (Biochemistry)

ز- تبيَّنَ من خلال الاطلاع على خريطة العلاقات الموجودة في برنامج الوافي للمخابر وجود بنية ضعيفة في تصميم العلاقات بين الجداول وخصوصاً جدولي المرضى (Patient) وتحاليل المرضى (PatTest)، حيث لا يوجد مخطط (كيان-علاقة) يحتوي جميع العلاقات في البرنامج مما أدى إلى حدوث فوضى في عمليات إدخال البيانات، فمثلاً يوجد أسماء لمرضى دون وجود تحاليل لهم، ويوجد تحاليل دون وجود أرقام هويات مرضى.. كما يوجد العديد من الجداول التي ليس لها حقول مفتاحية وإنما تعتمد على استخدام حقل (PatId) في الربط بين الجداول، انظر الشكل (5-2-5). ولن تجري المحاولة في البحث الحالي لإحداث تحسينات على العلاقات في قاعدة البيانات الموجودة[5]، ولكن سيتم رسم مخطط (كيان- علاقة) جديد ومبسط بهدف توضيح العلاقات المنطقية بين الجداول فقط، ويتكون فقط من أهم الجداول التي ستُستخدم في هذه الدراسة، فمثلاً تم الاستغناء عن عدد من جداول التحاليل الطبية الخالية من البيانات. وقد نشأت الضرورة أثناء رسم هذا العلاقات إلى استخدام نسخ خالية من البيانات عن بعض الجداول الموجودة التي تعذَّر تطبيق علاقة (1-∞) في ربطها معاً، وذلك بسبب وجود بيانات غير نظيفة أو مكررة أو لا يوجد لها بيانات مقابلة في الجداول الأخرى، ويبين الشكل (5-2-6) المخطط المقترح للعلاقات بين جداول مخبر التحاليل الطبية.



الشكل (5-2- 5): العلاقات بين الجداول في برنامج الوافي للتحاليل المخبرية

[5] لأن هذا البحث لا يهدف إلى التوصل إلى قاعدة بيانات بديلة للقواعد السابقة لاستخدامها في العمليات الجارية اليومية في المشفى.



الشكل (5-2- 6): مخطط كيان- علاقة، للجداول الرئيسة في قاعدة بيانات مخبر التحاليل الطبية المدروس

2•2•2•5 قاعدة بيانات مكتب القبول

تتكون قاعدة بيانات مكتب القبول من أربعة جداول أساسية هي:

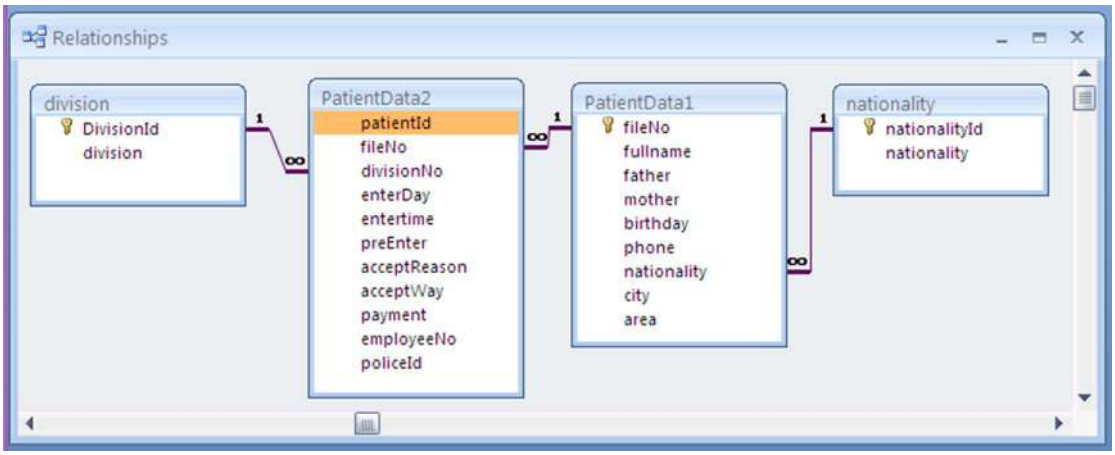
أ- جدول بيانات المرضى1 (PatientData1): ويتضمن البيانات الشخصية للمرضى، ويتألف من 43455 سجل.

ب- جدول بيانات المرضى2 (PatientData2): ويتضمن بيانات دخول المرضى إلى المشفى، ويتألف من 47728 سجل. ويرتبط هذا الجدول مع الجدول السابق، حيث يَفترض أن لكل مريض سجل واحد للبيانات الشخصية ويرتبط مع سجل واحد على الأقل من سجلات القبول في المشفى. أي يوجد في هذا الجدول توثيق لـــ 4273 (43455-47728) حالة قد تكرر فيها دخول المرضى المسجلين في الجدول (PatientData1) إلى المشفى.

ج- جدول الجنسية (nationality): ويتألف من 48 سجل، ويُعبر عن جنسيات المرضى الذين قبلوا في المشفى.

د- جدول أسماء الشعب الطبية التي يتم قبول المرضى فيها بحسب الاختصاصات الطبية اللازمة لعلاجهم (Division)، ويتألف من 27 سجل أو شعبة، منها على سبيل المثال (الشعبة الهضمية، شعبة الحواضن، شعبة الجراحة العامة، شعبة العناية القلبية..).

ترتبط الجداول السابقة في قاعدة بيانات مكتب القبول وفق مخطط الكيان–علاقة المبين في الشكل (5–2–7).



الشكل (5–2– 7): مخطط كيان–علاقة لقاعدة بيانات مكتب القبول

3•2•2•5 قاعدة بيانات مكتب السجلات الطبية

تتكون هذه القاعدة من جدولين أساسيين هما:

أ– جدول بيانات المرضى ويتضمن كافة بيانات المرضى الشخصية وبيانات القبول في المشفى وبيانات الخروج من المشفى، أي يوجد لكل عملية قبول في المشفى سجل في هذا الجدول. ويتكون هذا الجدول من 36582 سجل وذلك لغاية تاريخ 2008/5/28

ب– جدول أسماء الشعب الطبية في المشفى ويتكون من 27 شعبة، كما في قاعدة بيانات القبول.

وقد كانت بيانات هذه القاعدة مخزنة في جدول اكسل كما سبق وذُكر، ثم قمنا باستيرادها وحفظها في برنامج أكسس مع قاعدة بيانات مكتب القبول لكي يتم التعامل معها بصورة أفضل.

3•2•5 التعرف على بنى الجداول في قواعد البيانات المدروسة

تحتاج عملية اكتشاف المعرفة إلى التوسع أكثر وأكثر في التعرف على مكونات الجدول التي تحتاجها هذه العملية، وسيتم في الفقرات التالية التعرف على بنى الجداول والتركيز على الخاصيات التي تهم الدراسة.

1•3•2•5 التعرف على بنى الجداول في قاعدة بيانات مخبر التحاليل الطبية

1•1•3•2•5 بنية جدول المرضى (Patients)

يتضمن جدول المرضى كافة البيانات المتعلقة بخدمة إجراء التحاليل الطبية من حيث البيانات الضرورية لعملية الطباعة، ومن حيث بيانات استيفاء تكاليف إجراء التحاليل الطبية، انظر الشكل (5-2-8)، وقد تبين من خلال التعرف على البنية التصميمية (Design View) لجدول المرضى (Patients) ما يلي:

أ- يتألف جدول المرضى من 33 حقل، منها 4 حقول إلزامية وهي ترقيم المريض واسمه وكنيته وجنسه، ومنها مخصص لأمور طباعة النتائج مثل صفة المريض واسم الطبيب الطالب للتحليل وهي غير إلزامية، أما بقية الحقول فيختص معظمها بكيفية تسديد أجور التحاليل الطبية، وهي اختيارية وغير إلزامية مما سمح بامتلائها بالقيم المفقودة، وأدى إلى الاضطرار للاستغناء عنها في هذه الدراسة. ويمكن أن تُستخدم هذه الحقول – لو كانت تذخر بالبيانات – في إجراء عمليات تحليل وتنقيب واسعة ومفيدة عن الأمور المالية التي يتضمنها هذا الجدول، فمثلاً يمكن التعرف على صفات الزبائن وتوقع المتهربين من الدفع من خلال دراسة المبالغ المدفوعة والمبالغ المستحقة والمبالغ التي لن تُسدد أبداً (ديون معدومة) وأسماء الجهات التي تدفع التكاليف..الخ.

ب- قام مصمم برنامج الوافي باقتراح قيمة افتراضية أولية لعمر المريض عند إدخال بياناته إلى البرنامج الحاسوبي وهي 30 سنة، وتبقى هذه القيمة كما هي إلا إذا استبدلها مستخدم البرنامج بالعمر الحقيقي للمريض. وقد وجدنا أن 85% من المرضى الذين أجريت لهم تحاليل لهم عمر مساوي للعمر المقترح، مما يدل على إهمال العمر الحقيقي لمعظم المرضى، ويقود إلى ضَعف مدلولات خاصية العمر في هذه القاعدة ويجعلنا نقترح أن يتم حذف هذا الحقل وخصوصاً أنه يتوفر حقل بديل هو تاريخ الميلاد في قاعدة بيانات مكتب القبول.

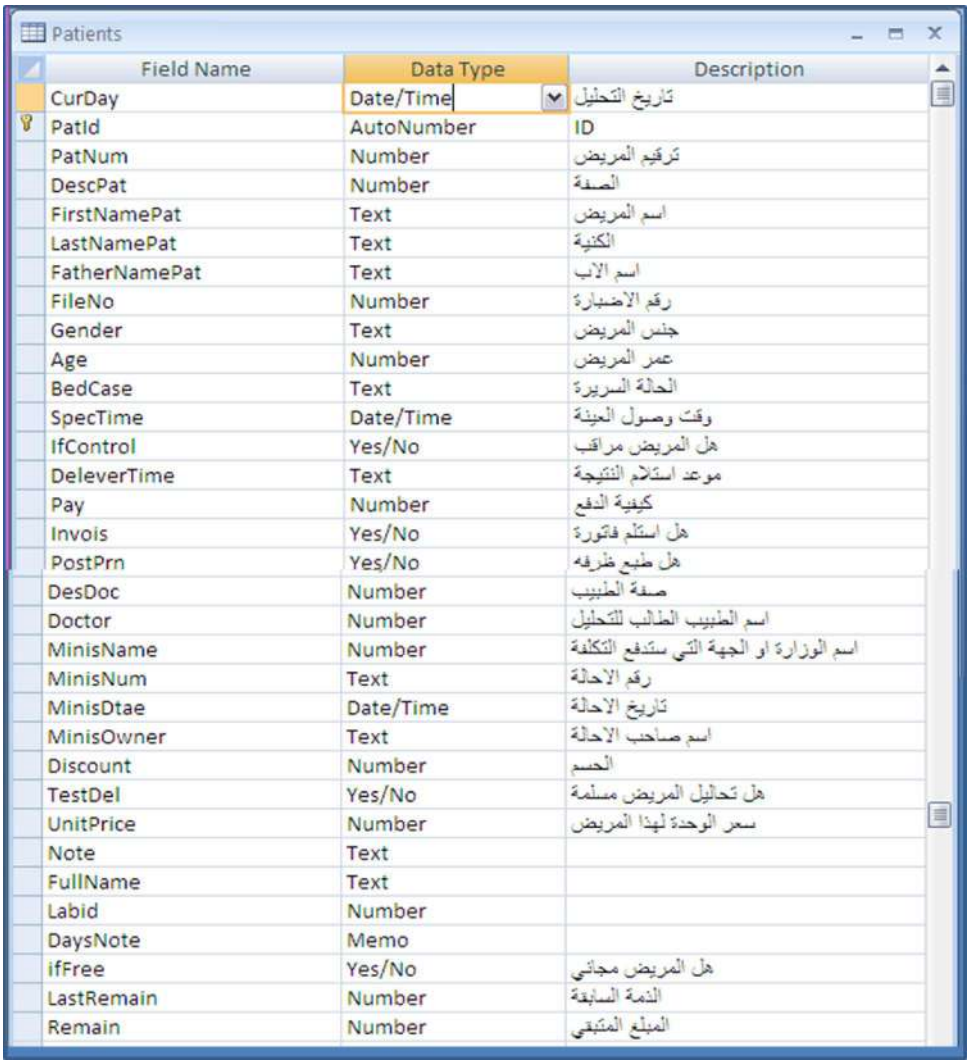
ج- يتضمن هذا الجدول حقل الحالة السريرية للمرضى (BedCase)، وهو من الحقول المفيدة في دراسة العلاقة الارتباطية بين السبب المرضي وبين نتائج التحاليل الطبية، ولكن

للأسف عثرنا في هذا الحقل على 9796 قيمة مفقودة من أصل 32455 قيمة أي بنسبة 30% تقريباً، والقيم الموجودة هي فقط أسماء الشُعب الطبية في المشفى التي طلبت إجراء التحاليل. وبناءً على ذلك لن يتم الاعتماد على هذا الحقل بل هنالك حقول بديلة عنه في قاعدتي بيانات مكتب القبول والسجلات الطبية.

د- يحتوي هذا الجدول على حقل اسمه (Tel) وقد خُصِّص هذا الحقل عند تصميم البرنامج لتسجيل رقم هاتف المريض، ولكن في الحقيقة استخدمه عاملو المخبر لتخزين رقم اضبارة المريض نظراً لخلو هذه القاعدة من حقل خاص برقم الاضبارة. ويحتوي هذا الحقل على الكثير من القيم المفقودة لأن التسجيل فيه اختياري، ولا يوجد تأكيد من إدارة المخبر على توثيق رقم الاضبارة الكترونياً. و من الملفت للانتباه أيضاً أن هذا الحقل من النمط النصي مما سمح بتخزين رموز غير رقمية فيه، وأدى إلى وجود أخطاء مطبعية عند كتابة رقم الاضبارة. كما وأدى النمط النصي لهذا الحقل إلى صعوبة ربط هذا الجدول مع جدول مكتب القبول حيث أن الحقل هناك رقمي وليس نصي، وتطلب الأمر الكثير من عمليات التنظيف والتنسيق التي سيرد ذكرها في المبحث القادم.

ه- يوجد حقل (IfFree) وله قيمتان (Yes/No) خاص للتمييز بين المرضى الذين يدفعون أجور التحاليل والذين لا يتوجب عليهم الدفع (المرضى المجَّانيِّين)، وفي الحقيقة كانت جميع قيم الحقل (No) أي لم يتم العثور في هذا الجدول على مرضى مجَّانيِّين في خدمة التحاليل الطبية.

و- يمثل حقل سعر الوحدة للمريض (UnitPrice) القيمة الترجيحية للسعر الذي يدفعه كل مريض لقاء التحاليل الطبية التي تُجرى له، وتقوم عادةً وزارة الصحة باقتراح حدين أعلى وأدنى لهذه التسعيرة، وتلتزم المخابر بأن تضع أسعارها ضمن هذين الحدين، وتختلف الأسعار بين المرضى بحسب وجهات نظر إدارة المخبر عن المرضى الذين ترغب بتخفيض الأسعار لهم. وتعتبر أسعار التحاليل الطبية في المخبر الحكومي المدروس أقل منها في المخابر الخاصة الأخرى، وقد كان يحاسب المرضى في عام 2007 بإحدى القيمتين (15، 20) للوحدة، وفي عام 2008 بإحدى القيمتين (20، 25) للوحدة.



الشكل (5-2- 8): بنية جدول المرضى في قاعدة بيانات مخبر التحاليل الطبية

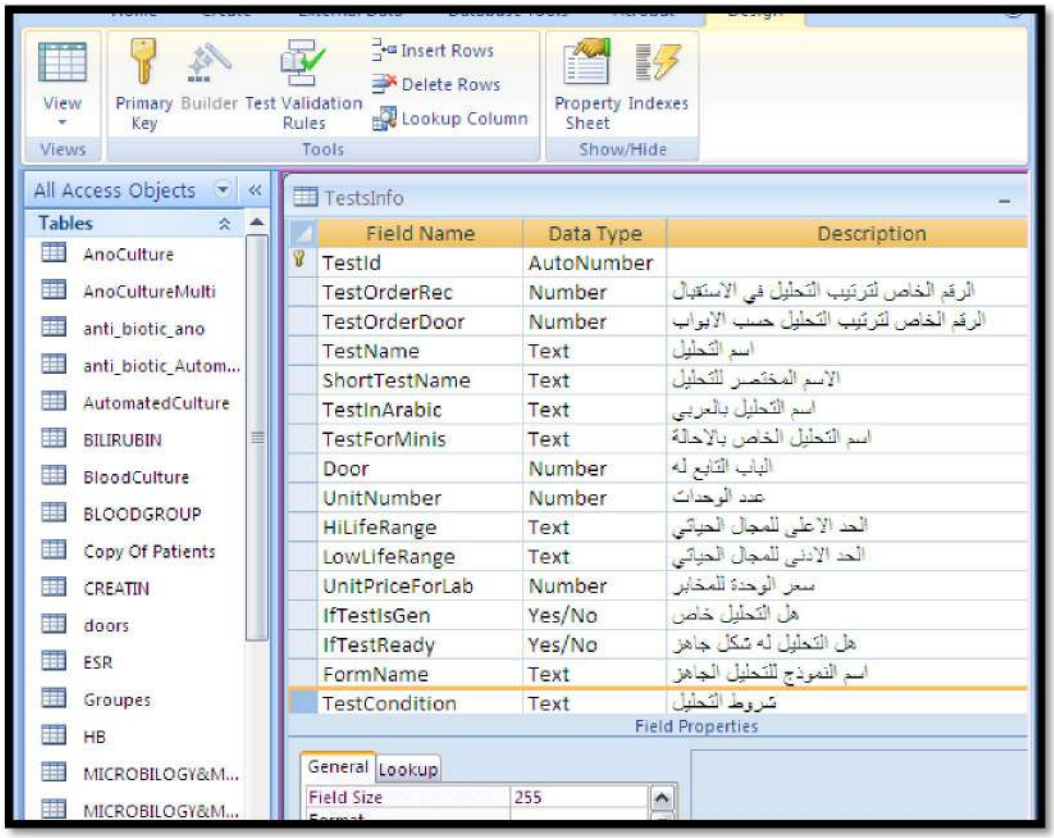
2•1•3•2•5 بنية جدول أسماء التحاليل المخبرية (TestsInfo)

يتكون هذا الجدول من 27 حقل، ويهتم هذا الجدول بتخزين خصائص كل نوع من أنواع التحاليل الطبية، انظر الشكل (5-2-9)، وسنعرض فيما يلي لأهم حقول الجدول:

أ- يوجد عدة حقول لعرض التسميات المختلفة لكل تحليل من التحاليل الطبية مثل التسمية الأجنبية والتسمية العربية والاسم المختصر، انظر الشكل (5-2-10).

ب- يرتبط هذا الجدول مع جدول الأبواب (Doors) من خلال الحقل (Door)، راجع الشكل (5-2-4).

ج- يتضمن هذا الجدول حقل الحد الأعلى للمجال الحياتي (HiLifeRange) وحقل الحد الأدنى للمجال الحياتي (LowLifeRange)، ويعبِّران عن الحدين الأعلى والأدنى الذين من غير المتوقع أن تقع نتائج التحليل الطبي خارجهما مهما تكن الظروف، وبغض النظر عن صفات المريض، وإلا اعتُبرت النتيجة خاطئة. ويمكن أن يفيد هذا الحقل في تتبع وضبط عمليات إدخال العاملين في المخبر للنتائج إلى البرنامج الحاسوبي، وذلك من خلال إظهار البرنامج لرسائل تحذير عند محاولة إدخال قيم للنتائج خارج هذين المجالين، ولكن للأسف لا وجود لمثل هذه التقنية في هذا البرنامج مما أدى إلى تزايد احتمال إدخال قيم خاطئة وغير منطقية للنتائج.



الشكل (5-2- 9): بنية جدول أسماء التحاليل المخبرية (TestsInfo)

د- يُستخدم حقل عدد الوحدات (UnitNumber) في هذا الجدول لحساب أجور التحاليل الطبية وفقاً للطريقة التالية:

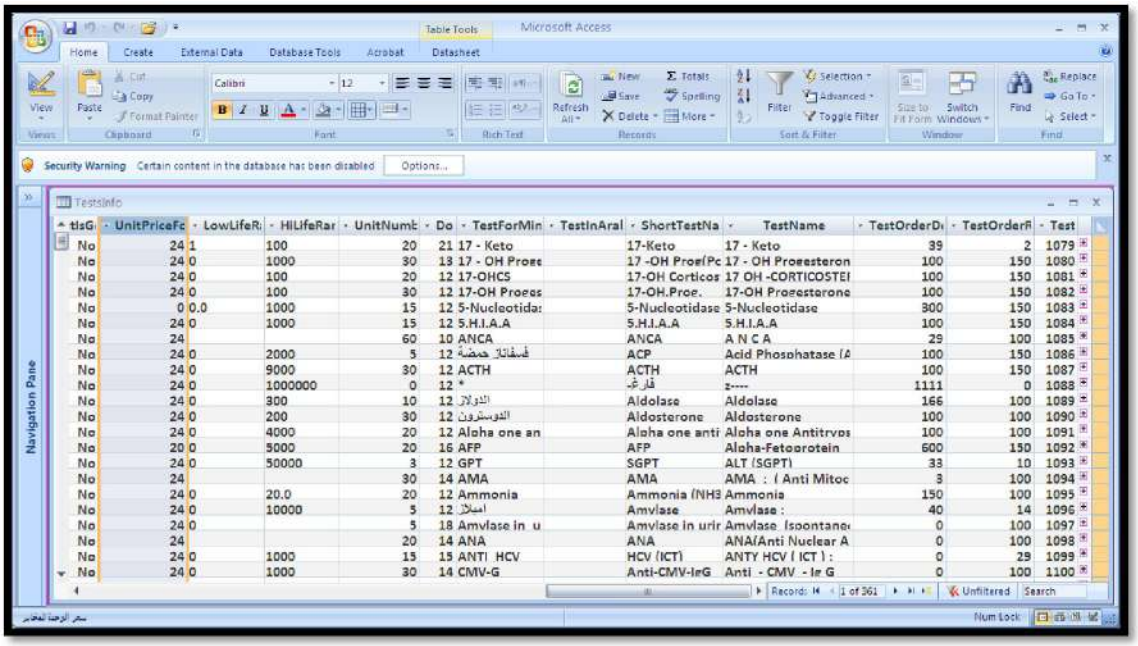
سعر التحليل = عدد الوحدات لهذا التحليل × سعر الوحدة

عدد الوحدات (UnitNumber) هو المقياس الذي يقيَّم به سعر هذا التحليل الطبي مقارنةً ببقية التحاليل، وتقوم وزارة الصحة عادةً بإصدار نشرة تتضمن عدد الوحدات الخاصة

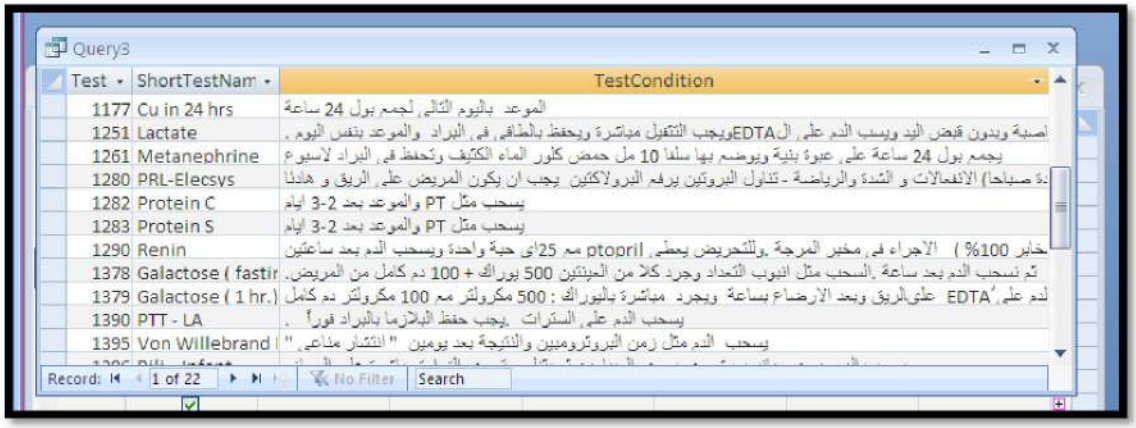
بكل تحليل، ومن المفترض أن تلتزم مخابر التحاليل الطبية بها، انظر الشكل (5-2-12).

سعر الوحدة (UnitPrice) هو القيمة الترجيحية للسعر الذي يدفعه كل مريض لقاء التحاليل الطبية التي تُجرى له (راجع الفقرة **5•2•3•1•1-و**).

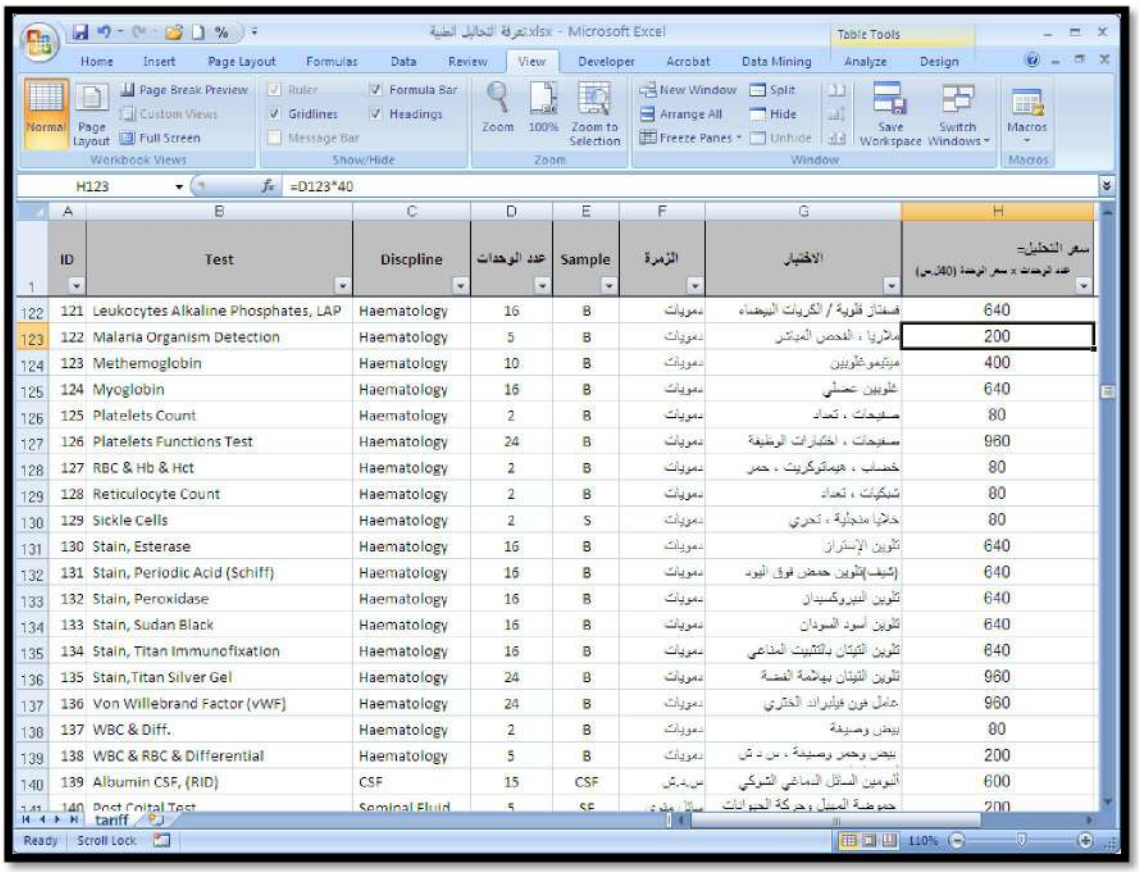
ه- تحتاج بعض التحاليل الطبية إلى شروط خاصة عند إجرائها، وتخزن هذه الشروط في حقل شروط التحليل (TestCondition)، ولكن لا يوجد في جدول أسماء التحاليل هذا سوى 22 تحليل له شروط مذكورة من أصل 367 تحليل موجود في هذه القاعدة. انظر الشكل (5-2-11).



الشكل (- -): مضمون الحقول في جدول أسماء التحاليل ( )



الشكل (5-2- 11): استعلام عن التحاليل التي لها شروط مدرجة في جدول أسماء التحاليل (TestInfo)



الشكل (5–2– 12): تعرفة التحاليل الطبية الصادرة من قبل وزارة الصحة حسب القرار / 26 ت تاريخ 11 آب2008 ، مع اضافة عمود لحساب سعر التحليل إذا كان سعر الوحدة 40 ل.س

3•1•3•2•5 بنية جدول القيم الطبيعية (NormalRanges)

يختلف عادةً تقييم نتائج التحاليل الطبية فيما إذا كانت طبيعية أم لا بحسب كل من العمر والجنس والمرأة الحامل. كما وقد تختلف بعض القيم الطبيعية من مخبر لآخر بحسب تقنيات ووسائل التحليل المتبعة. وتبعاً لذلك يحتوي برنامج التحاليل الطبية على جدول خاص بحساب القيم الطبيعية للمرضى، حيث يتم طباعة القيم الطبيعية لكل تحليل على ورقة النتائج إلى جانب نتيجة هذا التحليل الطبي، مما يساعد الطبيب المعالج في قراءة وفهم نتائج التحاليل بدقة. ويتكون هذا الجدول من الحقول المبينة في الشكل (5–2–13)، ونبين فيما يلي بعض الأمور المتعلقة بهذه الحقول:

أ- يوجد حقل (TestId) وهو حقل هوية التحليل الطبي ويرتبط هذا الحقل مع حقل هوية التحليل في جدول أسماء التحاليل الطبية.

ب- يتوفر في هذا الجدول حقل لجنس المريض (Gender)، ويتضمن ثلاثة قيم هي 1: ذكر، 2: أنثى، 3: امرأة حامل، ويُستخدم هذا الحقل عملياً لأهداف جلب وطباعة القيم الطبيعية

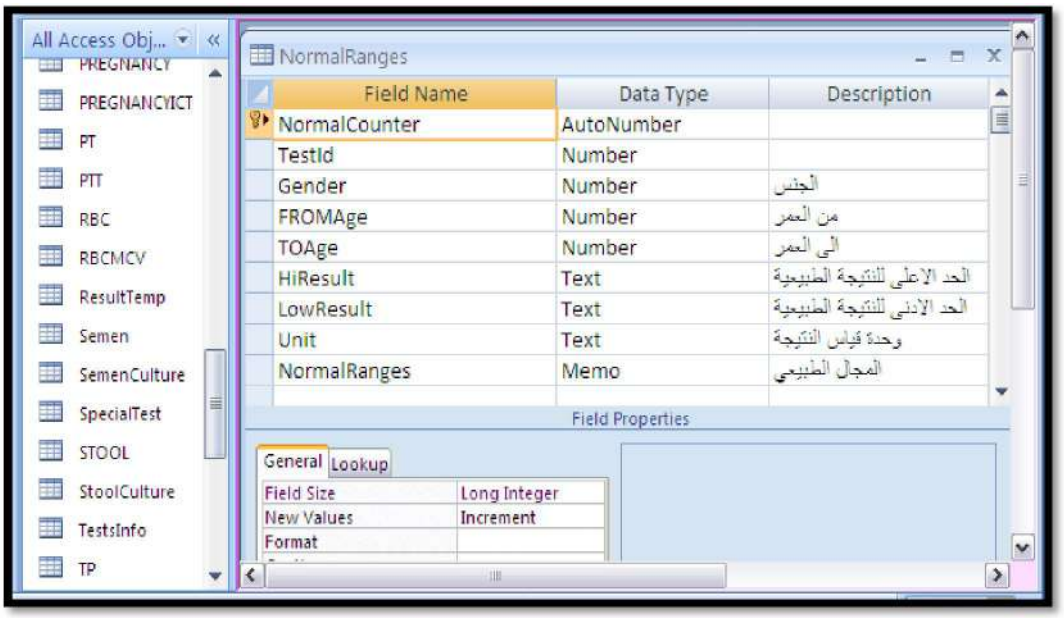
لكل تحليل طبي يُجرى للمريض. أما منطقياً تعتبر فكرة انضمام صفة المرأة الحامل إلى هذا الحقل خاطئة ومن الأفضل أن توضع هذه الصفة في حقل آخر، ولكن لم تسبب هذه الفكرة أية إشكالية للبحث الحالي لأنه لم يُعثر على بيانات للنساء الحوامل في التحاليل الطبية للمرضى الداخليين في المشفى.

ج- يقسم العمر في كل تحليل طبي إلى عدة مجالات حيث تتساوى القيم الطبيعية لكل تحليل طبي داخل كل مجال عمري للمرضى من نفس الجنس. وتبعاً لذلك يحتوي هذا الجدول على حقلين لحدي العمر الأدنى (FromAge) والأعلى (ToAge) للفئة العمرية التي تختلف عندهما تقييمات نتائج التحاليل الطبية، انظر الشكل (5-2-14).

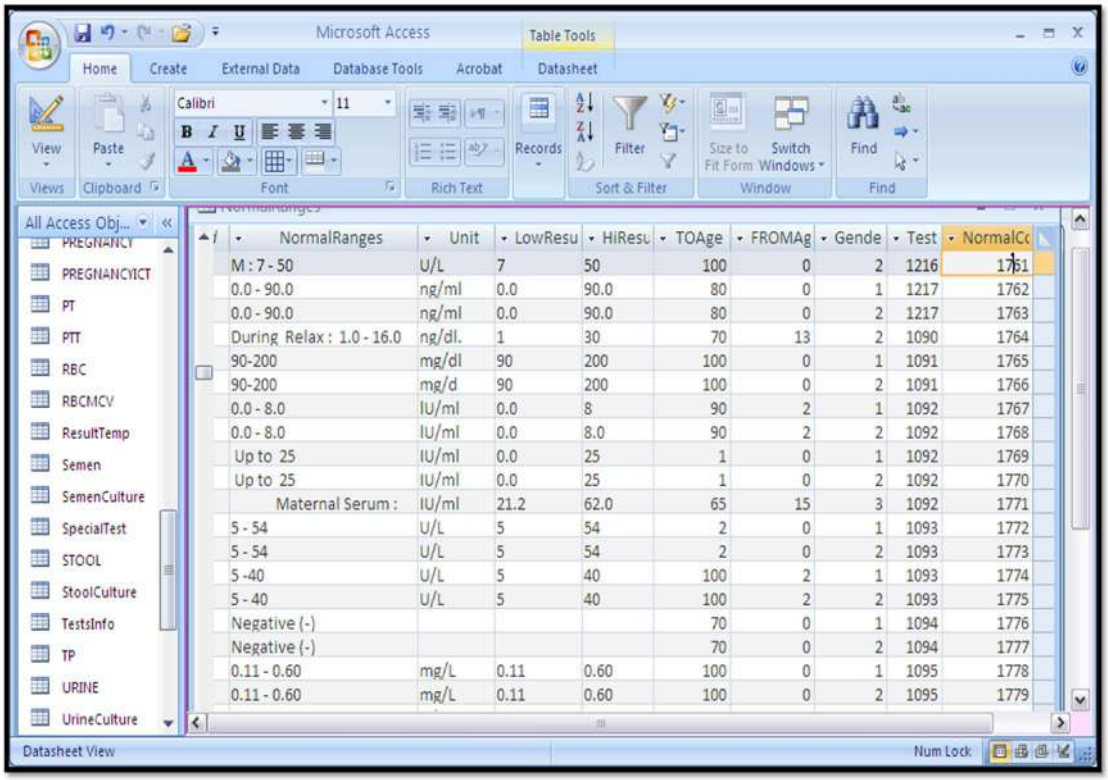
د- يتضمن هذا الجدول حقلي الحد الأدنى للنتيجة (LowResult) والحد الأعلى للنتيجة (HiResult)، وهما الحقلان المخصصان لتخزين الحديين الأعلى والأدنى الذين تقع ضمنهما النتيجة الطبيعية السليمة للتحليل الطبي بحسب الجنس والفئة العمرية الواردين في الحقول السابقة.

ه- أما حقل (NormalRanges) فهو حقل نصي ويتضمن المجال الطبيعي لكل تحليل طبي بحسب العمر والجنس، ويتضمن اجتماع محتويات حقلي الحد الأدنى والأعلى للنتيجة الطبيعية، ويُستخدم لأهداف طباعة النتائج، انظر الشكل (5-2-14).

ومن الملاحظات الهامة بالنسبة إلى هذا الجدول أنه لا يشتمل على كافة أنواع التحاليل الطبية الموجودة في جدول أسماء التحاليل (TestInfo)، وإنما فقط 303 تحاليل من أصل 367 تحليل موجود في جدول أسماء التحاليل ، ويعود ذلك إلى عدة أسباب منها بسبب التقصير في تعريف هذه القيم أثناء إنشاء هذا البرنامج، ومنها بسبب أن بعض التحاليل تتضمن عدة تحاليل فرعية مثل تحليل الكريات البيضاء (WBC)، والتي لم تؤخذ بعين الاعتبار أثناء تصميم البرنامج وخصوصاً أنه موجه لتخزين وطباعة نتائج التحاليل الطبية وليس لتقييم النتائج، حيث أن التقييم النهائي للنتائج يقع ضمن اختصاص الطبيب المعالج الطالب للتحاليل الطبية.



الشكل (– –): بنية جدول القيم الطبيعية للتحاليل الطبية



الشكل (5–2– 14): جدول القيم الطبيعية للتحاليل الطبية، ويُظهر اختلاف هذه القيم بحسب الجنس وبحسب الفئة العمرية

4•1•3•2•5 بنية جدول التحاليل المخبرية الأساسي (PatTest)

يتكون جدول التحاليل المخبرية (PatTest) من 15 حقل، وتتعلق معظم هذه الحقول بتفاصيل إجراء التحاليل المخبرية، انظر الشكل (5–2–15) للتعرف على أنماط هذه الحقول. وسنبين فيما يلي بعض التفاصيل المتعلقة بحقول هذا الجدول.

أ- يرتبط هذا الجدول مع جدول المرضى من خلال حقل هوية المريض (PatId)، ويرتبط مع جدول أسماء التحاليل الطبية عن طريق حقل هوية التحليل (TestId).

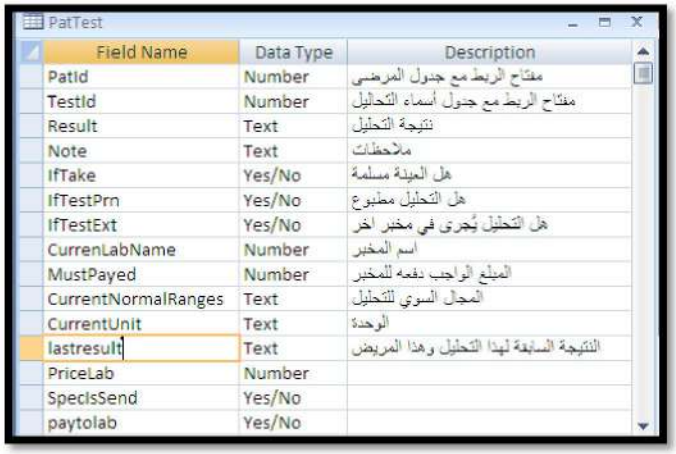
ب- يحتوي هذا الجدول على حقل النتيجة (Result)، ويُستخدم لتخزين نتائج التحاليل الطبية، وهو من نمط نصي (Text)، ولا توجد أية قيود على عمليات الإدخال إلى هذا الحقل. ويحتوي هذا الحقل على أشكال مختلفة من النتائج، فبعضها أرقام صحيحة وبعضها أرقام حقيقية، وبعضها مؤلف من عدة قيم يفصل بينها الرمز (-) مثل نتيجة تحليل الكريات البيضاء (11.5-84-8-0-0-0)، ويقصد بهذه النتيجة أن للتحليل سبعة تحاليل فرعية ولكل منها نتيجة. كما يوجد نتائج نصية مثل (Negative) وهي خاصة ببعض التحاليل، وتعني أن نتيجة التحليل المطلوب جيدة وأنه لا يوجد إصابة بالمرض الذي يكشفه هذا التحليل، والعكس بالعكس إذا كانت النتيجة (Positive) مثل تحليل العامل الاسترالي (HBS-Ag-ICT). كما يوجد في هذا الحقل بعض الرموز مثل (+، -، ++) وهي تُعبِّر أيضاً عن سلبية أو ايجابية النتيجة. وبذلك نجد أن هنالك صعوبة أولية في دراسة بيانات نتائج التحاليل الطبية الموجودة في حقل واحد والمتباينة في النمط، انظر الشكل (5-2-16).

ج- يمكن التعرف على أكثر التحاليل الطبية طلباً (أو أقلها طلباً) باجراء استعلام في أكسس عن أعداد كل من التحاليل الطبية التي تم اجرؤها خلال عامي 2007-2008، انظر الشكل (5-2-17). ويُستفاد من هذا الاستعلام في توجيه الاهتمام بالتحاليل الطبية في المراحل القادمة من اكتشاف المعرفة، وقد تبين أنه تم إجراء 124 نوع تحليل خلال عام 2007، و145 نوع تحليل في عام 2008.

د- يحتوي هذا الجدول على حقل (NormalRanges)، وهو الحقل المعبِّر عن المجال الطبيعي الذي تقع ضمنه النتائج الطبيعية لكل تحليل بحسب صفات المريض (العمر، والجنس)، ويستفاد من هذا الحقل في طباعة النتائج. ويقوم برنامج الوافي باستدعاء قيم هذا الحقل من حقل (NormalRanges) في جدول القيم الطبيعية (NormalRanges) بحسب اسم التحليل وعمر المريض وجنسه.

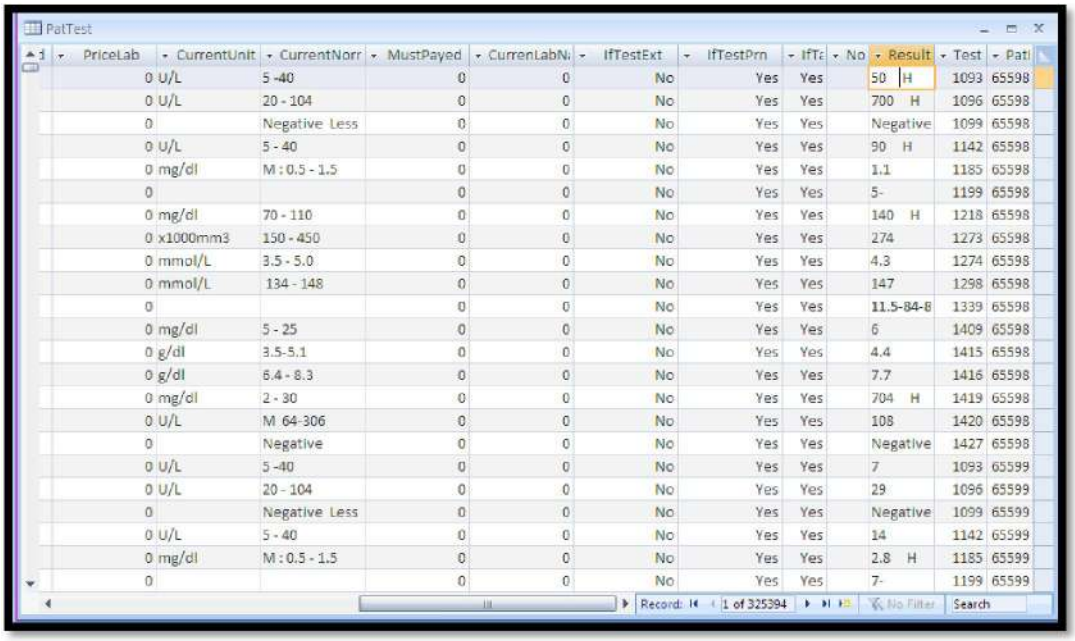
ه- يوجد عدة حقول متعلقة بقيام هذا المخبر أحياناً بإجراء بعض التحاليل في مخبر آخر، ولكنها خالية من البيانات ولن نتطرق لها في دراستنا.

و- يوجد حقل لتخزين النتيجة السابقة للتحليل الطبي في حال أن المريض قد أجري له التحليل الطبي سابقاً شريطة أن تنتمي النتيجة السابقة إلى نفس السنة وأنها ما تزال مخزنة في هذا البرنامج. وذلك لأنه يتم مسح جميع البيانات القديمة في بداية كل سنة بسبب تناقص سرعة أداء البرنامج عندما يزداد حجم البيانات المخبأة فيه عن حد معين.


PatTest

Field Name	Data Type	Description
PatId	Number	مفتاح الربط مع جدول المرضى
TestId	Number	مفتاح الربط مع جدول أسماء التحاليل
Result	Text	نتيجة التحليل
Note	Text	ملاحظات
IfTake	Yes/No	هل العينة مسلمة
IfTestPrn	Yes/No	هل التحليل مطبوع
IfTestExt	Yes/No	هل التحليل يُجرى في مخبر اخر
CurrenLabName	Number	اسم المخبر
MustPayed	Number	المبلغ الواجب دفعه للمخبر
CurrentNormalRanges	Text	المجال السوي للتحليل
CurrentUnit	Text	الوحدة
lastresult	Text	النتيجة السابقة لهذا التحليل وهذا المريض
PriceLab	Number	
SpecIsSend	Yes/No	
paytolab	Yes/No	

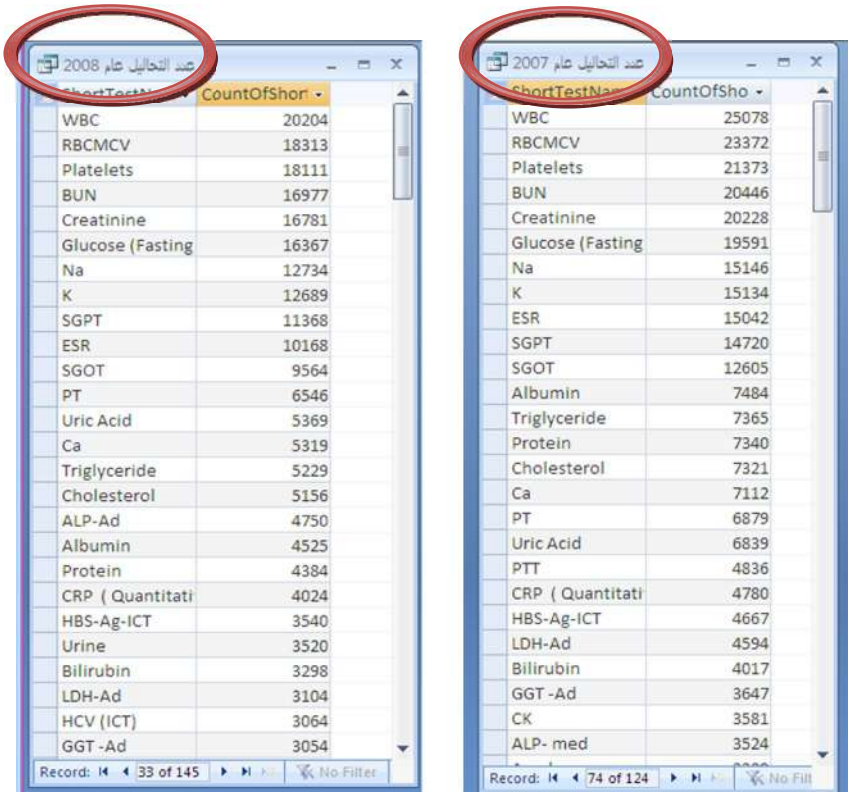
الشكل (- -): بنية جدول تحاليل المريض ( ) في برنامج الوافي للتحاليل الطبية


PatTest

PriceLab	CurrentUnit	CurrentNorr	MustPayed	CurrenLabN	IfTestExt	IfTestPrn	IfT	No	Result	Test	Pati
0	U/L	5 -40	0	0	No	Yes	Yes		50 H	1093	65598
0	U/L	20 - 104	0	0	No	Yes	Yes		700 H	1096	65598
0		Negative Less	0	0	No	Yes	Yes		Negative	1099	65598
0	U/L	5 - 40	0	0	No	Yes	Yes		90 H	1142	65598
0	mg/dl	M : 0.5 - 1.5	0	0	No	Yes	Yes		1.1	1185	65598
0			0	0	No	Yes	Yes		5-	1199	65598
0	mg/dl	70 - 110	0	0	No	Yes	Yes		140 H	1218	65598
0	x1000mm3	150 - 450	0	0	No	Yes	Yes		274	1273	65598
0	mmol/L	3.5 - 5.0	0	0	No	Yes	Yes		4.3	1274	65598
0	mmol/L	134 - 148	0	0	No	Yes	Yes		147	1298	65598
0			0	0	No	Yes	Yes		11.5-84-8	1339	65598
0	mg/dl	5 - 25	0	0	No	Yes	Yes		6	1409	65598
0	g/dl	3.5-5.1	0	0	No	Yes	Yes		4.4	1415	65598
0	g/dl	6.4 - 8.3	0	0	No	Yes	Yes		7.7	1416	65598
0	mg/dl	2 - 30	0	0	No	Yes	Yes		704 H	1419	65598
0	U/L	M 64-306	0	0	No	Yes	Yes		108	1420	65598
0		Negative	0	0	No	Yes	Yes		Negative	1427	65598
0	U/L	5 -40	0	0	No	Yes	Yes		7	1093	65599
0	U/L	20 - 104	0	0	No	Yes	Yes		29	1096	65599
0		Negative Less	0	0	No	Yes	Yes		Negative	1099	65599
0	U/L	5 - 40	0	0	No	Yes	Yes		14	1142	65599
0	mg/dl	M : 0.5 - 1.5	0	0	No	Yes	Yes		2.8 H	1185	65599
0			0	0	No	Yes	Yes		7-	1199	65599

Record: 1 of 325394 No Filter Search

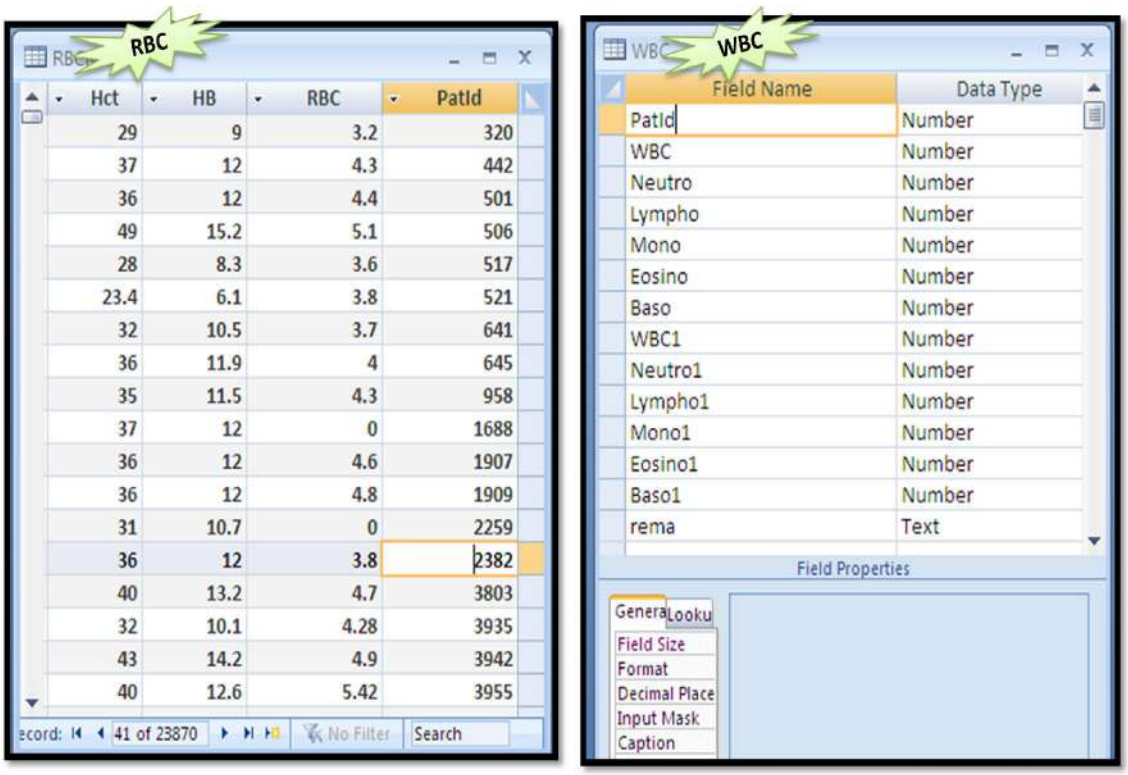
الشكل (5-2- 16): القيم المخزنة في جدول تحاليل المرضى (PatTest)، ويظهر حقل النتائج (result) ذي البنية النصية طرقاً مختلفة في كتابة النتيجة بحسب كل تحليل طبي



الشكل (5-2- 17): أعداد التحاليل التي تم اجراؤها في المخبر مرتبة تنازلياً بحسب عدد المرات التي أجري فيها كل تحليل طبي للمرضى خلال كل من عامي 2007 و 2008

5•1•3•2•5 بنية جداول التحاليل المخبرية المتعددة النتائج

ذكرنا سابقاً أنه يوجد 38 جدول للتحاليل الطبية التي لها أكثر من نتيجة واحدة. وتشترك معظم هذه الجداول في أن لها بنية متشابهة، مع اختلاف في عدد الحقول، فكل من هذه الجداول يحتوي على حقل (PatId) وهو الحقل الذي يحمل أرقام مقابلة لحقل (PatId) في كل من جدولي المرضى (Patients)، وجدول التحاليل المخبرية الأساسي (PatTest). كما خُصِّص في كل جدول حقل لكل تحليل فرعي تابع للتحليل الأساسي، فمثلاً يشتمل تحليل تعداد كريات الدم البيضاء (WBC) على عدة تحاليل فرعية تمثل أنواع كريات الدم البيضاء الموجودة في الدم وهي (WBC, Neutro, Lympho, Mono, Eosino, Baso,.. Baso1,..)، ويشتمل تحليل كريات الدم الحمراء (RBC) على ثلاثة تحاليل فرعية هي (RBC, HB, Hct). انظر الشكل (5-2-18).



الشكل (5–2– 18): أمثلة عن بنية جداول التحاليل التي تتضمن عدة نتائج

2•3•2•5 التعرف على بنى الجداول في قاعدة بيانات مكتب القبول

1•2•3•2•5 بنية جدول المرضى1 (PatientData1)

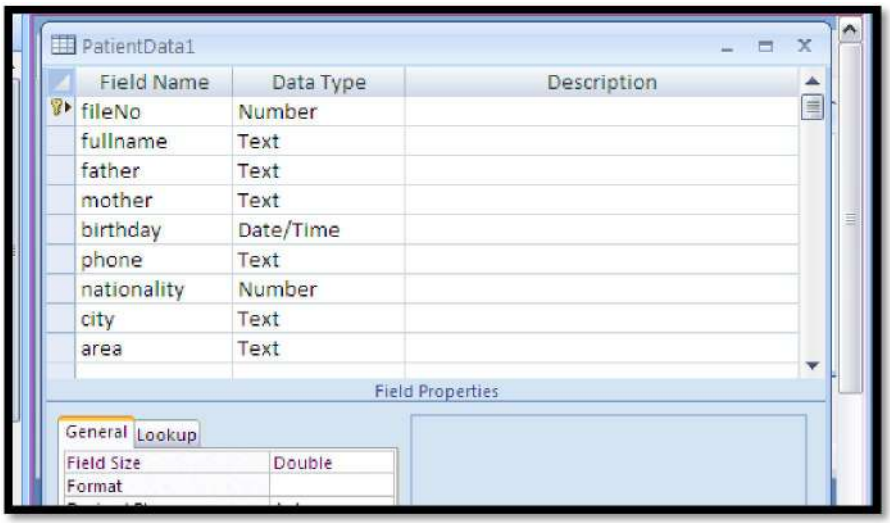
يتكون جدول بيانات المرضى1 من تسعة حقول كما هو مبين في الشكل (5–2–19)، وسنبين فيما يلي أهم النقاط المحيطة بهذا الجدول:

أ– يتضمن هذا الجدول الحقول التعريفية التي تُعتَبر ثابتة عند المرضى مثل الاسم والكنية واسم الأب واسم الأم (حقول نصية)، والجنسية (رقمي مرتبط مع جدول الجنسية).

ب– يوجد حقل لتاريخ الميلاد (نمط زمني)، ولكن لا يوجد أية قيود لضبط تاريخ الميلاد عند إدخال التاريخ إلى برنامج مكتب القبول، مما سمح بوجود عدد من القيم المشوهة في هذا الحقل، فمثلاً يوجد حقول يفوق فيها تاريخ الميلاد تاريخ القبول، وحقول تكون سنة الميلاد 1850م، 1730م.

ج– يَعتبر مبرمجي هذا البرنامج الحاسوبي أن حقلي المدينة والمنطقة من الحقول الثابتة عند المريض، وهذا غير دقيق حيث من الممكن أن يتغير مكان السكن مما يستوجب وجود جدول فرعي لهذه البيانات إذا كان التوثيق دقيقاً.

د- يحمل حقلي المدينة والمنطقة النمط النصي، وهذا من نقاط الضعف في هذه القاعدة، حيث يسمح البرنامج بإدخال أية أسماء يختارها المستخدم، وكان من الأفضل لو تم تخصيص جدول فرعي للمنطقة والمدينة ويرتبط بمفاتيح ربط أجنبي مع هذا الجدول. وتبعاً لذلك يحتوي هذان الحقلان على بعض الأخطاء المطبعية التي سيتم حلها في المرحلة القادمة.



الشكل (5-2- 19): بنية جدول (PatientData1) الموجود في قاعدة بيانات مكتب القبول

2•2•3•2•5 بنية جدول بيانات المرضى2 (PatientData2)

يتكون جدول بيانات المرضى2 من عدة حقول كما هو مبين في الشكل (5-2-20)، ونبين فيما يلي أهم النقاط المتعلقة بحقول هذا الجدول:

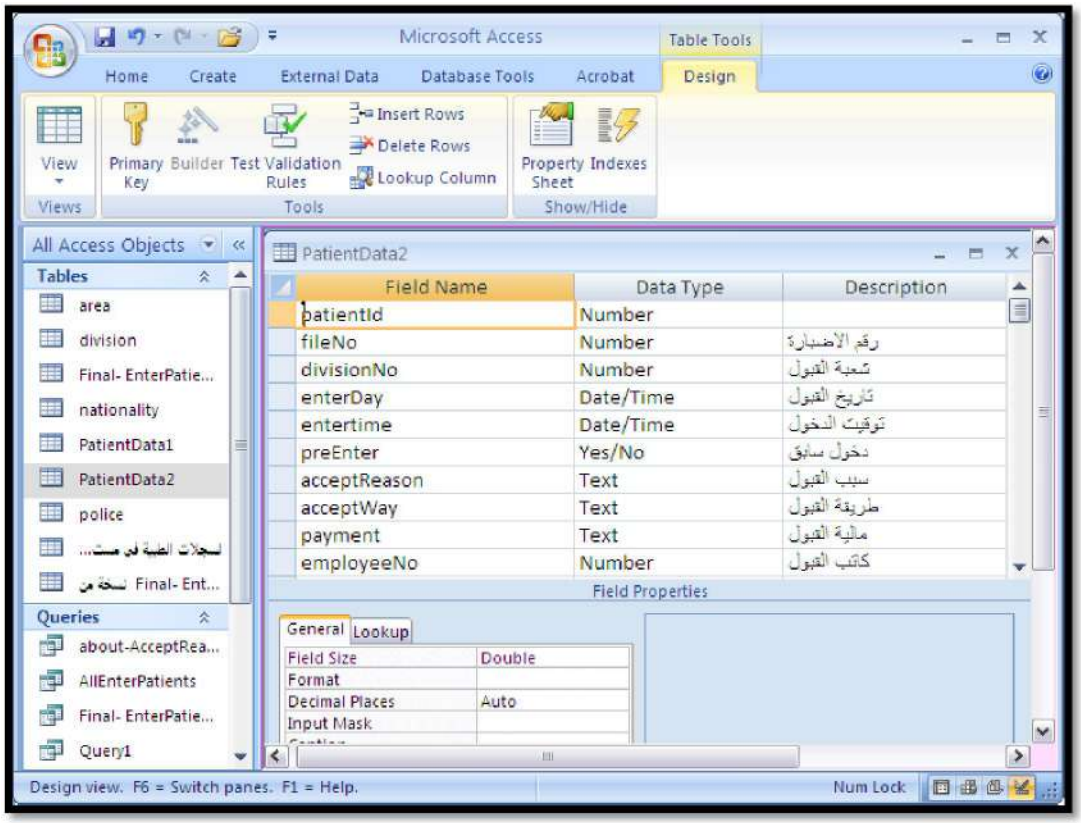
أ- لا يحتوي هذا الجدول على مفتاح أساسي، ولكن تم لغرض هذه الدراسة تحويل حقل (PatientId) إلى مفتاح أساسي، حيث لم يكن فيه أية أرقام مكررة.

ب- يرتبط حقل شعبة القبول الرقمي (DivisionNo) مع المفتاح الأساسي في جدول شعبة القبول.

ج- يتضمن حقل طريقة القبول توثيقاً لطرق القبول في المشفى، وقد عثرنا في هذا الحقل على الطرق التالية (إسعاف، عيادات المشفى، إجراء خارجي، مباشر، محال من مشفى آخر)، انظر الشكل (5-2-21)، وهي أكثر من الطرق الموجودة في الاستمارة الورقية للقبول الموجودة في الشكل (5-2-1).

د- تبين بالاستعلام من حقل مالية القبول (payment) أن 74% من المرضى تقريباً يعالجون بشكل مجاني، و24.5% يدفعون النفقات على حسابهم. وإن ثلثي المرضى المجَّانيين هم قبل عام 2008، حيث أصبح هذا المشفى مأجوراً بنسبة أكبر. انظر الشكل (5-2-22).

ه- يُعتبر حقل سبب القبول (AcceptReason) من أهم الحقول في هذه الدراسة، حيث يحتوي على الأسباب المرضية للقبول في المشفى، ويُعبِّر عن الحالة الطبية السريرية للمريض. وقد أظهر الاستعلام الأولي من هذا الجدول أنه يوجد 9490 سبب لقبول 47728 حالة مرضية في المشفى، ولكن لا يدل هذا الرقم الكبير على كثرة الحالات المرضية، وإنما يعود ذلك إلى أن حقل سبب القبول هو حقل نصي والنصوص المخزنة فيه ما هي إلا نقلاً عن التعبيرات الشخصية المختلفة التي يكتبها الأطباء الأخصائيون عن حالة المرضى، والتي قد تختلف حتى عند وصف الحالات المتشابهة، انظر الشكل (5-2-23).



الشكل (- -): بنية جدول ( ) الموجود في قاعدة بيانات مكتب القبول



الشكل (5-2- 21): استعلام عن طرق القبول في المشفى



الشكل (5-2- 22): استعلام عن الجهة المسددة لنفقات المشفى، ويشير الرقم إلى خطأ مطبعي أثناء التوثيق الالكتروني



الشكل (5-2- 23): استعلام عن أسباب القبول في المشفى من قاعدة بيانات القبول، ويبين الجدول أن اختلاف طرق التعبير عن الحالة المرضية يؤدي إلى تعدد الأسباب بشكل وهمي

3•3•2•5 التعرف على بنى الجداول في قاعدة بيانات قسم السجلات الطبية

تحتوي هذه القاعدة على جدول أساسي واحد يضم جميع بيانات المرضى، ويحمل هذا التصميم الكثير من نقاط الضعف. وبغض النظر عن هذه النقاط فإن المشكلة الجوهرية في هذا الجدول تكمن في النمط النصي الذي تتسم به معظم الحقول مهما كانت طبيعة بياناتها وذلك كما هو آت، انظر الشكل (5-2-24).



الشكل (5-2- 24): مكونات جدول بيانات المرضى من قاعدة بيانات السجلات الطبية

أ- تأخذ النتائج في حقل الجنس إحدى قيمتين (1:ذكر، 2:أنثى)، ولكن النمط النصي للحقل يستدعي ضرورة القيام في المرحلة القادمة بتنظيف هذه البيانات والتخلص من الرموز غير الرقمية فيها قبل التمكن من تحويله إلى حقل رقمي.

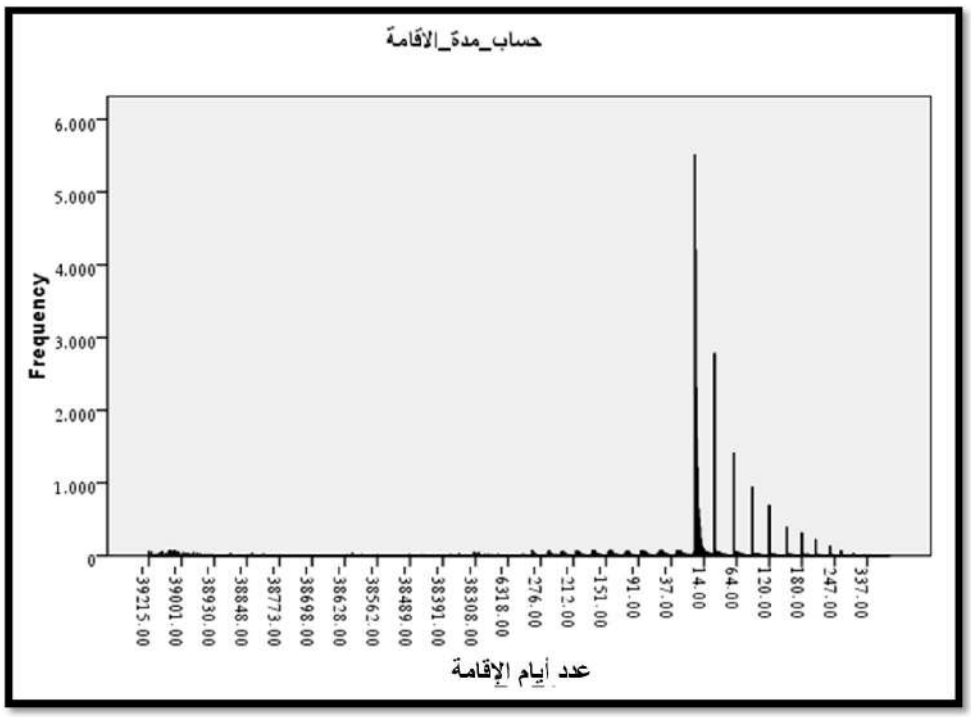
ب- يحمل تاريخ الدخول والخروج إلى المشفى النمط الحرفي وهو مخزن بشكل (dd/mm/yyyy)، وقد تم تحويله إلى نمط تاريخ ولكن من الشكل (mm/dd/yyyy) لأنه أنسب للتحليل ومتوافق مع جميع البرامج الحاسوبية (مثل SPSS) أكثر من الشكل القديم.

ج- يتم تخزين بيانات عمر المريض في حقلين، الأول للعمر ويُعبر عن عمر المريض بإحدى وحدات القياس التالية (الساعة، اليوم، الأسبوع، الشهر، السنة) الموجودة في الحقل الآخر الذي يحمل نفس هذا الاسم. ولكن لن يتم التعامل في هذه الدراسة مع هذين الحقلين لأن

عملية التخزين هذه تسبب صعوبةً في تحليل بيانات العمر، إضافةً إلى أنه يوجد بديل جيد له في قاعدة بيانات مكتب القبول وهو حقل تاريخ الميلاد.

د- أظهر الاستعلام عن أسماء المدن (أو الجهات) التي أتى منها المرضى أن عدد المدن يبلغ 145 مدينة، وهو رقم غير حقيقي وإنما ناتج عن الأخطاء المطبعية في كتابة أسماء المدن (مثل الحسكة، ابلحسكة، الحسكه، الحسطة، حسكة)، وهي من الأخطاء التي كان من الممكن تفاديها فيما لو صممت قاعدة البيانات بطريقة تسمح للمستخدم باختيار اسم المدينة من قائمة مُعرَّفة مسبقاً. ولكن لن يتم تصحيح هذه الأخطاء بسبب وجود حقل للمدينة جيد في قاعدة بيانات مكتب القبول.

ه- يحتوي حقل تاريخ خروج المرضى من المشفى على 3395 قيمة مفقودة من أصل 36582 قيمة أي بنسبة 9.3%. وقد تم العثور أثناء حساب عدد أيام إقامة المرضى في المشفى (تاريخ الخروج – تاريخ الدخول) على 8755 قيمة سالبة من أصل 36582 قيمة، أي ما نسبته 24%، منها بسبب القيم المفقودة في حقل تاريخ الخروج، ومنها بسبب أن تاريخ الخروج المخزن يسبق تاريخ الدخول المسجل. ويعود سبب هذا الخطأ إلى أنه لا يوجد أية قيود أو شروط على عمليات الإدخال إلى الحاسب، حيث كان من المفترض أن تتم برمجة قاعدة البيانات بما لا يسمح بأن يكون تاريخ الخروج سابق لتاريخ الدخول. كما تم العثور على 5261 قيمة تزيد فيها مدة الإقامة عن 150 يوم وقد تصل إلى 1441 يوم.. وهذا غير منطقي أبداً ويعود ذلك إلى كثرة الأخطاء أثناء إدخال قيم التاريخ إلى البرنامج، كأن يكتب المستخدم اليوم بدل الشهر أو العكس. انظر الشكل (5-2-25) الذي يعرض المدرج التكراري لعدد أيام الإقامة في المشفى. وعلى الرغم من الاهتمام الشديد في هذه الدراسة بمتغير مدة الإقامة إلا أنه للأسف من الصعب أن يُستفاد منه بشكله هذا، وإنما ستجري المحاولة لإجراء بعض التنظيفات على بياناته، ومن ثم يمكن استخدامه إذا نجحت هذه العملية.



الشكل (5-2- 25): المدرج التكراري لعدد أيام الاقامة في المشفى بحسب بيانات مكتب السجلات الطبية، باستخدام برنامج SPSS، ويلاحظ أن المدرج يذخر بالقيم غير المنطقية سواء السالبة أو القيم الكبيرة

و- يوجد في هذا الجدول حقل مخصص لتخزين التصنيف الدولي الأساسي للأمراض (International Class1)، ويعتبر وجود هذا التصنيف في قواعد البيانات الطبية على درجة من الأهمية والضرورة، لأن ذلك يساعد على الحصول على مؤشرات عالمية عن الوضع الطبي في المستشفيات ويساعد على ربط قواعد البيانات الطبية معاً طالما أنها تستخدم مصطلحات مشتركة. وقد تبين من خلال استطلاع هذا الحقل أنه يحتوي على 2141 تصنيف دولي ويشمل 33192 سجل، أما بقية السجلات فلم يكن فيها تصنيف دولي، أي فيها قيم مفقودة وعددها 3391 قيمة. وبحسب ما قاله العاملون في هذا القسم أنهم هم الذين يبحثون عن التصنيف الدولي المناسب ويضعونه في قاعدة بياناتهم.. ولكن لا يوجد تأكيد على صحة هذا التصنيف، وخصوصاً أن قاعدة بياناتهم تذخر بالأخطاء الأخرى التي تعتبر بسيطة ومن غير المفترض أن تُرتكب. ولكن سيتم الاستمرار بالاحتفاظ بهذا الحقل وبحقل التصنيف الدولي لأسباب الوفاة (International Class2)، فقد يكون هنالك فائدة غير مدركة من وجودهما في بداية هذه المرحلة. وقد ظهرت فعلاً في المرحلة الرابعة من مراحل اكتشاف المعرفة بعض الفوائد من وجود هذين الحقلين وسيتم التعرف عليها في حينها.

ز- يوجد ثلاثة حالات رئيسة لخروج المرضى هي 1: تحسن أو شفاء، 2: تحويل إلى مشفى آخر، 3: وفاة المريض. ويبين الشكل (5-2-26) عدد السجلات من كل حالة، مع العلم أنه يوجد 3 تسجيلات فيها أخطاء مطبعية (1، 200706، 0)، و1953 قيمة مفقودة، وقد أخبرنا العاملون في هذا القسم أن غالبية حالات الوفاة يتم توثيقها، ولذلك يمكن في حال الضرورة اعتبار أن القيم المفقودة هي خروج من المشفى دون وفاة.



الشكل (- -): استعلام عن الحالات الصحية للمرضى عند خروجهم من المشفى،

حيث 1: تحسن أو شفاء، : تحويل إلى مشفى آخر، : وفاة المريض، : قيم مفقودة، والبقية أخطاء مطبعية

ح- يوجد في هذه القاعدة حقل اسمه (الطبيب الاختصاصي المعالج)، ويعبر هذا الحقل بحسب تسميته عن اسم الطبيب الذي أشرف على معالجة كل مريض، ويحتوي هذا الجدول على 3767 قيمة مفقودة أي بنسبة % من العدد الكلي لسجلات هذا الجدول، ويبلغ عدد أسماء الأطباء الموجودين طبيب، انظر الشكل (- -). ويُستفاد من هذا الحقل في تقييم أداء الأطباء في المشفى بوجود المتغيرات المختلفة الأخرى التي يتوفر البعض منها في قواعد البيانات المدروسة في هذا البحث.



الشكل (5-2- 27): استعلام عن أسماء الأطباء المشرفين على المرضى وعدد المرضى لكل منهم بحسب قاعدة بيانات السجلات الطبية

قدمت مرحلة استيعاب البيانات شرحاً عن البيانات المتاحة للدراسة، حيث تم التعرف على مضمون كل من قاعدة بيانات مخبر التحاليل الطبية وقاعدة بيانات مكتب القبول في المشفى، وقاعدة بيانات مكتب السجلات الطبية، وقد تم التوسع أحياناً والاختصار أحياناً أخرى في عرض الأفكار المحيطة بالتعرف على البيانات، وذلك بحسب أهمية ومساهمة كل منها في عملية اكتشاف المعرفة.

تتشابك بعض الأفكار عن استيعاب البيانات بقوة مع الأفكار المرتبطة بتهيئة البيانات، بحيث يصعب فصلهما أحياناً عن بعض، ولذلك وتجنباً لتكرار الحديث في المبحثين عن نفس الفكرة فقد وجدنا أنه من الأنسب أن تُعرض مباشرة في مرحلة تهيئة البيانات.

الفصل الخامس

المبحث الثالث

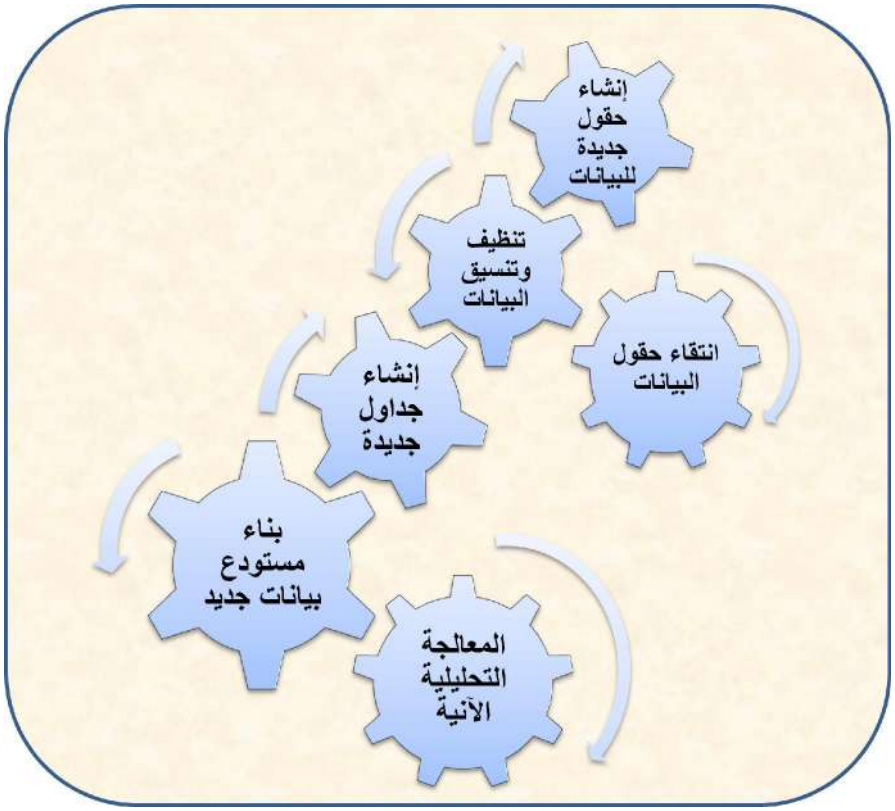
مرحلة تهيئة بيانات مخبر التحاليل الطبية

مقدمة

قَدَّمت مرحلة استيعاب بيانات مخبر التحاليل الطبية السابقة المدخلات الضرورية لمرحلة تهيئة هذه البيانات. وكما سبق وذكرنا في الفصل الثالث من هذه الدراسة بأن جزءاً فقط من هذه البيانات يمكن أن يُهيأ ليُستخدم في عمليات التحليل والتنقيب، فإنه سيتم في هذه المرحلة من التطبيق العملي انتقاء الحقول والبيانات المناسبة، وتنظيفها، وإعادة بناء قواعد البيانات الثلاث المتوفرة للدراسة بما يتناسب مع احتياجات عمليات التنقيب. ومن ثم سيجري العمل على دمجها معاً وتحويلها إلى مستودع بيانات لاستخدامه في عمليات المعالجة التحليلية الآنية وعمليات التنقيب. وعملياً لم يكن هنالك انفصال واضح بين تلك العمليات بل كانت متداخلة ومتكررة إلى أصبح مستودع البيانات صالحاً ومتطوراً بما يتناسب مع احتياجات التحليل والتنقيب، انظر الشكل (5-3-1).

نعرض في الفقرات التالية أهم الخطوات المتبعة في هذه المرحلة وذلك وفق الترتيب التالي:

- انتقاء الحقول التي تفيد في انجاز مراحل اكتشاف المعرفة
- تنظيف وتنسيق البيانات وإنشاء بيانات تركيب جديدة ودمج قواعد البيانات الثلاثة معاً
- تحويل البيانات إلى مستودع بيانات جديد في نظام SQL Server
- المعالجة التحليلية الآنية للبيانات



الشكل (5-3- 1): تداخل الخطوات المتبعة أثناء مرحلة تحضير البيانات

1•3•5 انتقاء الحقول التي تفيد في انجاز مراحل اكتشاف المعرفة

يعتمد تطبيق تقانة اكتشاف المعرفة على جزء من البيانات المتاحة، وتُقسم البيانات التي تُستخدم في عملية اكتشاف المعرفة إلى ثلاثة أقسام من حيث الدور الذي تقوم به في هذه العملية، وذلك على النحو التالي:

أ- البيانات التي ستُستخدم في عمليات التحليل والتنقيب (مثل مكان إقامة المريض).

ب- البيانات التي ستتحول إلى بيانات جديدة لاستخدامها في عمليات التحليل والتنقيب (مثل تاريخ ميلاد المريض الذي سيتحول إلى عمر المريض).

ج- البيانات التي تُستخدم فقط في تنظيف البيانات الأخرى (مثل رقم هاتف المريض)

وقد تلعبُ بعض البيانات أكثرَ من دور واحد في أثناء المرور بمراحل اكتشاف المعرفة، وبعض هذه الأدوار قد لا يكون ظاهراً في بداية العمل وإنما يظهر تباعاً. ونورد فيما يلي جدولاً بأسماء حقول البيانات التي سنستخدمها في دراستنا بغض النظر عن دور كل منها والذي سيظهر في حينه.

الجدول (5-3- 1): أسماء الجداول والحقول المنتقاة لانجاز مراحل اكتشاف المعرفة

اسم قاعدة البيانات	اسم الجدول	اسم الحقل	النمط الحالي للحقل
قاعدة بيانات مخبر التحاليل الطبية	Patient	PatId	Long Integer
		Tel	Double
		SpecTime	Date/Time
		CurDay	Date/Time
		UnitPrice	Byte
	PatTest	PatId	Long Integer
		TestId	Long Integer
		Result	Text
		CurrentUnit	Text
	TestInfo	TestId	Long Integer
		TestOrderRec	Integer
		TestOrderDoor	Integer
		TestName	Text
		ShortTestName	Text
		TestInArabic	Text
		Door	Long Integer
		UnitNumber	Single
		HiLifeRange	Text
		LowLifeRange	Text
	Doors	DoorId	Long Integer
		Door	Text
		DoorOrder	Byte
	NormalRanges	NormalCounter	Long Integer
		TestId	Long Integer
		Gender	Byte
		FromAge	Byte
		ToAge	Byte
		Unit	Text
		HiResult	Text
		LowResult	Text
	BILIRUBIN	PatId	Long Integer
		b1	Text
		b2	Text
	HB	PatId	Long Integer
		HB	Text
		Hct	Text
	PT	PatId	Long Integer
		ProthrombinT	Text
		Control	Text
		spac	Text
		INR	Text
	PTT	PatId	Long Integer
		PTT	Text
		Control	Text
	RBC	PatId	Long Integer
		RBC	Text
		HB	Text
		Hct	Text
	RBCMCV	PatId	Long Integer
		RBC	Single
		HB	Single
		Hct	Single
		rdw	Text
		rema	Text

اسم قاعدة البيانات	اسم الجدول	اسم الحقل	النمط الحالي للحقل
قاعدة بيانات مكتب القبول	URINE	PatId	Long Integer
		Color	Text
		Appearance	Text
		Specific Gravity	Text
		Reaction Ph	Text
		Protein	Text
		Sugar	Text
		Ketones	Text
		Nitrite	Text
		Urobilinogen	Text
		Bilirubin	Text
		Hemoglobin	Text
		Blood	Text
		Leucocytes	Text
		Erythrocytes	Text
		Epithelial Cells	Text
		Bacteria	Text
		CaOxalate	Text
		Uric_Acid	Text
	WBC	PatId	Long Integer
		WBC	Single
		Neutro	Single
		Lympho	Single
		Mono	Single
		Eosino	Single
		Baso	Single
	WIDAL	PatId	Long Integer
		TO	Text
		TH	Text
		AH	Text
		BH	Text
	WRIGHT	PatId	Long Integer
		A	Text
		M	Text
	PatientData1	FileNo	Double
		Fullname	Text
		Father	Text
		Mother	Text
		Birthday	Date/Time
		Phone	Text
		Nationality	Double
		City	Text
		Area	Text
	PatientData2	PatientId	Double
		FileNo	Double
		DivisionNo	Double
		EnterDay	Date/Time
		Entertime	Date/Time
		PreEnter	Yes/No
		AcceptReason	Text
		AccentWay	Text
		Payment	Text
		NationalityId	Double
	division	DivisionId	Double
		Division	Text
	nationality	NationalityId	Double
		Nationality	Text

اسم قاعدة البيانات	اسم الجدول	اسم الحقل	النمط الحالي للحقل
قاعدة بيانات مكتب السجلات الطبية	MedicalRecords	ID	Long Integer
		EnterDay	Date/Time
		FileNo	Number
		FullName	Text
		EnterDivision	Text
		FinalDiagnoses	Text
		Doctor	Text
		Gender	Text
		leavingDay	Date/Time
		LeavingStatus	Text
		DeathReason	Text
		LeavingDivision	Text
		International Class1	Text
		International Class2	Text

2•3•5 تنظيف وتنسيق البيانات وإنشاء بيانات تركيب جديدة

تحتوي قواعد البيانات الطبية عادةً على الكثير من القيم المفقودة والتي قد تصل نسبتها أحياناً إلى 50% من إجمالي القيم الموجودة[1]، وهذا ما تمت ملاحظته أيضاً في قواعد البيانات المتاحة للدراسة.

فرضت طبيعة بيانات مخبر التحاليل الطبية المدروسة تداخلاً بين مكونات هذه الخطوة بحيث يصعب الفصل بينها، وقد احتاجت هذه البيانات إلى إنشاء بيانات تركيب جديدة قبل أو بالتوازي مع عمليتي تنظيف وتنسيق البيانات، وسيتم لذلك عرض هذه الخطوات بحسب ما تم انجازه في كل قاعدة من قواعد البيانات المدروسة.

تعاني البيانات المدروسة من خمسة أنواع من المشاكل، والتي سنحاول حلها في هذه المرحلة، وهي:

أ- وجود قيم مفقودة، حيث تعاني قواعد البيانات المدروسة من نقص في توثيق بعض البيانات، مثل النقص في إدراج رقم اضبارة المريض في قاعدة بيانات مخبر التحاليل الطبية.

ب- وجود قيم مشوهة في بعض البيانات، مثل وجود تواريخ لميلاد المرضى بحيث يزيد معها أعمار هؤلاء عن الحدود الطبيعية.

[1] Cios, K.J.; Pedrycz, W.& Swiniarski, R. W. 2007, **Data Mining: A Knowledge Discovery Approach**, P. 41

ج- استخدام أنماط غير دقيقة لتعريف الخاصيات في الجدول، فمثلاً حملت بعض الحقول نمط النص (Text) على الرغم من أنها مخصصة لبيانات رقمية أو بيانات تاريخ، وقد تطلب حل هذه المشكلة أن يتم العمل على تصحيح القيم المشوهة الناتجة عن وجود رموز غير مرغوبة في البيانات، ومن ثم تحويل نمط الحقل إلى النمط المناسب للبيانات.

د- أدت البنية الضعيفة لجداول قاعدة بيانات مخبر التحاليل الطبية، إلى استحالة استخدامها في عمليات التنقيب ما لم يعاد تركيبها وفق قوالب جديدة.

ه- غياب الروابط الدقيقة بين الجداول في قاعدة بيانات مخبر التحاليل الطبية، وصعوبة إيجاد روابط دقيقة بين قواعد البيانات الثلاثة، وبما أن هذه الدراسة تحتاج إلى تصميم قاعدة أو مستودع بيانات جديد ترتبط فيه قواعد البيانات الثلاثة معاً، فإنه يترتب القيام بتوحيد أنماط الحقول التي ستُربط معاً. فمثلاً لوحظ أن حقل (FileNo) في قاعدة بيانات القبول وهو من نمط (Double)، يجب أن يرتبط مع حقل (Tel) في قاعدة بيانات المخبر والذي يتضمن أرقام أضابير المرضى ولكنه من نمط نص (Text).

وسنبين فيما هو آت الأعمال التي أُنجزت في كل من قواعد البيانات الثلاثة في سبيل تنظيف البيانات وتهيئتها للدخول في قاعدة البيانات الجديدة المتناسبة مع عمليات المعالجة التحليلية الآنية ومع عمليات التنقيب.

1•2•3•5 تنظيف وتنسيق البيانات وإنشاء بيانات تركيب جديدة في قاعدة بيانات مخبر التحاليل الطبية

قدَّمت مرحلة استيعاب البيانات وصفاً للحقول والجداول المكونة لقاعدة بيانات التحاليل الطبية، في حين أظهرت المرحلة الحالية من اكتشاف المعرفة أن هنالك عدد من الصعوبات التي تعيق استخدام جداول هذه القاعدة بوضعها الحالي في عمليات اكتشاف المعرفة، حيث تم التوصل في هذه المرحلة إلى أن تجاوز هذه العوائق يتطلب أولاً بناء جداول جديدة بديلة للجداول الحالية، ويكون فيها حقل نتائج التحاليل المخبرية هو حقل من نمط رقمي وليس نصي، وثانياً التوصل إلى فكرةٍ تتوحد فيها طريقة تقييم نتائج التحاليل المختلفة التي بعضها رموز وبعضها

أرقام، وثالثاً إيجاد حلول مناسبة لمشاكل البيانات المشوهة والبيانات المفقودة. وسنتعرف فيما يلي على الإجراءات التي قمنا بها في هذا الصدد.

1•1•2•3•5 إنشاء جدول جديد للتحاليل الفرعية (SubTests)

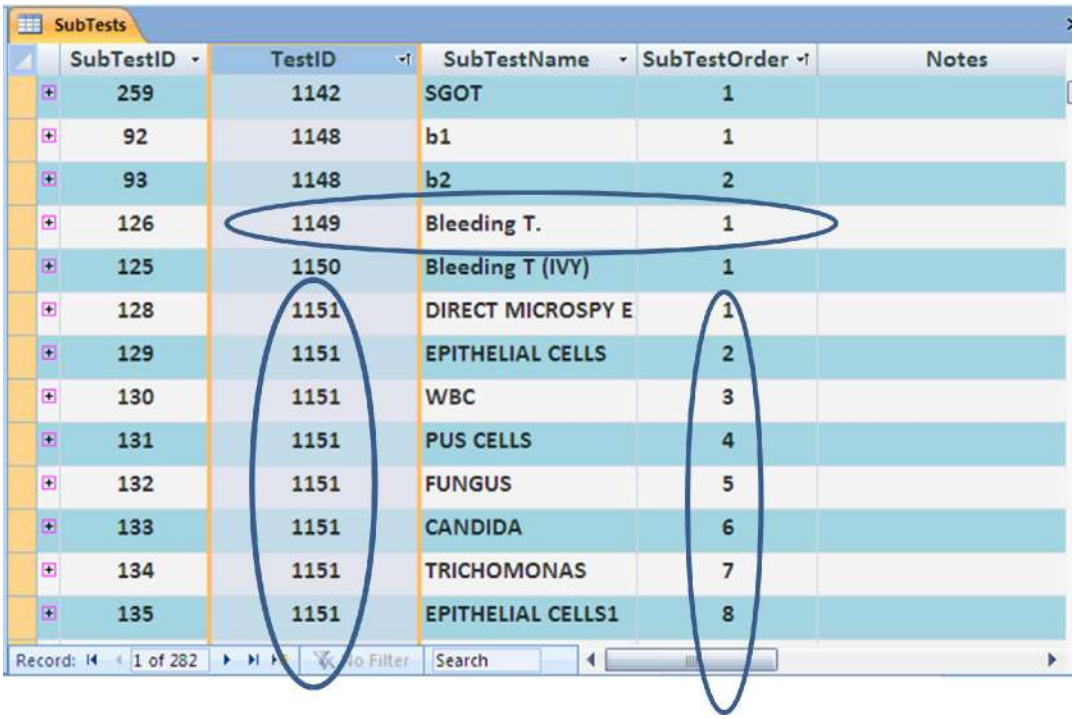
تتكون بعض التحاليل الطبية من عددٍ من التحاليل الفرعية، ويحتاج التعامل مع هذه التحاليل أن يتم توصيفها واعتبارها نقطة التركيز لعمليات التنقيب، وبالتالي بناء جداول جديدة للتعامل معها. وانطلاقاً من ذلك تم تصميم جدول جديد لتعريف التحاليل الفرعية، ويتفرع هذا الجدول عن جدول (TestsInfo)، ويتضمن أسماء كافة التحاليل الفرعية التابعة لتحليل رئيسي ما بالإضافة إلى أسماء التحاليل الرئيسة التي لها تحليل وحيد، انظر الشكل (5-3-2). وقد تمت تعبئة هذا الجدول بأسماء التحاليل الطبية الفرعية التي أجريت خلال عامي 2007-2008، ويتألف الجدول من الحقول التالية:

أ- حقل (SubTestID) وهو حقل تعريفي للجدول ويتم ترقيمه تلقائياً.

ب- حقل (TestId) وهو مفتاح الربط الأجنبي والذي يرتبط مع رقم التحليل الرئيس الموجود في جدول (TestsInfo).

ج- حقل (SubTestName) ويتضمن اسم التحليل الفرعي، وهو إما مستمد من عناوين حقول جداول التحاليل التي لها عدة نتائج فرعية مثل تحليل (RBC)، لاحظ الشكل (5-3-2)، أو يحمل نفس اسم التحليل الرئيس في حال أن للتحليل نتيجة واحدة.

د- حقل (SubTestOrder): وهو يعبر عن ترتيب ظهور نتيجة التحليل الفرعي كجزء من نتيجة التحليل الأساسي، وهو الترتيب المصطلح عليه طبياً لتسهيل قراءة نتائج التحاليل الطبية. فمثلاً تُكتب نتيجة تحليل تعداد كريات الدم الحمراء (RBC) على النحو التالي (4.7- 14.5-42) أي تتكون نتيجة تحليل كريات الدم الحمراء من ثلاثة نتائج فرعية الأولى (RBC) والثانية (HB) والثالثة (Hct)، أما عندما يكون للتحليل الرئيس نتيجة واحدة يكون هنالك رقم واحد لترتيب الظهور هو (1).



SubTests

SubTestID	TestID	SubTestName	SubTestOrder	Notes
259	1142	SGOT	1	
92	1148	b1	1	
93	1148	b2	2	
126	1149	Bleeding T.	1	
125	1150	Bleeding T (IVY)	1	
128	1151	DIRECT MICROSPY E	1	
129	1151	EPITHELIAL CELLS	2	
130	1151	WBC	3	
131	1151	PUS CELLS	4	
132	1151	FUNGUS	5	
133	1151	CANDIDA	6	
134	1151	TRICHOMONAS	7	
135	1151	EPITHELIAL CELLS1	8	

Record: 1 of 282 | No Filter | Search

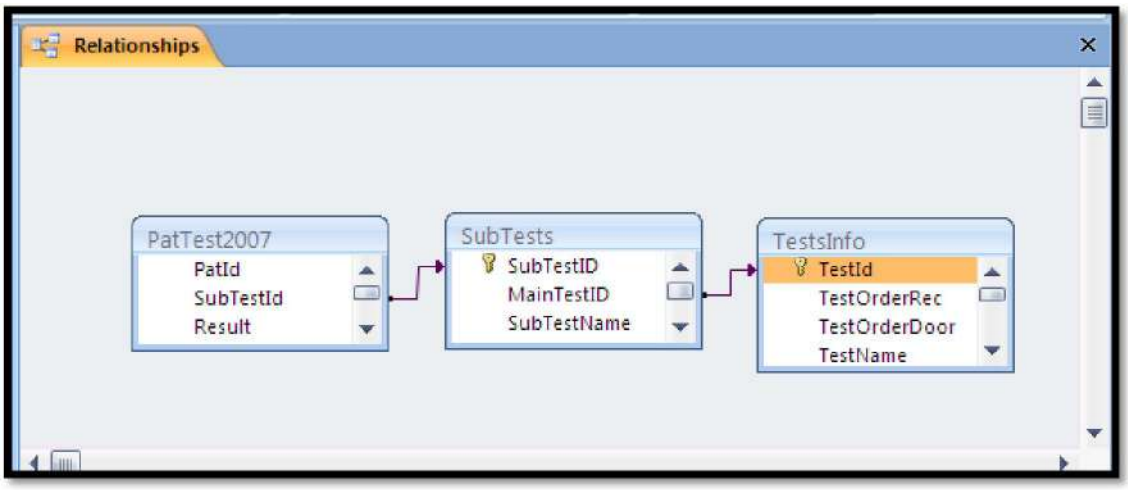
الشكل (5-3- 2): جدول التحاليل الفرعية (SubTests)، ويتبين فيه على سبيل المثال أن التحليل الذي رقمه 1151 له ثمانية تحاليل فرعية أما التحليل ذي الرقم (1149) فله نتيجة واحدة وبالتالي يكون التحليل الفرعي هو ذاته التحليل الرئيس

5•3•2•1•2 إنشاء جدول جديد لتخزين نتائج التحاليل الطبية (PatTest)

تعتمد الخطوة الثانية للتعامل مع نتائج التحاليل الفرعية على فكرة تصميم جدول جديد بديل عن الجداول القديمة التي تحتوي نتائج التحاليل الطبية للمرضى. ويستمد هذا الجدول بياناته من الجداول القديمة، بحيث يأخذ نتائج التحليل الذي له عدة نتائج فرعية من الجدول الخاص بهذا التحليل مثل تحليلي كريات الدم البيضاء (WBC) وكريات الدم الحمراء (RBC)، ويأخذ نتائج التحاليل الأخرى التي لها نتيجة واحدة مثل الكولسترول (Cholesterol) والكرياتينين (Creatinine) من الجدول الأساسي لنتائج التحاليل الطبية.

وكمرحلة انتقالية قبل لم شمل جداول قاعدة بيانات التحاليل الطبية لعام 2007 مع قاعدة بيانات التحاليل الطبية لعام 2008، تبين أنه من الأنسب والأسهل أن تُنظَّف وتُجهَّز البيانات في كل من هاتين القاعدتين ثم تُدمج بيانات الجداول النظيفة معاً فيما بعد. أي سيجري أولاً إنشاء الجدول (PatTest2007) الذي تُعرض فيه النتائج عند مستوى التحاليل الفرعية ويرتبط كل

تحليل فرعي مع ما يقابله في جدول (SubTests)، وهو بدوره يربط التحاليل الفرعية مع التحاليل الرئيسة التابعة لها في جدول (TestsInfo). انظر الشكل (5-3-3)



الشكل (5-3- 3): العلاقات الترابطية في برنامج أكسس بين الجداول (PatTest2007) و(SubTests) و(TestsInfo)

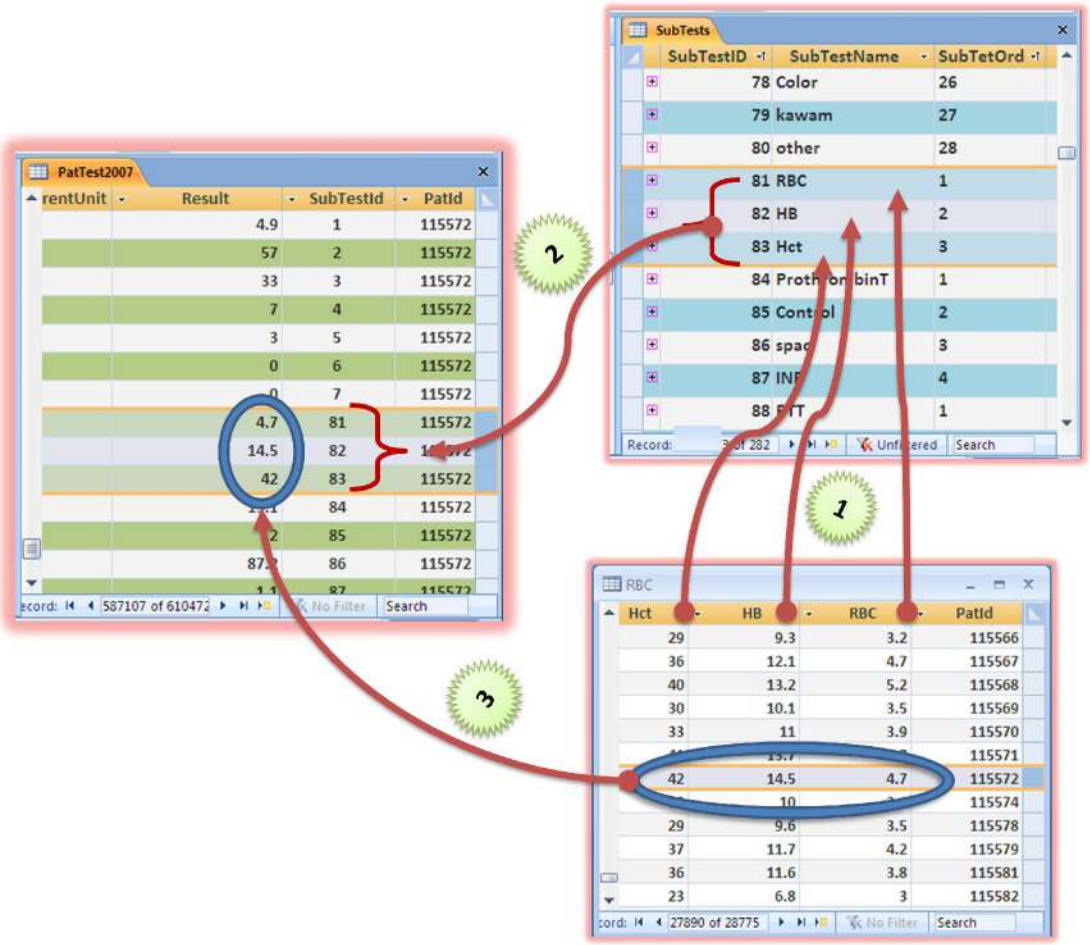
يتألف الجدول (PatTest2007) من ثلاثة حقول أساسية هي:

أ- حقل (PatId) وهو الحقل الذي يدل على إجراء التحاليل للمرضى، بحيث أن مجموعة التحاليل التي يجريها المريض في كل مرة تأخذ نفس هذا الرقم، وعندما يُجري المريض نفسه تحاليل أخرى في يوم آخر فإنه يأخذ رقم (PatId) جديد.

ب- حقل (SubTestId) وهو حقل رقمي ويرتبط بحقل رقم التحليل الفرعي الموجود في جدول (SubTests).

ج- حقل (Result) وهو حقل رقمي مخصص لتخزين نتائج التحاليل الطبية الفرعية بشكل رقمي.

يتحول كل سجل من سجلات التحاليل التي لها عدة نتائج فرعية إلى عدة سجلات بعدد التحاليل الفرعية المكونة لهذا التحليل. وكمثال على ذلك نجد أن تحليل الكريات الحمراء (RBC) يتألف من ثلاثة تحاليل فرعية كما هو مبين في الشكل (5-3-4)، وقد تحول كل سجل في هذا الجدول إلى ثلاثة سجلات في جدول نتائج التحاليل الجديد (PatTest2007).



الشكل (5-3- 4): تحويل كل سجل في جدول تحليل كريات الدم الحمراء (RBC) إلى ثلاثة سجلات في الجدول الجديد (PatTest2007)، وذلك لأن تحليل كريات الدم الحمراء يتكون من ثلاثة تحاليل فرعية هي (RBC, HB, Hct) والتي تم ترقيمها سابقاً في جدول (SubTests) بالأرقام (81، 82، 83)

وقد جرت مراعاة الأمور التالية أثناء عملية نقل البيانات من الجداول القديمة إلى جدول النتائج الجديد (PatTest2007):

أ- يُعبَّر عن نتائج بعض التحاليل بطريقة غير رقمية أي نصية (مثل Positive, N, Negative clear، Bloody، Turbid، Crystal)، أو رمزية (مثل –، +، ++، +++)، أو قيم منفصلة (مثل 1/80، 1/160، 1/320)، وبالتالي يصعب على البرامج الحاسوبية فهم وتقييم ومقارنة هذه النتائج بوضعها الحالي. وللتغلب على هذه المشكلة جرى إتباع أسلوب تحويل التعابير غير الرقمية في حقل النتائج إلى أرقام تُعبِّر عن هذه النتائج وعن درجاتها، فمثلاً تم استبدال الرقم (0) بالقيم التي تدل على سلبية التحليل، والرقم (1) بالنتائج الايجابية، والرقم (2) بالنتائج التي تحمل إشارتين ايجابيتين، والرقم (3) بالنتائج ذات الثلاث إشارات موجبة، أي كلما زادت شدة الإصابة ازداد الرقم المعبر عنها.

ولذلك تم إنشاء جدول النتائج النصية (TextResult) وهو جدول تعريفي للتعابير التي جرى استبدالها ومدلول كل منها، انظر الشكل (5-3-5). وقد تزود الجدول بالبيانات النصية لنتائج التحاليل الطبية من البيانات المتوفرة في عامي 2007 و2008، وأصبح بالتالي مؤلفاً من 161 سجلاً، ومن الممكن رفده في حال الرغبة مستقبلاً بالمزيد من التقييمات في حال ظهور نتائج نصية جديدة غير واردة في البيانات الحالية.

واستبدلت بعد ذلك القيم النصية الموجودة في جدول نتائج التحاليل الطبية القديم بالقيم الرقمية الجديدة المقترحة في هذا الجدول، ثم نُقلت القيم الرقمية الجديدة إلى جدول النتائج الجديد (PatTest2007).

TextResult

TextResultID	ResultCode	TextResult	SubTestID
138	0	N	47
139	1	+	47
140	2	++	47
141	3	+++	47
142	4	++++	47
143	5	+++++	47
144	0	N	48
145	1	+	48
146	2	++	48
147	3	+++	48
80	0	Neg.	161
79	1	Pos.	161
26	0	Neg.	165
12	1	Pos.	165
47	0	clear	166
43	1	Crystal-clear	166
44	2	Crystal	166
8	3	little Turbid	166
41	4	Turbid	166
52	5	Bloody	166

Record: 99 of 161 Filtered Search

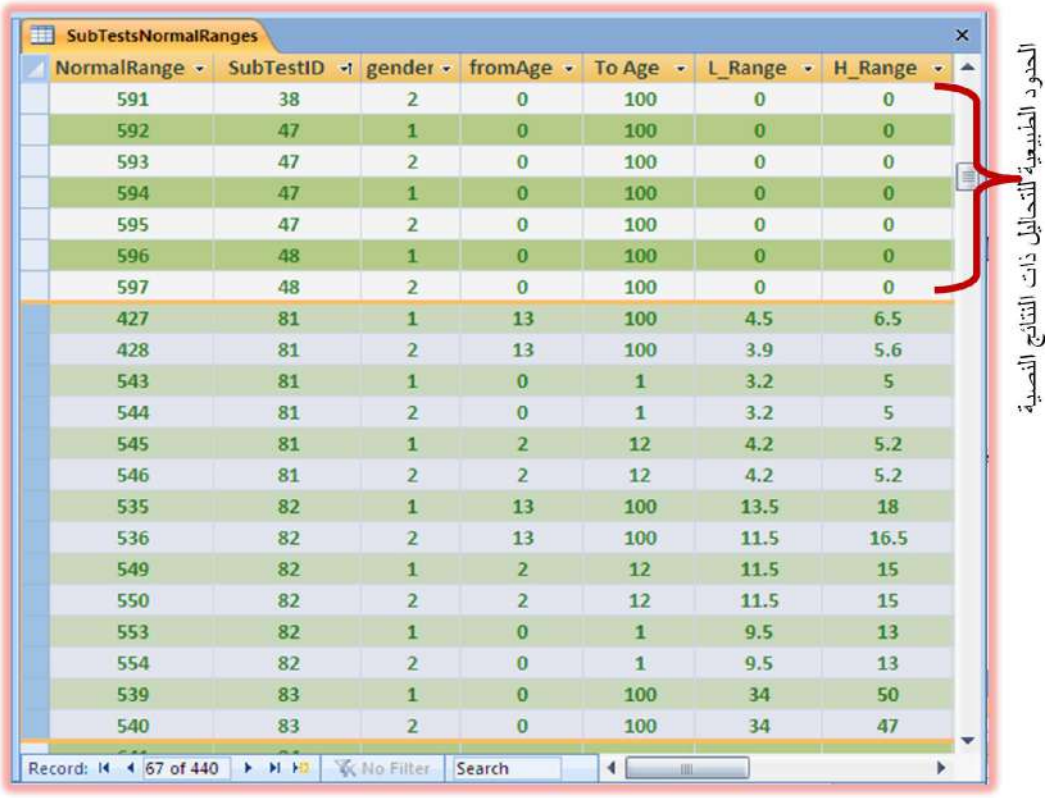
الشكل (5-3- 5): جدول تحويل التعابير غير الرقمية لنتائج التحاليل الطبية إلى قيم رقمية، وقد كان يُعبَّر سابقاً عن القيم السالبة مثلاً بعدة رموز مثل (-, Negative, N, Neg.)، بينما أصبح يُعبر عن أي منها حالياً بالقيمة (0)

وقد ساعد تحويل التعابير غير الرقمية إلى رقمية في إيجاد طريقة جيدة لتعريف الحدود الطبيعية لهذه النتائج في جدول القيم الطبيعية للتحاليل الفرعية الجديد، وبالتالي مساعدة البرامج الحاسوبية على التمييز بين النتائج الطبيعية وغير الطبيعية بغض النظر عن اسم التحليل.

ب- استدعى الاعتمادُ على نتائج التحاليل الفرعية الحاجةَ إلى وجود تعريف للقيم الطبيعية لهذه التحاليل، وبالتالي تم إنشاء جدول جديد للقيم الطبيعية المقابلة للتحاليل الفرعية (SubTestsNormalRanges) الواردة في جداول التحاليل التي أجريت للمرضى. واستمد هذا الجدول جزءاً من بياناته من الجدول الأساسي (NormalRanges) المخصص لتقييم التحاليل الرئيسة التي لها نتيجة وحيدة، والذي احتوى على بعض الأخطاء والقيم المفقودة التي استوجبت التصويب، واستمد الجزء الآخر من عدة مصادر أهمها المراجع العلمية ومواقع الانترنت التخصصية.

كما تمت إضافة تقديرات للقيم الطبيعية للتحاليل الطبية ذات النتائج النصية، فمثلاً إذا كان لدينا تحليل ما وكان الترميز الرقمي لنتائجه كما يلي (-:0، +:1، ++:2، +++:3) فإنه يمكن اعتبار أن كل من الحد الأعلى والأدنى للقيم الطبيعية لنتائج هذا التحليل هو 0، بمعنى أن أي نتيجة تقع خارج مجال الصفر ستكون غير طبيعية وكلما زادت القيمة كلما دل على شدة الإصابة المرضية التي يدل عليها هذا التحليل، انظر الشكل (5-3-6).

وقد بلغ عدد سجلات هذا الجدول 440 سجل حيث أنه جرى فقط إدراج القيم الطبيعية للتحاليل التي وردت لها نتائج في العامين المدروسين..



الشكل (5-3- 6): الجدول الجديد للقيم الطبيعية (SubTestsNormalRanges) ويتضمن القيم الطبيعية للتحاليل الفرعية

ج- تبين من المقارنة بين القيم الطبيعية في عامي 2007، 2008 أن هنالك اختلافات بين العامين في وحدات قياس بعض التحاليل، فمثلاً قُدِّر الحد الأعلى للقيم الطبيعية للبروتين في عام 2007 بـــ 0.150 (g/24 hrs) بينما وجدناه في عام 2008 مقدراً بقيمة 150mg/24 hrs، كما في الشكل (5-3-7). وقد أدى هذا التعديل في مقياس التقييم إلى تشوه بيانات نتائج التحاليل الطبية، حيث وجدنا أن بعض النتائج خُزِّنت بمراتب عشرية ونتائج أخرى بالأعداد الطبيعية. ولتوحيد وحدة القياس والنتائج قمنا باعتماد وحدة قياس عام 2008، وعدلنا نتائج التحاليل المقيمة على أساس وحدة القياس القديمة وذلك بضربها بالرقم 1000 لكي تتسجم مع القيم الجديدة.

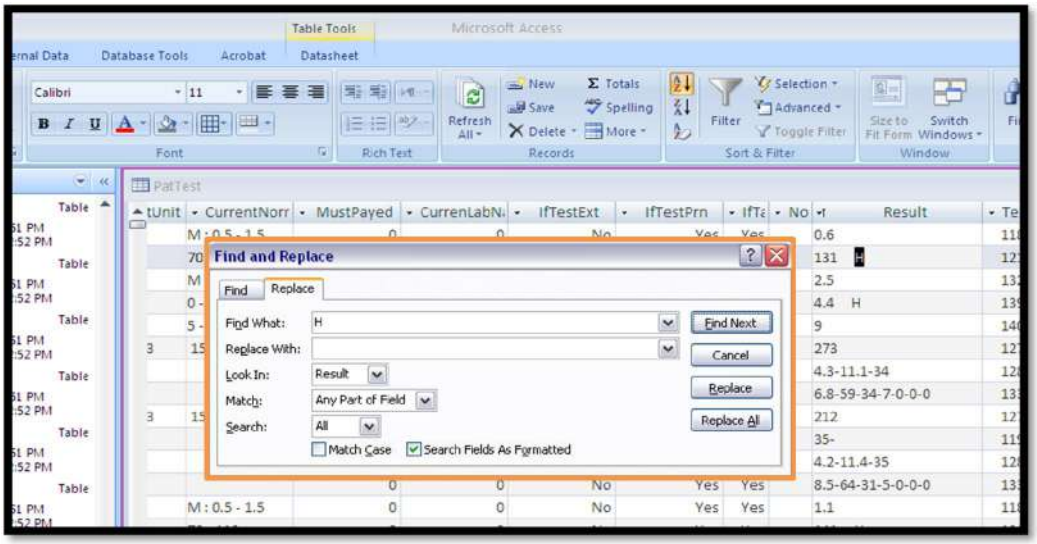
SubTestsNormalRanges									
NormalRangesID	SubTestID	SubTestName	gender	fromAge	To Age	Unite	L_Range	H_Range	Note
511	284	U/Protein in 24 hrs	1	0	100	mg/24 hrs	0	150	
512	284	U/Protein in 24 hrs	2	0	100	mg/24 hrs	0	150	

الشكل (5-3- 7): القيم الطبيعية لتحليل البروتين مقاسة بالملي-غرام في عام 2008، بينما كانت تقاس بالغرام في عام 2007

د- تحتوي بعض حقول نتائج التحاليل الطبية على الرمز H وهو رمز متعارف عليه أثناء طباعة النتيجة للمريض، ويعني النتيجة مرتفعة عن الحد الطبيعي (L تعني منخفضة عن الحد الطبيعي)، وقد قام عاملو المخبر أحياناً بكتابة أحد هذين الرمزين يدوياً وليس آلياً كتقييم للنتائج. ولابد لهدف الدراسة من إزالة كاملة لأي من هذين الرمزين من حقل النتائج بهدف التوصل إلى حقل رقمي للنتائج خالٍ من الرموز، ولأنهما يعبران عن الرأي الشخصي لموظف المخبر أثناء كتابة النتيجة إلى برنامج نتائج التحاليل الطبية، كما أنه يوجد نسبة كبيرة من القيم المرتفعة والمنخفضة التي لم يكتب بجانبها أي رمز. وتمت عملية الحذف هذه ببساطة من خلال تعليمة استبدال كل من هذين الرمزين بلا شيء أينما وردا في هذا الحقل، انظر الشكل (5-3-8).

ه- أظهر حقل النتائج (Result) وجود بعض الرموز المدخلة بطريق الخطأ أو أنها تعني عدم وجود نتيجة مثل (*،***، - /، إإ، ttt،SSK .%،#.)، وقد تمت إزالة كافة الرموز والأخطاء المطبعية من حقل نتائج التحاليل الطبية، وبالتالي أصبحت حقول بعض النتائج

التي لا تحتوي على قيم رقمية فارغة من البيانات أي قيمتها (Null).



الشكل (5–3– 8): طريقة حذف الرموز غير المرغوبة من حقل النتائج في جدول تحاليل المرضى

و- عُثر على بعض السجلات التي تحتوي على عبارات نصية عند النتائج مثل (اعادة، لا يوجد مصل، غير نظامي، ملغى، الدم غير كاف، الدم منحل، غير نظامي، لا يوجد مواد..) وقد تم حذف هذه السجلات، بسبب عدم فائدتها.

ز- تعاني قواعد البيانات المدروسة من عدم وجود أية قيود على القيم المدخلة إلى النظام، مما أدى إلى ظهور أعداد كبير من القيم المشوهة، فمثلاً لا يوجد قيود – حد أعلى وحد أدنى- على قيم نتائج التحاليل الطبية التي يتم إدخالها. ولذلك جرى في هذه الدراسة وبطرق شبه يدوية التأكد من أن نتائج التحاليل تقع ضمن الحدين الأدنى والأعلى المقبولين كنتائج، وإيجاد حلول مناسبة للقيم الخاطئة، وهنا تمت الاستعانة بقيم حقلي الحدين الأدنى والأعلى للمجال الحياتي للتحليل – إن وجِدت – وبالمراجع الطبية المناسبة لتقويم معظم هذه الأخطاء.

ح- تبيَّن وجود بعض النتائج المقيّمة بمراتب عشرية مثل (1.300)، في حين أن بقية النتائج لهذا التحليل كانت تقع بين (800–4000)، وبالبحث عن أسباب القيم المشوهة تبين أن ذلك يعود إما إلى خطأ في إدخال الفاصلة إلى هذه النتائج، أو إلى تَغيُّر وحدة القياس من الغرام إلى الميلي غرام لهذه النتائج، والتي ذكرت قبل قليل، وبكل الأحوال تم إيجاد الحلول المناسبة لتصحيح هذه التشوهات في البيانات.

ط- تبيَّن من خلال استعراض البيانات أن هنالك بيانات عن نتائج تحاليل دون وجود ارتباط لها مع بيانات المرضى الذين أجريت لهم هذه التحاليل، وقد جرت المحاولة في البداية للإبقاء على مثل هذه البيانات علها تفيد في هذه الدراسة، مثل دراسة التوزيعات التكرارية لنتائج التحاليل الطبية، أو اكتشاف أسباب فقدان البيانات الشخصية لهؤلاء المرضى.. ولكننا وبعد الدوران بين مراحل اكتشاف المعرفة المختلفة وجدنا أنه من الأفضل الاستغناء عنها لأن هذه البيانات لا تحمل أية فائدة لعمليات التحليل والتنقيب، كما أنها تُسيء إلى قاعدة بيانات التحاليل الطبية حيث تسمح بوجود بيانات عن تحاليل طبية غير تابعة للمرضى ولا تحمل أية خاصيات مفيدة.

ي- يحتوي جدول نتائج التحاليل الطبية على الكثير من القيم المفقودة مسبقاً أو بسبب عمليات التنظيف، وعلى اعتبار أنه لا يوجد دراسات سابقة عن هذه البيانات فقد كان من الصعب أن نجد طريقة مناسبة لتوقع القيم المفقودة لبعض التحاليل من خلال نتائج التحاليل الأخرى فمثلاً لا يوجد لدينا معادلات أو علاقات رياضية بين نتائج التحاليل الطبية المختلفة كالعلاقة بين ارتفاع السكر وارتفاع الكولسترول بحسب الحالة المرضية.. ولذلك كان من الأفضل ومنعاً للوقوع في أخطاء جديدة أن تُحذف السجلات التي تحتوي القيم المفقودة.

وبعد الانتهاء من إزالة كافة الرموز والأخطاء المطبعية واستبدال كافة النتائج النصية في كل من جدولي النتائج لعامي 2007- 2008 بقيم رقمية أصبح بالإمكان دمج الجدولين في جدول واحد هو (PatTest)، كما جرى أخيراً تحويل نمط حقل النتائج (Result) من النصي إلى الرقمي، وهو أحد الأهداف المنشودة من الإجراءات السابقة.

3•1•2•3•5 تنظيف جدول بيانات المرضى في مخبر التحليل (Patient)

يقوم جدول المريض (Patient) في قاعدة بيانات المخبر بدور الجدول الخاص بخدمة التحليل، حيث أن مفتاحه الأساسي هو حقل (PatId)، ويُستخدم كمفتاح أجنبي في جدول نتائج التحاليل، وتحمل التحاليل الطبية التي تُجرى للمريض بآن واحد نفس المفتاح الأجنبي، كما يتغير المفتاح الأجنبي لنفس المريض إذا أجريت له تحاليل أخرى في وقت جديد. ويمكن أن يطلق على هذا

الجدول اسم جديد مناسب وهو (PatientTestsService) لاستخدامه في قاعدة البيانات الجديدة التي ستجمع بين قواعد البيانات الثلاثة المتاحة للدراسة، وقد أجريت على هذا الجدول بعض عمليات التنظيف على النحو التالي:

أ- تم اعتبار – عند بداية العمل في مرحلتي استيعاب وتهيئة البيانات- أن حقل اسم المريض وكنيته هما الحقلان المناسبان لكي يُستخدما في الربط بين قاعدة بيانات مخبر التحاليل الطبية وقاعدة بيانات مكتب القبول، وذلك لأن حقل رقم الاضبارة يحتوي على العديد من القيم المفقودة، بينما لا يوجد قيم مفقودة في الحقول المتعلقة بالاسم. وبناءً على ذلك جرى تكريس الكثير من الجهد في التفكير بطرق التخلص من عيوب وأخطاء اللغة العربية التي ارتُكبت أثناء إدخال البيانات، مثل أخطاء الهمزات في حرف الألف (أ، إ، آ).. ولكن تبين فيما بعد أن هذا لن يفيد في البيانات المدروسة، وأنه مهما ارتفعت نسبة السجلات التي أمكن ربطها معاً من خلال عملية تنظيف البيانات، فإن الأخطاء الموجودة تفوق القدرة البشرية والآلية على حلها لأنه لا يوجد استراتيجيات واضحة أثناء كتابة الأسماء سواء في برنامج مخبر التحاليل أو في برنامج مكتب القبول، فأحياناً كثيرة يتضمن الاسم الكامل اسم الأب، أو يتم استبدال الكنية باسم الأب.. وبناءً على ذلك تبين أنه من المجدي أكثر أن يُستخدم رقم الاضبارة في ربط الجداول بعد أن تم الاهتمام بتصحيح أخطائه. وقد وَرَدَ سابقاً أن رقم اضبارة المريض يُخزن في جدول المرضى في الحقل المخصص لرقم الهاتف وهو من النمط النصي، ولذلك تم تغيير اسمه إلى (FileNo)، وتغيير نمطه إلى الرقمي ليتوافق ويرتبط مع رقم اضبارة المريض في كل من جدولي المرضى في قاعدتي بيانات مكتب القبول ومكتب السجلات الطبية، بعد تنظيف البيانات المشوهة فيه والتي تحتوي على رموز وكتابات غير رقمية. كما جرى فحص سجلات المرضى يدوياً وتمت إضافة أو تصحيح رقم الاضبارة في سجلات مرضى قاعدة بيانات التحاليل الطبية التي تحتوي اسم المريض الذي أمكن التعرف عليه بسهولة بالاستعانة ببيانات سجلات القبول.

ب- يحتوي جدول المرضى كما سبق وذكرنا على حقل سعر الوحدة (UnitPrice)، وقد أظهر الاستعلام الذي أجري على بيانات مخبر التحاليل الطبية لعام 2008 أن 28048 مريض يدفعون بسعر 20 ليرة للوحدة، وأن 4355 مريض يدفعون بسعر 25 ليرة للوحدة، في

حين أن 43 سجل مريض فيهم أخطاء في كتابة أسعار الوحدات. وأظهر الاستعلام عن بيانات عام 2007 أن 35300 مريض يدفعون بسعر 20 ليرة للوحدة، و62 مريض يدفعون 15 ليرة للوحدة، في حين أن 35 سجل فيها قيم خاطئة متنوعة... وقد جرى تقويم القيم الخاطئة باستبدالها بسعر الوحدة الأكثر تكراراً وهو 20 ليرة للوحدة.

2•2•3•5 تنظيف بيانات القبول في المشفى

1•2•2•3•5 تنظيف حقل رقم الاضبارة (FileNo)

أ- مشكلة القيم المفقودة أو المشوهة

ذكرنا سابقاً بأنه سيتم ربط قاعدة بيانات مكتب القبول بقاعدة بيانات مخبر التحاليل الطبية باستخدام حقل رقم الاضبارة، ولذلك تم بذل أقصى جهد ممكن لتصحيح الأخطاء، وتنظيف قيم هذا الحقل في كل مكان ترد فيه، وتم العمل على تنظيف بيانات حقل رقم الاضبارة في قاعدة بيانات القبول بالتوازي مع حقل رقم الاضبارة في قاعدة بيانات مخبر التحاليل الطبية، وبالاستعانة أحياناً ببيانات السجلات الطبية. وقد وصلت في نهاية عمليات التنظيف نسبة المرضى في قاعدة بيانات التحاليل الطبية الذين لهم أرقام أضابير مقابلة لأرقام الأضابير في قاعدة بيانات مكتب القبول إلى 58%. وكانت مسألة مصير بيانات التحاليل الطبية التي لا تحمل رقم اضبارة تجول في الذهن أثناء التنقل الأولي بين مراحل عملية اكتشاف المعرفة.. وقد تقرر في النهاية التخلص من هذه البيانات، وذلك بعدما تبين أنها لن تخدم هذه العملية ولن تزودها بأية معرفة، وحتى أنها تعيق ربط جداول قاعدة البيانات بروابط (1-∞).

ب- مشكلة التكرار في سجلات بعض المرضى

برزت أثناء تنظيف البيانات مشكلة أخرى في حقل رقم الاضبارة حيث لوحظ وجود عدد من المرضى يحملون أكثر من رقم للاضبارة (FileNo)، ويعود ذلك إلى عدم اهتمام موظفي القبول في التدقيق بمسألة أن يأخذ المريض نفس رقم الاضبارة في كل مرة يتم قبوله في المشفى. ويؤدي هذا الأمر من ناحية دراسة البيانات إلى تكرار في بيانات المريض وإلى ضياع التاريخ المرضي للمريض. وبناءً على ذلك وتمسكاً بأهمية توحيد أرقام الأضابير

للمرضى في هذه الدراسة، فقد بُذل الكثير من الجهد في البحث شبه اليدوي عن هذه الأخطاء وبالتالي انتقاء أحد أرقام الأضابير للمرضى وغالباً الرقم الأخير، واستبدال الأرقام الأخرى للمريض به، ومن ثم البحث ضمن سجلات التحاليل الطبية (في كل من عامي 2007، 2008) عن أرقام الأضابير التي ألغيت من سجلات القبول، واستبدالها أيضاً بالرقم النهائي المعتمد لاضبارة المريض. وقد جرت الاستعانة هنا للتحقق من أن البيانات تنطبق على نفس المريض بحقول (اسم الأب، اسم الأم، تاريخ الميلاد، مكان الإقامة، رقم الهاتف) حيث كان يوجد أحياناً بعض الاختلافات في بيانات نفس المريض سواء في حقل اسم المريض (الذي قد يحمل اسم الأب كجزء من الاسم الكامل) أو في حقل تاريخ الميلاد (الذي يختلف التصريح به من قبول لآخر والذي جرى أيضاً توحيده بانتقاء آخر تاريخ مصرح به غالباً)، أو باسم الأب أو بمكان الإقامة..الخ من التفاصيل التي أخذت حيزاً زمنياً طويلاً من الوقت لتصحيحها والذي كان من المفترض أن يُكرس فقط جزء منه لمعالجة أخطاء في البيانات أهم بكثير من هذه الأخطاء التي تنم عن الضعف الشديد في أنظمة قواعد البيانات المستخدمة في هذه الدراسة.

2•2•2•3•5 تنظيف حقول مكان الإقامة

أ- تنظيف حقل المناطق (Area) وتوحيد انتماءات المناطق إلى المدن (City)

بيَّنت الاستعلامات عن أسماء المدن والمناطق التابعة لها وجود العديد من الأخطاء المطبعية والهجائية، وأخطاء تبعية المناطق للمدن فمثلاً وُجد أن بعض المناطق تم نسبها أحياناً إلى ريف دمشق وأحياناً أخرى إلى دمشق وأحياناً أخرى إلى درعا.. وقد جرى تصحيح هذه الأخطاء بطريقة شبه آلية من خلال أدوات الاستعلام والانتقاء والبحث والاستبدال في برنامج أكسس، فوصَل عدد المناطق (التابعة للمدن) إلى 2208 منطقة بعد أن كان 4013 منطقة قبل التنظيف، انظر الشكل (5-3-9).



الشكل (– –): الاستعلام عن مناطق ومدن إقامة المرضى قبل وبعد عمليات التنظيف

ب- إنشاء جدول جديد لمكان الإقامة

جرى إنشاء جدول مستقل لمكان إقامة المرضى ( ) لاستخدامه لاحقاً كبعد جغرافي في مستودع البيانات، ويرتبط مع جدول بيانات المرضى بمفتاح أجنبي، وتم تزويده بحقل للمنطقة الجغرافية التي ينتمي إليها مكان إقامة المريض ( )، ويتألف هذا الحقل من خمس مناطق جغرافية كما في الشكل (– –).



الشكل (5-3- 10): جدول مكان الإقامة الجديد وقد أضيف إليه حقل المناطق الجغرافية

3•2•2•3•5 تنظيف حقل تاريخ الميلاد

أظهر حقل تاريخ ميلاد المريض، أن هنالك عدد قليل من المرضى تزيد أعمارهم عن المائة، وقد يصل إلى 2000 سنة بسبب الأخطاء المطبعية أحياناً أو عدم تـوخي الدقـة فـي أثنـاء تحصيل بيانات المرضى. وقد تم تصحيح هذه القيم واستبدال معظمها بحيث لا تزيد الأعمـار عن المائة عام. كما ظهرت الحاجة أثناء المعالجة التحليلية للبيانات إلى ضرورة وجود حقـل خاص بأعمار المرضى، وتم لذلك احتساب أعمار المرضى في يوم القبول في المشفى، ومـن ثم جرى إضافة حقل خاص بأعمار المرضى إلى جدول بيانات المرضى الجديد الـذي يُتـابع العمل على بنائه خلال هذه المرحلة.

4•2•2•3•5 تنظيف حقل أسباب القبول

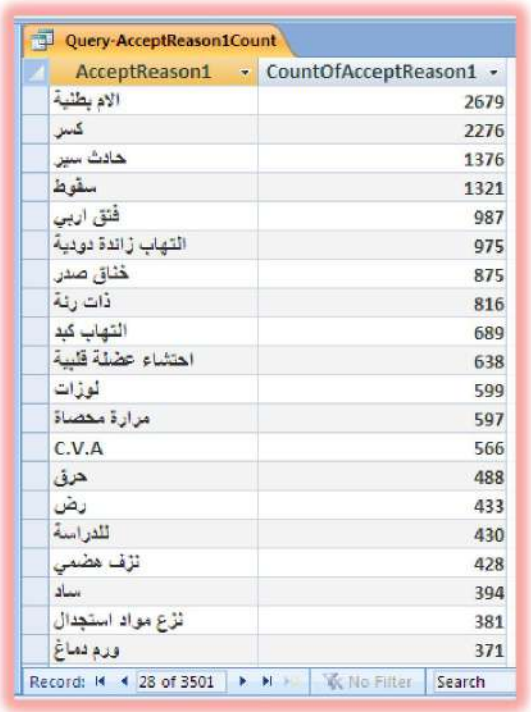
أوضح الاستعلام الأولي الذي ورد في الفقرة **(2•4•2•5)** أن هنالك 9490 سبب لقبول المرضى في المشفى، وأن هذا العدد الكبير لا يعود فقط إلى تنوع أسباب المرض، وإنما يعود أيضاً إما إلى أخطاء مطبعية أو إلى اختلاف طرق التعبير عن الحالات المرضية باختلاف الأطباء المشرفين على تشخيص الأمراض. وقد جرى ضمن مرحلة تنظيف البيانات تكريس الكثير من الجهد لحل هذه المشكلة وتقليص عدد الأسباب بهدف زيادة عدد الأشخاص الذين ينتمون إلى نفس السبب المرضي. وقد اتبعت عدة إجراءات للقيام بذلك، حيث تم أولاً التخلص من بعض التشابهات الهجائية باستبدال جميع الأحرف المتشابهة بحرف موحد شريطة أن لا يؤثر على تغيير مضمون الكلمات، فمثلاً استبدل حرف "ا" بأحرف الألف "أ، إ، آ".

وجرى إتباع طريقة تجزئة البيانات باستخدام أدوات الانتقاء (Filter) حيث كانت تُنتقى البيانات التي تحتوي على كلمة معينة أو أكثر ثم تُرتب هجائياً، ثم يجري التعرف على مضمون الخلايا المتشابهة المعنى ويُستعاض عنها بعبارة موحدة توضع في حقل إضافي أطلق عليه اسم (AcceptReason1). وفي حال احتوت الخلايا المتشابهة على كلمات تدل على بعض التفاصيل مثل شدة الإصابة أو نوعها أو مكانها (حادة، مزمن، ساعد، فخذ) فيتم وضع هذه الكلمات التي تدل على التفاصيل الإضافية للإصابة في حقل إضافي آخر هو (AcceptReason2). وكمثال على ذلك فقد أُحرز اختصارٌ في أسباب القبول التي تحتوي على كلمة رئة من (226) سبب إلى (156) سبب، انظر الشكل (5-3-11).

وتحقق نتيجة لجميع عمليات الاختصار التي أجريت على حقل الأسباب المرضية انخفاضٌ للعدد الكلي للأسباب المرضية إلى 3501 سبب، أي إلى ما يقارب ثلث عدد أسباب القبول الموجودة في قاعدة بيانات مكتب القبول الأساسية، انظر الشكل (5-3-12).



الشكل (- -): استعلام عن الحالات المرضية التي تحتوي كلمة (رنة) قبل وبعد تنظيف البيانات، وقد أدت عملية التنظيف إلى اختصار عبارات الأسباب المرضية التي تتضمن كلمة (رنة) من ( ) حالة إلى ( ) حالة



الشكل (5-3- 12): تقلص عدد الأسباب المرضية من (9490 سبب) إلى (3501 سبب) بعد عمليات التنظيف

5•2•2•3•5 جدول الشعبة الطبية

تنجح عمليات المعالجة التحليلية الآنية في التعامل مع البنى الهرمية للبيانات (Hierarchy)، أي تلك التي تستطيع أن تنشرها أو تضغطها بحسب المستوى التصنيفي المرغوب عرضه، وبناءً على ذلك فقد جرت المحاولات لتعريف مجموعات تصنيفية جديدة للبيانات مثل المناطق الجغرافية التي أضيفت إلى جدول المدن. أما في جدول الشعب الطبية فنظراً لأنه يوجد 27 شعبة طبية في المشفى فقد تم اقتراح تكوين مجموعات من هذه الشعب عددها 13 مجموعة، انظر الشكل (5–3–13)، والتي يُمكن أن تُستخدم في تصنيف البيانات أثناء المراحل القادمة من اكتشاف المعرفة.

division

Divisior	division	Divisior	DivisionGroup
27	الحواضن	3	أطفال أوحواضن
18	جراحة اطفال	1	جراحة
26	جراحة صدرية	1	جراحة
11	جراحة عامة	1	جراحة
12	جراحة عصبية	1	جراحة
20	جراحة اوعية	1	جراحة
4	جراحة بولية	1	جراحة
16	تجميل	5	حروق أوتجميل
15	حروق	5	حروق أوتجميل
10	داخلية صدرية	7	داخلية صدرية
24	رقابة دوائية	10	رقابة دوائية
23	عناية عصبية	7	عصبية عناية أو داخلية
2	داخلية عصبية	7	عصبية عناية أو داخلية
22	مفاصل	4	عظمية أومفاصل

Record: 7 of 27 | No Filter | Search

الشكل (5–3– 13): جدول الشعب الطبية وقد أضيف إليه حقل مجموعات الشعب الطبية.

6•2•2•3•5 تنظيف حقل اسم الأب

يُفيدُ –بشكل عام– وجودُ اسم الأب كثيراً فيما لو كانت قاعدة البيانات تحتوي على جدول بأسماء المرضى له بنية شجرية ويحمل فيها حقل الأب رقماً يرتبط مع المفتاح الأساسي للجدول الدال على الرقم الخاص بكل مريض، بحيث يسمح باستخراج علاقات القربى بين الأشخاص.. ومنه يمكن التعرف مثلاً على الانتشار الوراثي لبعض الأمراض.. ولكن لن يُتاح تطبيق ذلك في هذه الدراسة، مما يقلص من أهمية تصويب أخطاء حقل اسم الأب إلا من

ناحية إزالة التكرار في سجلات المرضى نتيجة اختلاف تسمية الأب لنفس المريض كما جاء في فقرة تنظيف حقل رقم الاضبارة.

3•2•3•5 تنظيف بيانات مكتب السجلات الطبية

تعاني قاعدة بيانات مكتب السجلات الطبية من وجود عدد كبير من القيم المفقودة بحيث يصعب تعبئة هذه القيم، وخصوصاً أن معظم القيم المفقودة تتعلق بالنواحي الطبية التي يصعب تخمينها واستنتاجها من خلال الخاصيات القليلة المتوفرة في هذه القاعدة. وسنعرض فيما يلي لبعض المحاولات التي أجريت في سبيل الاستفادة من هذه البيانات.

1•3•2•3•5 الحقول المتعلقة بقبول المرضى في المشفى

تعرفنا في المرحلة الثانية من مراحل اكتشاف المعرفة على عيوب البيانات في قاعدة بيانات مكتب السجلات الطبية، وبما أن بعض الحقول كانت متعلقة بدخول المرضى للمشفى وأنها مشابهة لتلك الموجودة في قاعدة بيانات القبول والتي تتمتع بجودة أكبر فقد تم تجاهل تنظيف هذه البيانات.

2•3•2•3•5 تنظيف حقل حالة خروج المرضى (OutState)

يحتوي حقل حالة الخروج من المشفى على ثلاثة قيم خاطئة و1953 قيمة مفقودة، وقد جرى تحويل هذه القيم إلى (تحسن أو شفاء المريض)، وأصبح عدد المرضى في حقل حالة الخروج (وبعد عمليات التنظيف) يتوزع على النحو المبين في الشكل (5-3-14).

Query-حالة المرضى الصحية عند الخروج	
Out State	Count Of Out State
تحسن أو شفاء المريض	34897
تحويل إلى مشفى آخر	331
وفاة المريض	1355

الشكل (5-3- 14): نتائج الاستعلام عن حالة خروج المرضى بعد تنظيف البيانات في جدول السجلات الطبية

3•3•2•3•5 تنظيف حقل تاريخ الخروج

تعرفنا في الفقرة **(5•2•5)** على عيوب تسجيل تاريخ خروج المرضى من المشفى، وقد جرت محاولة تشذيب هذه البيانات وحساب مدة الإقامة في المشفى، ولكن للأسف كان الموضوع

معقداً للغاية، واتضح أنه من الأفضل أن تتوقف هذه المحاولة غير المجدية، وخصوصاً أنه عندما رُبط هذا الجدول مع جدول بيانات المرضى في قاعدة بيانات مكتب القبول ازدادت أعداد القيم المفقودة إلى أكثر من نصف عدد السجلات. وتبين نتيجةً لذلك أن البيانات السيئة تبقى سيئة ومن الصعب تحسينها، وهي من العيوب الأساسية التي تشوب معظم قواعد البيانات في البلدان غير المتقدمة.

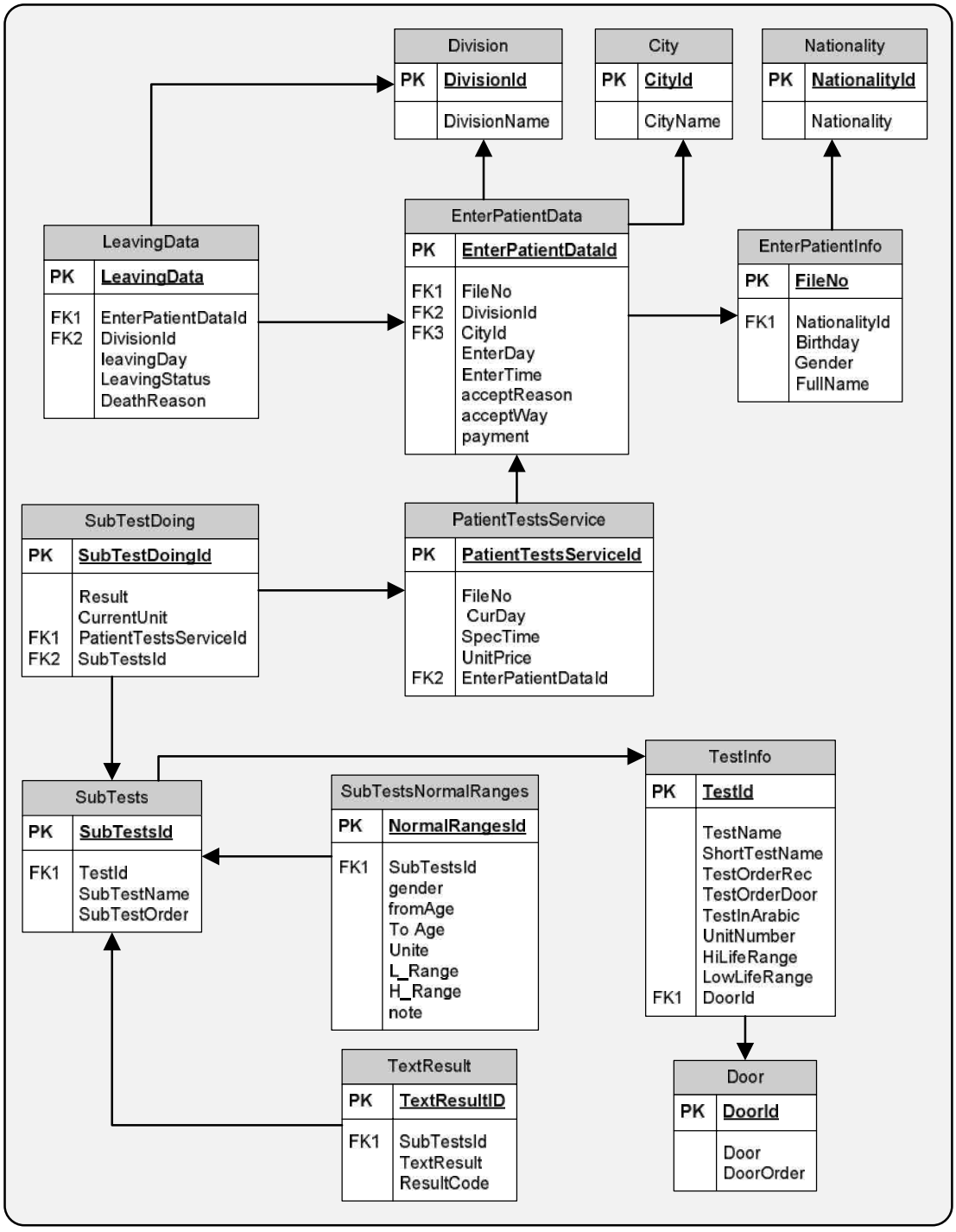
أدت عملية الاستعلام عن أرقام الأضابير المشتركة بين جدول السجلات الطبية وجدول المرضى في مكتب القبول أنه يوجد فقط 24779 سجل من أصل 36583 سجل في السجلات الطبية يشترك برقم الاضبارة مع سجلات مكتب القبول، ويعود ذلك إلى سببين الأول هو عدم الحصول على بيانات من مكتب السجلات الطبية إلا لغاية الشهر الخامس من عام 2008، والسبب الثاني هو وجود نقص أو أخطاء في توثيق بيانات السجلات الطبية. ولذلك ستُعتبر خلايا سجلات القبول التي لا يقابلها بيانات عن الخروج من مكتب السجلات الطبية بأنها خلايا فارغة أي قيمها (Null)، وستجري المحاولة للاستفادة قدر الإمكان من البيانات المتوفرة.

4•2•3•5 دمج قواعد البيانات الثلاثة ضمن قاعدة بيانات واحدة في برنامج أكسس

تم جمع جداول البيانات النظيفة الناتجة عن العمليات السابقة من كل من قواعد بيانات مخبر التحاليل الطبية ومكتب القبول ومكتب السجلات الطبية معاً في مكان تخزيني واحد في برنامج أكسس، فتشكلت تقريباً قاعدة بيانات واحدة تحتوي فقط على الجداول والحقول التي تهم الدراسة الحالية. وهي قاعدة بيانات تأسيسية لمستودع البيانات الذي سيتم بناؤه في الخطوات القادمة اعتماداً على الاستعلامات المستمدة منها.

ولكن هذه القاعدة ما تزال تفتقد لوجود مفتاح ربط منطقي بين جدول قبول المريض في المشفى وجدول خدمة تحاليل المريض، حيث من المفترض أن يوجد جدول وسيط بين بيانات المريض وبين خدمة إجراء التحاليل وهو جدول القبول، وأن يأخذ المريض رقم اضبارة تدل على بياناته الشخصية، ورقم خاص بكل عملية قبول في المشفى (EnterPatientDataID) وهو الرقم الذي يجب أن يستخدمه مخبر التحاليل الطبية للتعرف على هذا المريض عند إجراء التحاليل له وليس رقم الاضبارة المُطبَّق حالياً.. ولكن سيتم في فقرات قادمة ولأهداف استكمال هذه الدراسة استخدام حقل رقم الاضبارة (FileNo) في ربط مثل هذه الجداول.

وقد جرى تعريف الروابط المنطقية بين جداول هذه القاعدة وفق مخطط الكيان–علاقة المبين في الشكل (5–3–15)، وتمثل هذه القاعدة الشكل النظري الذي يمكن أن ترتبط به قاعدة بيانات مخبر التحاليل الطبية مع بيانات قبول وخروج المرضى من المشفى.



الشكل (5–3– 15): مخطط كيان– علاقة النظري لقاعدة البيانات الجديدة (Access Clinical Laboratory Study.accdb) المستنتجة من قاعدة بيانات مخبر التحاليل الطبية وقاعدة بيانات مكتب القبول وقاعدة بيانات مكتب السجلات الطبية، وصمم المخطط باستخدام برنامج (Microsoft Visio)

3•3•5 بناء مستودع البيانات الجديد (Clinical Laboratory Study) في نظام (Microsoft SQL Server 2008)

يستطيع برنامج SQL Server وغيره من برامج التنقيب أن يقوم بعمليات التحليل والتنقيب على أي نوع من الملفات مثل ملفات أكسس أو اكسل، ولكن من الأفضل أن يتم نقل وتخزين كافة البيانات التي يحتاجها هذا البحث إلى برنامج SQL Server، وذلك لعدة أسباب منها التعامل المباشر والأسهل مع البيانات، وإمكانيات التخزين الواسعة للبيانات فيSQL Server، كما أن الربط بين الجداول يكون أدق وخصوصاً أنه يراعي التطابق الدقيق بين أنماط حقول البيانات المراد ربطها، وذلك على العكس من برنامج أكسس حيث أنه بمجرد أن يأخذ الحقل صفة Number يمكن ربطه مع أي حقل رقمي آخر مهما كان نوع الرقم (Integer, Long Integer, Double, Single)، بينما يجب أن يتحقق التماثل الدقيق في نوع الرقم في (SQL Server).

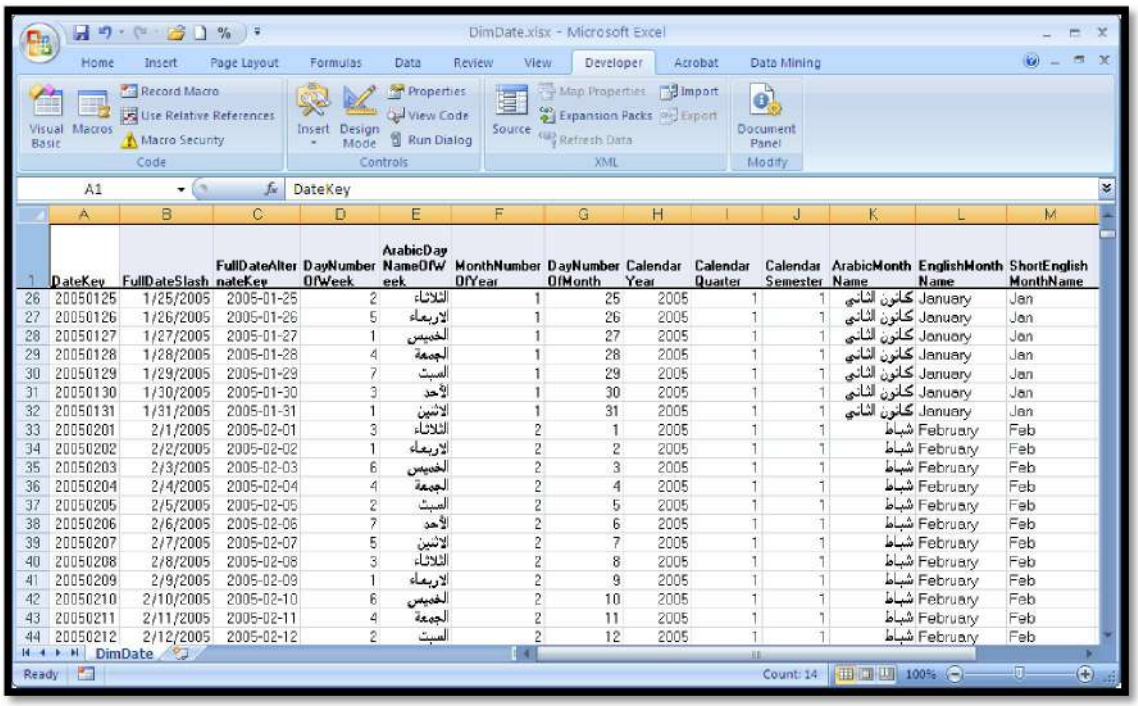
1•3•3•5 تجهيز البيانات المراد نقلها

تتمحور خطوة العمل على تجهيز البيانات حول تعديل المعطيات في الحقول والبيانات بشكل يتلاءم مع بناء مستودع بيانات قادر على تأمين الاحتياجات من البيانات لخدمة عمليات المعالجة التحليلية الآنية وعمليات التنقيب. وقد تم تحقيق جزء من هذه المتطلبات أثناء تنسيق وبناء بيانات التركيب الجديدة المعروضة في الفقرات السابقة، وسيتم في الفقرات التالية استكمال بناء هذا المستودع.

ترتكز عمليات المعالجة التحليلية الآنية المرغوبة في هذا البحث على إيجاد بنية مكعب بيانات متناسبة مع البيانات المتوفرة، وقد وُضع في هذه المرحلة تصميم أولي لمكونات مكعب البيانات المطلوب لعمليات المعالجة التحليلية الآنية وهي جدول الحقيقة وما يتضمنه من مقاييس (Measures) وجداول الأبعاد وبُناها الهرمية.

أظهرت النظرة الأولية للبيانات المتوفرة وجود مقياسين أساسين مباشرين هما أسعار وأعداد التحاليل الطبية التي أجريت للمرضى، وبناءً على ذلك جرى اقتراح جدول حقيقة أساسي هو جدول التحاليل الطبية الرئيسة المنفذة (FactMainTestDoing)، وأربعة جداول أبعاد هي بعد بيانات التحاليل الطبية، وبُعد بيانات المرضى والبعد الزمني، والبعد الجغرافي.

يتألف البعد الزمني من حقل مفتاح أساسي (DateKey) وحقل التاريخ وحقول أخرى يصف كل منها انتماء هذا التاريخ إلى إحدى التقسيمات الزمنية المختلفة (السنة، نصف السنة، ربع السنة، الشهر، الأسبوع، اليوم)، ويساعد هذا الجدول في تصفح البيانات نسبةً إلى أي من التقسيمات المختلفة للزمن. وحيث أنه لا يتوفر هذا الجدول في قواعد البيانات العلاقاتية فقد تم إنشاؤه باستخدام الدوال المتوفرة في برنامج اكسل (Excel) لاستخدامه كبُعد زمني (DimDate) يمتد بين عامي 2005 و2009[1]، انظر الشكل (5-3-16). ثم جرى تصدير هذا الجدول إلى قاعدة البيانات في أكسس وتم استخدامه في تحويل الحقول المتعلقة بالزمن في هذه القاعدة إلى حقول من نمط رقمي قابل للارتباط مع الحقل الأساسي لجدول البعد الزمني (DateKey).



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	DateKey	FullDateSlash	FullDateAlternateKey	DayNumberOfWeek	ArabicDayNameOfWeek	MonthNumberOfYear	DayNumberOfMonth	CalendarYear	CalendarQuarter	CalendarSemester	ArabicMonthName	EnglishMonthName	ShortEnglishMonthName
26	20050125	1/25/2005	2005-01-25	2	الثلاثاء	1	25	2005	1	1	كانون الثاني	January	Jan
27	20050126	1/26/2005	2005-01-26	5	الاربعاء	1	26	2005	1	1	كانون الثاني	January	Jan
28	20050127	1/27/2005	2005-01-27	1	الخميس	1	27	2005	1	1	كانون الثاني	January	Jan
29	20050128	1/28/2005	2005-01-28	4	الجمعة	1	28	2005	1	1	كانون الثاني	January	Jan
30	20050129	1/29/2005	2005-01-29	7	السبت	1	29	2005	1	1	كانون الثاني	January	Jan
31	20050130	1/30/2005	2005-01-30	3	الأحد	1	30	2005	1	1	كانون الثاني	January	Jan
32	20050131	1/31/2005	2005-01-31	1	الاثنين	1	31	2005	1	1	كانون الثاني	January	Jan
33	20050201	2/1/2005	2005-02-01	3	الثلاثاء	2	1	2005	1	1	شباط	February	Feb
34	20050202	2/2/2005	2005-02-02	1	الاربعاء	2	2	2005	1	1	شباط	February	Feb
35	20050203	2/3/2005	2005-02-03	6	الخميس	2	3	2005	1	1	شباط	February	Feb
36	20050204	2/4/2005	2005-02-04	4	الجمعة	2	4	2005	1	1	شباط	February	Feb
37	20050205	2/5/2005	2005-02-05	2	السبت	2	5	2005	1	1	شباط	February	Feb
38	20050206	2/6/2005	2005-02-06	7	الأحد	2	6	2005	1	1	شباط	February	Feb
39	20050207	2/7/2005	2005-02-07	5	الاثنين	2	7	2005	1	1	شباط	February	Feb
40	20050208	2/8/2005	2005-02-08	3	الثلاثاء	2	8	2005	1	1	شباط	February	Feb
41	20050209	2/9/2005	2005-02-09	1	الاربعاء	2	9	2005	1	1	شباط	February	Feb
42	20050210	2/10/2005	2005-02-10	6	الخميس	2	10	2005	1	1	شباط	February	Feb
43	20050211	2/11/2005	2005-02-11	4	الجمعة	2	11	2005	1	1	شباط	February	Feb
44	20050212	2/12/2005	2005-02-12	2	السبت	2	12	2005	1	1	شباط	February	Feb

الشكل (5-3- 16): تصميم جدول البعد الزمني باستخدام التوابع الموجودة في برنامج اكسل

يتكون النموذج الأولي لمكعب البيانات من الجداول التالية والتي ترتبط وفق المخطط المعروض في الشكل (5-3-17):

[1] يشتمل البعد الزمني على بيانات من عام 2005 وذلك بسبب وجود بعض سجلات المرضى في عام 2005، وذلك كما ورد في الفقرة **(2•1•2•5)**

أ- **بُعد المريض**

- مفتاح المريض
- الاسم الكامل
- رقم الاضبارة
- الجنس
- تاريخ الميلاد
- الجنسية
- مفتاح البعد الجغرافي

ب- **البُعد الجغرافي ويرتبط ببعد المريض**

- مفتاح البعد الجغرافي
- اسم المنطقة
- اسم المدينة
- الموقع الجغرافي

ج- **بُعد الزمن**

- مفتاح بعد الزمن
- التاريخ
- رقم ترتيب اليوم من الأسبوع
- اسم يوم الأسبوع بالعربي
- رقم ترتيب اليوم من الشهر
- رقم ترتيب الشهر من السنة
- رقم السنة
- رقم ربع السنة
- رقم نصف السنة
- اسم الشهر بالعربي
- اسم الشهر بالانكليزي
- الاسم المختصر للشهر بالانكليزي

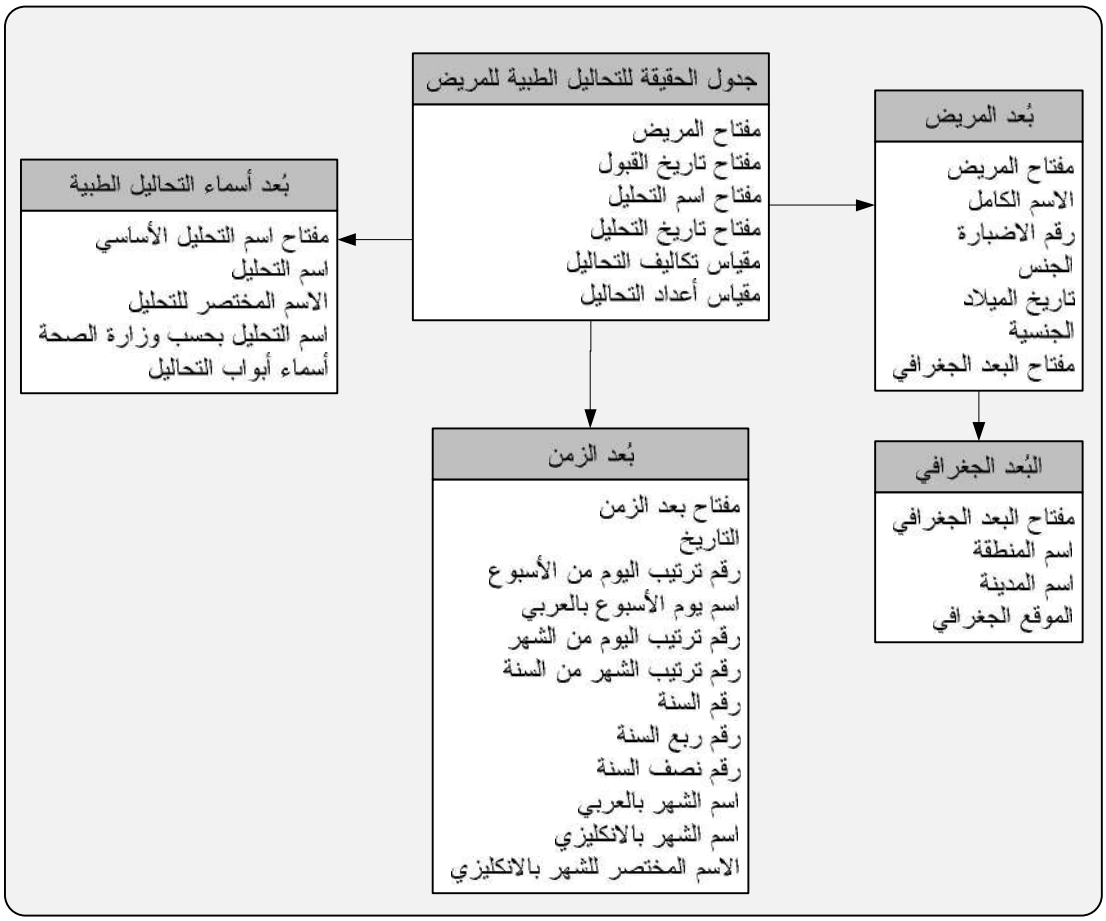
د- **بُعد أسماء التحاليل الطبية**

- مفتاح اسم التحليل الأساسي

- اسم التحليل الطبي
- الاسم المختصر للتحليل الطبي
- اسم التحليل بحسب وزارة الصحة
- أسماء أبواب التحاليل

ه- جدول الحقيقة للتحاليل الطبية للمريض

- مفتاح المريض
- مفتاح تاريخ القبول
- مفتاح اسم التحليل الطبي
- مفتاح تاريخ إجراء التحليل الطبي
- مقياس أسعار التحاليل
- مقياس أعداد التحاليل



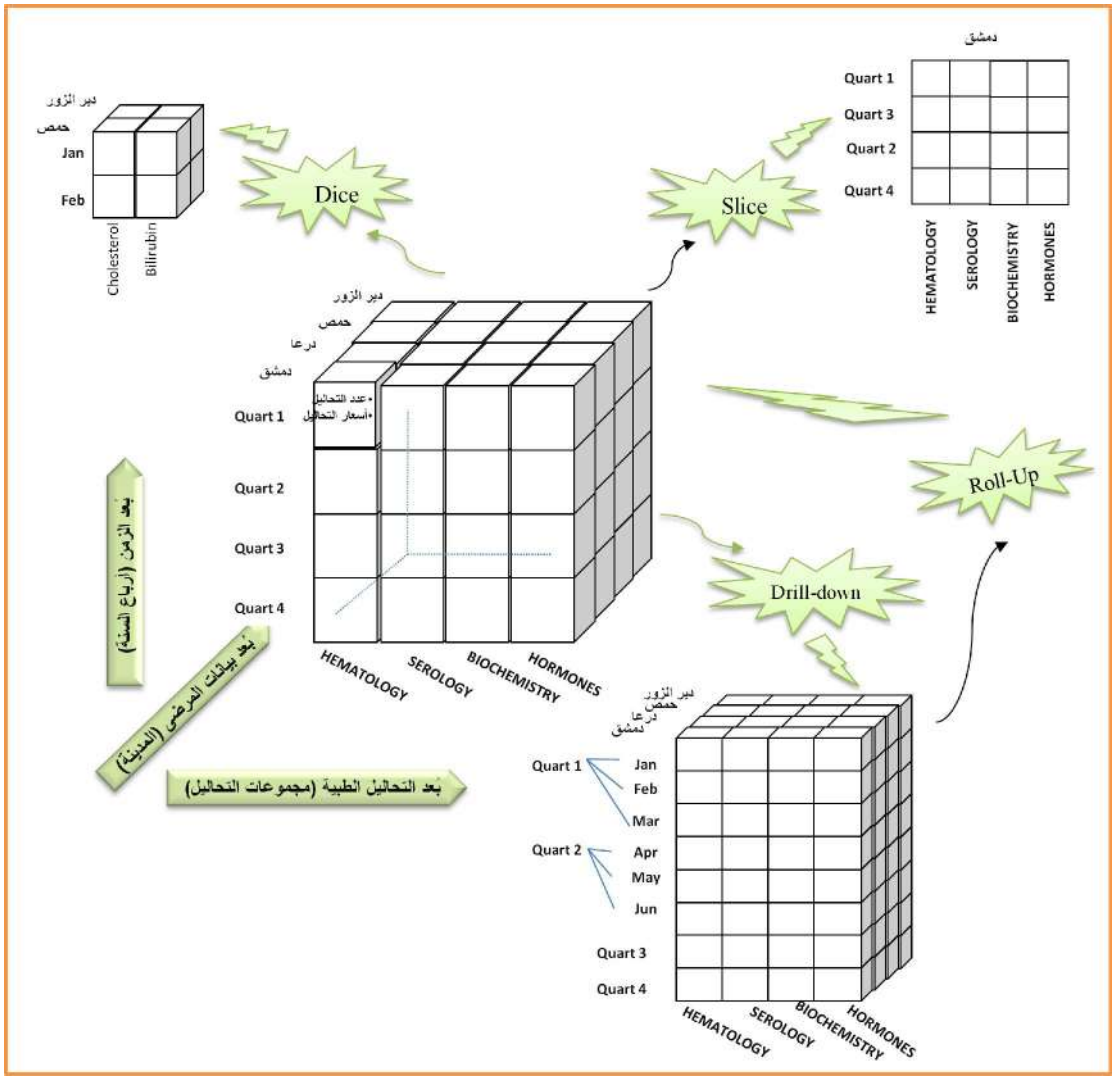
الشكل (5-3- 17): النموذج الأولي لمكعب البيانات المخصص لاستعراض تكاليف وأعداد التحاليل الطبية بحسب الأبعاد المختلفة

يحتاج تحقق أو تنفيذ مكعب البيانات السابق إلى ربط جدول بُعد المريض مع جدول الحقيقة للتحاليل الطبية التي تُجرى له بعلاقة من النوع (1-∞)، ولكن لم تقدِّم البيانات المتاحة للدراسة حقول الربط الدقيقة بين هذين الجدولين بسبب الانفصال التام لقاعدة بيانات التحاليل الطبية عن قاعدة بيانات مكتب القبول، كما سبق وذُكر. وكان الحل الأنسب لبناء هذه العلاقة هو من خلال رقم الاضبارة، والذي تطلَّب التخلي عن جزء من سجلات المرضى الذين تكرر دخولهم إلى المشفى، فأصبح عدد السجلات المتبقية في بيانات القبول (43063) سجل، بعد أن كان عددهم (47731) سجل، وقد تم الإبقاء على السجل الأخير لكل مريض تكرر دخوله إلى المشفى وحذفت بقية سجلاته. وقادت عملية الحذف هذه إلى حصد الكثير من الايجابيات والفوائد لعمليتي التحليل والتنقيب والتي لم تكن لتتحقق لولاها. ومن ثم أُعطيت بيانات المرضى الذين تكرر دخولهم إلى المشفى اهتماماً خاصاً بمعزل عن بقية البيانات، وذلك في المبحث السادس من هذا الفصل.

يبين الشكل (5-3-18) بعض اقتراحات التصفح متعدد الأبعاد على مكعب البيانات السابق، وقد تم بناء الجداول المناسبة لهذا المكعب في مستودع البيانات كما سيظهر في الفقرات القادمة، ثم جرت تجربة تطبيق هذا النموذج المصغر لمكعب البيانات واستعراضه عبر أدوات المعالجة التحليلية الآنية، مما ساعد في إصلاح عيوبه وتطويره إلى أن تم التوصل إلى بناء مكعب بيانات أوسع وأفضل منه وسيظهر لاحقاً بعد التعرف على خطوات بناء مستودع البيانات.

تختلف بنى الجداول في مستودع البيانات عنها في قاعدة البيانات العلاقاتية (راجع الفصل الأول)، وقد تم التوصل إلى الجداول المناسبة للمستودع من خلال بناء عدد من الاستعلامات المتتالية والمتراكبة في قاعدة البيانات في أكسس[1].

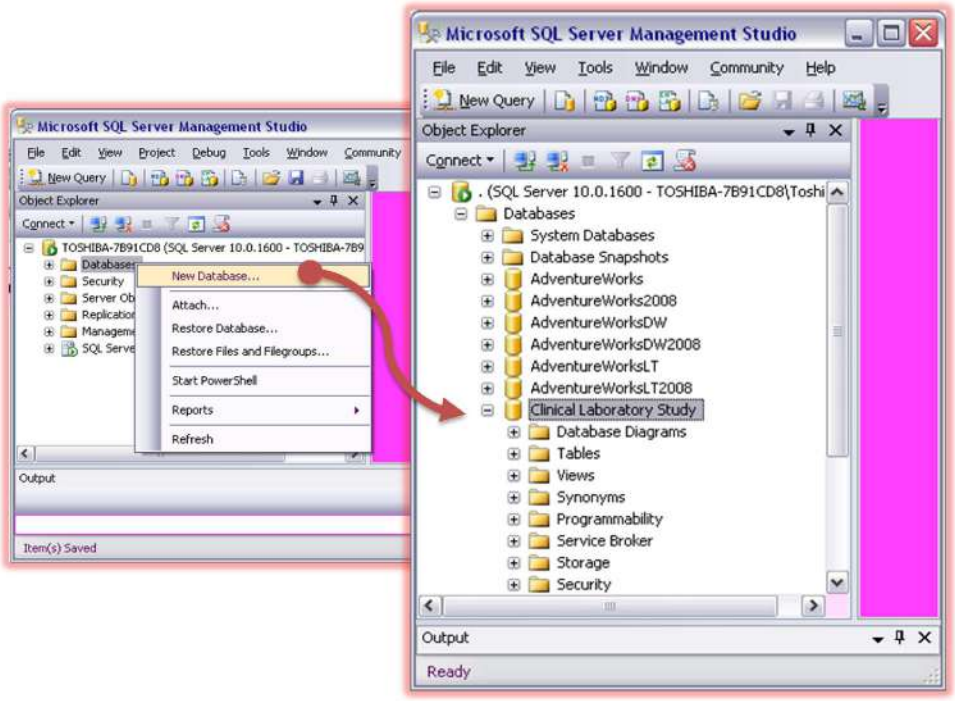
[1] لن يتم الخوض في تفاصيل هذه الاستعلامات تجنباً للإطالة في هذا البحث وخوفاً من ضياع التركيز على النقاط الأساسية فيه.



الشكل (5–3– 18): التصميم الأولي لمكعب البيانات الخاص بالتحاليل الطبية، وبعض عمليات التقطيع والاستعلام من هذا المكعب

2•3•3•5 إنشاء بنية مستودع البيانات الجديد في SQL Server

يحتاج بناء مستودع البيانات أولاً إلى إنشاء البنية الأساسية التي ستنتقل إليها البيانات، وقد تم إنشاء هذه البنية من خلال استدعاء إجراء إنشاء قاعدة (مستودع) بيانات جديدة فارغة في برنامج (SQL Server Management Studio)، وأطلق اسم (Clinical Laboratory Study) على هذا المستودع الخاص ببيانات مخبر التحاليل الطبية، ويحتوي على الهياكل العامة لبناء جميع المكونات الضرورية فيه كالجداول والمرائي والدوال والمخططات البيانية، انظر الشكل (5–3–19).

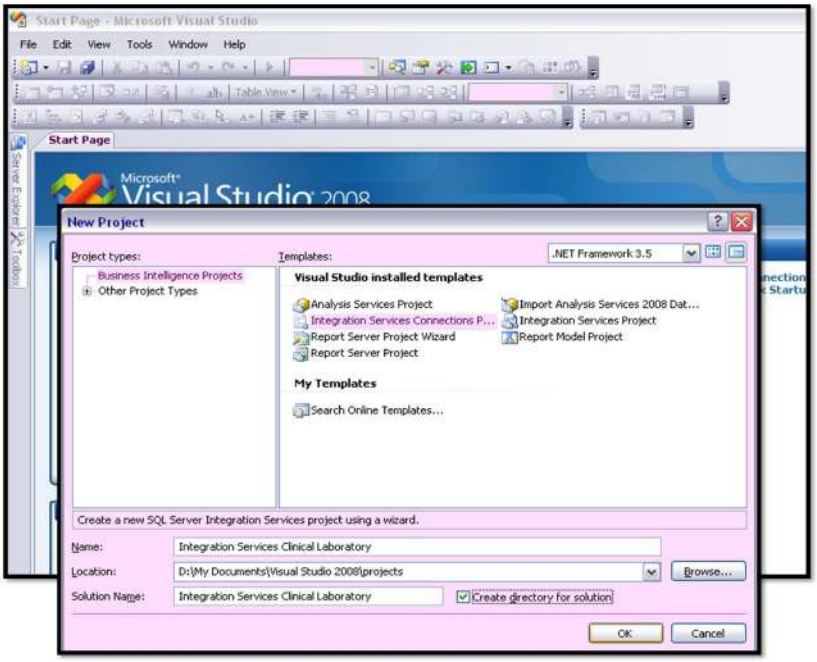


الشكل (5-3- 19): إنشاء بنية مستودع بيانات مخبر التحاليل الطبية (Clinical Laboratory Study) في برنامج (SQL Server Management Studio)

3•3•3•5 إجراءات سحب البيانات من قاعدة البيانات إلى مستودع البيانات

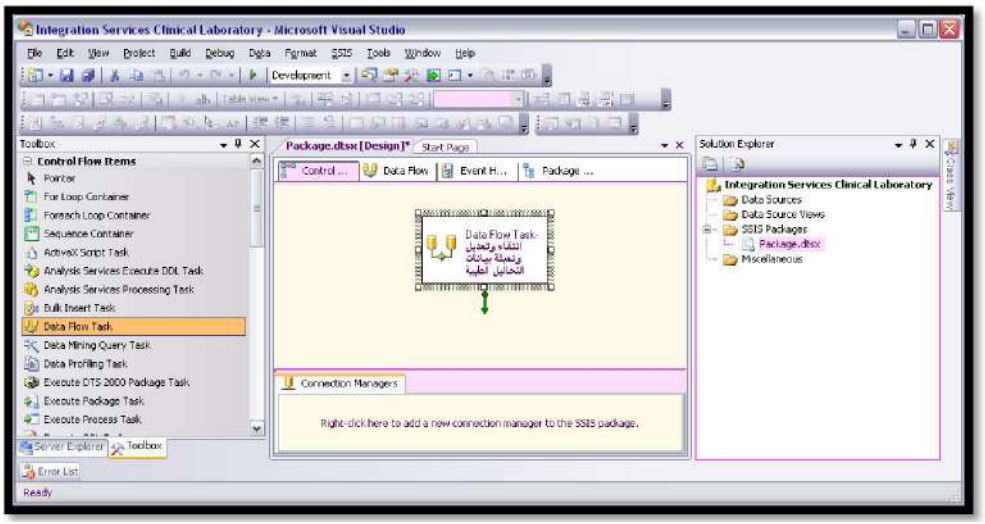
تتزود البنية الأساسية لمستودع بيانات التحاليل الطبية (Clinical Laboratory Study) بالبيانات من قاعدة البيانات النظيفة الموجودة في أكسس، وقد استُعين في هذه الدراسة ببرنامج (Microsoft Visual Studio) لأداء هذه العملية، وهو أحد الأدوات البرمجية التي تنتمي إلى عائلة (Microsoft SQL Server)، والذي سيُستخدم أيضاً في بناء مكعب البيانات وفي عمليات التنقيب في البيانات. ويُعتبر مستودع البيانات المراد إنشاؤه في هذه الدراسة بسيطاً حيث أنه ناتج عن دمج ثلاث قواعد بيانات متجانسة البنية ومبنية كلها في أكسس وحجمها النسبي صغير. ونبين فيما يلي الخطوات التي اتُبعت لتزويد مستودع البيانات الجديد بالجداول المناسبة.

أ- تم في البداية إنشاء مشروع جديد في هذا البرنامج لتلبية خدمات تكامل البيانات (Integration Services)، باسم (Clinical Laboratory-Integration Services)، حيث تساعد هذه الخدمات في التعامل مع البيانات المراد نقلها بين الأنظمة المختلفة لتخزين البيانات، انظر الشكل (5-3-20).



الشكل (– –): إنشاء مشروع لنقل البيانات من قاعدة البيانات إلى مستودع البيانات
باستخدام خدمات التكامل في برنامج ( )

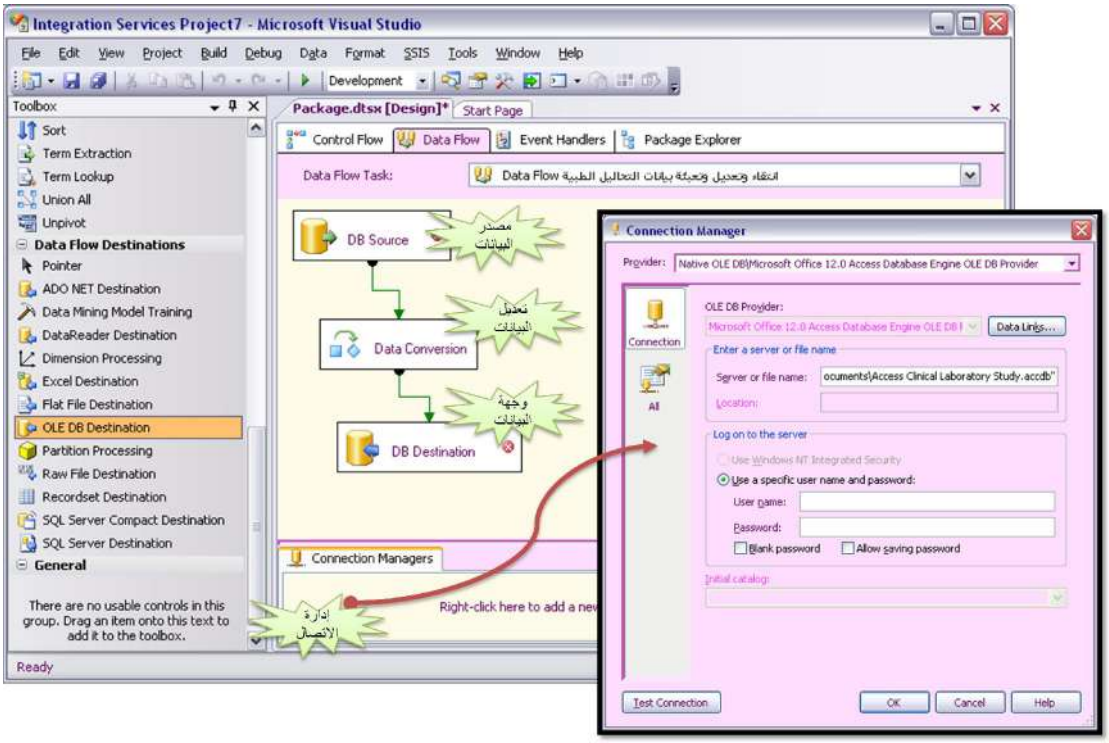
ب- إدراج مهمة لتدفق البيانات من قائمة ( ) وأطلق عليها تسمية مناسبة للمهمة الموكلة إليها، وهي (انتقاء وتعديل وتعبئة بيانات التحاليل الطبية) كما هو معروض في الشكل (– –).



الشكل (5-3- 21): إدراج مهمة لتدفق البيانات في برنامج (Microsoft Visual Studio)

ج- تعريف مراحل تنفيذ هذه المهمة، والتي تتألف من ثلاثة مراحل: الأولى جلب البيانات من المصدر (DB Source)، والثانية تعديل أنماط البيانات عند الضرورة (Data Conversion)، والثالثة تعبئة البيانات في مكان قاعدة البيانات المقصودة (DB Destination).

د- تحديد الاتصال بمصدر الحصول على البيانات وهو قاعدة البيانات الشاملة التي أنشئت سابقاً في أكسس (Access Clinical Laboratory Study.accdb)، والاتصال بوجهة ذهاب البيانات إلى مستودع البيانات المدروس (Clinical Laboratory Study)، كما في الشكل (5-3-22).



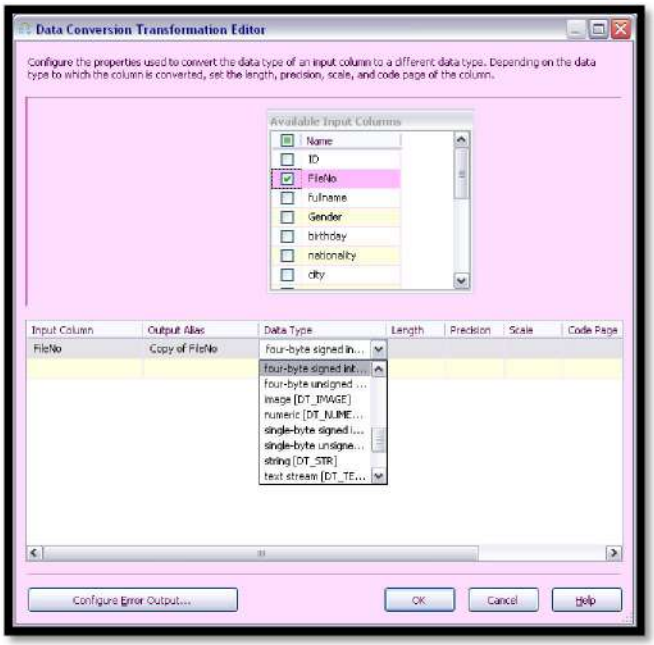
الشكل (5-3- 22): تعريف الاتصال بمصدر البيانات المرغوب نقلها وتعريف مراحل التنفيذ باستخدام خدمات التكامل في برنامج (Visual Studio)

ه- تحديد الجدول المراد نقله باستخدام إجراء (DB Source)، حيث تستطيع هذه العملية أن تنقل جدولاً واحداً في كل مرة. ثم تُجرى التعديلات على أنماط بعض البيانات بما يتناسب مع عمليات الربط بين الجداول ومع ما يناسب عمليات المعالجة التحليلية الآنية وذلك باستخدام إجراء (Data Conversion) كما في الشكل (5-3-23)، وقد احتاجت العديد من البيانات إلى تعديل في نمطها (من float إلى Integer)[1] بسبب التصميمات الضعيفة لقواعد البيانات الأصلية المتاحة للدراسة، والتي قام مصمموها في أحيانٍ كثيرةٍ بالإبقاء على الأنماط الأولية للبيانات التي يقترحها برنامج تصميم قواعد البيانات أكسس دون مراعاة الطبيعة المختلفة للبيانات، ودون معرفة أن هذه القواعد ستُستخدم يوماً ما لأغراض جديدة مثل التنقيب في البيانات. وأخيراً يتم تعريف الجدول الهدف في مستودع البيانات ضمن

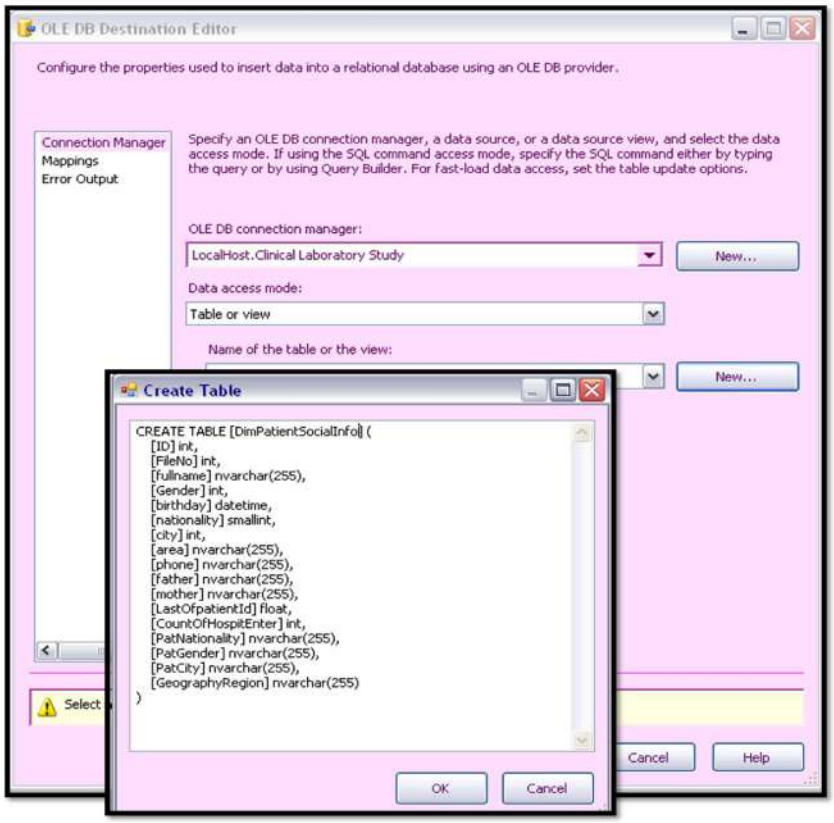
[1] ازدادت الحاجة لاستخدام SQL Server في تغير أنماط البيانات خصوصاً في حالة الجداول الكبيرة التي يفوق عدد سجلاتها الـ 65000 سجل، حيث لا تسمح الإمكانيات الموجودة في برنامج أكسس بإجراء مثل هذه التعديلات.

إجراء (DB Destination) كما في الشكل (5-3-24).

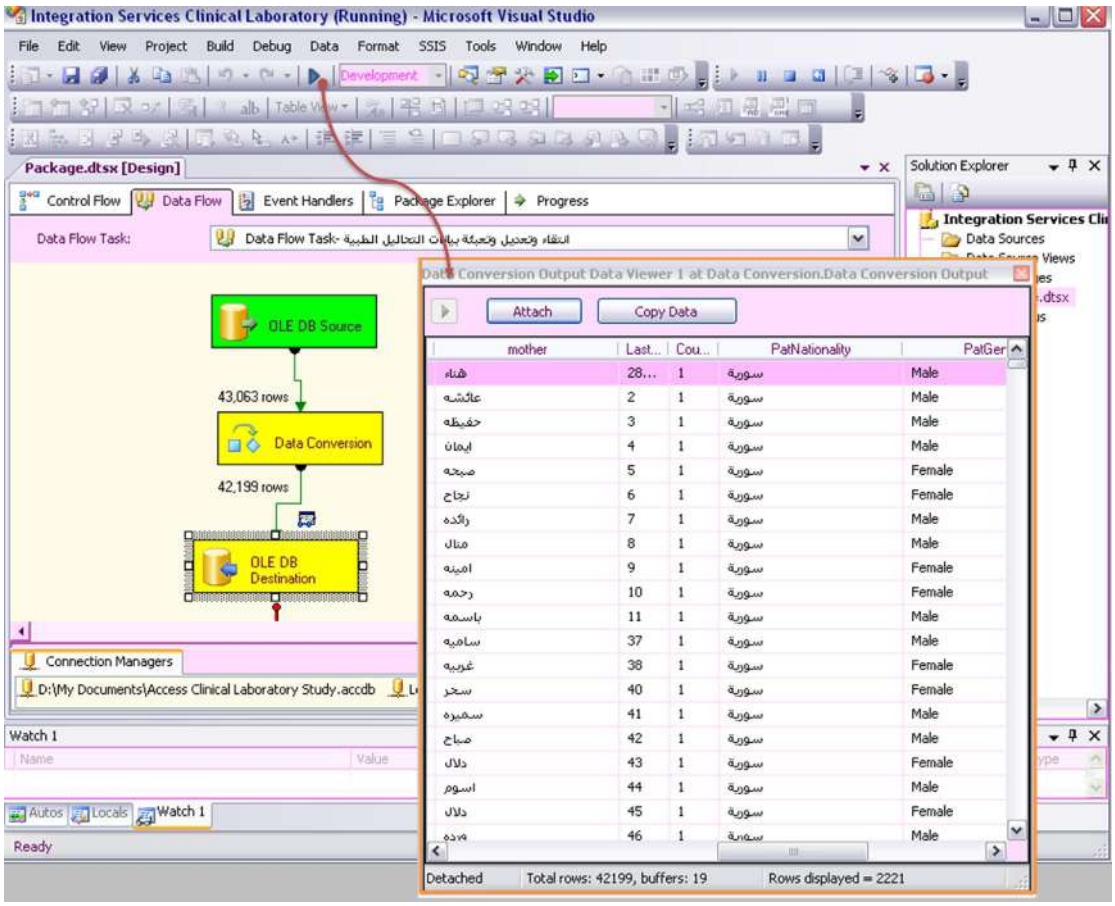
و- يُستدعي بعد ذلك إجراء فحص وتنفيذ عملية النقل (Start Debugging)، كما في الشكل (5-3-25).



الشكل (5- -): عملية تغيير أنماط بعض البيانات المنقولة من خلال إجراء تحويل أنماط البيانات ((Data



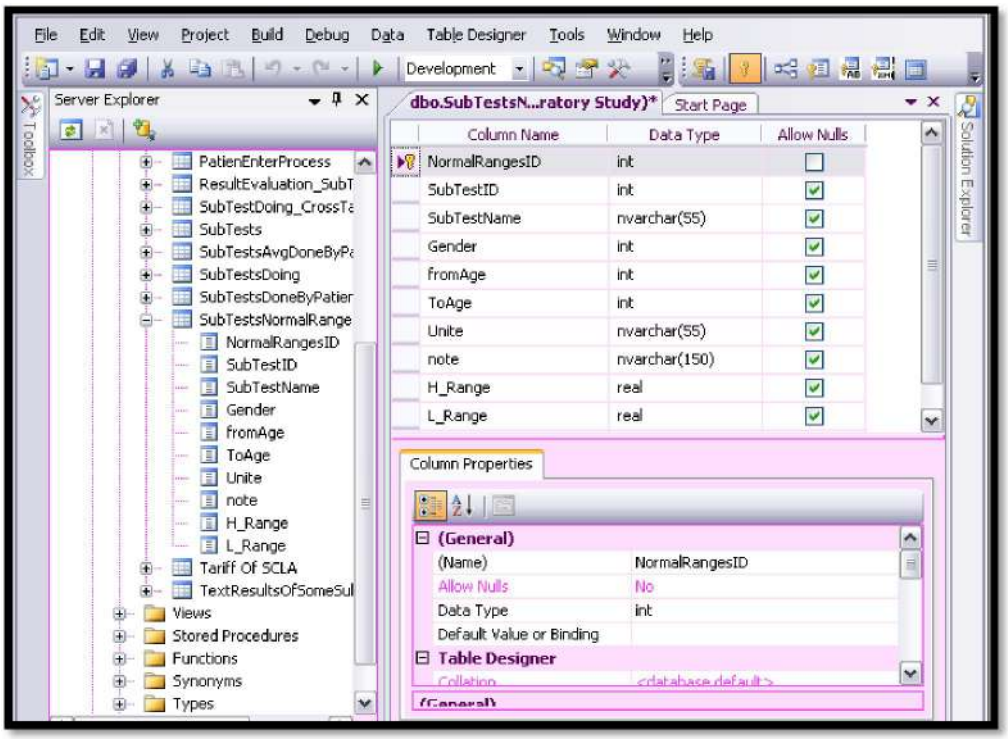
الشكل (5-3- 24): تعريف اسم جدول البيانات الجديد والحقول التي سيتكون منها، وذلك ضمن إجراء مكان قاعدة البيانات المقصودة (DB Destination)



الشكل (5–3– 25): مشهد تم التقاطه أثناء أداء عملية نقل سجلات البيانات من أكسس إلى SQL Server، وتستغرق هذه العملية مدداً زمنيةً متفاوتةً بحسب حجم الجدول وبحسب الوقت الذي يحتاجه تعديل أنماط البيانات

ز- وبعد الــتأكد من إتمام عملية النقل بنجاح يجب تحديث مستودع البيانات (Refresh) الذي يمكن الوصول إليه من نافذة (Server Explorer) في هذا البرنامج، ومن ثم يجري تعريف المفتاح الأولي للجدول الذي نُقل، كما يمكن إجراء بعض التعديلات على خصائص حقول البيانات (Column Properties) بحسب الضرورة. انظر الشكل (5–3–26).

ح- وقد جرى تكرار الخطوات الثلاثة الأخيرة (ه، و، ز) عدة مرات إلى حين تم الانتهاء من نقل جميع الجداول المراد نقلها من قاعدة البيانات إلى مستودع البيانات.

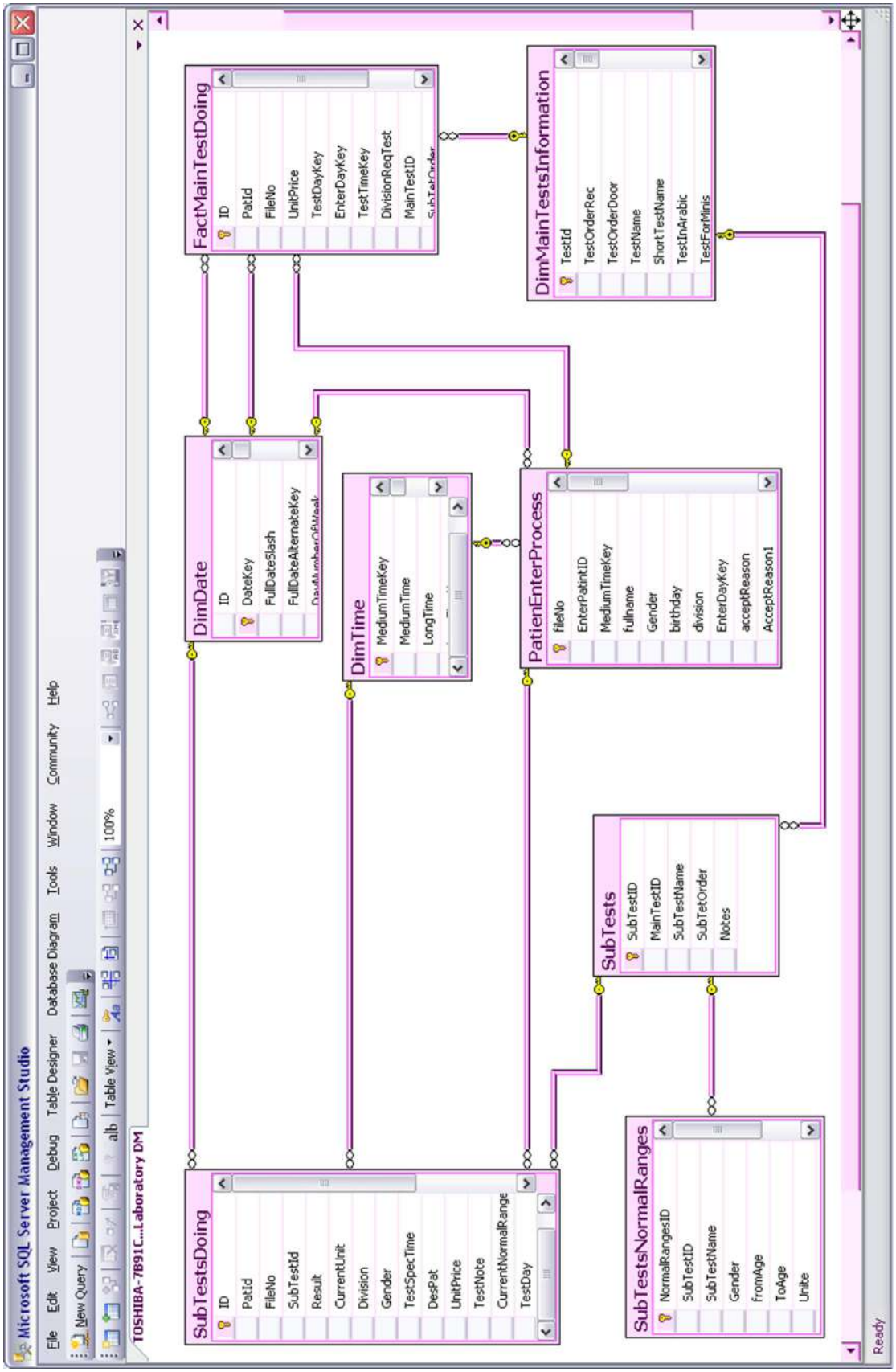


الشكل (5-3- 26): تعريف المفتاح الأولي للجدول، وتعديل خصائص الحقول بحسب الضرورة

4•3•3•5 تعريف العلاقات بين الجداول في مستودع البيانات

أُطلقت تسميات جديدة على عدد من الحقول والجداول، فالبعض منها أصبح يبدأ بكلمة (Dim) لأنه صمم ليكون أحد أبعاد مكعب البيانات لاحقاً مثل البعد الزمني (DimDate)، وجرى بناء جدول حقيقة أساسي يبدأ اسمه بكلمة (Fact) وهو جدول التحاليل الأساسية التي أجريت للمرضى (FactMainTestDoing)، وتم دمج كافة بيانات المرضى الشخصية والطبية في جدول واحد هو (PatientEnterProcess). كما كان هنالك عددٌ من الإضافات المختلفة التي أغنت هذا المستودع، والتي ستظهر آثارها عند استخدامها في عمليات التحليل والتنقيب.

وتم تعريف العلاقات بين أهم الجداول التي نقلت إلى مستودع البيانات من خلال استدعاء إجراء لتشكيل رسم بياني جديد (Diagram) في برنامج (Management Studio)، حيث أضيفت إليه الجداول المرغوب عرضها، ثم جرى تعريف علاقات الارتباط المناسبة بينها. وبهذا تم التوصل إلى التصميم الأساسي لمستودع بيانات مخبر التحاليل الطبية والمعروض في الشكل (5-3-27).



الشكل (5-3- 27): مخطط العلاقات الترابطية بين الجداول الأساسية في مستودع بيانات التحاليل الطبية

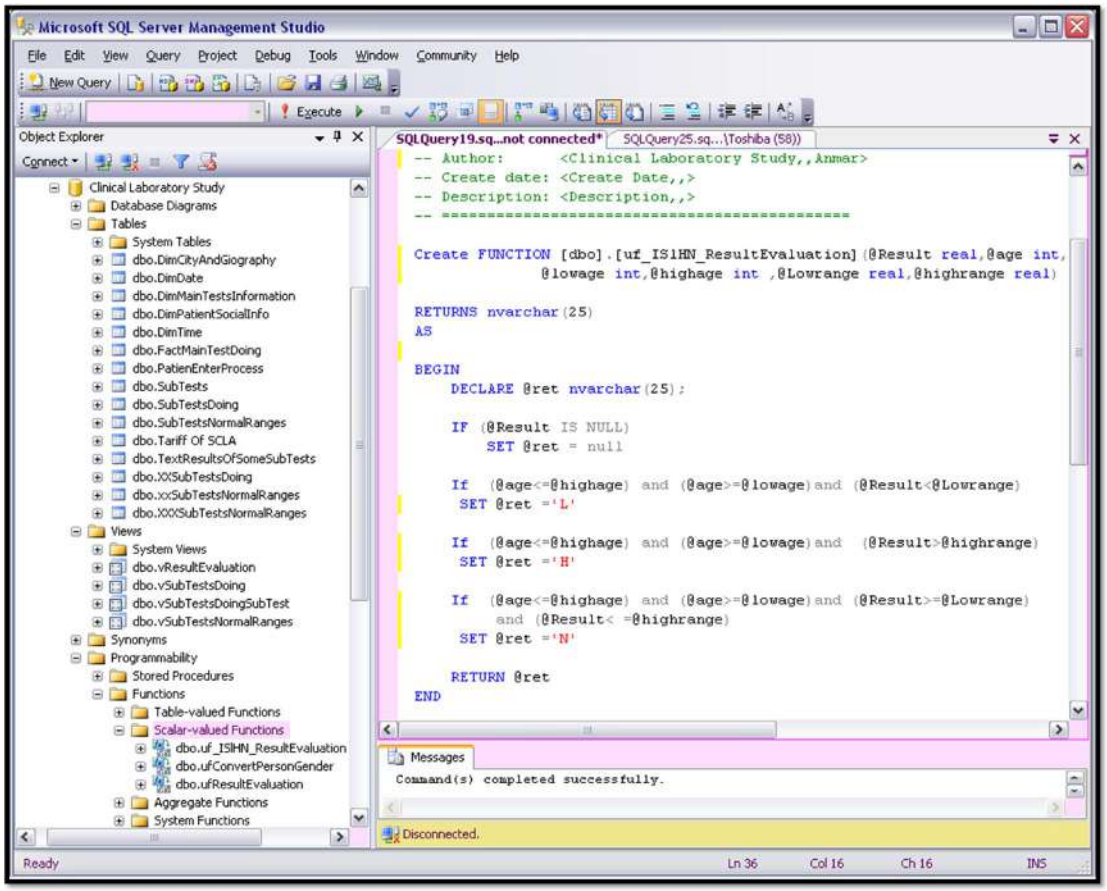
5•3•3•5 تطوير مستودع البيانات (Clinical Laboratory Study)

جرى خلال مراحل اكتشاف المعرفة تطوير مستودع البيانات عدة مرات لتلبية الاحتياجات المتوالية التي كانت تظهر أثناء استخدام البيانات في عمليات التحليل والتنقيب، وقد صُمم عدد من التوابع (Function) للقيام بالعمليات الحسابية والحصول على متغيرات جديدة، كما أنشئ عدد من المرائي (Views) عن البيانات، ونبين فيما يلي أهم هذه التطويرات:

أ- إنشاء تابع ومرأى لتقييم نتائج التحاليل الطبية

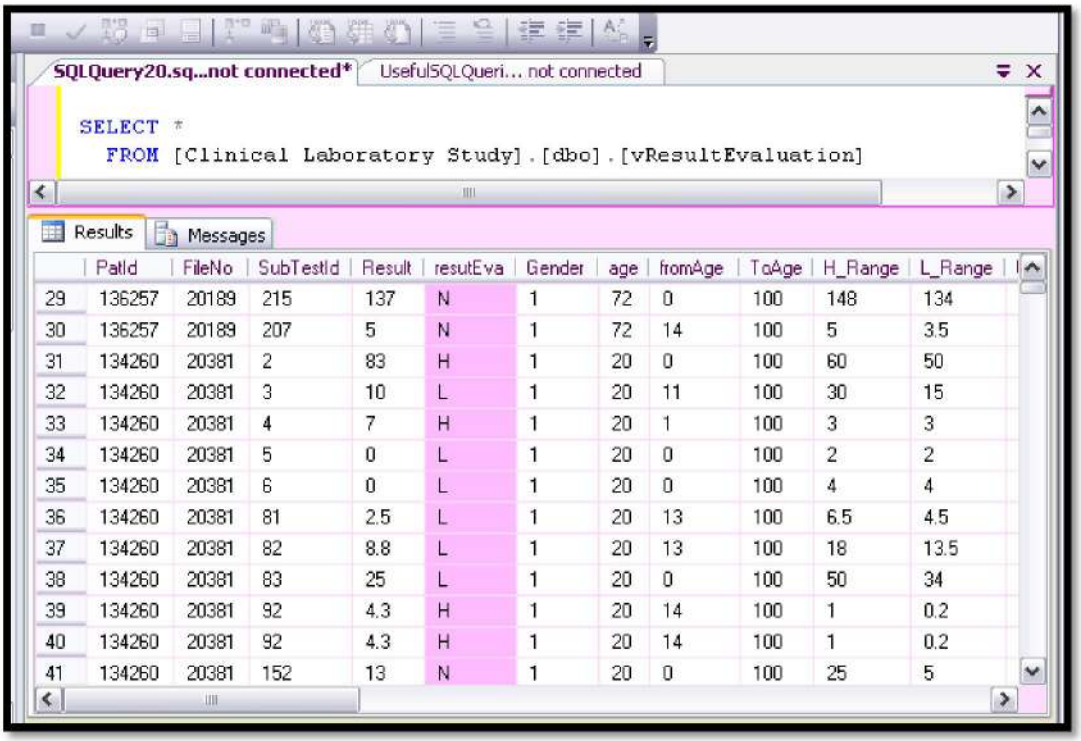
يُعتبر من أهم ما تم انجازه في خدمة عمليات التنقيب في هذا البحث هو إنشاء تابع أُطلق عليه اسم (uf_IsLHN_ResultEvaluation) لتقييم نتائج التحاليل الطبية، وإرجاع نتائج التحاليل الطبية إلى الرموز (N, H, L)، حيث يقوم هذا التابع باختبار مكان وقوع نتيجة التحليل الطبي بالنسبة للحدين الأدنى والأعلى للقيم الطبيعية التي تختلف بحسب العمر والجنس واسم التحليل الطبي، ومن ثم يُسجل النتيجة في حقل خاص يأخذ إحدى ثلاثة رموز هي ("N" وتعني أن نتيجة التحليل طبيعية وتقع ضمن الحدين الأدنى والأعلى للقيم الطبيعية لكل من التحاليل الطبية وبحسب جنس وعمر المريض، و"L" وتعني أن نتيجة التحليل أقل من الحد الأدنى الطبيعي، و"H" وتعني أن نتيجة التحليل أعلى من الحد الأعلى الطبيعي). وبهذا أُمكن التوصل إلى طريقة موحدة لقياس قيم نتائج التحاليل الطبية باختلاف أنواعها. وأصبح بالإمكان التعرف على جميع التحاليل التي أظهرت نتائجها قيماً طبيعية، أو قيماً غير طبيعية دالةً على وجود حالة مرضية معينة عند المريض، وذلك بغض النظر عن اسم التحليل أو طبيعة نتائجه، لاحظ الشكل (5-3-28).

تم بعد ذلك إنشاء مرأى أول (View) يجمع بيانات نتائج التحاليل الطبية مع ما يقابلها من بيانات شخصية عن عمر وجنس المريض، ثم أنشئ مرأى آخر (vResultEvaluation) يأخذ بياناته من المرأى الأول، ومن جدول القيم الطبيعية للتحاليل الفرعية، ويتضمن حقل جديد لتقييم نتائج التحاليل الطبية والذي يستمد قيمه عن طريق التعليمات الموجودة في تابع تقييم النتائج (uf_IsLHN_ResultEvaluation)، انظر الشكل (5-3-29). وقد تطور تصميم هذا المرأى عدة مرات واشتُقت منه مرائي أخرى بحسب الاحتياجات التي كانت تتطلبها عمليات التنقيب في البيانات أحياناً.



الشكل (– –): إنشاء دالة لتقييم نتائج التحاليل الطبية

(: النتيجة تقع ضمن الحدود الطبيعية، : النتيجة أقل من الحد الأدنى الطبيعي، : النتيجة أعلى من الحد الأعلى الطبيعي)

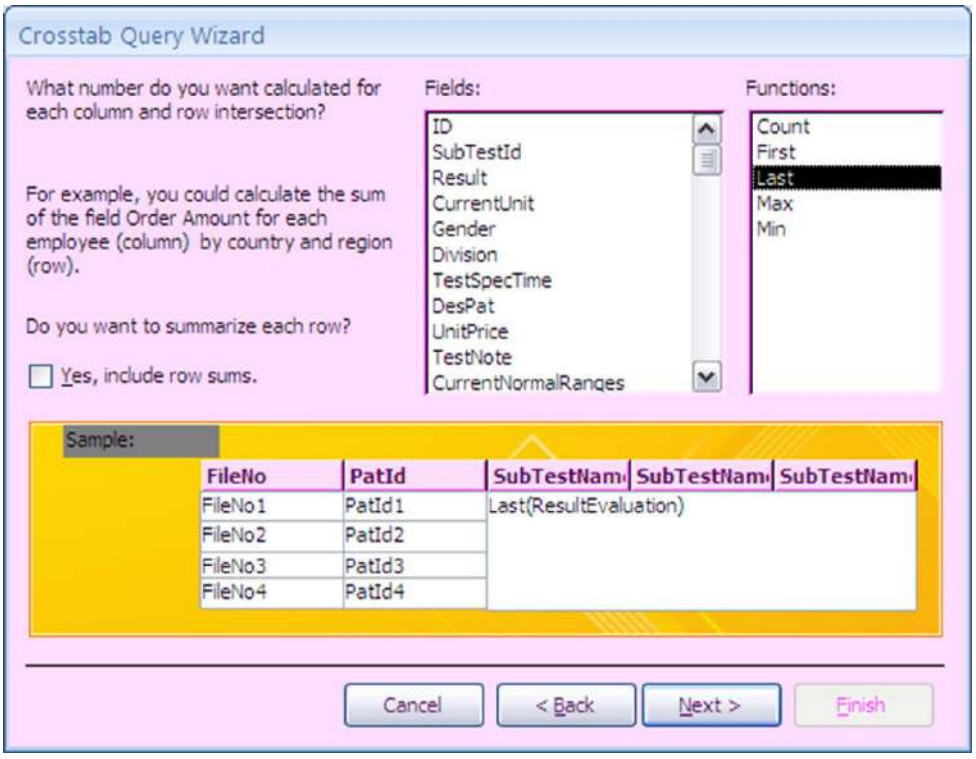


	PatId	FileNo	SubTestId	Result	resutEva	Gender	age	fromAge	ToAge	H_Range	L_Range
29	136257	20189	215	137	N	1	72	0	100	148	134
30	136257	20189	207	5	N	1	72	14	100	5	3.5
31	134260	20381	2	83	H	1	20	0	100	60	50
32	134260	20381	3	10	L	1	20	11	100	30	15
33	134260	20381	4	7	H	1	20	1	100	3	3
34	134260	20381	5	0	L	1	20	0	100	2	2
35	134260	20381	6	0	L	1	20	0	100	4	4
36	134260	20381	81	2.5	L	1	20	13	100	6.5	4.5
37	134260	20381	82	8.8	L	1	20	13	100	18	13.5
38	134260	20381	83	25	L	1	20	0	100	50	34
39	134260	20381	92	4.3	H	1	20	14	100	1	0.2
40	134260	20381	92	4.3	H	1	20	14	100	1	0.2
41	134260	20381	152	13	N	1	20	0	100	25	5

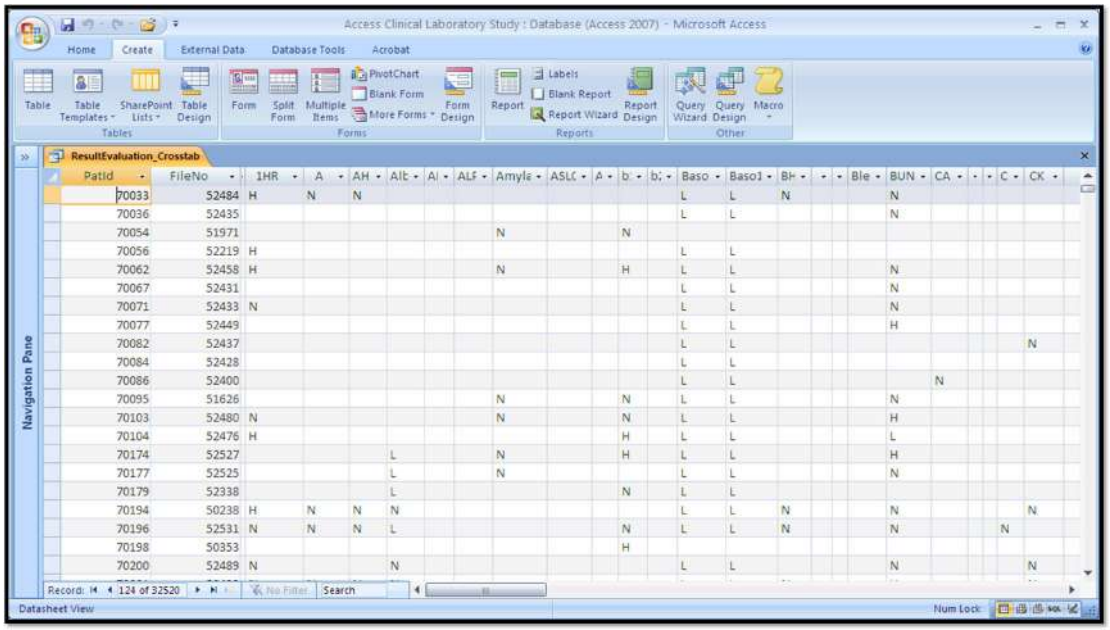
الشكل (5-3- 29): مرأى (vResultEvaluation) الخاص بتقييم نتائج التحاليل الطبية

ب- تحويل أسطر نتائج التحاليل الطبية إلى أعمدة

ظهرت الحاجة أثناء عمليات التنقيب في البيانات إلى مقارنة نتائج التحاليل الطبية مع بعضها، وذلك للتعرف مثلاً على التأثير المتبادل لنتائج كل من التحاليل الطبية مع بعضها ومع الخاصيات المختلفة للمرضى.... والعديد من الاحتياجات الأخرى التي ستظهر في الأماكن المخصصة لها لاحقاً. وقد وقف حقل نتائج التحاليل الطبية الوحيد كحجر عثرة أمام القيام بعمليات تنقيب فعالة في بيانات مخبر التحاليل الطبية، وازدادت الحاجة إلى تخصيص حقل لكل تحليل من التحاليل الطبية بدلاً من الحقل الذي يجمع بداخله كافة نتائج التحاليل الطبية. وقد تم التمكن من حل هذه القضية باستخدام أداة الاستعلام التصالبي في أكسس والتي طُبِّقت على محتويات جدول نتائج التحاليل الطبية الفرعية ذي النتائج المقيَّمة في الفقرة السابقة، كما في الشكل (5-3-30)، وأطلق على الجدول الناتج اسم (ResultEvaluation_Crosstab)، ويعتبر هذا الجدول من وجهة نظر برنامج أكسس بمثابة جدول تقارير، ويفتقد بالتالي إلى بنية الخلايا المرنة ولا يمتلك القدرة على التصرف كاستعلام أو كجدول عادي في أكسس. ولحل هذه المشكلة تم التحايل عليه وبُني استعلام جديد في أكسس يستمد بياناته من كامل بيانات هذا الجدول. وأُمكن بعد ذلك تحويله إلى جدول عادي كما هو معروض في الشكل (5-3-31).



الشكل (5-3- 30): استخدام الاستعلام التصالبي في تحويل أسماء التحاليل الطبية إلى خاصيات



الشكل (5–3– 31) الجدول التصالبي الذي تحولت فيه أسماء التحاليل الطبية إلى خاصيات، ويتضمن تقييم نتائج التحاليل الطبية ويتضمن (109 خاصيات) أي بعدد أسماء التحاليل الطبية

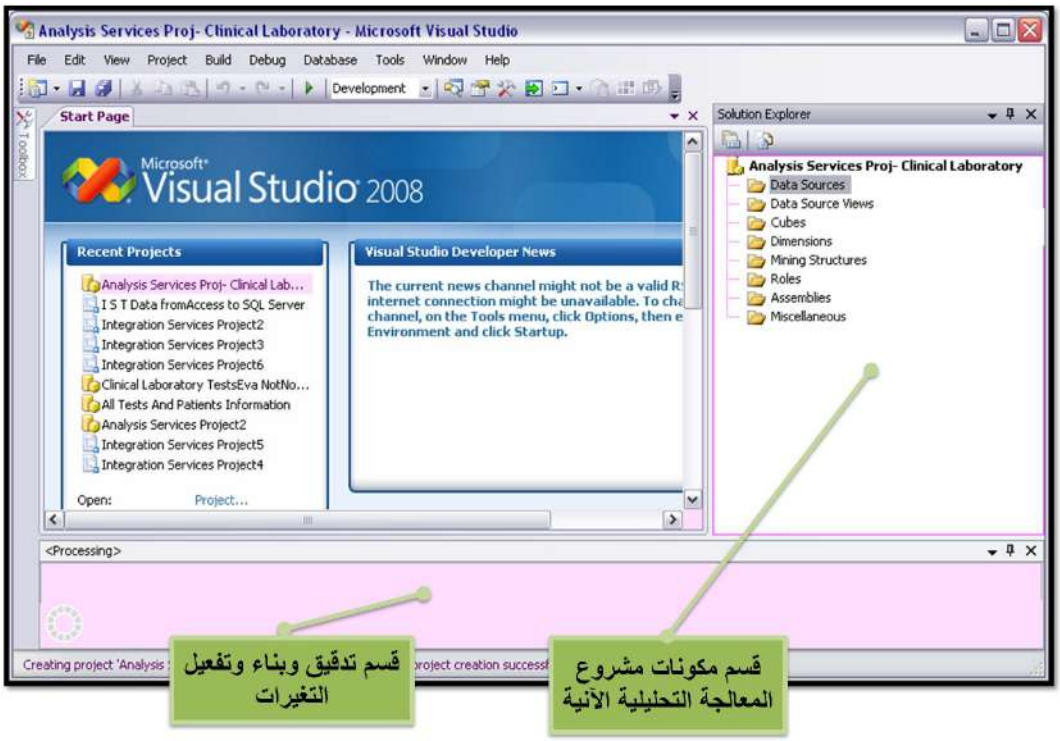
4•3•5 المعالجة التحليلية الآنية للبيانات (OLAP)

تَستمدُ تقنية المعالجة التحليلية الآنية في هذه الدراسة البياناتَ من مستودع بيانات التحاليل الطبية الذي تم تصميمه في الفقرات السابقة. ويوفر البرنامج الحاسوبي (Visual Studio) البيئة المناسبة لعمل هذه التقنية، وقد جرى في البداية تطبيق هذه التقنية على النموذج المصغر لمكعب البيانات (راجع الفقرة 5•3•3•1)، ثم تم التوصل إلى مكعب بيانات مؤلف من ثلاثة جداول للحقيقة هي جدول إجراء التحاليل الطبية الرئيسة وجدول إجراء التحاليل الطبية الفرعية وجدول بيانات قبول المرضى في المشفى، وخمسة جداول أبعاد هي بُعد الزمن وبُعد التوقيت[1]، وبُعد معلومات التحاليل الطبية الأساسية وبُعد معلومات التحاليل الطبية الفرعية، ويُعتبر جدول بيانات قبول المرضى في المشفى على أنه جدول حقيقة وجدول بُعد وذلك بحسب الأدوار التي يمكن أن يؤديها أثناء تطبيق عمليات المعالجة التحليلية الآنية، ونبين فيما يلي الخطوات التي أُتّبعت في هذا المجال.

[1] تم بناء جدول بُعد التوقيت بنفس طريقة بناء جدول بُعد الزمن.

1•4•3•5 إنشاء مشروع المعالجة التحليلية الآنية

تم تعريف مشروع جديد للمعالجة التحليلية الآنية باستخدام إجراء (New Project) في برنامج (Visual Studio)، وذلك ضمن مشاريع خدمات التحليل (Analysis Services Project)، وأطلق عليه اسم (Analysis Services Proj- Clinical Laboratory). وتَشكَّلت بذلك البنية التأسيسية لهذا المشروع الذي سيبنى فيه مكعب البيانات، وستُطبق فيه عمليات المعالجة التحليلية الآنية، انظر الشكل (5–3–32).

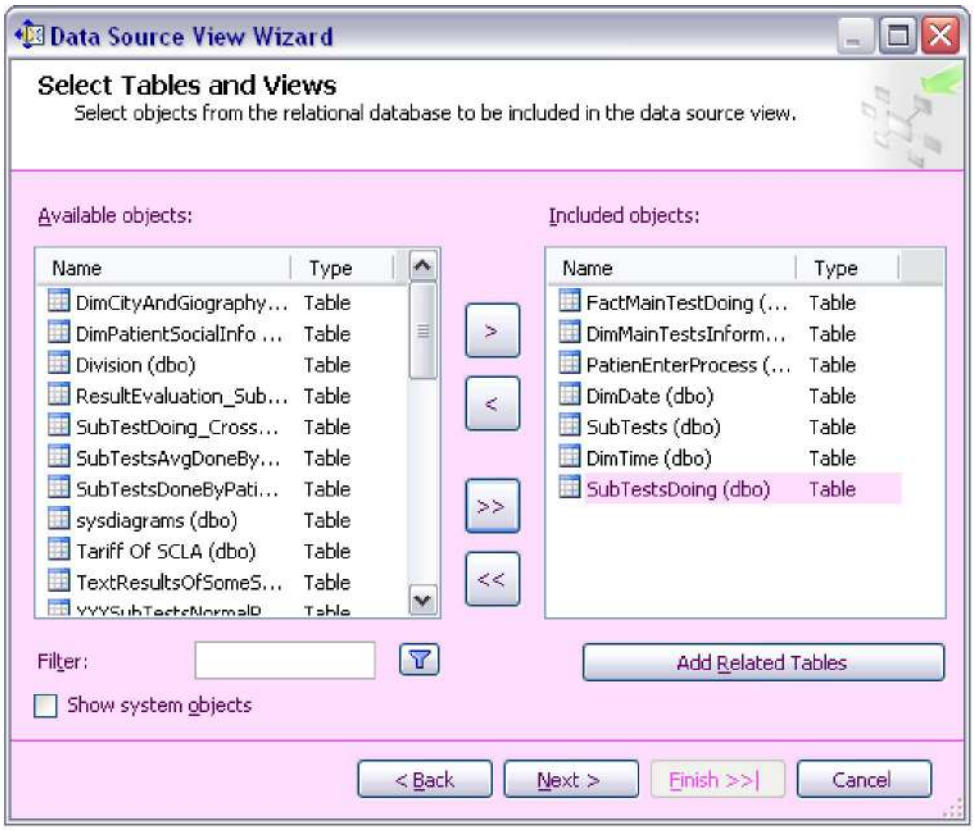


الشكل (5–3– 32): البنية التأسيسية لمشروع خدمات المعالجة التحليلية الآنية لبيانات مخابر التحاليل الطبية في برنامج (Microsoft Visual Studio)

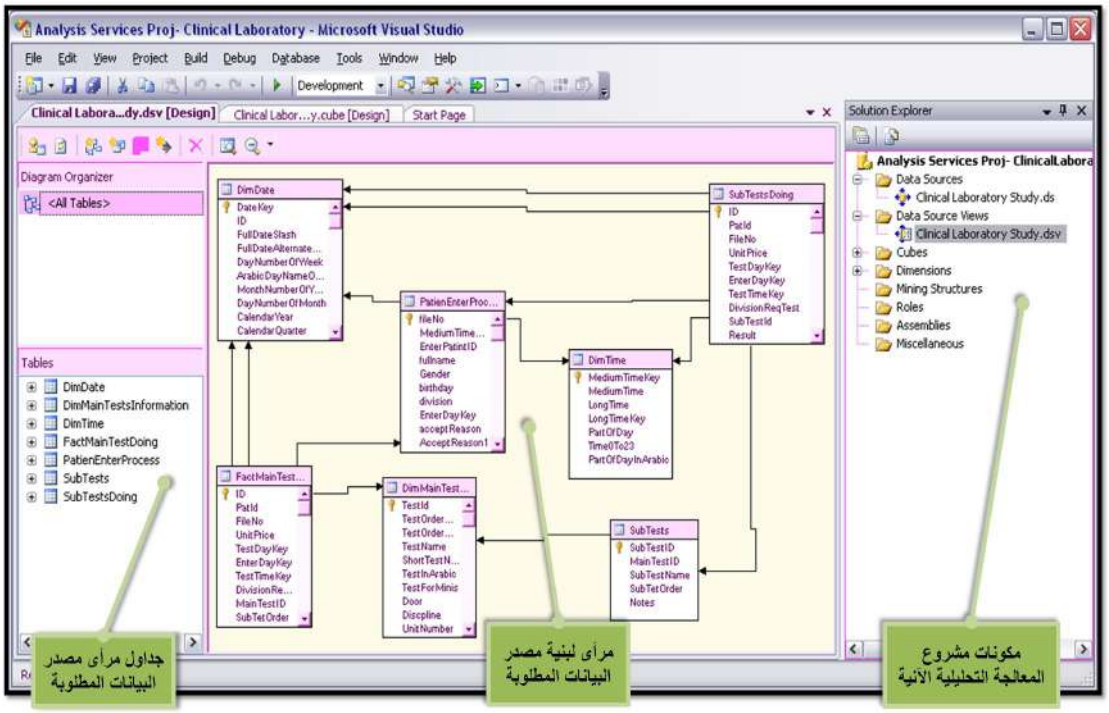
2•4•3•5 تعريف مصدر البيانات

جرى تعريف مستودع البيانات (Clinical Laboratory Study) على أنه مصدر الحصول على البيانات (Data Source) في قسم مكونات مشروع المعالجة التحليلية الآنية، ثم استدعي مرأىً عن بعض جداول هذا المصدر ضمن (Data Source views)، وهي من الجداول الأساسية الموجودة في مخطط العلاقات الترابطية السابق تمثيله في الشكل (5–3–27). ويُبين الشكل (5–3–33) نافذة انتقاء هذه الجداول، في حين يُظهر الشكل (5–3–34) مرأى الجداول التي

تم اختيارها والعلاقات الترابطية بينها، والتي تعتبر بمثابة المواد الأولية اللازمة لبناء مكعب البيانات.



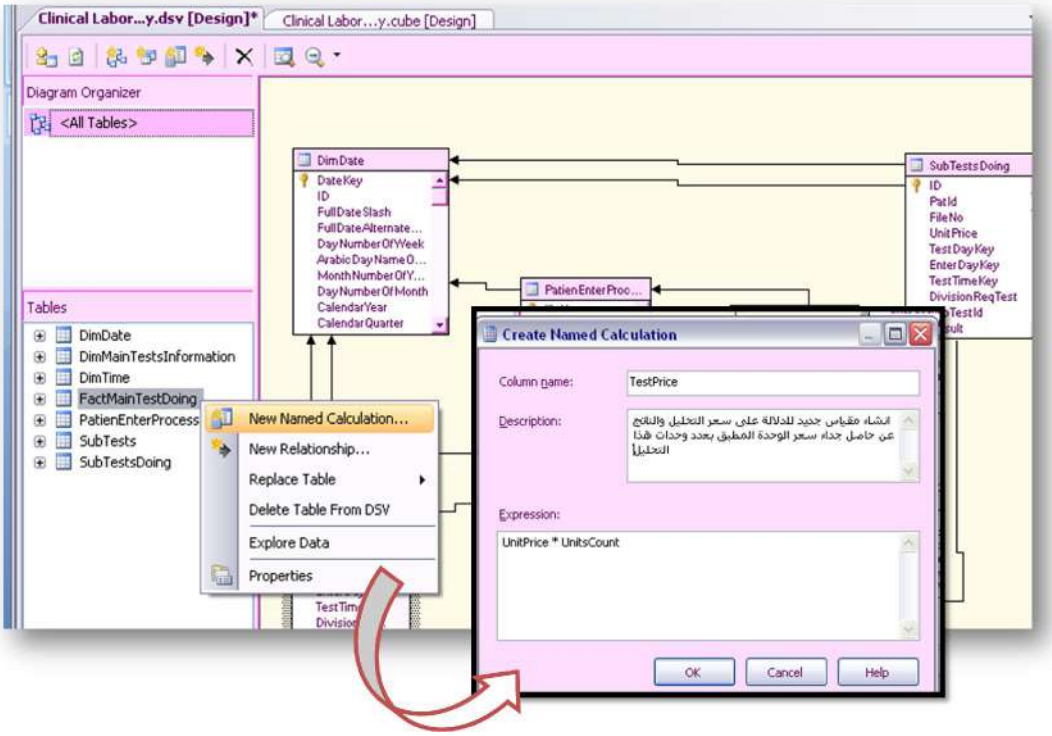
الشكل (– –): اختيار جداول مرأى مصدر البيانات في برنامج ( )



الشكل (5–3– 34): مرأى بنية مصدر البيانات في مشروع المعالجة التحليلية الآنية لبيانات مخبر التحاليل الطبية المدروس

3•4•3•5 تعريف مقاييس جديدة في مرأى مصدر البيانات

يوفر برنامج (Visual Studio) أدوات خاصة للحصول على مقياس جديد مستمد من المقاييس الموجودة، وذلك باستدعاء إجراء (Create Named Calculation) الذي يَظهر بالضغط على زر الفأرة الأيمن فوق اسم الجدول الذي يحتوي على المقاييس الموجودة، وكمثال على ذلك تم استدعاء هذا الإجراء على جدول الحقيقة (FactMainTestDoing) الموجود في قائمة جداول مرأى مصدر البيانات، وكُتبت العملية الحسابية المبينة في الشكل (5–3–35)، فنتج عنها مقياس جديد أطلقت عليه تسمة "سعر التحليل" (TestPrice) وهو يعبر عن المبلغ المترتب على المريض دفعه لقاء خدمة التحليل التي قدمت له في المشفى.

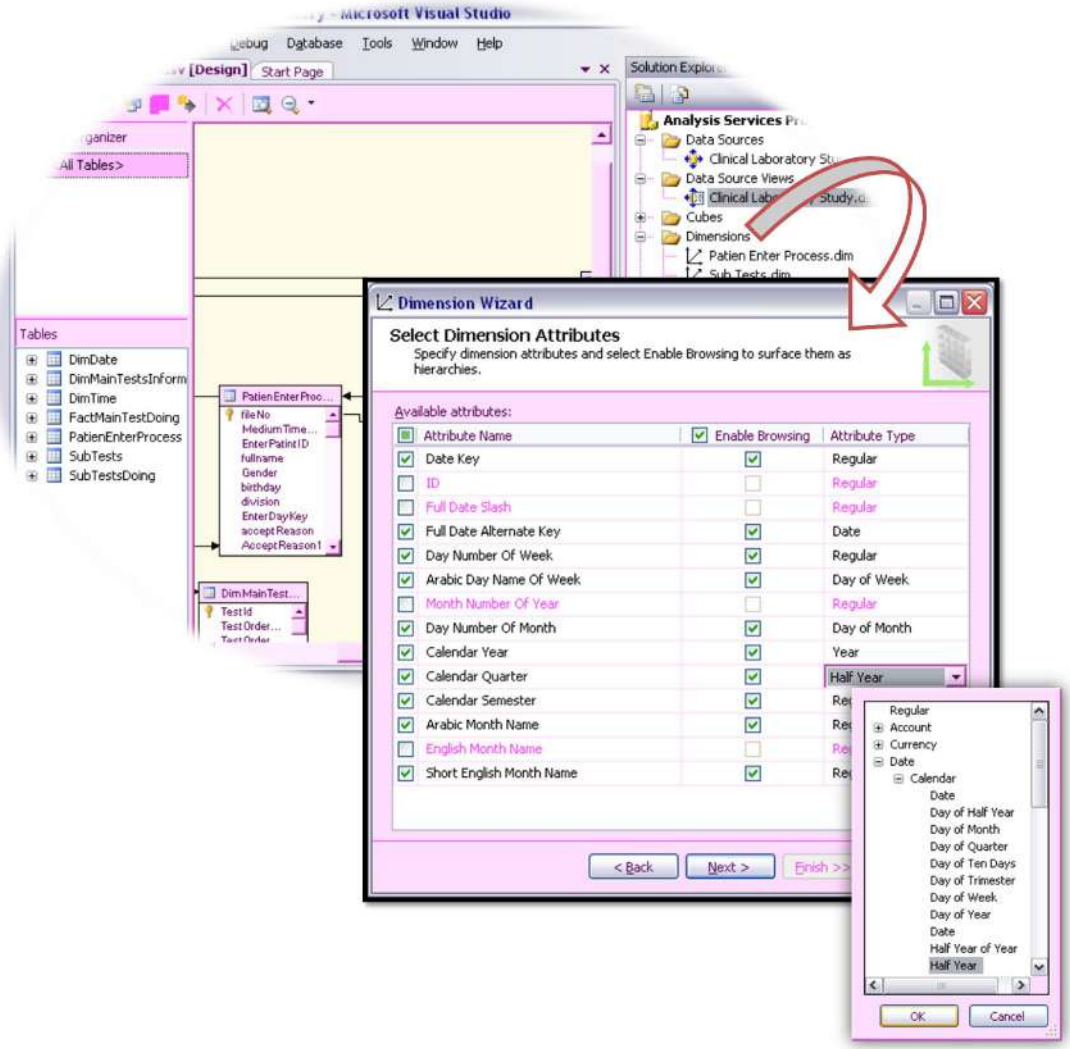


الشكل (5–3– 35): تعريف مقياس سعر التحليل وهو مقياس جديد مشتق من المقاييس الموجودة في جدول الحقيقة

4•4•3•5 تعريف أبعاد المكعب

يتكون كل بعد من أبعاد مكعب البيانات من عددٍ من الخاصيات (الحقول) التي يتم اختيارها من بين الخاصيات الموجودة في أحد جداول مرأى مصدر البيانات، ويتم تعريف مكونات

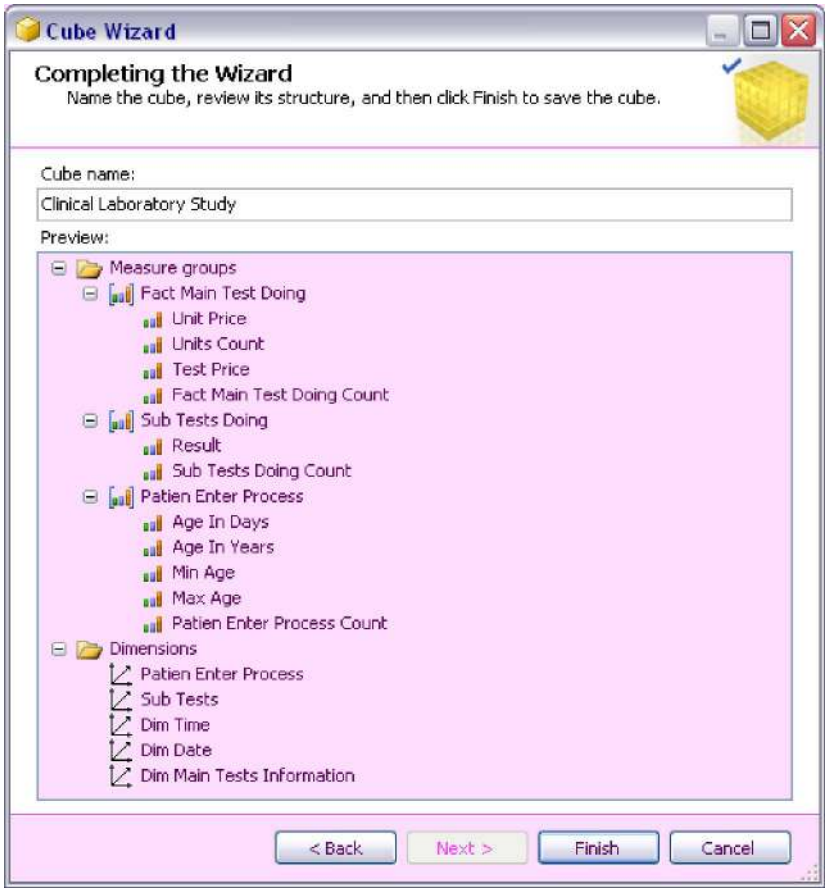
البعد من خلال إجراء إضافة بُعد في قسم مكونات مشروع المعالجة التحليلية الآنية، ويساعد هذا الإجراء أيضاً على تعريف أنماط بعض خاصيات البيانات التي يتألف منها هذا البُعد. فمثلاً يتألف البُعد الزمني من عدة خاصيات كالتاريخ وأسماء الأيام والسنوات وأسماء الأشهر وأرباع السنة، ولا تحمل هذه الخاصيات أي دلالات زمنية بالنسبة للمشروع الذي أُنشئ ما لم يحدد لها النمط المناسب من الأنماط المقترحة في البرنامج أثناء تعريف البعد الزمني، أو لاحقاً من خلال الوصول إلى قائمة خواص الحقول. وقد وُجِد أنه من الأنسب تحديد أنماط الخاصيات في هذا البعد في أثناء تعريفه، انظر الشكل (5-3-36). كما وتم تعريف أربعة أبعاد أخرى لاستخدامها في مكعب البيانات المرغوب.



الشكل (5-3- 36): اختيار الأنماط المناسبة للخاصيات في البعد الزمني أثناء إجراء تعريف هذا البعد

5•4•3•5 تعريف مكعب البيانات

تم في هذه الدراسة الاعتماد على فكرة تعريف كافة أبعاد مكعب البيانات قبل تعريف المكعب، وبالتالي سيجري في هذه الخطوة تعريف جداول الحقيقة (Fact Tables) والمقاييس (Measures) التي يحتويها كل منها، ومن ثم سيضاف إليها الأبعاد التي تحددت في الخطوة السابقة، وبهذا تكون كافة مكونات مكعب البيانات قد تحددت، ويُظهر الشكل (5-3-37) المرحلة الأخيرة من مراحل بناء مكعب بيانات التحاليل الطبية وتتضمن قائمة مكونات هذا المكعب من جدول الحقيقة والأبعاد.

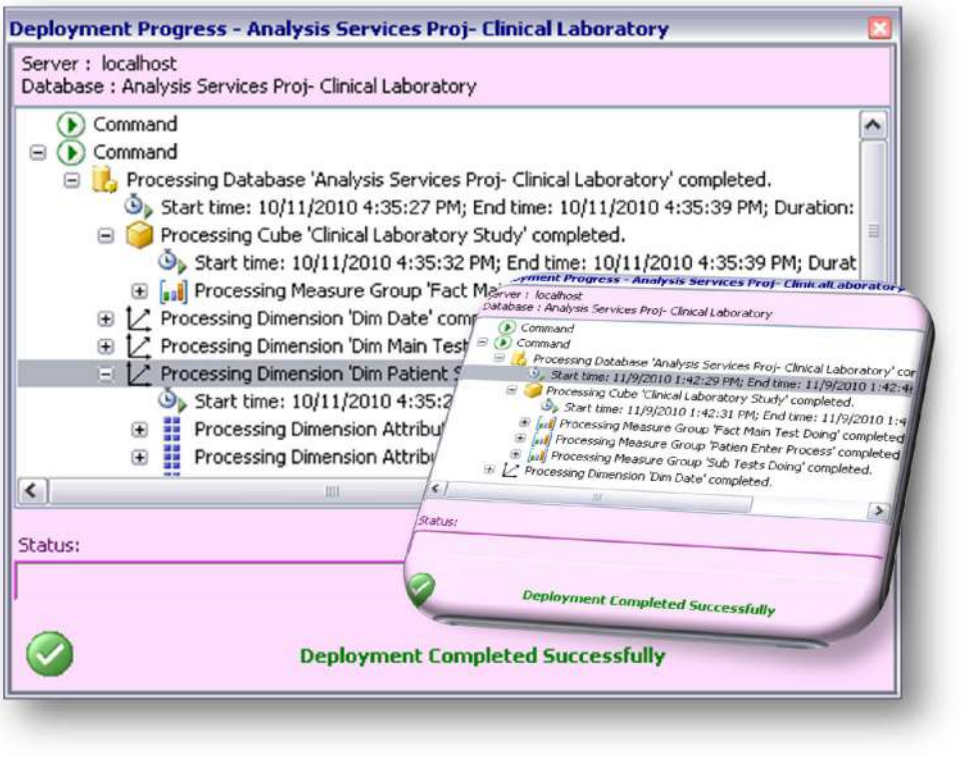


الشكل (5-3- 37): مكونات مكعب البيانات الخاص بدراسة بيانات مخبر التحاليل الطبية

6•4•3•5 تدقيق وبناء وتفعيل مكعب بيانات التحاليل الطبية

تختص هذه الخطوة باستدعاء إجراء (Start Debugging)، وهو الإجراء الذي يتأكد فيه البرنامج من صحة بناء المكعب بكل ما يحمله من تفاصيل، وفي حال وجود أخطاء يعطي البرنامج تقريراً بالأخطاء لكي يتم تصحيحها، وبعد القيام بأية إصلاحات أو إضافات يُعاود

استدعاء هذا الإجراء مرة أخرى، ويبين الشكل (5–3–38) نماذج عن التقارير الصحيحة التي أصدرها البرنامج عقب بناء مكعب بيانات التحاليل الطبية، وبعد إجراء بعض التغييرات على هذا المكعب.

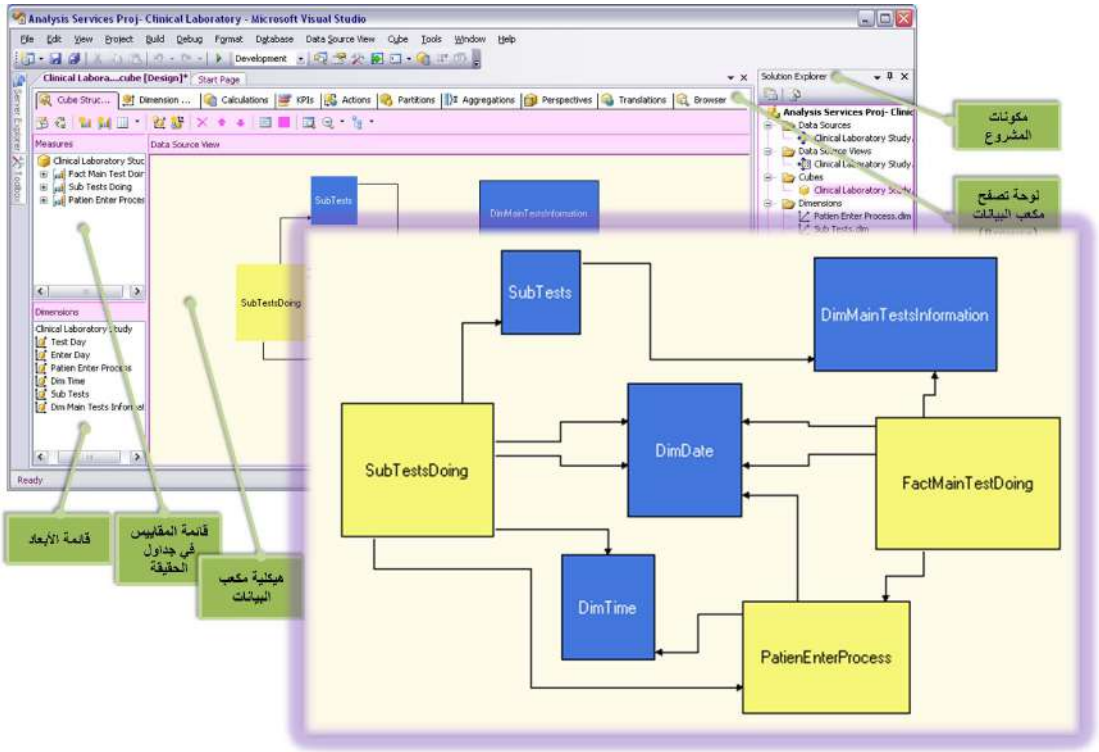


الشكل (5–3– 38): تقارير نجاح معالجة مكعب البيانات أو أية تغيُّرات تطرأ عليه

وبعد الانتهاء من العمليات الضرورية لبناء وتعديل مكعب البيانات، ظهرت عدة نوافذ جديدة في مشروع المعالجة التحليلية الآنية للمساعدة في تطوير مكعب البيانات الناتج، وكانت أولى هذه النوافذ هي نافذة بنية مكعب البيانات (Cube Structure) التي تألَّفت من ثلاثة جداول حقيقة وخمسة أبعاد، ويطلق على شكل هذا المكعب الناتج اسم مخطوطة مجموعة النجوم (constellation schema)[1] لأنه مكون من عدة جداول حقيقة وعدة جداول أبعاد وتربطها علاقات متداخلة، انظر الشكل (5–3–39). وقد كان بالإمكان إضافة جداول أبعاد فرعية إلى هذا المكعب مثل البعد الجغرافي الذي كان يرتبط ببعد بيانات المرضى في النموذج الأولي لمكعب البيانات الموجود في الشكل (5–3–17)، ولكن تبين بعد التجريب أنه يمكن الاستغناء عن مثل هذا البعد في المكعب المدروس بإضافة مكوناته إلى جدول بيانات المرضى، حيث

[1] راجع الفقرة (6•1•1).

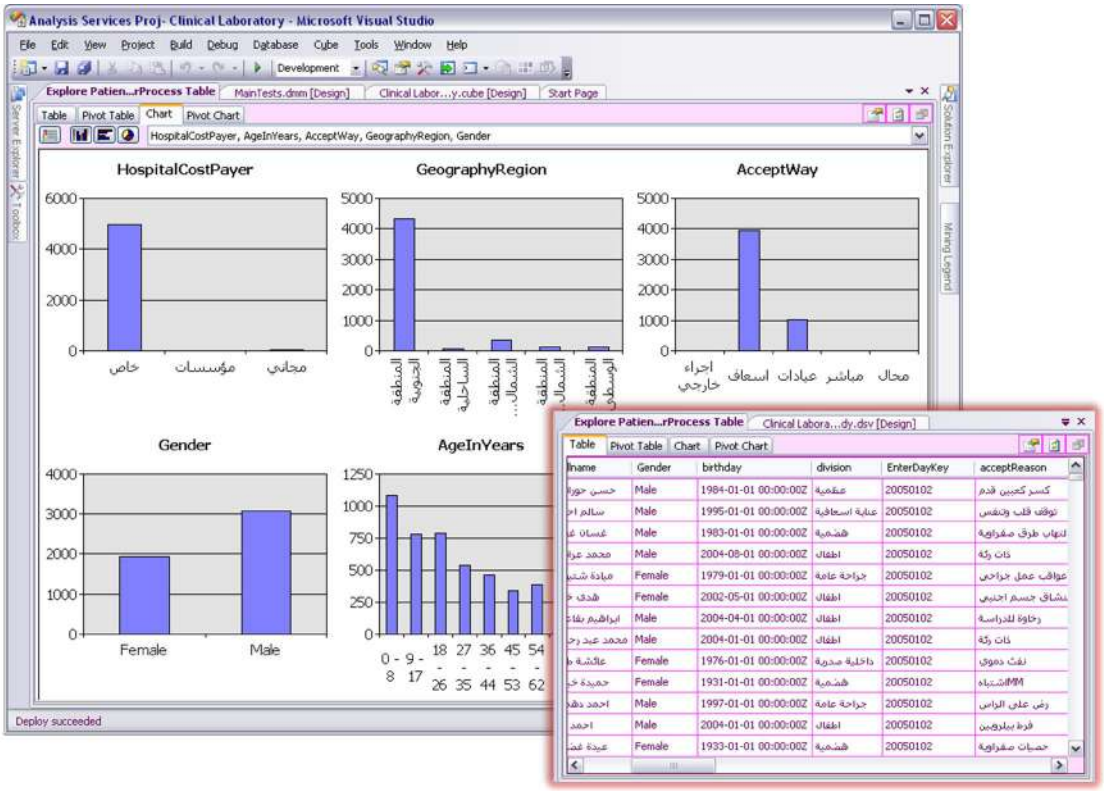
يساعد الاستغناء عن الأبعاد الفرعية في استعراض بعض المقاييس بطريقة أوضح وأفضل، وخصوصاً أنه لم يعد يستطع البعدُ الجغرافي أن يقدم أية فائدة إضافية للمكعب بعدما حُذفت البيانات المكررة للمرضى والتي كان فيها تغيرات قليلة في أماكن إقامة بعض المرضى.



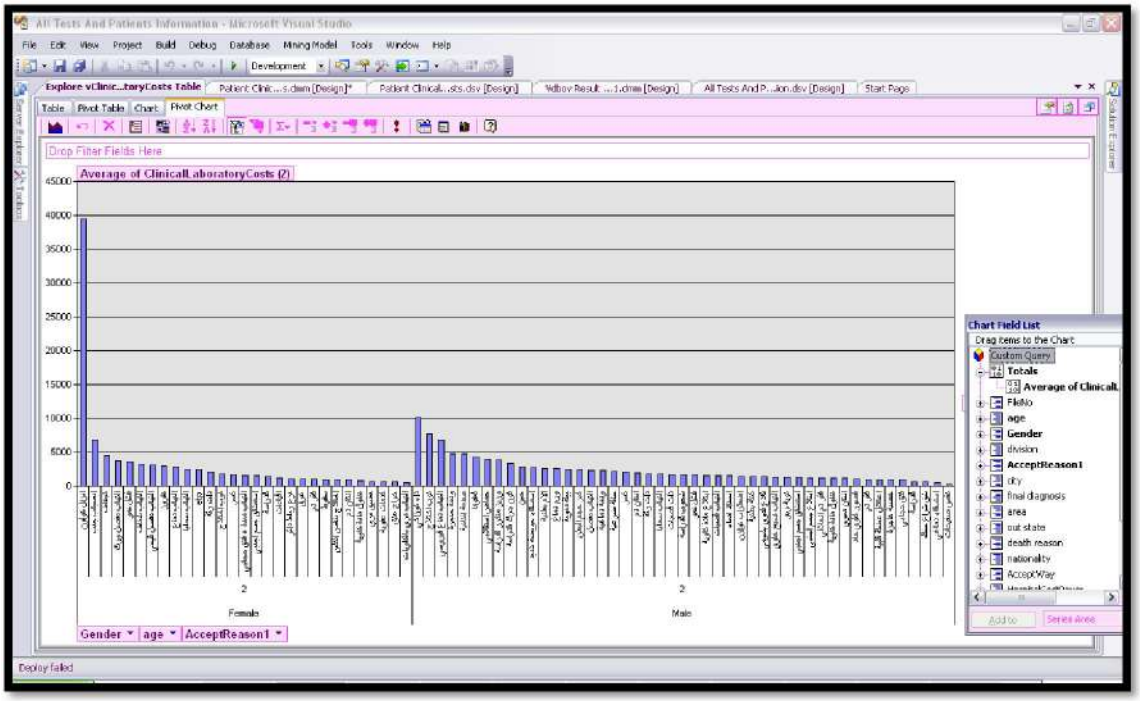
الشكل (5-3- 39): مكعب بيانات التحاليل الطبية المخبرية المستخدم في عمليات المعالجة التحليلية الآنية

5•3•4•7 استطلاع البيانات (Explore Data)

يساعد الضغط بالزر الأيمن من الفأرة على أي من جدول مكونات مكعب البيانات السابق على طلب استعراض عينة من بياناته (منتقاة آلياً) بعدة أشكال يختارها المستخدم (جدول بيانات، مخطط بياني، جدول محوري (Pivot Table)، مخطط محوري (Pivot Chart)). ويمكن أن تُستخدم هذه الخدمة بشكل مستمر أثناء العمل في هذا المشروع، وخصوصاً عند تطبيق عمليات المعالجة التحليلية الآنية، حيث تكثر الحاجة إلى تذكُّر بعض المعلومات البسيطة عن البيانات بسرعة ودون الحاجة إلى الاستعلام من الجداول الأساسية لقاعدة البيانات في برنامج (Management Studio)، انظر الشكلين (5-3-40) و (5-3-41).



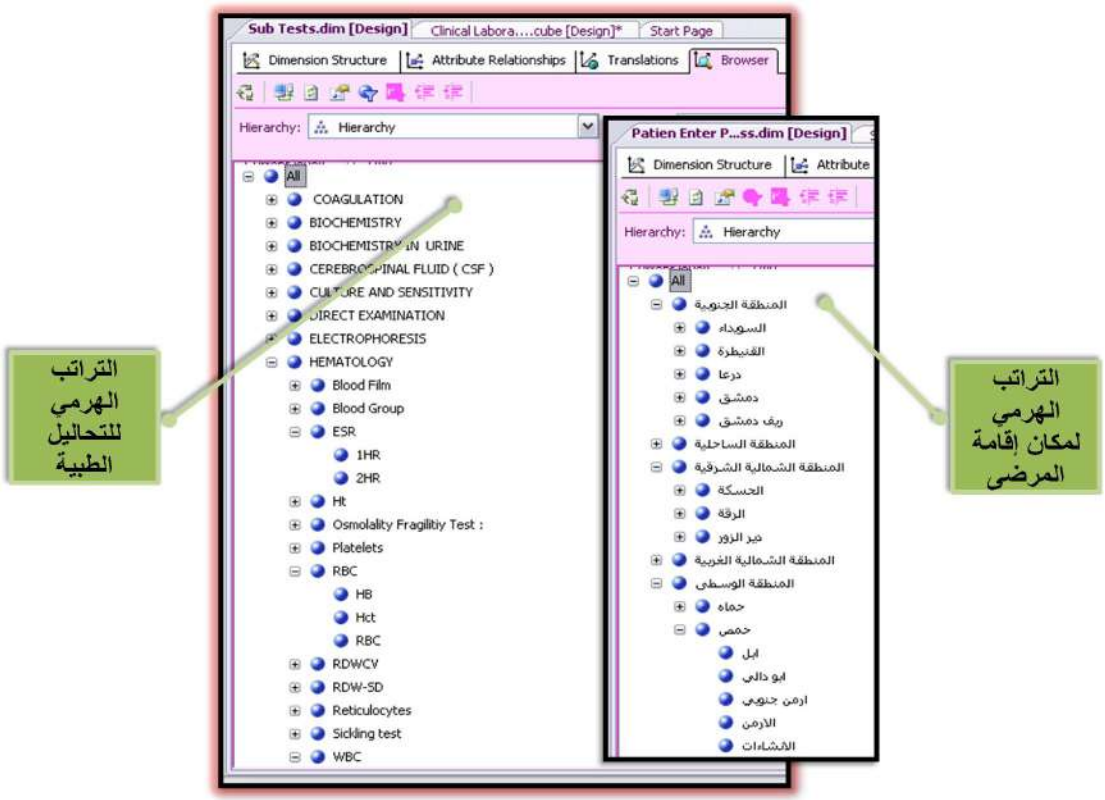
الشكل (– –): استعراض عينة من البيانات المسحوبة من جدول بيانات دخول المرضى إلى المشفى باستخدام ميزة استطلاع البيانات في هيكل مكعب البيانات



الشكل (5–3– 41): تصفح وسطي تكاليف التحاليل الطبية للمرضى بعمر السنتين ولكن بحسب الجنس والأمراض باستخدام (Pivot Char).

8•4•3•5 بناء تراتبات هرمية (Hierarchy) للأبعاد

يحتاج المكعب إلى التعرف على البنى الهرمية أو علاقات الانتماء بين بعض الخاصيات المكونة للأبعاد، وتدعى هذه العلاقات بالتراتبات الهرمية، ويُستفاد منها في الحصول بسهولة على تدرج هرمي أثناء تصفح هذا المكعب، وقد جرى بناء عدد من التراتبات الهرمية في مكعب بيانات التحاليل الطبية، مثل تراتب التحاليل الطبية الذي يتكون من ثلاث طبقات، وتتضمن الأولى أسماء مجموعات التحاليل (Discipline)، وتشتمل الثانية على أسماء التحاليل الرئيسة المندرجة تحت كل من هذه المجموعات (MainTests)، وتحتوي الثالثة على أسماء التحاليل الفرعية (SubTests)، ويمكن تصفح هذه الهرمية بعد بنائها ومعالجتها في صفحة البعد ضمن قسم (Browse) انظر الشكل (5-3-42).



الشكل (5-3- 42): التراتب الهرمي للتحاليل الطبية (التحاليل الفرعية ⊃ التحاليل الرئيسة ⊃ مجموعات التحاليل) والتراتب الهرمي لمكان إقامة المرضى (المنطقة ⊃ المدينة ⊃ المنطقة الجغرافية في سورية)

9•4•3•5 إضافة مقاييس جديدة داخل مكعب البيانات

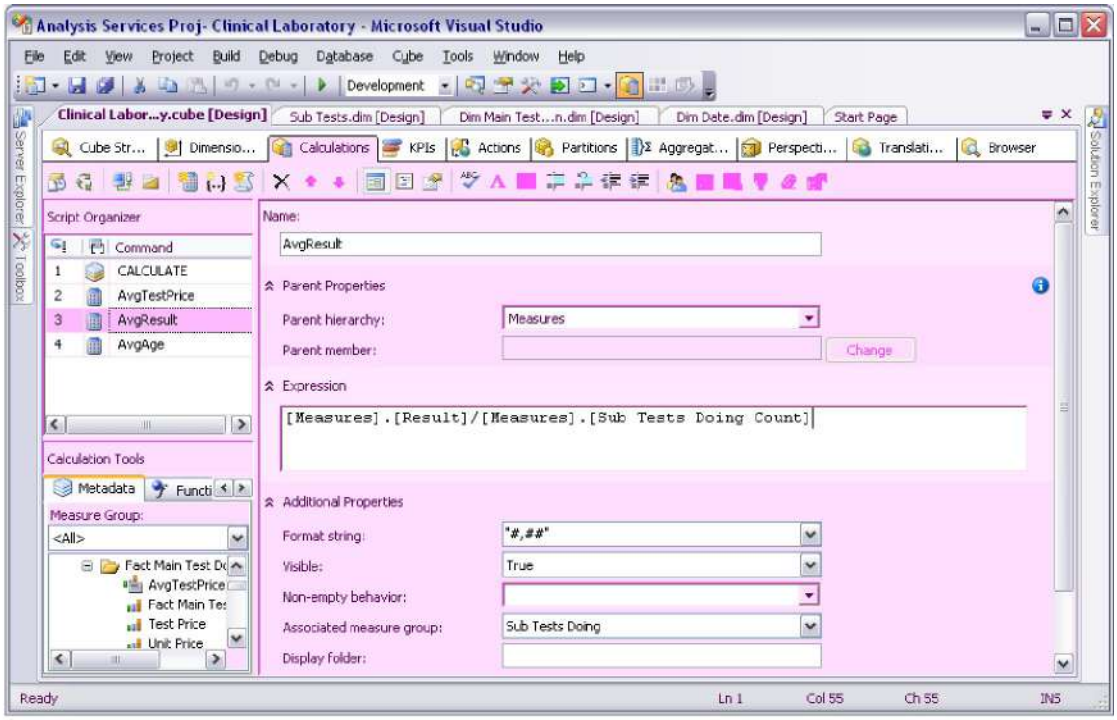
استدعت الضرورة عدة مرات أثناء التطبيق العملي لعمليات المعالجة التحليلية الآنية على مكعب البيانات إلى إيجاد مقاييس أخرى غير متوفرة في جداول مرأى مصدر البيانات، ويمكن التوصل إلى هذه المقاييس من خلال نافذة العمليات الحسابية (Calculations) التابعة لمكعب البيانات. فمثلاً تم إنشاء مقياس لحساب القيمة الوسطية لنتائج التحاليل الطبية[1] بالنسبة للمرضى، وتنتج هذه القيمة من خلال تقسيم مجموع نتائج التحاليل الطبية (Result) على عدد هذه التحاليل (Sub Tests Doing Count) كما في الشكل (5–3–43)، حيث أن عدد التحاليل الطبية هو مقياس خاص بمكعب البيانات وقد جرى احتسابه آلياً أثناء تعريف جدول الحقيقة، وأن مقياس (Result) يأخذ الصفة التجميعية في قائمة خواصه (Properties).

كما جرت إضافة أسماء لمقاييس جديدة تعادل قيمها قيمة المقاييس القديمة في نافذة مرأى مصدر البيانات، ثم أعطيت لها صفات جديدة في قائمة خواصها في لوحة مكعب البيانات، وكمثال على ذلك أضيف مقياس العمر الأكبر والعمر الأصغر إلى جدول بيانات المرضى في مرأى مصدر البيانات، بحيث أن قيمة كل منهما تساوي مبدئياً قيمة مقياس العمر في هذا الجدول، ثم استُدعيت خواص هذين المقياسين في نافذة مكعب البيانات واختير لكل منهما أحد النمطين (Min, Max) كما في الشكل (5–3–44).

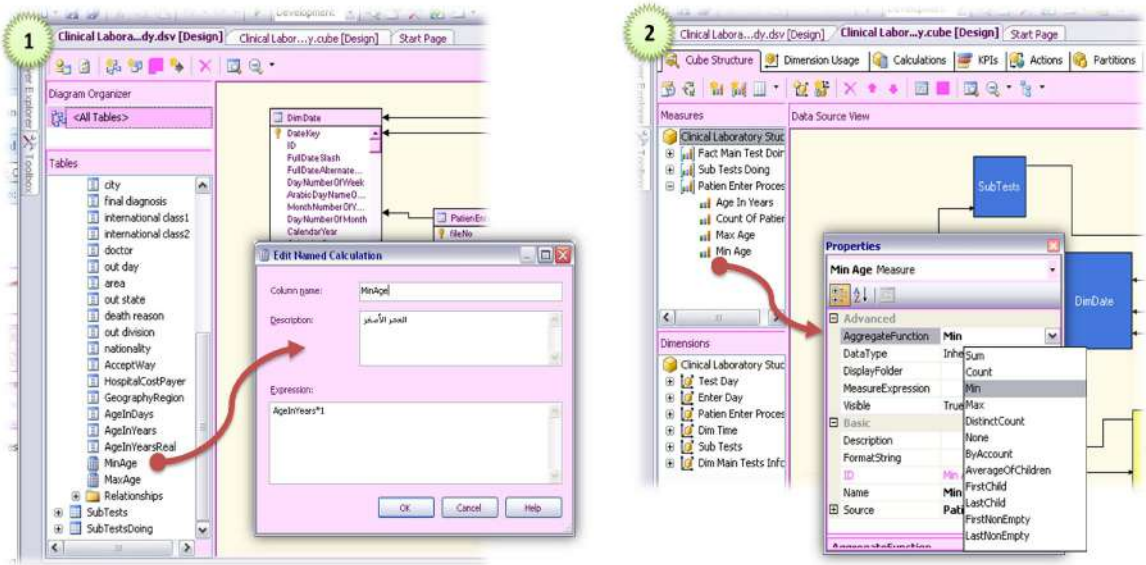
وقد تم تطوير تصميم مكعب البيانات بالاستعانة بالخدمات المختلفة المتاحة في نوافذ هذا المكعب، وأصبح مكعب بيانات التحاليل الطبية جاهزاً للاستعراض من قبل أدوات المعالجة التحليلية الآنية.

---

[1] يُستفاد طبياً من القيمة الوسطية لنتائج التحاليل الطبية في التعرف مثلاً على القيمة الوسطية لتحليل طبي ما عند مريض ما إذا أجري له هذا التحليل عدة مرات، أو في التعرّف على القيمة الوسطية لتحليل طبي ما ومقارنتها بين عدة فئات من المرضى (مرضى الشعبة القلبية، المرضى الذكور، مرضى الأعمار الصغيرة..)، كما يُمكن أن يُستفاد من القيم الوسطية للتحاليل الطبية في البيانات ذات الأحجام الكبيرة في وضع حدود جديدة للقيم الطبيعية عند شعب من الشعوب أو فئة من المرضى، أو إجراء المقارنات الإحصائية بين القيم الوسطية النظرية والقيم الوسطية الفعلية..الخ.



الشكل (– –): تطبيق العمليات الحسابية في إيجاد مقاييس جديدة داخل مكعب البيانات،
وهو الوسط الحسابي لنتائج التحاليل الطبية للمرضى



الشكل (5–3– 44): إضافة مقاييس جديدة إلى مكعب البيانات تساوي المقاييس القديمة ولكنها تختلف عنها في الخصائص

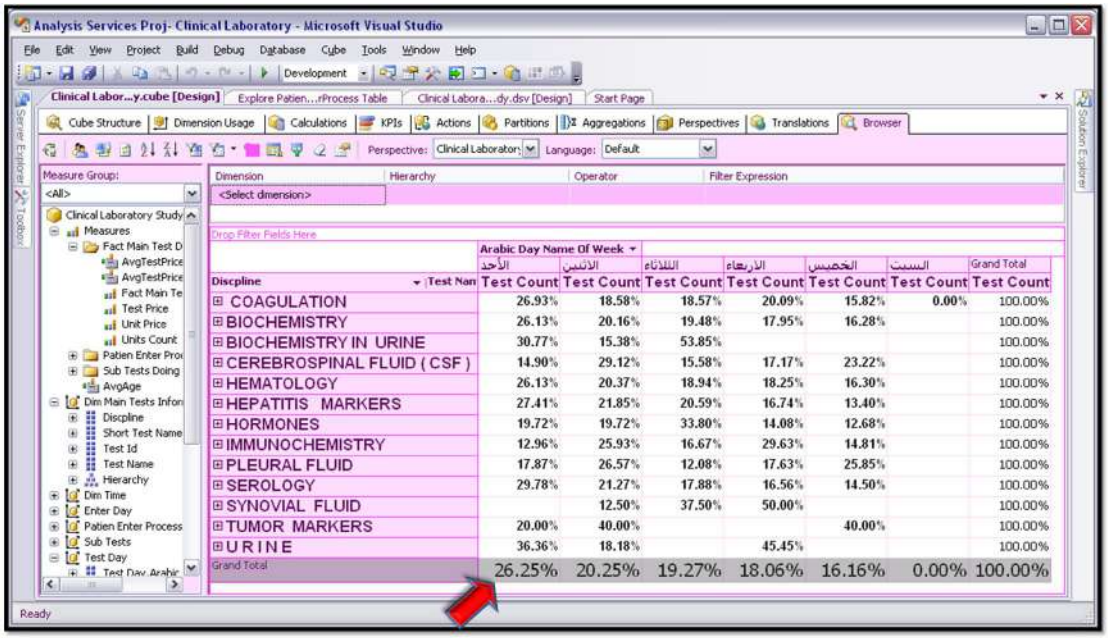
10•4•3•5 تصفح مكعب البيانات في برنامج (Microsoft Visual Studio)

تدعى عملية تصفح مكعب البيانات بالمعالجة التحليلية الآنية أو بالاستعلام متعدد الأبعاد، وتختص هذه العملية بتصفح المقاييس الموجودة في مكعب البيانات من زوايا مختلفة بحسب خاصيات الأبعاد أو التراتبات الهرمية التي يتم اختيارها، ويمكن عمل عدد كبير جداً من الاستعلامات متعددة الأبعاد على هذا المكعب.

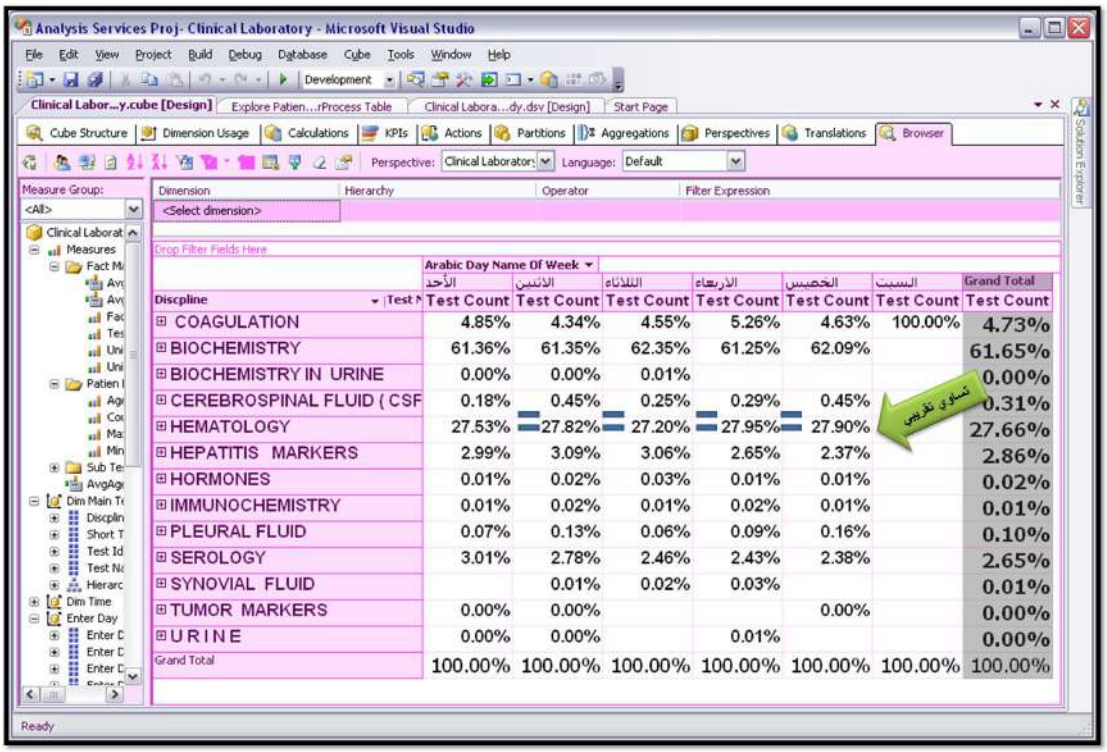
تختص نافذة التصفح (Browse) التابعة لنوافذ مكعب البيانات في الاستعراض المباشر لمكونات هذا المكعب، وهي أسرع أداة تصفح بسبب وجودها ضمن برنامج تصميم المكعب، كما أنها أداة تفاعلية مع عملية التصميم حيث أنها لا تستطيع أن تعمل ما لم يكن تصميم المكعب صحيحاً، وتُستخدم أيضاً في التحقق المباشر من نتائج إجراء أية تعديلات في مكعب البيانات. ونبين فيما يلي بعضاً من الأمثلة عن المعلومات التي استخلصتها عمليات المعالجة التحليلية الآنية من مكعب البيانات السابق، وقد تم تطبيق العمليات الأساسية التي تطرقنا لها في الفقرة (1•1•7).

أ- توزع أعداد التحاليل الطبية بحسب مجموعات التحاليل وأيام الأسبوع

أظهرت نتائج الاستعلام أنه لم يتم إجراء أية تحاليل طبية في يومي الجمعة والسبت لأنهما يومي العطلة الأسبوعية لموظفي مخبر التحاليل الطبية في المشفى، وقد تبين أن أكبر نسبة من التحاليل الطبية يتم إجراؤها في يوم الأحد وتصل إلى 26.25% من مجمل التحاليل الطبية خلال أيام الأسبوع، انظر الشكل (5-3-45)، وبعيداً عن السبب المنطقي لهذه النتيجة وهو وجود يومي العطلة قبل هذا اليوم، فلربما تساعد هذه النتيجة في التقصي أكثر عن أنواع التحاليل الطبية التي يكثر طلبها في هذا اليوم وعن أعدادها الحقيقية والمتوقعة، وبالتالي الاستعداد مسبقاً من قبل كوادر المشفى والمواد الأولية لمقابلة الطلب الإضافي على هذه الخدمة. وللمساعدة قليلاً في جلب المزيد من الاستعلامات في هذا الاتجاه جرى طلب النسب المئوية الشاقولية بحسب مجموعات التحاليل فتبين أن توزع نسب التحاليل اليومية في كل مجموعة لا تختلف كثيراً بين أيام الأسبوع، أي أن الذي يختلف هو أعداد المرضى الذين تُجرى لهم التحاليل، وأن الزيادة أو النقصان تطرأ على أعداد كل أنواع التحاليل بنفس نسب تغير أعداد المرضى تقريباً. انظر الشكل (5-3-46).



الشكل (– –): استعلام من مكعب بيانات التحاليل الطبية عن التوزع النسبي لأعداد التحاليل الطبية بحسب نوعها خلال أيام الأسبوع (حساب النسب أفقياً)، وتُعرض البيانات المتعلقة بالتحاليل الطبية وقد طويت للأعلى ( ) بحسب مجموعات التحاليل الطبية، حيث توجد إشارة الجمع إلى جانبها وتدل على أنها مطوية للأعلى وقابلة للنشر للأسفل ( ) لإظهار أسماء التحاليل الرئيسة التابعة لكل مجموعة من مجموعات التحاليل الطبية



الشكل (5-3- 46): استعلام من مكعب بيانات التحاليل الطبية عن التوزع النسبي لأعداد التحاليل الطبية اليومي بحسب مجموعات التحاليل (تم حساب النسب شاقولياً)

ب- النتيجة الوسطية لكل من التحاليل الطبية الفرعية وموزعة بحسب الجنس في بعض الشعب الطبية

أُجري الاستعلام عن أكثر التحاليل الطبية طلباً في الشعب الطبية المتعلقة بالجراحة في المشفى (جراحة عصبية، جراحة عامة، جراحة صدرية، جراحة بولية، جراحة أوعية)، ومن ثم طُلب الاستعلام عن وسطي نتائج التحاليل الطبية عند مرضى الجراحة وبحسب الجنس، انظر الشكل (5-3-47). ويعتبر هذا النوع من الاستعلامات اقتطاعاً لجزء من مكعب البيانات (Dice) المتعلق بفئة مرضى الشعب الطبية الجراحية، ويساعد هذا النوع من الاستعلامات متعددة الأبعاد في التعرف على المقاييس لدى فئة معينة من المرضى، ومقارنة النتائج بين الجنسين، فمثلاً نجد أن معظم التحاليل الطبية تكون نتائجها الوسطية مرتفعة عند الذكور أكثر من الإناث. وبغض النظر عن الفوائد الدقيقة لهذا الاستعلام فهو يُعتبر نموذجاً للاستعلامات الممكن إجراؤها على البيانات المتاحة.

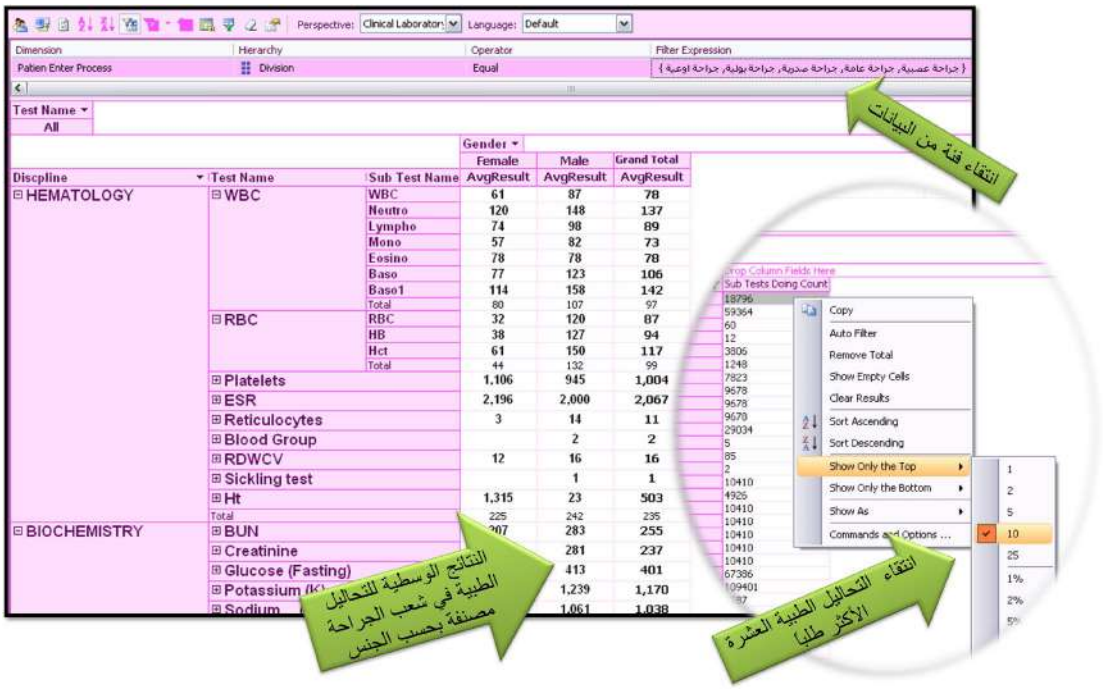
وقد تم التمكن من إجراء مثل هذا الاستعلام بفضل الطريقة التي اتبعت في إعادة بناء جدول نتائج التحاليل الطبية بحيث تظهر نتائج التحاليل (Results) عند مستوى التحاليل الفرعية (SubTest) وليس التحاليل الرئيسة (MainTests).

كما أصبح بالإمكان تصفح النتائج الوسطية للتحاليل الطبية لأنه تم تحويل حقل النتائج إلى حقل رقمي (Number) بدل النصي (Text) الذي كان متبعاً في قاعدة بيانات المخبر الأساسية.

وظهرت في هذا الاستعلام البنية الهرمية التي تتدرج من التحاليل الفرعية إلى التحاليل الرئيسة إلى مجموعات التحاليل. ويمكن مستقبلاً تشكيل بنى هرمية جديدة للتحاليل الطبية، بحيث ترتبط التحاليل بالأمراض التي تحتاج إليها، وبالخاصيات المختلفة للمرضى (وذلك بالاستعانة بأدوات التنقيب كتلك التي سيرد ذكرها لاحقاً في الفقرة **5•4•2•3•6**).

ومن الممكن تطوير هذا الاستعلام بشكل أعمق واستخراج مثلاً الأسباب المرضية الأكثر ترافقاً مع ارتفاع قيم نتائج بعض التحاليل الطبية. كما يمكن أخذ نتائج مثل هذه الاستعلامات وتحويلها إلى البرامج الإحصائية لدراستها. وبكل حال من الأحوال فإن توجيه مثل هذه الطلبات أو الاستعلامات عن البيانات يمكن أن يتم وفقاً لاهتمامات ورغبات الباحثين في الحقل

الطبي، وتزداد جودة وقدرة هذه الاستعلامات بوجود خاصيات بيانات الكترونية إضافية أخرى عن المرضى مثل تلك الموجودة في نظم المعلومات الطبية العالمية.



الشكل (5-3- 47): الاستعلام المتعدد الأبعاد عن وسطي نتائج التحاليل الطبية لأكثر التحاليل الطبية طلباً في الشعب الطبية التي تُعنى بالجراحة في المشفى، وموزعة حسب الجنس، وتظهر عملية الاقتطاع (Dice) لفئة من البيانات، وقد نثرت للأسفل (Drill Down) مجموعات التحاليل الطبية وبعض التحاليل الرئيسة حيث تظهر إلى جانبها إشارة الطرح (-)

ج- الاستعلام عن مقاييس العمر (العمر الوسطي، والعمر الأدنى والعمر الأقصى) وأعداد المرضى بحسب الجنس والأسباب المرضية

ظهرت الرغبة في التعرف على الأعمار العليا والدنيا والوسطية للإصابة بالأمراض بحسب الجنس، ويستفاد من ذلك في مراقبة التعرض لأنواع معينة من الأمراض عند الجنسين، وقد أجري هذا الاستعلام ورتبت فيه النتائج تنازلياً بحسب أعداد المرضى في كل من الأمراض. وقد أظهر هذا الاستعلام المعروض في الشكل (5-3-48) أن هنالك تقارباً كبيراً في الأعمار الوسطية عند كل من الجنسين، ولكن تختلف الأعمار الدنيا بينهما في بعض الأمراض، فمثلاً تعرَّض الذكور لاحتشاء العضلة القلبية ابتداءً من عمر السنتين في حين تعرَّضت أصغر أنثى لذلك في عمر 26. وهنا يمكن أن تَظهر حدود إمكانيات المعالجة التحليلية الآنية التي لا تستطيع بمفردها أن تقيِّم الظواهر بل تحتاج إلى أدوات أخرى كالأدوات الإحصائية التي تعمل

على تقييم الفرضيات واكتشاف فيما إذا كان فعلاً لا يوجد فارق حقيقي بين الجنسين في العمر الوسطي للتعرض للأمراض.

Drop Filter Fields Here

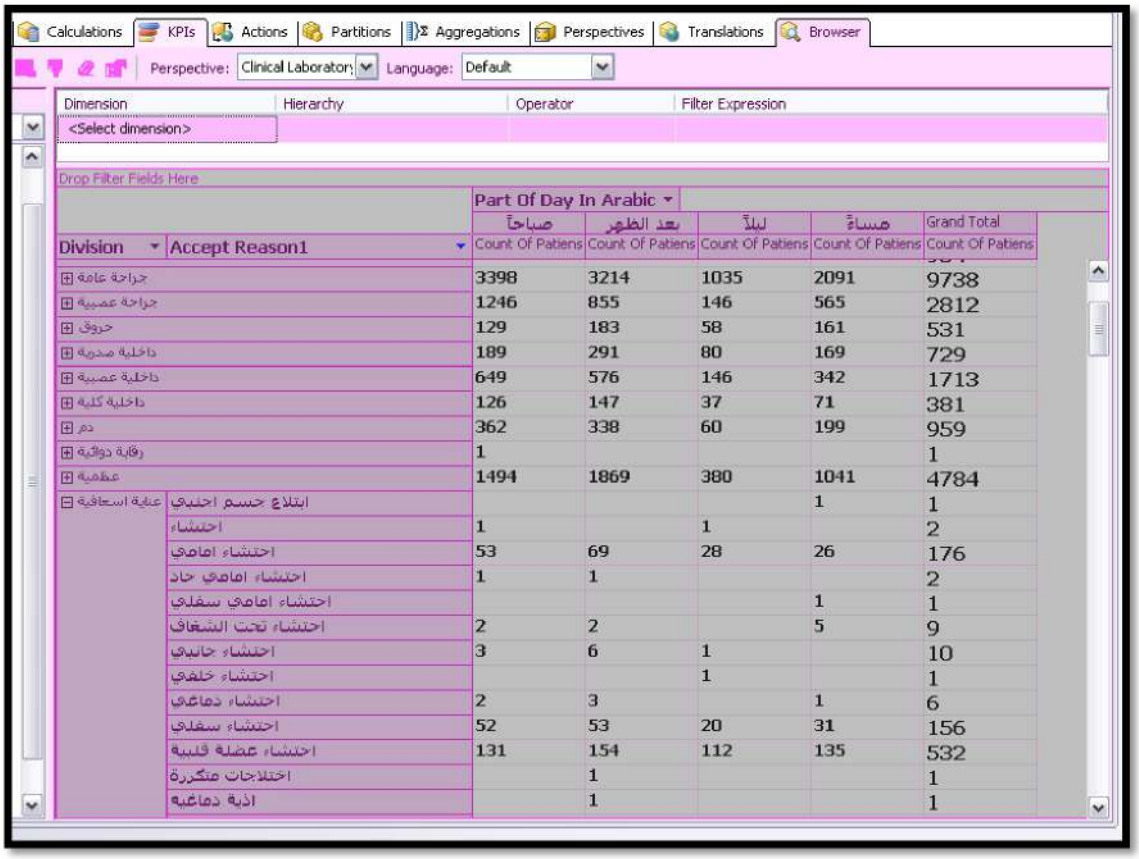
	Gender							
	Female				Male			
Accept Reason1	Min Age	Max Age	AvgAge	Count Of Patients	Min Age	Max Age	AvgAge	Count Of Patients
الام بطنية	0	94	25	1366	0	88	25	1313
كسر	0	99	39	593	0	96	29	1683
حادث سير	0	88	29	304	0	97	28	1072
سقوط	0	98	26	357	0	93	21	964
فتق اربي	0	78	13	173	0	95	27	814
التهاب زائدة دودية	1	72	21	452	0	87	21	523
خناق صدر	1	89	56	260	3	94	53	615
ذات رئة	0	94	15	312	0	86	17	504
التهاب كبد	1	74	40	233	0	82	37	456
احتشاء عضلة قلبية	26	85	59	115	2	94	53	523
لوزات	0	47	11	298	1	43	10	301
مرارة محصاة	0	89	43	422	8	83	44	175
C.V.A	4	99	62	244	7	93	59	322
حرق	0	75	14	173	0	90	18	315
رض	0	82	20	110	0	86	23	323
للدراسة	0	87	44	192	0	84	43	238
نزف هضمي	0	98	53	136	0	94	50	292
سداد	1	86	61	166	2	91	60	228
نزع مواد استجداب	1	85	25	110	0	82	24	271
ورم دماغ	0	80	35	164	0	70	36	207
ERCP								

الشكل (5-3- 48): الاستعلام من مكعب بيانات التحاليل الطبية عن مقاييس العمر (العمر الوسطي، العمر الأعلى، العمر الأدنى) وأعداد المرضى في كل من الجنسين وبحسب الأسباب المرضية (المرتبة تنازلياً بحسب أعداد المرضى)[1]

د- توزع أعداد المرضى القادمين إلى المشفى بحسب شعبة القبول والسبب المرضي وتوقيت القدوم إلى المشفى

تختلف أوقات قدوم المرضى إلى المشفى باختلاف الأسباب المرضية التي يتعرضون لها، وبشكل عام تقل أعداد المرضى القادمين ليلاً، وبناءً على ذلك تقل الكوادر البشرية الطبية في الليل. ولكن هل يوجد اختلافات عن هذه القاعدة في أمراض معينة أو في شعب معينة؟ وهل يمكن أن يتم تخطيط تواجد الكوادر البشرية والمعدات والمواد بصورة متناسبة مع احتياجات كل فئة من فئات الأمراض في كل وقت من أوقات اليوم؟. تساعد أدوات المعالجة التحليلية الآنية في التعرف على هذه التفاصيل والإجابة عن تلك التساؤلات. ويبين الشكل (5-3-49) مثالاً بسيطاً عن مثل هذا النوع من الاستعلامات بحسب بُعد توقيت القدوم إلى المشفى وبحسب الشعب الطبية والأسباب المرضية في كل منها. ويمكن أن يتم توجيه نقاط الاهتمام في مثل هذا الاستعلام بحسب رأي الجهة المستفيدة من المعلومات التي يُقدمها.

[1] يُعبِّر العمر 0 في حقل العمر عن الأطفال الرضع دون السنة من العمر.



Division	Accept Reason1	صباحاً	بعد الظهر	ليلاً	مساءً	Grand Total
جراحة عامة		3398	3214	1035	2091	9738
جراحة عصبية		1246	855	146	565	2812
حروق		129	183	58	161	531
داخلية صدرية		189	291	80	169	729
داخلية عصبية		649	576	146	342	1713
داخلية كلية		126	147	37	71	381
دم		362	338	60	199	959
رقابة دوائية		1				1
عظمية		1494	1869	380	1041	4784
عناية اسعافية	ابتلاع جسم اجنبي				1	1
	احتشاء	1		1		2
	احتشاء امامي	53	69	28	26	176
	احتشاء امامي حاد	1	1			2
	احتشاء امامي سفلي				1	1
	احتشاء تحت الشغاف	2	2		5	9
	احتشاء جانبي	3	6	1		10
	احتشاء خلفي			1		1
	احتشاء دماغي	2	3		1	6
	احتشاء سفلي	52	53	20	31	156
	احتشاء عضلة قلبية	131	154	112	135	532
	اختلاجات متكررة		1			1
	اذية دماغية		1			1

الشكل (5-3- 49): توزع أعداد المرضى القادمين إلى المشفى بحسب شعبة القبول والسبب المرضي وتوقيت القدوم إلى المشفى

ه- التوزع النسبي للإيراد الإجمالي للمخبر من التحاليل الطبية[1] بحسب أرباع السنة

يُمكن الاستدلال على إيراد المخبر بشكل غير مباشر -في البيانات المتوفرة للدراسة- باستخدام مجموع أسعار التحاليل الطبية التي دفعها المرضى والمتمثلة بالمقياس (TestPrice) خلال أرباع السنة. ويُلاحظ من خلال هذا الاستعلام أن إيراد المخبر يبلغ ذروته خلال الربع الثالث من العام، انظر الشكل (5-3-50). ويدل هذا الارتفاع من جهة أخرى على تزايد الحالات المرضية خلال هذا الربع مثل التهاب الكبد (حيث تزايدت تحاليل علامات التهاب الكبد (Hepatitis Markers) كثيراً في الربع الثالث). ويساعد التعرُّف المسبق على مثل هذه الأمور الإدارةَ الطبية لكي تُوفِّر الأجهزة والمواد والكوادر البشرية المتناسبة مع تقلبات أعداد المرضى والتحاليل الطبية ومتطلبات علاجهم، كما أنه قد يلفت الانتباه نحو إجراء الدراسات الطبية الداعمة حول أسباب هذه الزيادات، والتي لا تندرج طبعاً ضمن إطار البحث الحالي

[1] جرى طلب استعراض البيانات بحسب النسب المئوية من أجل تبسيط وتسريع قراءة البيانات، ولأنه من غير المستهدف هنا القيام بدراسة محاسبية أو مالية للمخبر.

المتعلق بتجريب أدوات حديثة لتصفح ودراسة البيانات الطبية.

Drop Filter Fields Here

	Calendar Quarter				
	1	2	3	4	Grand Total
Discpline	Test Price	Test Price	Test Price	Test Price	Test Price
BIOCHEMISTRY	19.38%	27.60%	29.18%	23.85%	100.00%
HEMATOLOGY	19.19%	27.85%	29.06%	23.89%	100.00%
HEPATITIS MARKERS	18.85%	23.03%	31.14%	26.99%	100.00%
COAGULATION	18.46%	26.41%	27.72%	27.41%	100.00%
SEROLOGY	20.39%	27.42%	30.14%	22.05%	100.00%
CEREBROSPINAL FLUID (CSF)	34.34%	36.38%		29.27%	100.00%
HORMONES	2.82%	8.45%	38.03%	50.70%	100.00%
IMMUNOCHEMISTRY	25.93%	20.37%	40.74%	12.96%	100.00%
PLEURAL FLUID	34.30%	51.21%	3.86%	10.63%	100.00%
TUMOR MARKERS			80.00%	20.00%	100.00%
SYNOVIAL FLUID	25.00%	25.00%		50.00%	100.00%
U R I N E	63.64%		18.18%	18.18%	100.00%
BIOCHEMISTRY IN URINE	15.38%	23.08%	46.15%	15.38%	100.00%
Grand Total	19.35%	27.52%	29.04%	24.08%	100.00%

الشكل (– –): التوزع النسبي لإيرادات المخبر من التحاليل الطبية على أرباع السنة وبحسب مجموعات التحاليل

أثارت النتائج السابقة الفضول نحو التعمق أكثر والتعرف على تفاصيل الأرقام السنوية خلال أرباع السنة، والتأكد فيما إذا كانت الزيادات في إيرادات المخبر خلال الربع الثالث هي زيادة خاصة بهذا الربع أم أنها تعود إلى تزايد كبير للإيرادات في عام معين.

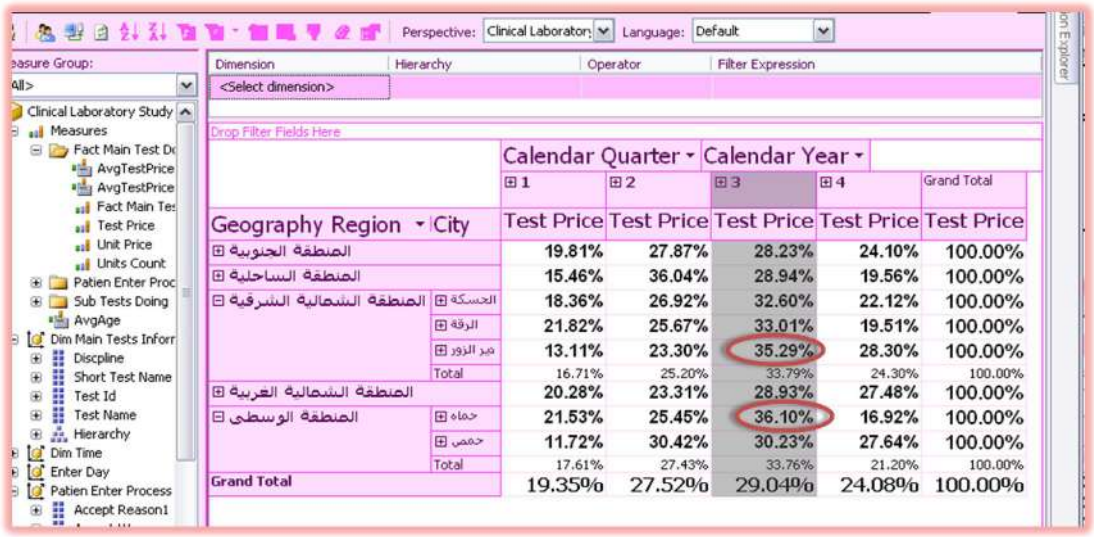
وتبعاً لذلك تم استدعاء الاستعلام السابق على مستوى الربع ومنه على مستوى العام، كما هو ظاهر في الشكل (– –)، فأظهر هذا الاستعلام أن هنالك تقارباً كبيراً بين نسب الإيرادات المحققة في كل من العامين – في الربع الثالث، مما يدل على أن التزايد الحاصل في هذا الربع هو تزايد خاص به ويحصل في كل عام.

Drop Filter Fields Here

	Calendar Quarter		Calendar Year						
	⊞1	⊞2	⊟3			⊟4			Grand Total
			2007	2008	Total	2007	2008	Total	
Discpline	Test Price	Test Price	Test Price	Test Price	Test Price	Test Price	Test Price	Test Price	Test Pric
BIOCHEMISTRY	19.38%	27.60%	14.42%	14.76%	29.18%	12.78%	11.06%	23.85%	100.00%
HEMATOLOGY	19.19%	27.85%	14.75%	14.31%	29.06%	12.67%	11.22%	23.89%	100.00%
HEPATITIS MARKERS	18.85%	23.03%	14.65%	16.49%	31.14%	14.11%	12.88%	26.99%	100.00%
COAGULATION	18.46%	26.41%	15.05%	12.67%	27.72%	14.84%	12.57%	27.41%	100.00%
SEROLOGY	20.39%	27.42%	15.82%	14.32%	30.14%	10.55%	11.50%	22.05%	100.00%
CEREBROSPINAL FLUID (CSF)	34.34%	36.38%				29.27%		29.27%	100.00%
HORMONES	2.82%	8.45%	9.86%	28.17%	38.03%		50.70%	50.70%	100.00%
IMMUNOCHEMISTRY	25.93%	20.37%	20.37%	20.37%	40.74%	3.70%	9.26%	12.96%	100.00%
PLEURAL FLUID	34.30%	51.21%	3.86%		3.86%	10.63%		10.63%	100.00%
TUMOR MARKERS				80.00%	80.00%		20.00%	20.00%	100.00%
SYNOVIAL FLUID	25.00%	25.00%				50.00%		50.00%	100.00%
U R I N E	63.64%		18.18%		18.18%		18.18%	18.18%	100.00%
BIOCHEMISTRY IN URINE	15.38%	23.08%		46.15%	46.15%		15.38%	15.38%	100.00%
Grand Total	19.35%	27.52%	14.53%	14.51%	29.04%	12.88%	11.20%	24.08%	100.00%

الشكل (5-3- 51): التوزع النسبي لإيرادات المخبر من مجموعات التحاليل الطبية في الربع الثالث وبحسب السنوات

كما وأدت نتائج الاستعلام السابق إلى التفكير بالتفتيش أكثر عن مصادر هذه الزيادات في التحاليل، وأجري لذلك الاستعلام عن التوزع الجغرافي للمرضى الذين أجريت لهم التحاليل الطبية بحسب أماكن سكنهم التي أتوا منها، وقد وُجد أن الارتفاع في إيرادات التحاليل الطبية خلال الربع الثالث كان بنتيجة ارتفاع الطلب على التحاليل الطبية للمرضى القادمين من المنطقة الوسطى خصوصاً حماه، ومن المنطقة الشمالية الشرقية وخصوصاً دير الزور، انظر الشكل (5-3-52).



الشكل (5-3- 52): التوزع النسبي لإيرادات المخبر من التحاليل الطبية على أرباع السنة وبحسب أماكن إقامة المرضى

وبهذا نجد أن تقنية المعالجة الآلية تدعم إجراءَ المزيد من الاستعلامات عن التوزع الجغرافي للمرضى والأمراض التي يعانون منها، مما يساعد في التعرف على الخدمات الطبية التي لا تتوفر في الأماكن الجغرافية المختلفة، وتجعل المرضى يقطعون مسافات طويلة للحصول على الخدمات الطبية المطلوبة. ويساعد مثل هذا النوع من الاستعلام – في حال تم تطبيقه في كل المشافي– على التخطيط لدعم الرعاية الصحية في أماكن الطلب عليها[1].

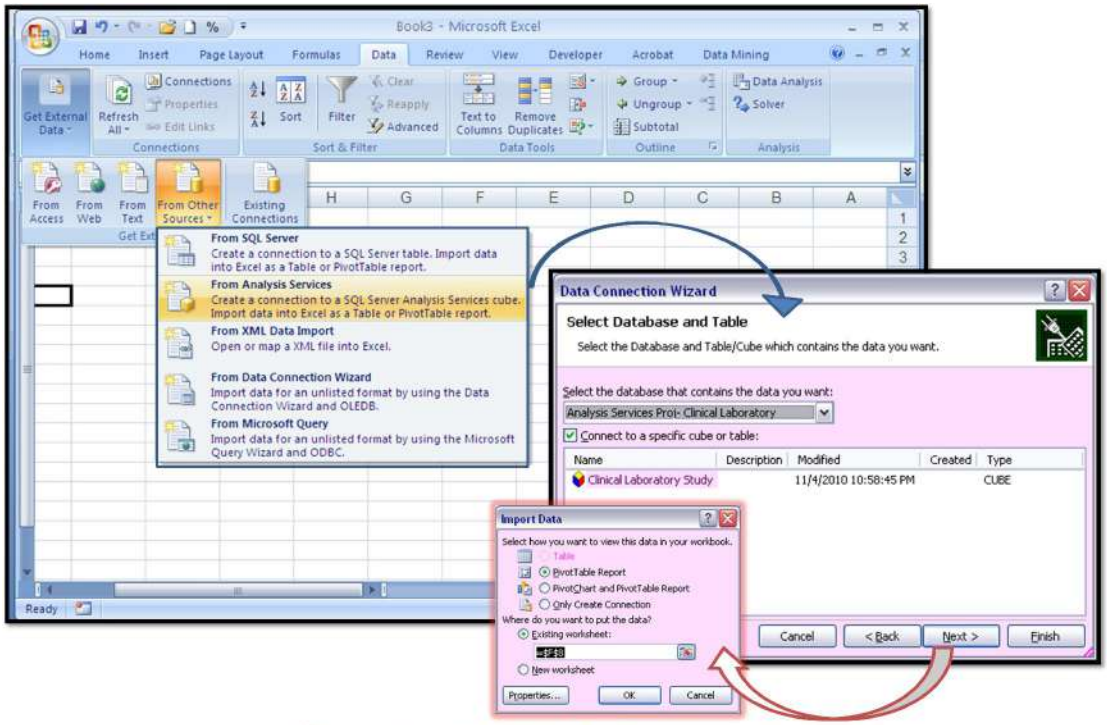
[1] أظهرت دراسة أجريت في سورية وجود زيادة كبيرة في عدد حالات الإصابة بمرض التهاب الكبد في جميع المحافظات بالنسبة للنمط ب، وتوضّع بؤري في المحافظات الشرقية والشمالية للنمط ج، مما دعا لتشكيل لجنة وطنية للعمل على وضع أسس لمكافحة هذا المرض، وإحداث مراكز صحية إضافية خاصة بهذه الإصابات وذات توزع جغرافي يضمن وصول الخدمة الصحية لجميع الأفراد المصابين مجاناً مع التركيز على ضرورة تخفيف عبء الانتقال على المريض لضمان حسن المتابعة لحالته، وسيتم العمل خلال عام 2011 على تطبيق نظام أتمتة الكترونية لهذه المراكز وإعطاء المرضى بطاقات الكترونية ذكية تمكنهم من تلقي الخدمة من أي من هذه المراكز المتاحة [صحيفة الثورة، 2011/2/17].

11•4•3•5 تصفح مكعب البيانات في برنامج (Microsoft Excel 2007)

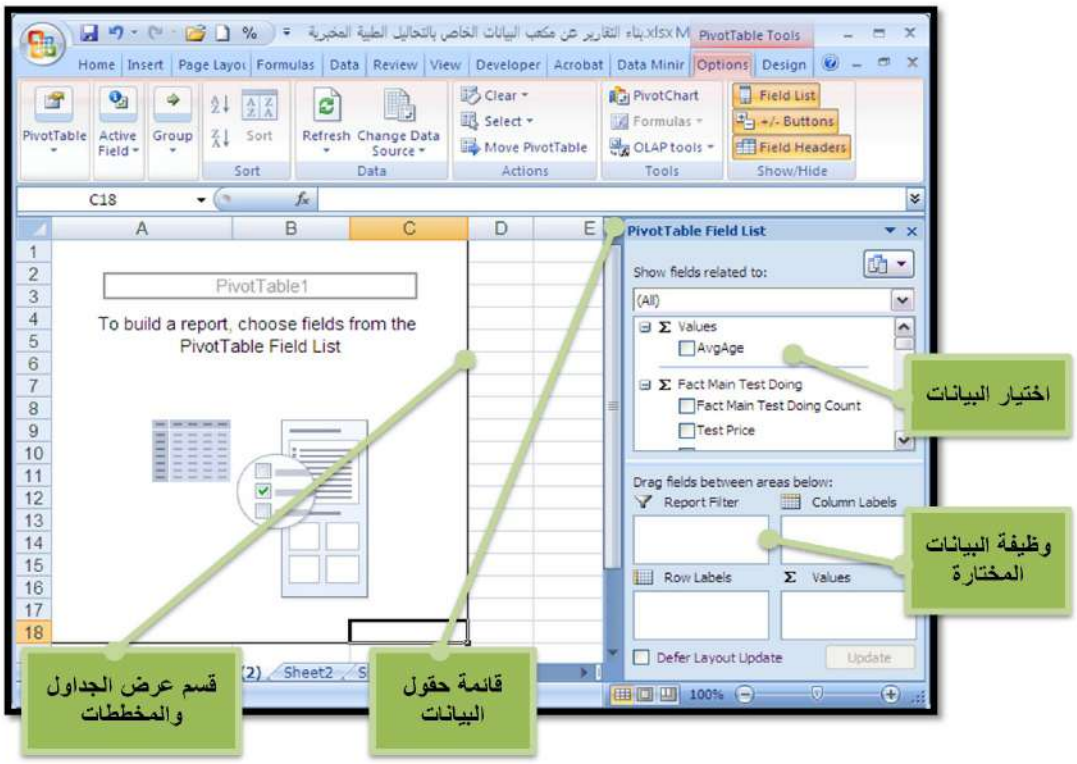
تتصف الاستعلامات التي يتم طلبها في برنامج (Microsoft Visual Studio) بأنها غير قابلة للتخزين بشكلها الجدولي الذي تظهر فيه، وإنما يجب استدعاؤها في كل مرة يحتاج فيها المستفيد إلى المعلومات، وهي ليست بالعملية الصعبة إذا تم التدرب عليها، ولكن تكمن المشكلة بأن صانعي القرار أو الأشخاص المهتمين بنتائج المعالجة التحليلية الآنية يحتاجون إلى عرض النتائج على شكل تقارير الكترونية أو ورقية وأن تكون متاحة بين أيديهم بشكل دائم. كما يحتاج هؤلاء إلى عرض النتائج بشكل جداول ورسوم بيانية جذابة وواضحة مما لا يتوفر في هذا البرنامج. ولتحقيق ذلك يمكن الاستعانة ببعض البرامج الالكترونية المختلفة مثل برنامج اكسل، حيث يستطيعُ هذا البرنامج أن ينشئ ارتباطاً أو اتصالاً بخدمات التحليل (Analyses Services) التي أجريت في برنامج (Microsoft Visual Studio)، فتصبح بذلك مكونات مكعب البيانات متاحة في اكسل لإجراء أية استعلامات منها وبعدة أشكال للاستعراض كالجداول والرسوم البيانية.

وقد جرى في هذه الدراسة استخدام اكسل لهذه الغاية، وتم بدايةً تعريف الاتصال ببيانات مكعب بيانات التحاليل الطبية في مشروع المعالجة التحليلية الآنية السابق باستخدام إجراء (Get External Connection from Other Sources) الموجود ضمن قائمة (Data) في برنامج اكسل، انظر الشكل (5-3-53).

ويُنتج الاتصالُ في اكسل ببيانات من نمط خدمات التحاليل ورقةَ عمل تفاعلية تتيح للمستخدم انتقاء حقول البيانات التي يرغب بعرضها في شكل جدول محوري أو في مخطط بيانات محوري، وتتألف هذه الورقة من قسمين أساسيين الأول ويتضمن قائمة البيانات في الجهة اليمنى وتتكون بدورها من جزأين الأول يتضمن البيانات الموجودة في مكعب البيانات والثاني يتضمن وظيفة (موقع ظهور) البيانات المرغوبة في الجدول المحوري (في السطر، أو العمود، أو كقيم). ويختص القسم الأساسي الثاني باستعراض الجدول (المخطط) المحوري، الذي يُظهر البيانات التي تُطلب في قائمة البيانات، ويمكن تطوير المخرجات بمرونة كبيرة بحسب احتياجات المستفيدين منها، انظر الشكل (5-3-54).

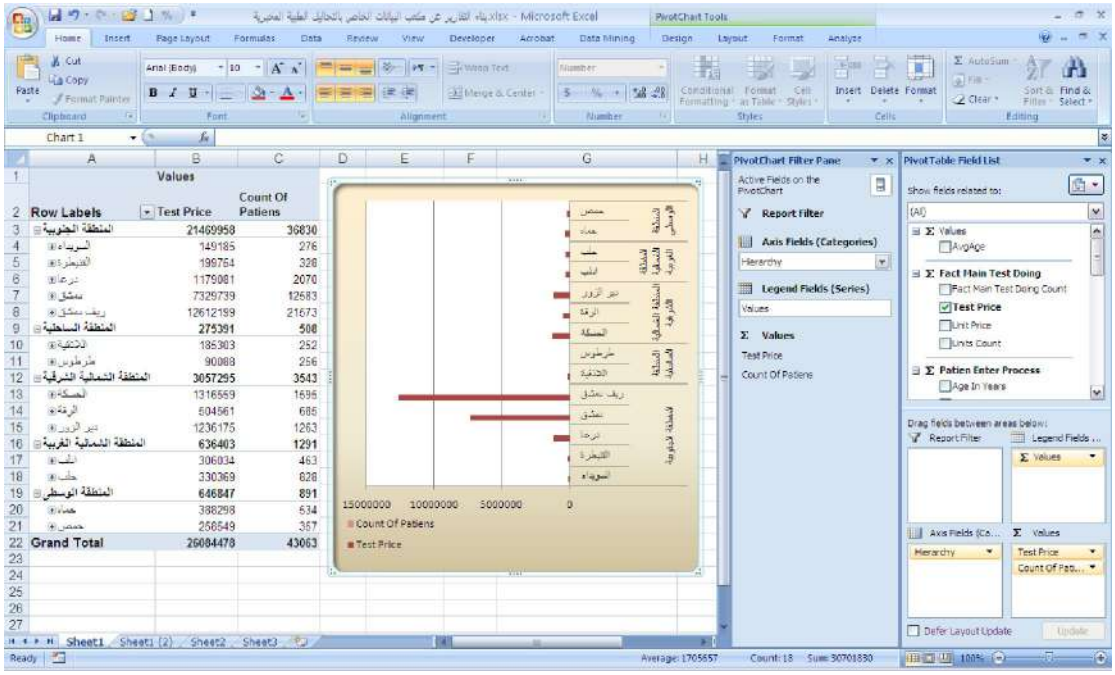


الشكل (– –): تعريف الاتصال في برنامج اكسل مع مكعب بيانات التحاليل الطبية الذي تم إنشاؤه سابقاً في برنامج ( )، واختيار طريقة عرض تقرير البيانات (جدول محوري، مخطط محوري، اتصال فقط)



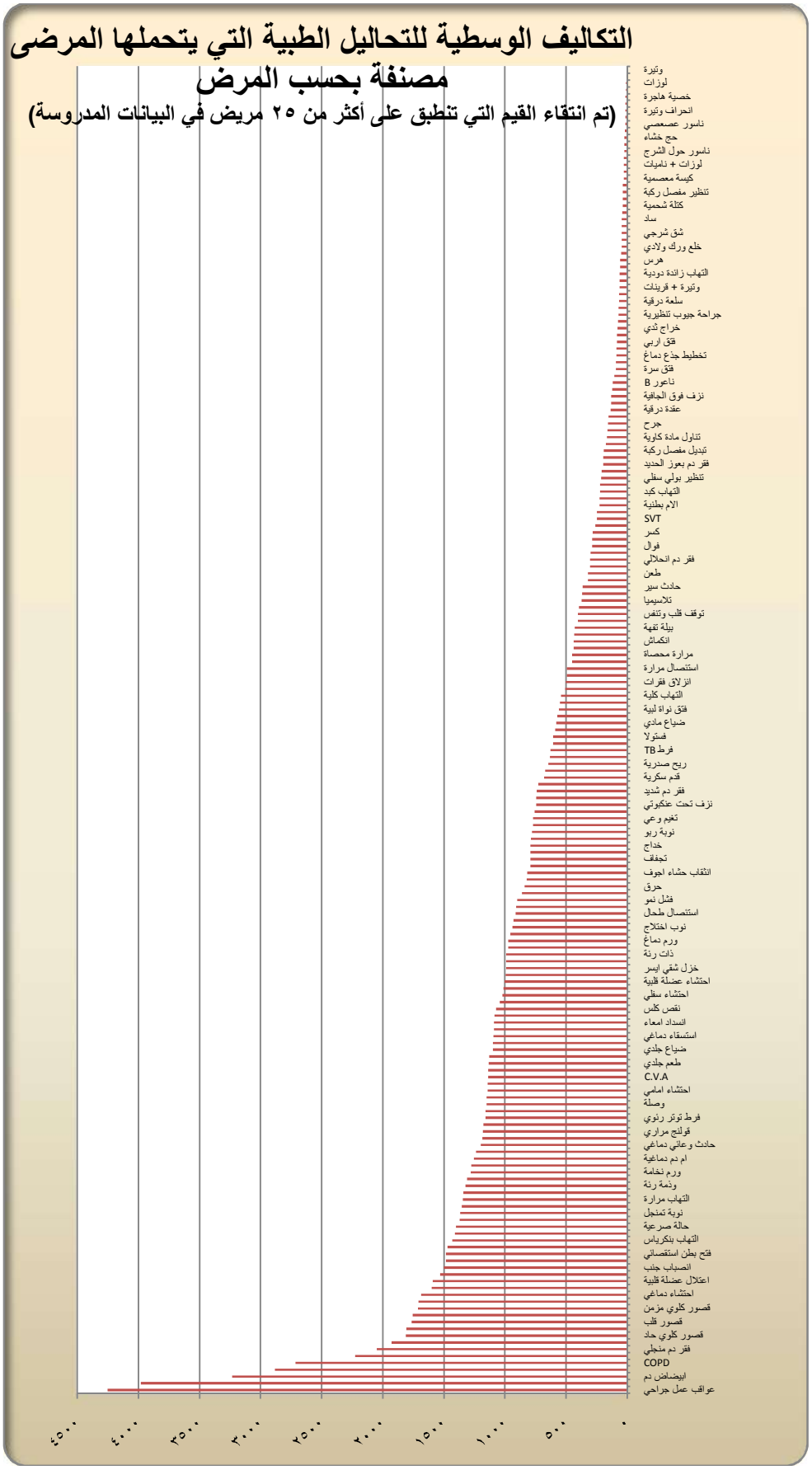
الشكل (5–3– 54): مكونات ورقة العمل الناتجة عن اتصال اكسل مع مكعب بيانات التحاليل الطبية في مشروع المعالجة التحليلية الآنية

وقد تم على سبيل المثال استعراض التوزع الجغرافي (وفق التراتب الهرمي للموقع الجغرافي) لأعداد المرضى وتكاليف التحاليل التي دفعوها إلى المشفى، وذلك بشكل جدول محوري وبشكل مخطط بياني كما في الشكل (5-3-55).



الشكل (5-3- 55): مثال عن استخدام اكسل في استعراض مكعب بيانات التحاليل الطبية

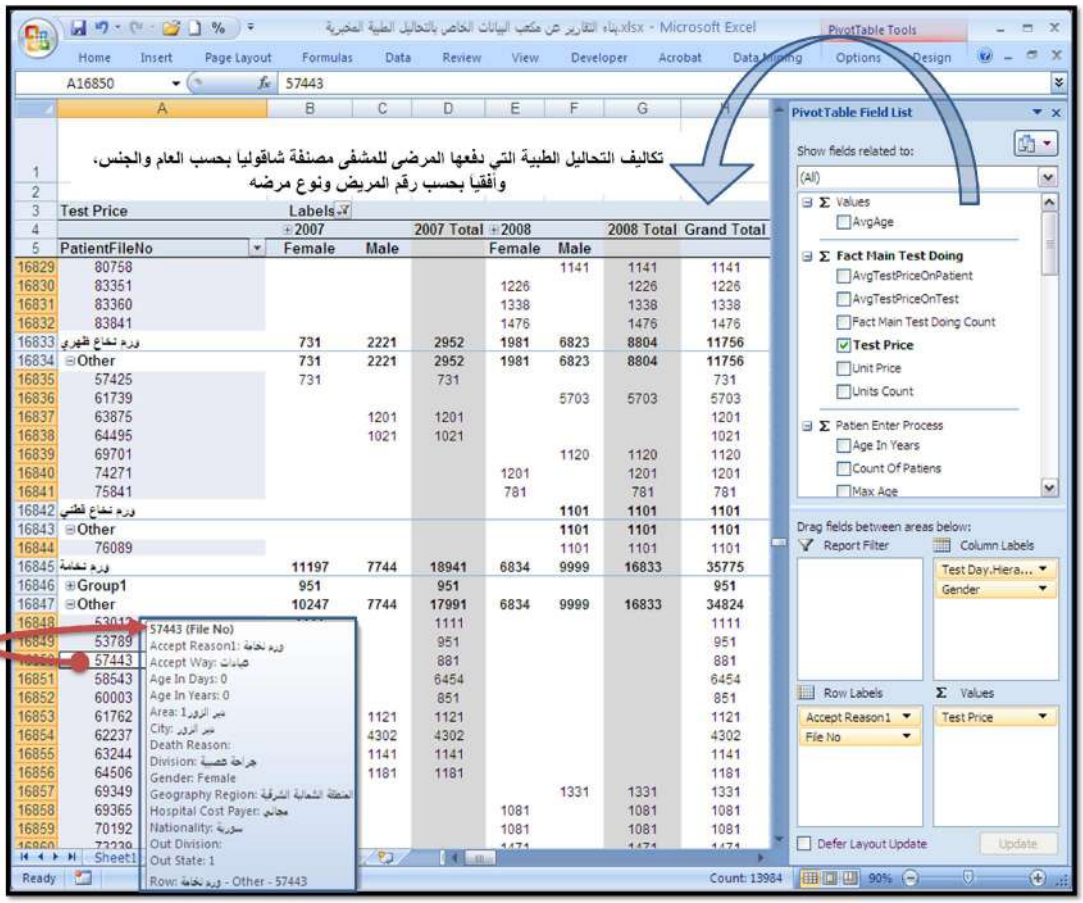
وتعتبر مخرجاتُ عمليات المعالجة التحليلية الآنية إما كمعلومات مفيدة بشكل مباشر للجهات المستفيدة منها، أو كمواد أولية يمكن إضافة بعض العمليات والمعالجات عليها وتحويلها بعد ذلك إلى معلومات مفيدة في برنامج اكسل أو في البرامج الإحصائية الأخرى، كأن يتم مثلاً حساب وسطي تكاليف التحاليل الطبية التي تحملها كل مريض بحسب المرض وعرض النتائج في مخطط بياني كما في الشكل (5-3-56). كما يمكن أن تُستخدم مخرجات هذه التقارير في إجراء بعض عمليات التنقيب فيها.



الشكل (5-3- 56): الاستفادة من مخرجات المعالجة التحليلية الآنية في حساب التكاليف الوسطية للتحاليل الطبية التي يتحملها المرضى مصنفة بحسب المرض

تبين من خلال البيانات المدروسة أن وسطي تكاليف مجموعة التحاليل الطبية التي تُطلب من المرضى في كل مرة هي 605 ل.س، أما وسطي تكاليف التحاليل الطبية للمريض خلال فترة الإقامة في المشفى هي 1851 ل.س.

وتساعد المعالجة التحليلية الآنية أيضاً في إظهار جداول المقارنات المتعددة الأبعاد، فمثلاً يتضمن الاستعلام متعدد الأبعاد في الشكل (5-3-57) تكاليف التحاليل الطبية التي دفعها المرضى في كل عام وحسب الجنس وحسب نوع المرض. ويلاحظ أن برنامج اكسل يقوم بعرض المحتويات التفصيلية للخلايا بمجرد المرور بمؤشر الفأرة فوق أي خلية من خلايا الجدول المحوري.



الشكل (5-3- 57): الاستعلام متعدد الأبعاد عن تكاليف التحاليل الطبية التي دفعها المرضى للمشفى مصنفة شاقولياً بحسب العام والجنس، وأفقياً بحسب نوع المرض وأرقام أضابير المرضى الذين تعرضوا للإصابة بكل من هذه الأمراض

نلاحظ مما سبق أن الاستعلامات متعددة الأبعاد هي عمليات غير منتهية على البيانات، ويمكن إجراء عدد كبير جداً من الاستعلامات بحسب احتياجات الشخص المستفيد من هذه البيانات.
وتفيد تقانة المعالجة التحليلية الآنية في التعرف على البيانات واستعراضها من الأبعاد المختلفة ومن ثم تحويلها إلى معلومات تلخيصية مفيدة عن البيانات، كما يمكن انتقاء وتصفح وتحليل أي جزء من أجزاء مكعب البيانات باستخدام أدوات هذه التقنية.

وتعتبر عمليات المعالجة التحليلية الآنية كإحدى الركائز الأساسية التي يمكن الاستعانة بها في فهم البيانات، وتحويلها إلى بيانات أكثر جودة وأكثر فاعلية في خدمة المستفيدين منها، كما يمكنهم إخضاع مخرجات هذه الاستعلامات إلى المعالجات الإحصائية اللاحقة التي تساعدهم في حل المسائل التي لا تستطيع المعالجة التحليلية الآنية أن تجيبهم عليها. والأهم من ذلك أن عمليات المعالجة التحليلية الآنية تساعد في التعرف على بعض المسائل التي يُحتاج فيها إلى استخدام تقانة التنقيب في البيانات كامتداد لما قدمته أدواتها من حلول، وسيظهر هذا الأمر بشكل أوضح أثناء مرحلة التنقيب في البيانات.

الفصل الخامس

المبحث الرابع

مرحلة التنقيب في بيانات مخبر التحاليل الطبية

مقدمة

تعرَّفنا في الفصل الرابع من هذه الدراسة على الأسلوب التقليدي للأبحاث الطبية التي تعتمد على نتائج التحاليل الطبية في إتمام بعض جوانبها، ويعتمد هذا الأسلوب باختصار على وجود فرضيات طبية ووجود مرضى مناسبين لهذه الدراسة، وتُجرى لهم الفحوصات السريرية وبعض التحاليل الطبية المحددة، ويتم تتبع حالاتهم وتعبئة بياناتهم في استمارات إحصائية ومن ثم يتم اختبار الفرضيات وعرض النتائج. ولكن هل يمكن اكتشاف جميع العلاقات بين الأمراض والتحاليل، أو بين التحاليل مع بعضها بالأساليب التقليدية؟؟ إن الإجابة على هذا السؤال في العصر الحالي هي باختصار.. لا لا يمكن!! أو قد يطول الأمر كثيراً!! ولكن وجود البيانات الهائلة المخزنة في الحواسيب أدى إلى تطور أساليب التفكير وإلى تطور أساليب معالجة البيانات بغية التعرف على المعلومات الكامنة في كل هذه البيانات.. وتعتبر أدوات التنقيب في قواعد البيانات إحدى هذه الأساليب التي تستطيع أن تساعد على أداء هذه المهمة، وأن تزودَ أيضاً الأساليبَ التقليدية بالمعلومات والفرضيات.

تُعتبر المراحل السابقة من عملية اكتشاف المعرفة المخبأة في بيانات مخبر التحاليل الطبية مراحل تمهيدية لمرحلة التنقيب في البيانات، ويُنظر إلى مرحلة التنقيب في البيانات على أنها لب عملية اكتشاف المعرفة لأنها المرحلة التي تجعل البيانات المنتقلة من المراحل السابقة تستيقظ وتخرج للحياة والعمل وترى النور أخيراً بعد أن كانت تنام في أعماق الحواسيب وتحديداً أعماق قواعد البيانات.

وقد وقع الاختيار في هذا البحث على تطبيق خوارزميات مايكروسوفت للتنقيب في البيانات من خلال استخدام أدوات شركة مايكروسوفت البرمجية أي (SQL Server 2008) و(Microsoft Excel 2007)، وذلك لأنها أدوات متطورة بشكل جيد وأيضاً متوفرة للاستخدام في أسواقنا المحلية بأحدث نسخها، وأيضاً لأنها أدوات صديقة للمستخدمين المختصين وغير المختصين بالهندسة المعلوماتية. ويختص برنامج (Microsoft Visual Studio 2008) باستخدام أدوات مايكروسوفت لانجاز عمليات التنقيب في البيانات، ويُساعد هذا البرنامج في إجراء عمليات التنقيب ضمن مجموعة مشاريع خدمات التحليل (Analysis Services Projects) التابعة لنمط مشاريع ذكاء الأعمال (Business Intelligence Projects) والتي اُستخدمت سابقاً في عمليات المعالجة التحليلية الآنية لبيانات التحاليل الطبية.

توزعت عملياتُ التنقيب في مستودع بيانات التحاليل الطبية ضمنَ مشروعين أساسيين للتنقيب، يختص المشروع الأول بالتنقيب في البيانات الشخصية والمرضية للمرضى، ويختص المشروع الثاني بالتنقيب في بيانات التحاليل الطبية المخبرية. وقد تم تنفيذ كل من هذه المشاريع وفق خمس خطوات وهي خطوات مرحلة التنقيب في البيانات المعروضة في الفصل الثالث من هذه الدراسة. كما وتمت إضافة نموذجين إلى المشروع الثاني يختصان بالتنقيب في البيانات الشخصية والمرضية ونتائج التحاليل الطبية باستخدام واجهات برنامج اكسل.

1•4•5 مشروع التنقيب في البيانات الشخصية والمرضية لمرضى المشفى

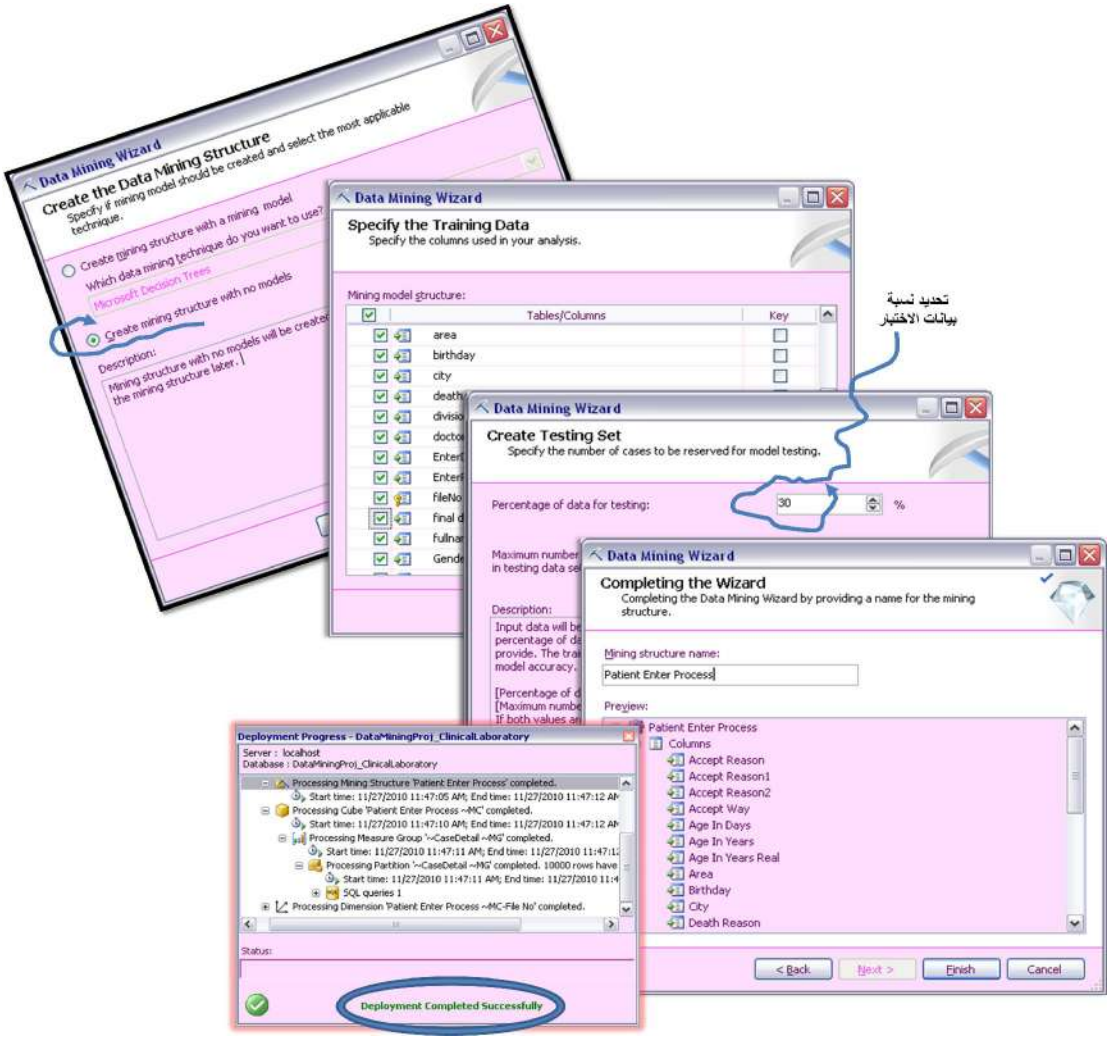
يدور هذا المشروع حول استخدام البيانات الشخصية والصحية للمرضى في عمليات التنقيب، وهو يُمثل صورة مصغرة عن عمليات التنقيب في السجلات الطبية أو الصحية للمرضى، ولذلك سيتم التطرق إلى الكثير من التفاصيل حول بناء هذا المشروع وشرح مخرجاته. وقد جرت محاولات كثيرة في إطار هذه الدراسة للحصول على نماذج مقبولة ومرضية من هذه البيانات إلى أن تم التوصل إلى مجموعتين من النماذج، وتختص المجموعة الأولى باكتشاف أسباب الوفاة الأكثر تأثراً بالخاصيات المختلفة للمرضى من خلال دراسة خصائص المرضى المرافقة لكل سبب من أسباب الوفاة، وتختص المجموعة الثانية بالتنبؤ بحالة خروج المرضى من المشفى. وسنعرض فيما يلي مراحل بناء هذا المشروع وأهم التفاصيل المرتبطة به.

1•1•4•5 إنشاء المشروع وتحديد مصدر البيانات

تم بدايةً استدعاء إجراء بناء مشروع جديد للتنقيب في البيانات في برنامج (Visual Studio)، وأُطلق على هذا المشروع اسم (DataMIningProj_ClinicalLaboratory)، ثم تم تعريف مصدر البيانات ضمن نافذة مكونات المشروع على أنه مستودع بيانات التحاليل الطبية (LocalHost.Clinical Laboratory Study) الذي أُنشئ سابقاً في مرحلة تحضير البيانات، وجرى بعد ذلك استدعاء الجدول (PatientEnterProcess) الموجود ضمن هذه القاعدة لاستخدامه كمرأى لبيانات التنقيب، ويشتمل هذا المرأى على بيانات المرضى الشخصية وبيانات الدخول والخروج من المشفى، ويبلغ عدد سجلاته 43063 سجل.

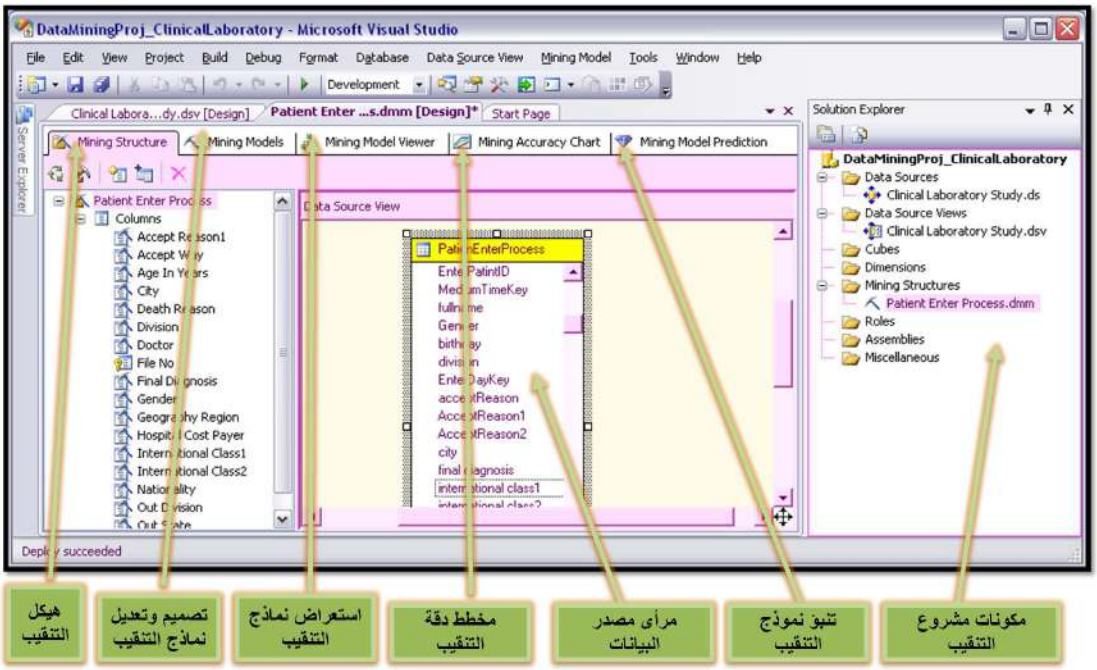
2•1•4•5 بناء هيكل التنقيب

تألفت عمليات بناء هيكل التنقيب من عدة خطوات متسلسلة، وقد تم أولاً استدعاء إجراء بناء هيكل التنقيب الموجود ضمن قائمة مكونات مشروع التنقيب، وجرى اختيار إنشاء هيكل عام للتنقيب في البيانات دون التخصص بنموذج تنقيب معين، فظهرت قائمة الجداول المعرفة في مرأى البيانات لاختيار ما يناسب منها، وقد احتوت هنا فقط على جدول واحد هو جدول البيانات الشخصية والمرضية للمرضى، تلتها قائمة بخاصيات هذا الجدول، حيث جرى الانتقاء الأولي للحقول (خاصيات البيانات) التي ستُستخدم في بناء نماذج التنقيب، علماً أنه يمكن لاحقاً إجراء أي تعديل على هذه الحقول (كالحذف، والإضافة، وتغيير الخواص..) بحسب الضرورات التي تظهر أثناء تطبيق نماذج التنقيب المختلفة. كما وقدِّرت في إحدى هذه الخطوات نسبة بيانات الاختبار (Testing Data) على أنها مساوية إلى (30%) من البيانات الموجودة في مرأى بيانات التنقيب، وتُشكل بالتالي بيانات التجريب (Training Data) ما نسبته (70%). وبعد الانتهاء من بناء الهيكل الأساسي تم استدعاء الإجراء الخاص بتدقيق وبناء وتفعيل الطلبات السابقة (Debug, Build & Deploy)، وهو الإجراء الذي يجب استدعاؤه باستمرار عقب أية إضافات أو تعديلات في العمل، انظر الشكل (5-4-1).



الشكل (5-4-1): إجراءات بناء هيكل للتنقيب في بيانات التحاليل الطبية باستخدام البرنامج الحاسوبي (Microsoft Visual Studio)، وتقرير نجاح عملية التدقيق والبناء والتفعيل

ونذكر كمثال على التعديلات التي أجريت على خواص البيانات ذلك التعديل على خاصية العمر التي أخذت نمط المتغير المستمر (Continues) آلياً عند بناء هيكل التنقيب والتي لا تتلاءم مع نماذج التنقيب ولذلك تم تغيرها – في نافذة خواص حقل الاسم في بنية التنقيب– إلى نمط متغير متقطع ومجزأ إلى فئات (Discretized_ClustersMethod)... وبهذا تكون قد بُنيت الأساسات الأولية التي سترتكز عليها نماذج التنقيب في هذا المشروع، ويُبين الشكل (5-4-2) هيكلَ التنقيب الناتج، بالإضافة إلى النوافذ الأخرى من البرنامج التي تختص بالمساعدة في متابعة الخطوات القادمة مثل تصميم النماذج، وتقييمها.



الشكل (5-4- 2): هيكل التنقيب في بيانات مرضى مخبر التحاليل الطبية، ومضمون النوافذ المختلفة المتعلقة بعملية التنقيب

3•1•4•5 تصميم واستعراض نماذج التنقيب التي تهدف إلى اكتشاف أسباب الوفاة الأكثر تأثراً بالخاصيات المختلفة للمرضى

اتصفت سجلات المرضى المتوفين في المشفى باكتمال معظم البيانات عنها وخصوصاً المستمدة من قاعدة بيانات مكتب السجلات الطبية، أي لا يوجد سوى عدد قليل جداً من القيم المفقودة، وذلك على عكس الصفة الغالبة على سجلات هذه القاعدة كما سبق وذكرنا في مرحلة استيعاب البيانات. وبناءً على ذلك فإنه سيتم التركيز في نماذج التنقيب الحالية على بيانات المرضى المتوفين فقط، حيث تساعد قلة القيم المفقودة هنا على الاستفادة من أغلبية الحقول المتعلقة بخروج المرضى من المشفى، مثل التصنيف الدولي للأمراض المسمى في قاعدة البيانات المدروسة بـــ (International Class1)، والتصنيف الدولي لأسباب الوفاة المسمى بـــ (International Class2).

تم في هذه الخطوة تصميم واختبار عدد من نماذج التنقيب، ولكن ثلاثةً منها كانت الأفضل نسبياً في معالجة القضية المطروحة. وقد تم تصميم كل من نماذج التنقيب باستخدام إجراء (New Mining Model) ضمن النافذة الخاصة بتصميم نماذج التنقيب (Mining Models)، ويختلف كل نموذج من النماذج الثلاثة المعتمدة عن غيره من النماذج باختلاف خوارزمية

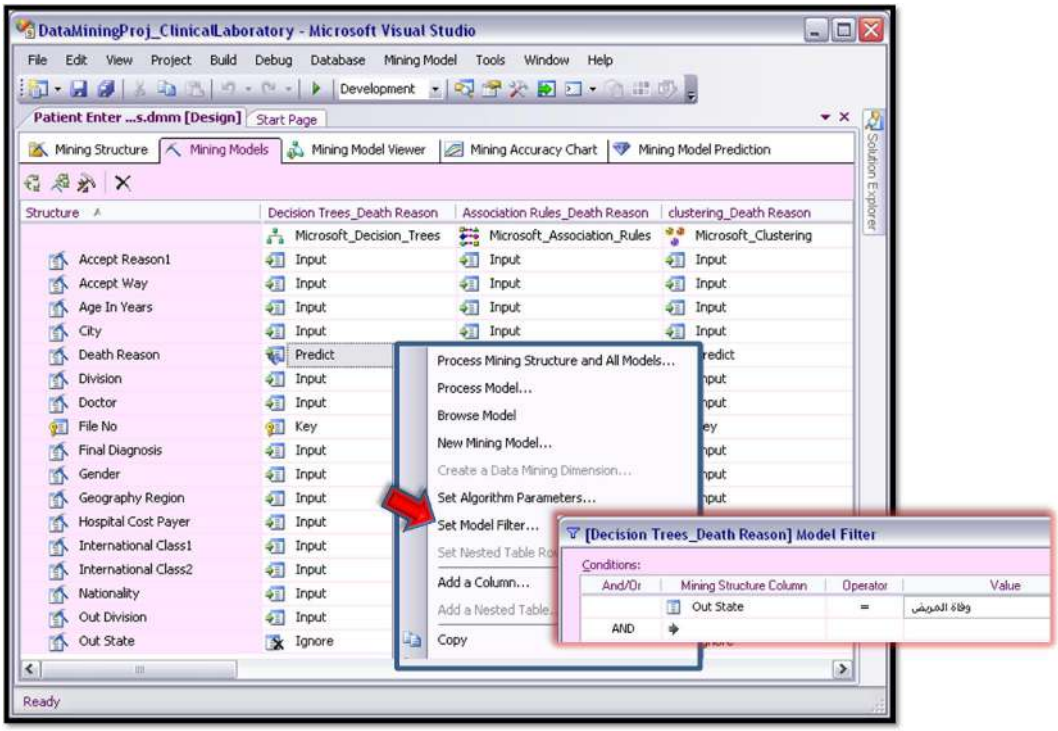
التنقيب المستخدمة، ويتشابه مع غيره في مهام حقول البيانات التي يشتمل عليها. وتتوزع مهام حقول البيانات على الشكل التالي:

1- أخذ حقل سبب الوفاة (Death Reason) دور المتغير المتنبأ به (Predict)

2- حمل حقل حالة الخروج من المشفى صفة التجاهل (Ignore)، وقد استُخدم هذا المتغير لانتقاء فقط الحقول التي تكون فيها حالة الخروج هي الوفاة، انظر الشكل (5-4-3).

3- أخذت بقية حقول هيكل التنقيب دور متغيرات الإدخال (Input)، مثل حقل سبب القبول.

وبهذا بلغ عدد السجلات المنتقاة للمرضى الذين توفوا (814) سجل ولها 28 حقل (خاصية)، وقد قام البرنامج الحاسوبي بتقسيم هذا العدد آلياً وداخلياً إلى مجموعتين هما مجموعة التجريب التي تُستخدم في استنتاج النموذج وتشكل 70% تقريباً من هذه السجلات (566) سجل، ومجموعة الاختبار التي تُستخدم في التحقق من أداء النموذج وتشكل 30% تقريباً (248) سجل.

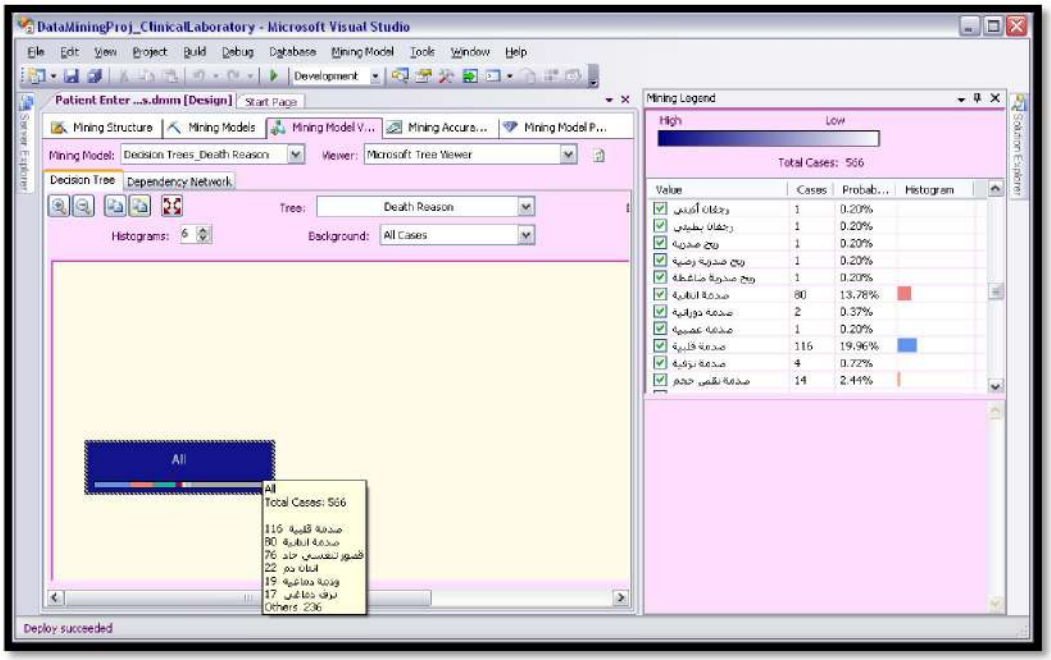


الشكل (5-4- 3): نافذة نماذج التنقيب في بيانات المرضى للتنبؤ بأسباب الوفاة، وقائمة انتقاء فقط البيانات المتعلقة بوفيات المرضى

وبعد الانتهاء من كل أو بعض الاختيارات السابقة كان يتم استدعاء إجراءات تدقيق وبناء وتفعيل هذه النماذج (Debug, Build & Deploy)، وذلك للتمكن من استعراض النتائج، ونبين فيما يلي نماذج التنقيب الناتجة.

1•3•1•4•5 نموذج أشجار القرار (DecisionTrees_DeathReason)

يعتمد نموذج التنقيب الأول على خوارزمية شجرة القرار، وقد تبين في نافذة عرض هذا النموذج أن البيانات المتوفرة لم تستطع أن تقدم سوى مستوى واحد في شجرة القرار، ويتكون هذا المستوى من أسباب الوفاة بحسب نسبة المتوفين في كل حالة، انظر الشكل (5-4-4). وبناءً على ذلك نجد أن هذا النموذج قليل الفائدة في تقديم معارف تفصيلية جديدة بالنسبة للمسألة المدروسة.



الشكل (5-4- 4): نموذج أشجار القرار (DecisionTrees_DeathReason) للتعرف على خصائص المرضى المتوفين بالأسباب المختلفة للوفاة

2•3•1•4•5 نموذج قواعد الاقتران (AssociationRules_DeathReason)

يستند نموذج التنقيب الثاني إلى خوارزمية قواعد الاقتران، ويتألف هذا النموذج من ثلاث نوافذ لاستعراض المعرفة المكتشفة، وهي:

أ- نافذة (Rules) وتتضمن قواعد الاقتران المكتشفة، واحتمال كل منها (Probability) وأهمية كل قاعدة (Importance)، انظر الشكل (5-4-5)

ب- نافذة شبكة التبعية (Dependency Network) التي تُظهر الروابط بين قيم الخاصيات في أهم القواعد المكتشفة وشدة هذه الروابط، لاحظ الشكل (5-4-6).

ج- نافذة مجموعات الاقتران (Itemsets) التي تَعرض مجموعات البيانات بحسب عدد عناصرها (Size) وبحسب تكرارها في البيانات (Support)، كما في الشكل (5-4-7).

تُقاس أهمية القواعد المكتشفة في برنامج (Visual Studio) باستخدام مقياس الأهمية (Importance) إضافةً إلى مقياسي الاحتمال والدعم المذكورين في الفصل الثالث. وينتُج هذا المقياس من خلال حساب نسبة الحالات التي تتضمن جميع عناصر طرفي القاعدة إلى الحالات التي تتضمن عنصر النتيجة في الطرف الأيمن ولا تتضمن عناصر الطرف الأيسر، ثم تسوى (normalizes) النسبة باستعمال ترجيح لوغاريتمي. وبشكل عام فإن هذا المقياس يدل على الأهمية المقارنة للقواعد، حيث يمكن التعرف على أهمية كل قاعدة من مقارنتها بغيرها.

ونورد فيما يلي مجموعة من الحقائق المكتشفة في هذا النموذج والمستنتجة من النوافذ الثلاثة.

1- أظهرت شبكة التبعية أن الأسباب الرئيسة الأولى في الوفاة، والتي يمكن استنتاجها من خصائص المرضى في البيانات المدروسة هي القصور التنفسي الحاد والصدمة الإنتانية والصدمة القلبية.
2- تتعلق قواعد الاقتران الستة الأكثر احتمالاً (نافذة (Rules)) بالخصائص المقترنة مع سبب الوفاة بالقصور التنفسي الحاد، وتتلخص هذه القواعد الستة بأن مشكلة داء الأغشية الهلامية ذات التصنيف الدولي الأساسي (P22.0) والتصنيف الدولي للوفاة (4-029)[1] والتي تحدُث للمرضى في شعبة الحواضن والقادمين من المنطقة الجنوبية تؤدي إلى الوفاة بالقصور التنفسي الحاد بالتأكيد (الاحتمال=1)، وتقدر أهمية هذه القواعد بـــ (0.878) كما في الشكل (5-4-5). وقد تبين في نافذة (ItemSets) أن عدد المرضى الذين انطبقت عليهم هذه القواعد يتراوح بين 9-10 مرضى، انظر الشكل (5-4-6).
3- أظهرت شبكة الاقتران أن سبب الوفاة بالقصور التنفسي الحاد يقترن مع أسماء اثنين من الأطباء في المشفى، ولكن لا يمكن مطلقاً أن يُقيِّم حالياً أداء الأطباء في ضوء البيانات المتاحة حتى لو ارتبط اسمهم بوفاة المرضى.. ولكن جرى اختيار دراسة تأثير حقل اسم

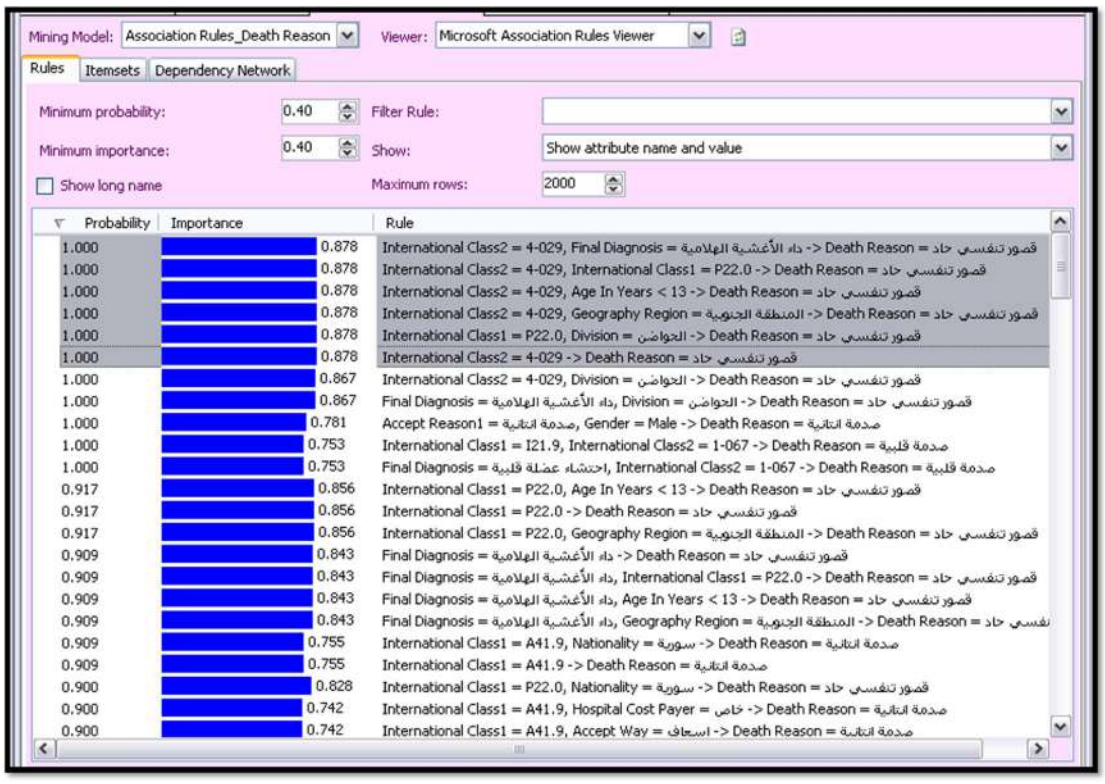
[1] وفق الترميز المتبع في قاعدة بيانات السجلات الطبية في المشفى، ولكن لم يتم العثور على مقابل اسمي له في قائمة نظام ترميز أسباب الوفاة [www.mohp.gov.eg (ICD-10)]. ويقع التحقق من مدلولات وصحة الرموز خارج نطاق هذه الدراسة.

الطبيب المعالج لإظهار أنه يمكن أن تساعد أدوات التنقيب في البيانات على تقييم أساليب العلاج وأداء الأطباء والممرضين ومختلف الكوادر التي تتعامل مع المرضى، ولكن يحتاج الأمر إلى المزيد من خاصيات البيانات التي لا تدعمها قاعدة البيانات المتاحة لدراستنا.

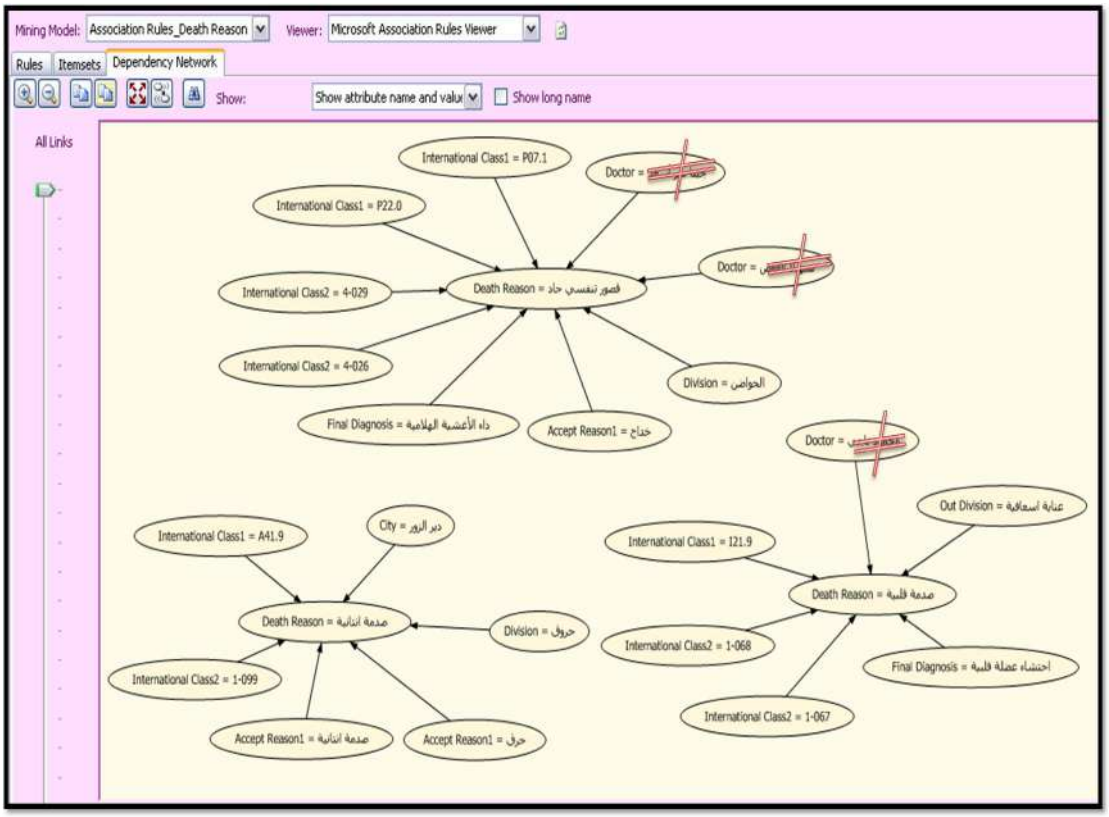
4- كشفت قواعد الاقتران أنه إذا كان سبب القبول في المشفى هو الصدمة الإنتانية، وكان جنس المريض ذكر وتوفي هذا المريض فإن سبب الوفاة هو الصدمة الإنتانية باحتمال قدره 1، وتقدر أهمية هذه القاعدة بـــ 0.781.

5- كشفت قواعد الاقتران أنه عندما يكون جنس المريض أنثى، ويزيد عمرها عن السبعين عاماً، وقُبلت في المشفى في شعبة الداخلية الصدرية، فإن احتمال أن يكون سبب وفاتها هو القصور التنفسي الحاد يتراوح بين (0.941–0.944). وتتراوح أهمية هذه النتيجة بين (0.925–0.931)، أي أنها نتيجة هامة وهي الأكثر أهمية مقارنةً ببقية القواعد المكتشفة.

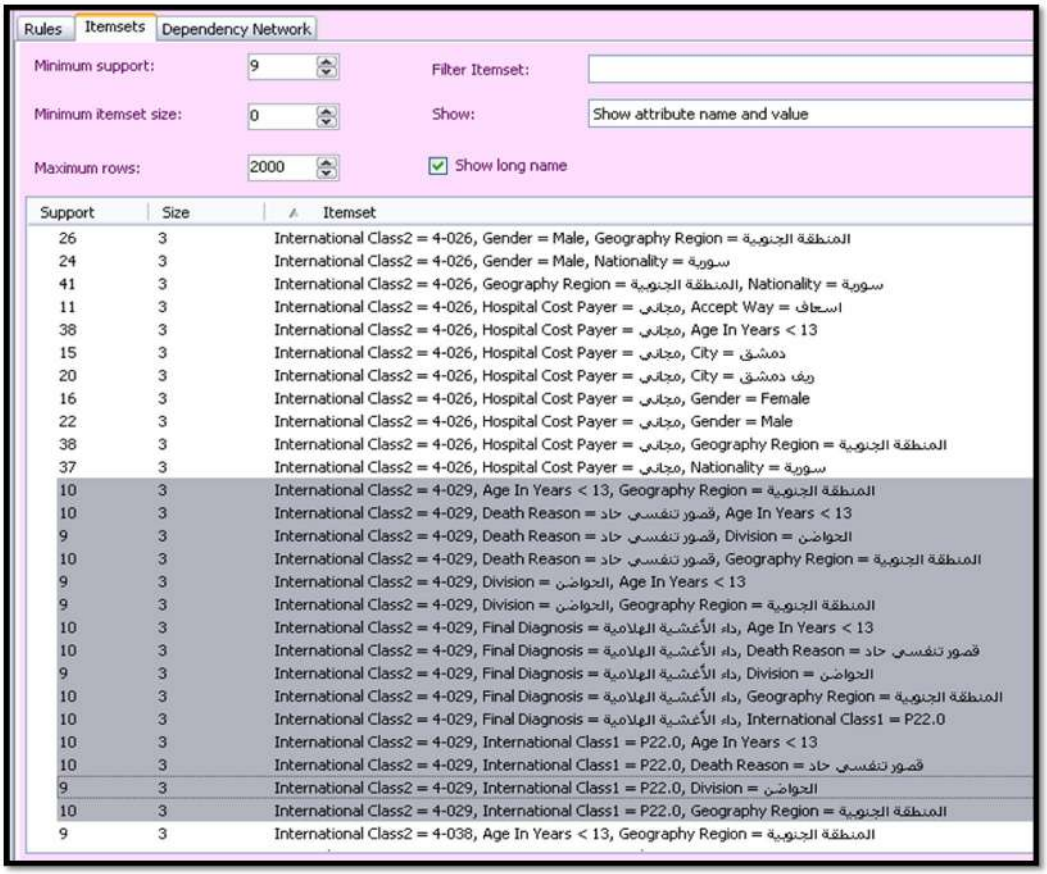
6- أظهرت شبكة التبعية أن سبب الوفاة بالصدمة الإنتانية يتأثر بسبب القبول بالحروق وعند المرضى القادمين من مدينة دير الزور.



الشكل (5–4– 5): نافذة قواعد الاقتران في نموذج قواعد الاقتران (AssociationRules_DeathReason)



الشكل (– –): نافذة شبكة التبعية في نموذج قواعد الاقتران ( )



الشكل (5-4- 7): نافذة مجموعات الاقتران في نموذج قواعد الاقتران (AssociationRules_DeathReason)، ويظهر فيها هنا قيم خاصيات البيانات المترافقة في المجموعات المؤلفة من 3 خاصيات (Size=3)، وتكرار كل مجموعة في بيانات الاختبار (Support)

3•3•1•4•5 نموذج التجزئة العنقودية (Clustering_DeathReason)

يستند نموذج التنقيب الثالث إلى خوارزمية التجزئة العنقودية، ولا تشترط هذه الخوارزمية وجود حقل متنبأ به لأنها خوارزمية تصنيفية، وقد قامت هذه الخوارزمية بتقسيم البيانات داخلياً في هذا النموذج إلى عشر مجموعات مختلفة الأحجام بحيث تمتلك العناصر في كل مجموعة صفات متشابهة، ويحتوي هذا النموذج على أربعة نوافذ لاستعراض النتائج، وذلك على النحو التالي:

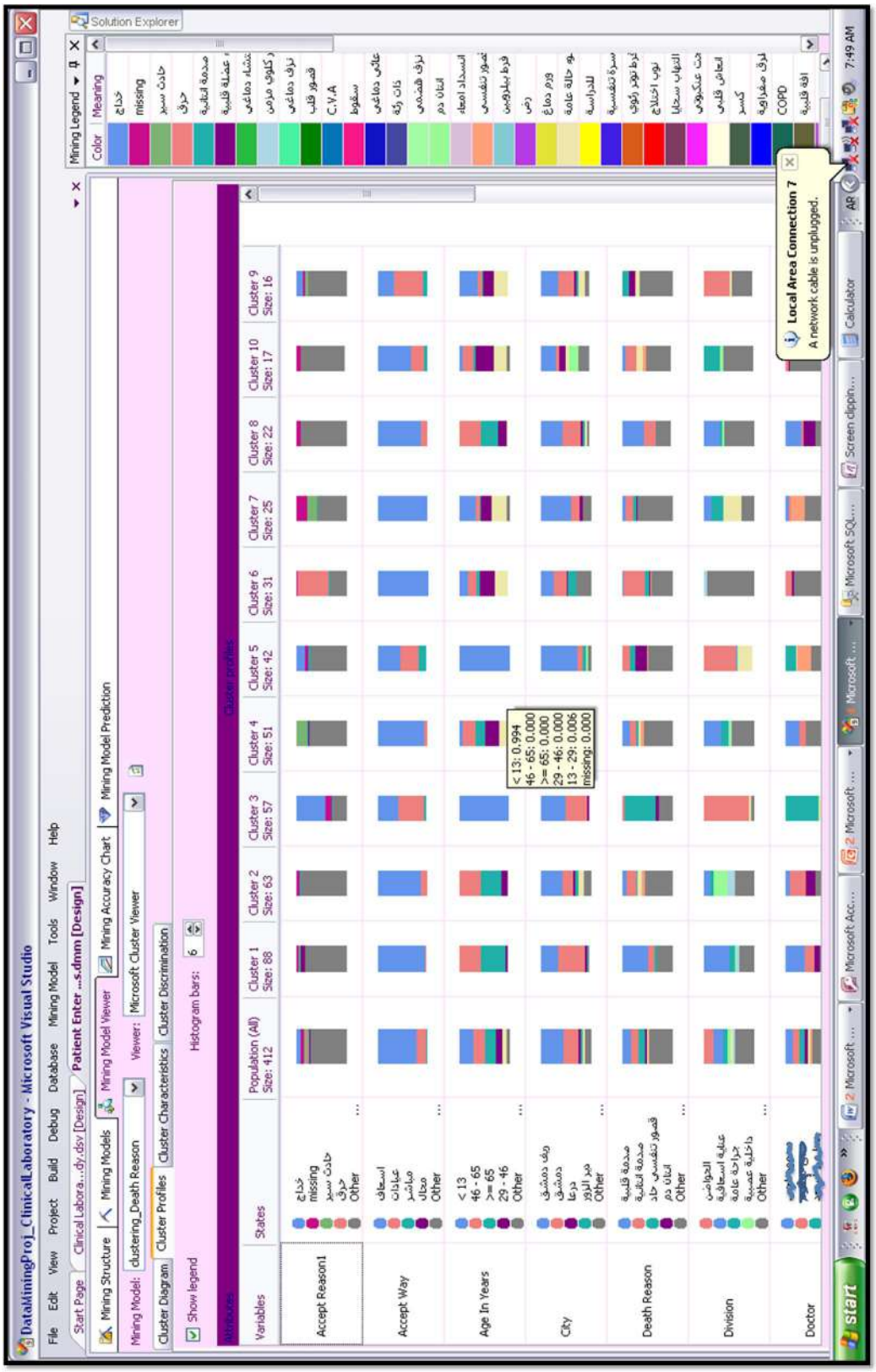
أ- نافذة خصائص عقد العنقود (Cluster Profile)

تفيد نافذة خصائص العقد في الاطلاع السريع على مكونات كل مجموعة أو عقدة، واحتمال ظهور قيم الخاصيات في كل منها، فمثلاً يُقدَّر في المجموعة الثالثة احتمال أن يكون عمر المريض أقل من 13 سنة بقيمة (1)، ويُقدَّر احتمال أن يكون مقيماً في دمشق بقيمة (0.568)، وأن يكون مقيماً في ريف دمشق (0.388)، ويُقدَّر احتمال أن يكون سبب وفاته هو القصور التنفسي الحاد بقيمة (0.501)، كما ويُقدَّر احتمال أن يكون في شعبة الحواضن هو (0.911).. وهكذا، انظر الشكل (5-4-8).

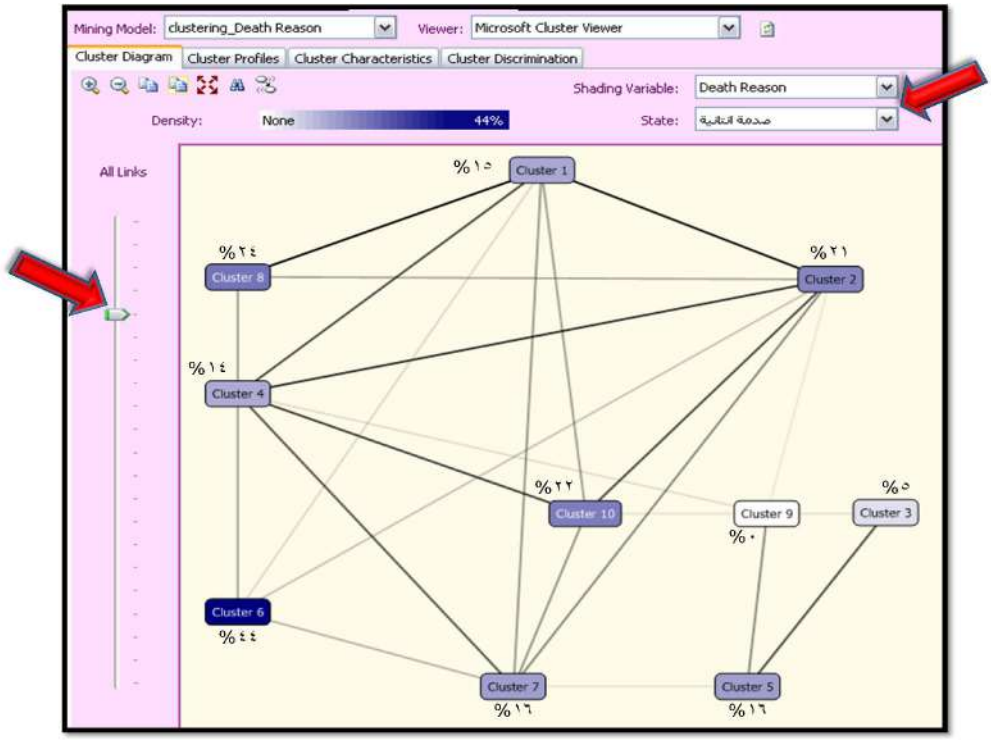
ب- نافذة الرسم البياني العنقودي (Cluster Diagram)

تتضمن نافذة الرسم البياني العنقودي (Cluster Diagram) شبكة العقد التي يتكون منها نموذج التجزئة العنقودية. وترتبط هذه العقد معاً بخطوط مستقيمة، وتُعبِّر درجة لون الخط الواصل بين عقدتين في هذه النافذة عن قوة التقارب أو التشابه بين العقدتين، فكلما كان لون الخط غامقاً كلما دل على أن التقارب أقوى، وكلما كان لون الخط فاتحاً دل على ضَعف التقارب بين العقدتين، فمثلاً نجد أن هنالك تقارباً بين العقدتين الأولى والثانية، وبين الأولى والثامنة. كما وتساعد هذه النافذة في التعرف على نسبة كثافة قيمة معينة للمتغير المتنبأ به في كل عقدة من العقد. فمثلاً تم اختيار التعرف على نسبة المتوفين بالصدمة الإنتانية في الشكل (5-4-9)، فتتبين أن 44% من أسباب الوفاة في العقدة السادسة هي بسبب الصدمة الإنتانية، بينما لا تحتوي العقدة التاسعة على هذا السبب ويمكن التعرف على شدة الكثافة من خلال درجة لون العقدة بالنسبة إلى بقية العقد، فكلما كان لون العقدة غامقاً كلما دل على ارتفاع الكثافة في هذا العقدة. ونشير هنا إلى أن نسب الكثافة في عقدة معينة تظهر فقط عندما يمر مؤشر الفأرة من

فوقها وتختفي بعد ذلك، وقد تمت لاحقاً كتابة النسب على النافذة الناتجة لضرورة إظهار اختلافات النسب بين العقد.

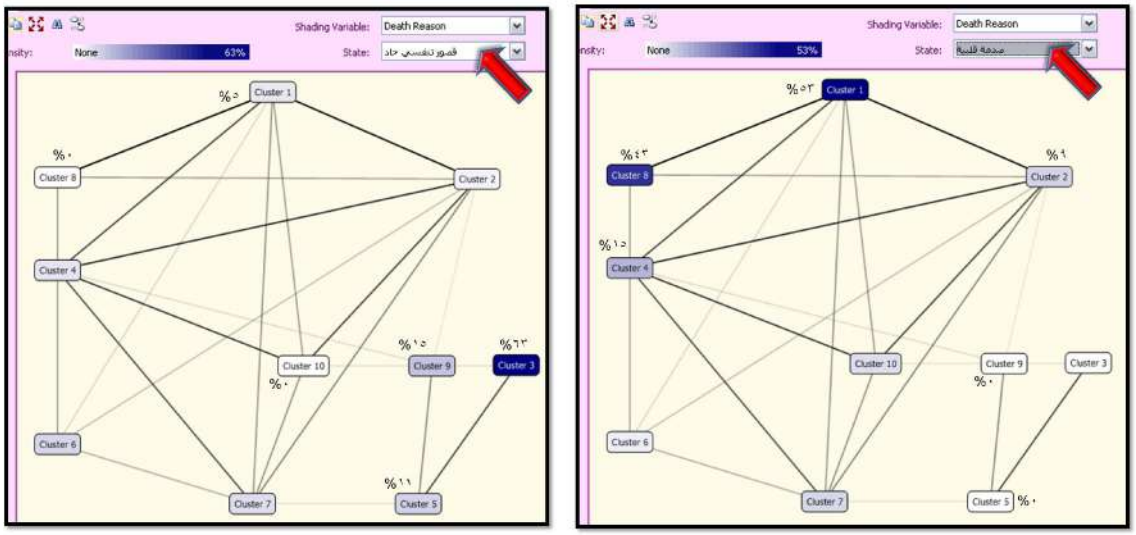


الشكل (5-4- 8): نافذة خصائص عقد العنقود (Cluster Profiles) في نموذج التجزئة العنقودية (Clustering_DeathReason)، وقد رُتِّبت المجموعات آلياً بحسب حجم كل منها



الشكل (– –): نافذة الرسم البياني العنقودي ( ) في نموذج التجزئة العنقودية ( )،
ويُظهر في هذا الرسم كثافة سبب الوفاة بالصدمة الإنتانية في كل عقدة من عقد العنقود بالنسبة إلى بقية الأسباب في ذات العقدة

ويتبين أيضاً أن سبب الوفاة بالصدمة القلبية يتواجد بنسبة %، % في كل من العقدتين الأولى والثامنة على التوالي، وأن سبب الوفاة بالقصور التنفسي الحاد يتواجد بنسبة % في العقدة الثالثة، انظر الشكل (– –).



الشكل (5–4– 10): نافذة الرسم البياني العنقودي (Cluster Diagram) في نموذج التجزئة العنقودية (Clustering_DeathReason)، ويظهر إلى اليمين كثافة سبب الوفاة بالصدمة القلبية والذي يتواجد بأعلى كثافة في العقدة 1، ويظهر إلى اليسار كثافة سبب الوفاة بالقصور التنفسي الحاد الذي يتجمع بأعلى كثافة في العقدة 3

أ- نافذة صفات العقدة (Cluster Characteristics)

تختص هذه النافذة بعرض صفات كل عقدة، وهذه الصفات هي عبارة عن قيم الخاصيات التي تتواجد في العقدة واحتمال ظهور كل قيمة بالنسبة إلى بقية قيم الخاصية. وتتصف هذه النافذة بالمرونة بحيث يمكن ترتيب الجدول فيها تنازلياً أو تصاعدياً بحسب قيم أي عمود فيها. وقد وقع الاختيار كمثال على هذه النافذة أن تُستعرض صفات العقدة 1 وهي أكبر عقدة من حيث عدد البيانات فيها (عودة إلى الشكل (5-4-8))، ويبين الشكل (5-4-11) صفات هذه العقدة والتي عُرضت بطريقتين، الأولى بحسب التسلسل التنازلي لاحتمالات صفات المرضى في العقدة، والطريقة الثانية بحسب الترتيب الأبجدي لأسماء الصفات، ونقرأ في نافذة صفات العقدة 1 المعلومات التالية:

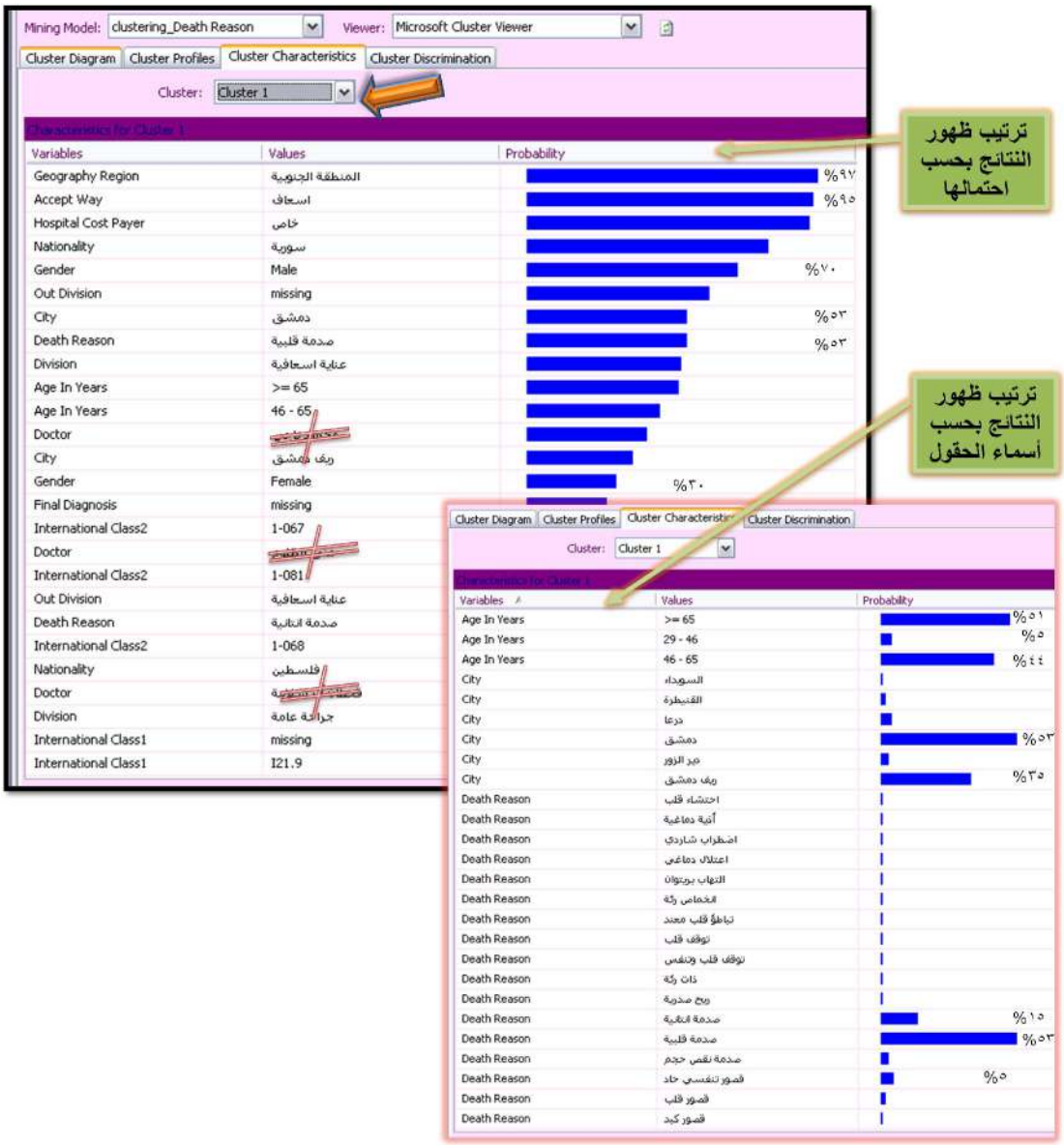
1- يُشكل سبب الوفاة بالصدمة القلبية 53% من أسباب الوفاة في هذه المجموعة.

2- يُشكل الذكور 70% من المرضى في هذه العقدة.

3- تتراوح أعمار 44% من عناصر هذه العقدة بين 46-65 سنة[1]، ويزيد عمر 51% منهم عن عمر 65 سنة، أي أن معظم المرضى في هذه المجموعة تزيد أعمارهم عن 46 سنة.

4- إن معظم مرضى هذه المجموعة (95%) أتوا بحالة إسعافية إلى المشفى من مناطق قريبة في دمشق (53%) وريف دمشق (35%).

[1] يعمل برنامج (Visual Studio) آلياً على تقسيم الأعمار إلى فئات عمرية بحسب ما يراه مناسباً لعمليات التصنيف، ويصطلح هذا البرنامج على أن الحد الأعلى لاسم كل فئة عمرية غير مشمول فيها بل ينتمي إلى الفئة التالية لها حيث يمثّل حدها الأدنى.



الشكل (5-4- 11): نافذة صفات العقدة (Cluster Characteristics) في نموذج التجزئة العنقودية (Clustering_DeathReason)، ويُظهِر الشكل صفات العقدة الأولى مرتبة تنازلياً بحسب قيمة الاحتمال في النافذة الأعلى، وبحسب أسماء الحقول في النافذة الأسفل

ب- نافذة تمييز العقد (Cluster Discrimination)

تسمح هذه النافذة بمقارنة قيم الخاصيات بين عقدتين، وهي تعرض أهم الاختلافات بين العقد من حيث قيم الخاصيات، وتُعطي من أجل كل قيمة للخاصية ترجيحاً للعقدة الأنسب إلى هذه القيمة، ويُوجد وراءَ كل ترجيحٍ قوةٌ يُعبَّر عنها بنسبة مئوية تتناسب مع طول الشريط المقابل للقيمة المرجَّحة لإحدى العقدتين (Favors Cluster (#)). ويُبين الشكل (5-4-12) مقارنةً بين العقدتين 1 و2، وينتج عن هذه المقارنة المعلومات التالية:

1- إذا كان سبب الوفاة هو الصدمة القلبية فإن المريض المتوفى ينتمي إلى مجموعة المرضى في العقدة 1 بنسبة ترجيحية 93% مقارنةً بانتمائه إلى العقدة 2.

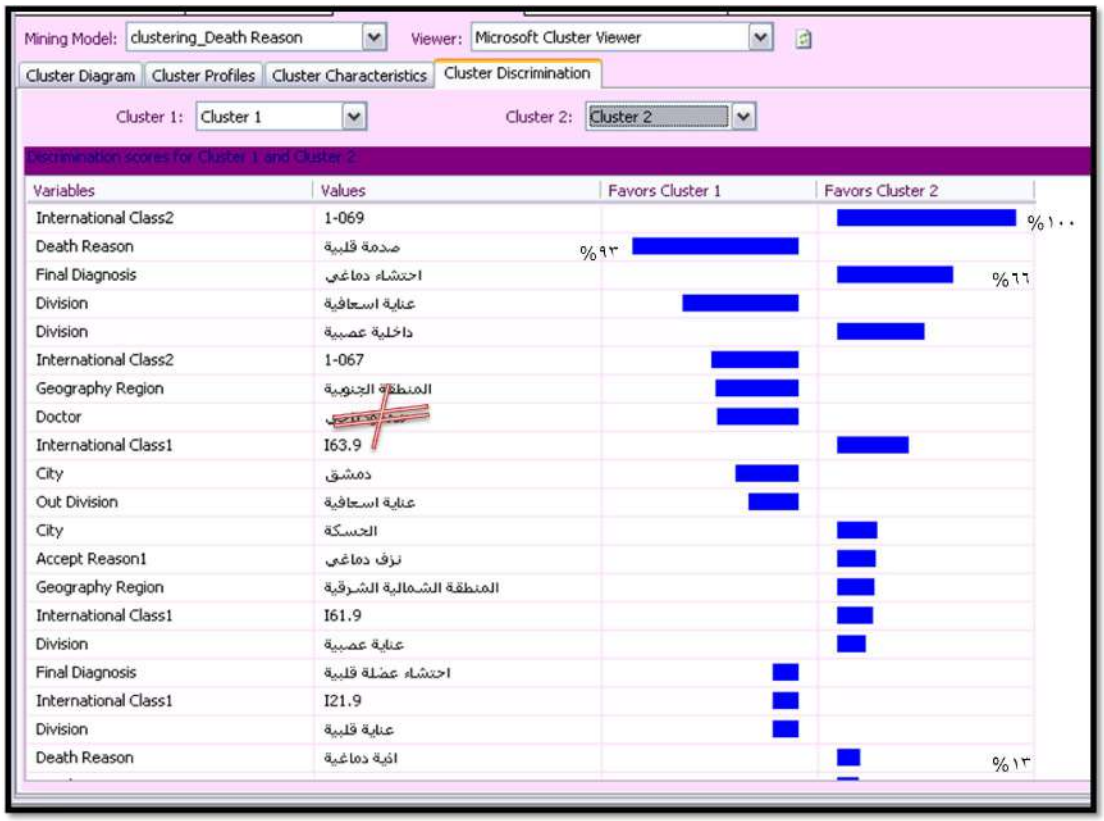
2- إذا كان التصنيف الدولي لسبب الوفاة هو (069-1) فإنه من المرجح بقوة 100% أن ينتمي المريض المتوفى إلى العقدة 2 وليس الأولى.

3- تساعد نتائج نماذج التنقيب[1] أحياناً في التحقق من بعض البديهيات أو الكشف عن أخطاء إدخال البيانات، أو التناقضات بين الحقول التي تحمل مدلولات متقاربة، وهي من الأمور المفيدة جداً في التحقق من جودة البيانات المدخلة إلى أنظمة عمل قواعد البيانات، وخصوصاً عند التعامل مع قواعد بيانات لا يدعم مصمموها الترابط المنطقي بين البيانات أثناء عمليات الإدخال. فمثلاً يَفرض المنطق أن يكون التصنيف الدولي لأسباب الوفاة في الخاصية (International Class2) متوافقاً مع أسباب الوفاة في الحقل (DeathReason)، ومتوافقاً تقريباً مع التشخيص النهائي للمرض ومع الشعبة الطبية..الخ. فإذا كان كذلك فإنه يدل على جودة قاعدة البيانات، وإلا اعتُبرت هذه القاعدة ضعيفة.

وتطبيقاً على هذه الفكرة نجد في المثال المطروح في الشكل (5-4-12) أن التصنيف الدولي ذا الرمز (069-1) يُرجح انتماء المريض المتوفى إلى العقدة 2، ويُشير هذا التصنيف إلى الوفاة بالأمراض الدماغية وأمراض أوعية المخ[2]، ونلاحظ أيضاً في هذا الشكل أن التشخيص النهائي بالاحتشاء الدماغي يُرجِّح بقوة 66% أن ينتمي المريض المتوفى إلى العقدة 2، وأن شعبة الداخلية العصبية تُرجِّح بقوة 50% أن ينتمي هذا الشخص إلى العقدة 2... وبذلك نجد أن هنالك توافق مقبول نسبياً بين الصفات المرجحة للعقدة 2، ولكن لا يمكن تعميم هذه النتيجة والحكم على بقية الإدخالات أنها متوافقة ما لم يتم التأكد من ذلك.

[1] تساعد معظم خوارزميات التنقيب على التحقق من التوافق بين قيم الخاصيات، وتم اختيار هذا النموذج كمثال على هذه الفكرة.

[2] www.mohp.gov.eg (نظام ترميز أسباب الوفاة ICD-10)



الشكل (5-4- 12): نافذة تمييز العقد (Cluster Discrimination) في نموذج التجزئة العنقودية (Clustering_DeathReason)، وقد جرى اختبار المقارنة بين العقدة 1 والعقدة 2

4•1•4•5 تصميم واستعراض نماذج التنقيب الخاصة بالتنبؤ بحالة الخروج من المشفى (الشفاء أو الوفاة)

يستند تصميم نماذج التنقيب الخاصة بالتنبؤ بحالة الخروج من المشفى إلى هيكل التنقيب الموجود في هذا المشروع، وقد تم انتقاء بيانات المرضى الذين يخرجون فقط بحالة شفاء أو وفاة من المشفى وعددهم 15430 مريض وذلك لأن تجربة إدخال بقية الحالات (التحويل إلى مشفى آخر، والقيم المفقودة) أدى إلى التقليل من دقة وفعالية النماذج في عمليات التنبؤ[1]. وتتوزع مهام حقول البيانات في هذه النماذج على النحو التالي:

1- أخذ حقل حالة الخروج من المشفى (OutState) دور المتغير المتنبأ به (Predict).

2- حمل كل من حقل سبب الوفاة (Death Reason) وحقل التصنيف الدولي لسبب الوفاة (International Class2) صفة التجاهل (Ignore)، لأنهما لا يُعتبران من الحقول المؤثرة

[1] تُعبِّر هذه الحالة عن العودة الراجعة من مرحلة تقييم المعرفة إلى مرحلة بناء نماذج التنقيب.

في التنبؤ الصحيح، وأن وجودهما يؤدي إلى تشويه النتائج[1].

3- أخذت بقية حقول هيكل التنقيب دور متغيرات الإدخال (Input).

وسنناقش فيما يلي أهم النماذج التي تم التوصل إليها للتنبؤ بحالة الخروج من المشفى وهي النماذج المستندة إلى كل من خوارزميات الشبكات العصبونية والانحدار المنطقي ونايف بايز.

1•4•1•4•5 نموذج نايف بايز (NaiveBayes_OutStae)

يساعد النموذج المستند إلى خوارزمية نايف بايز في التعرف على البيانات من خلال الاحتمالات البسيطة لقيم كل خاصية من خاصيات المدخلات المؤثرة في الخاصيات المتنبأ بها. وقد عُرِضت نتائج هذا النموذج بطرق عرض مختلفة في النوافذ التالية:

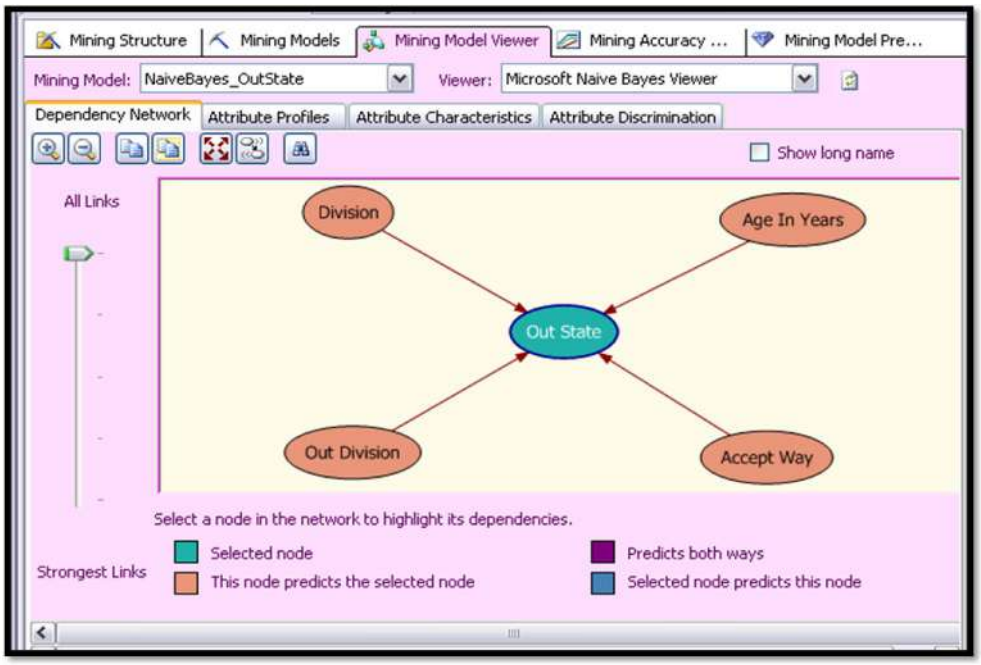
أ- نافذة شبكة التبعية (Dependency Network)

أظهرت نافذة شبكة التبعية أن حالة خروج المرضى من المشفى تعتمد على أربعة خاصيات فقط من خاصيات المدخلات وهي عمر المريض وطريقة قبوله في المشفى والشعبة التي قُبل فيها والشعبة الطبية التي تخرج منها، انظر الشكل (5-4-13).

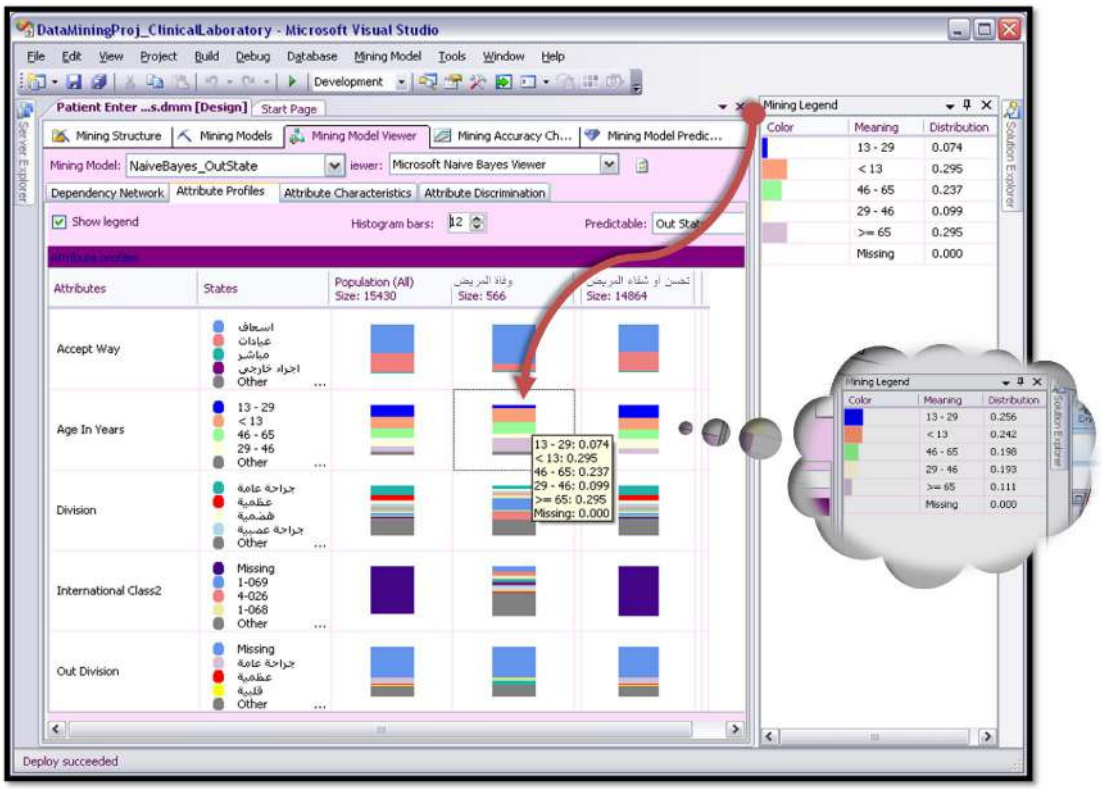
ب- نافذة اللمحة المختصرة للخاصيات (Attribute Profiles)

تُقدم هذه النافذة الاحتمالات المقابلة لكل قيمة من قيم الخاصيات في كل حالة من حالات تخرج المرضى من المشفى، فمثلاً يُقدر احتمال أن يكون عمر المريض المتوفى أكبر من 65 سنة بـــ 29.5% مقارنةً ببقية أعمار المتوفين، ونجد بالمقابل أن احتمال أن يكون عمر المريض المعافى أكبر من 65 سنة هو 11.1% كما في الشكل (5-4-14). أي أن المرضى المقبولين في المشفى وتزيد أعمارهم عن 65 سنة يكونون أكثر عرضة للوفاة في المشفى من بقية الأعمار.

[1] جرت تجربة إدخال هذين الحقلين إلى هذه المجموعة من نماذج علها تقدم بعض الفوائد في التحقق من توافق بعض البيانات، ولكنها أدت إلى التوصل إلى نماذج خاطئة، ولا يمكن تفسير نتائجها بشكل صحيح.



الشكل (– –): نافذة شبكة التبعية ( ) في نموذج ( )

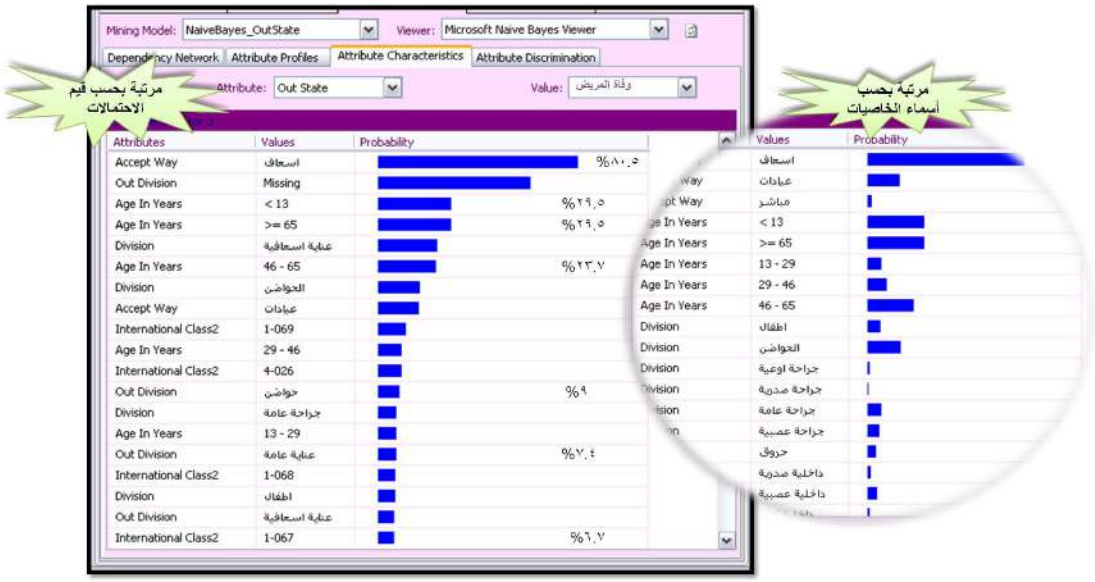


الشكل (5–4– 14): نافذة اللمحة المختصرة للخاصيات (Attribute Profile) في نموذج (NaiveBayes_OutStae)

ج– نافذة صفات الخاصيات (Attribute Characteristics)

تُساهم نافذة صفات الخاصيات في التعرف بالتفصيل على الاحتمالات المرافقة لقيم الخاصيات بالنسبة لبقية صفات الخاصية، أي أنها تتوسع بعرض المعارف التي تقدمها نافذة اللمحة

المختصرة للخاصيات، وتتمتع هذه النافذة بالمرونة بحيث تتيح عرض النتائج بترتيبات مختلفة كما في الشكل (5-4-15).



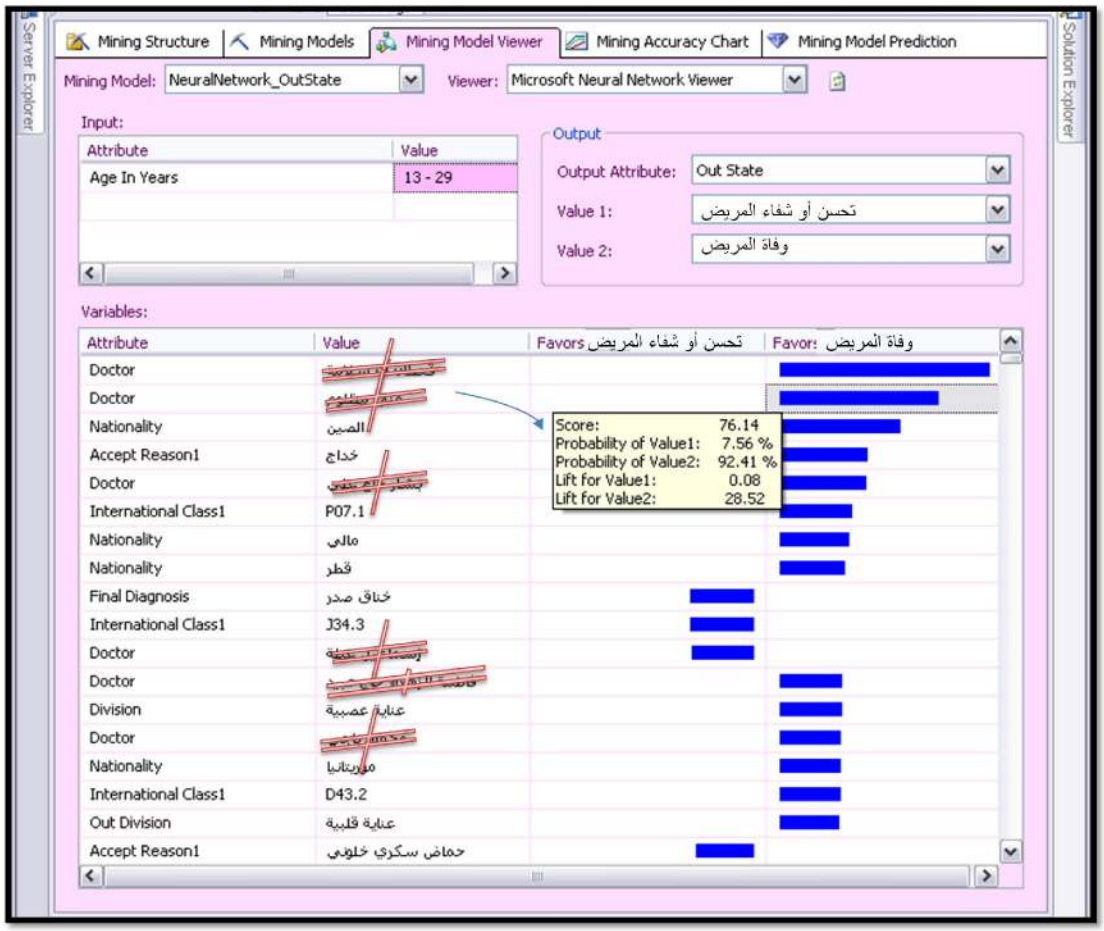
الشكل (5-4- 15): نافذة صفات الخاصيات (Attribute Characteristics) في نموذج (NaiveBayes_OutStae)

2•4•1•4•5 نموذج الشبكات العصبونية (NeuralNetworks_OutState)

يعتمد هذا النموذج على استخدام خوارزمية الشبكات العصبونية في عمليات التنقيب، وتتألف نافذة عرض النموذج من عدة أقسام لانتقاء المخرجات المرغوب بعرضها، كما هو مبين في الشكل (5-4-16)، حيث تَظهرُ ترجيحاتُ هذا النموذج لإحدى حالتي الخروج من المشفى (شفاء المريض أو وفاته) بحسب قيم خاصية العمر – التي تم اختيارها كمدخلات أساسية (Input) في الزاوية اليُسرى العليا في صورة النافذة – ووفقاً لكل من قيم خاصيات المدخلات الأخرى المدرجة في جدول المتغيرات (Variables) لهذه النافذة.

يقوم هذه النموذج بإعطاء علامات (Score) بين [0-100] تُعبر عن قوة تأثير قيمة المدخلات على ترجيح إحدى القيمتين المتنبأ بهما. وتُحتسب هذه التقييمات بواسطة عمليات حسابية داخلية تستند إلى ترجيح الاحتمالات المقابلة لكل قيمة من قيم الخاصيات بالقيم الترجيحية التي تعطى لهم في الطبقات الخفية من الشبكة العصبونية. وكلما ارتفعت العلامة كلما دلت على أهمية هذه القيمة في التأثير على القيمة المتنبأ بها. ويَعرض نموذجُ التنقيب النتائجَ الأولية

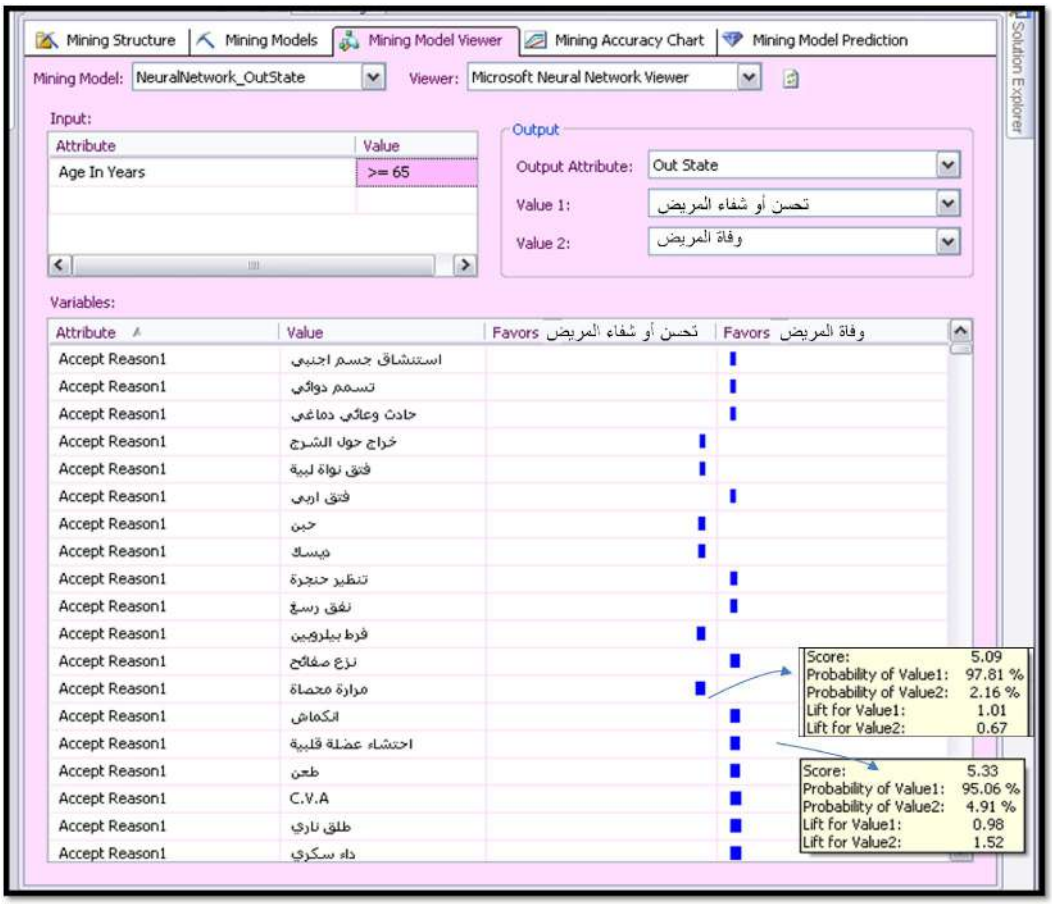
مرتبةً تنازلياً بحسب العلامات التي حصلت عليها كل حالة كما في الشكل (5-4-16). ويمكن بعد ذلك استعراض النموذج بطرق وترتيبات مختلفة بحسب الاستفسارات المرغوب بها.



الشكل (5-4- 16): نافذة نتائج نموذج الشبكات العصبونية (NeuralNetworks_OutState) المتعلقة بتقييم تأثير كل من قيم خاصيات المرضى على حالة الخروج من المشفى للفئة العمرية 13-29 سنة، ومرتبة تنازلياً بحسب العلامات المحسوبة داخلياً لكل حالة

ومن زاويةٍ أخرى يَعرض الشكل (5-4-17) تأثير كل من أسباب القبول على حالة الخروج من المشفى في الأعمار التي تزيد عن 65 سنة. ويلاحظ في هذا الشكل أن سبب المرارة المحصاة يُرجح خروج المريض بحالة تحسن (بعلامة تأثير 5.09)، أما سبب احتشاء العضلة القلبية فيرجح خروج المريض بحالة وفاة (بعلامة تأثير 5.33). ولكن نلاحظ أن 97.81% من المرضى الذين عانوا من المرارة المحصاة (Probability of Value1=97.81%) قد خرجوا بحالة تحسن، وأن 95.06% من المرضى الذين تعرضوا لاحتشاء العضلة القلبية قد خرجوا بحالة تحسن أيضاً وليس بحالة وفاة. أي أن احتمالي التحسن كبيرين والنسبتين متقاربتين، ولكن تقييم النموذج لكل من هاتين الحالتين كان مختلفاً، ويعود ذلك إلى أن هذا النموذج له

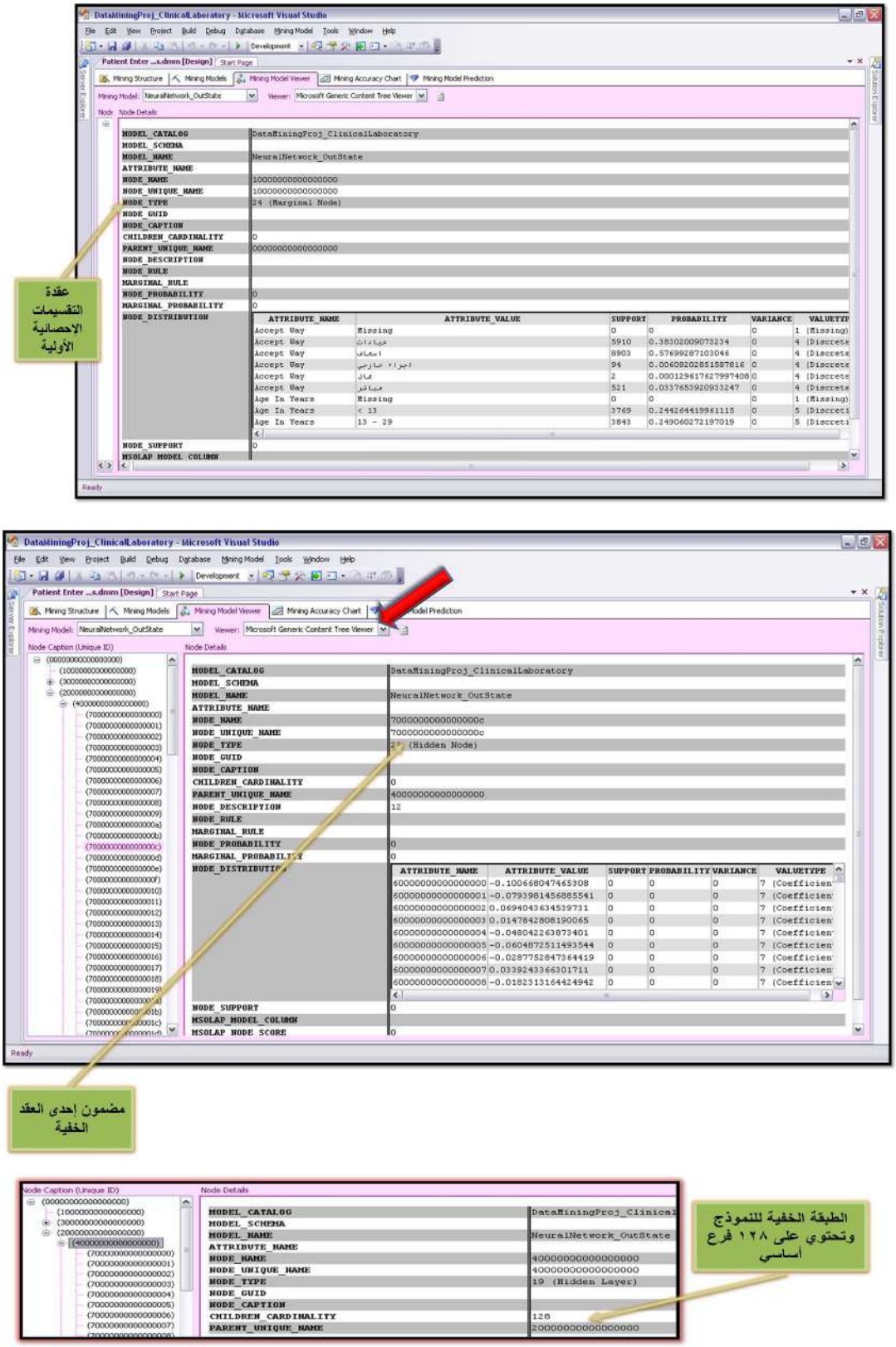
حساباته وترجيحاته الخاصة به والتي يتم على أساسها تقييم النتيجة، وتظهر محصلة الترجيحات الداخلية في هذا النموذج تحت اسمي (Lift for Value1) و(Lift for Value2)، وإن قيمة الترجيح الأكبر هي التي تُحدد حالة الخروج المتنبأ بها (المُرجَّحة).



الشكل (5-4- 17): نافذة نتائج نموذج الشبكات العصبونية (NeuralNetworks_OutState) المتعلقة بتقييم تأثير كل من قيم خاصية "سبب قبول المرضى" على خاصية حالة الخروج من المشفى في الأعمار التي تزيد عن 65 سنة

وعلى اعتبار أن هذا النموذج من النماذج المعقدة والتي تحتوي على الكثير من العمليات الآلية الداخلية فسنعرض على سبيل الاطلاع لقطات من هذه العمليات في الشكل (5-4-18)، حيث يمكن ملاحظة التجزئة التي تُجرى على بيانات المدخلات، ويمكن مشاهدة مضمون أي من الخلايا العصبونية الداخلية لهذا النموذج (خلايا الإدخال، الخلايا المخفية، المخرجات).. ولكن لن يتم الخوض أكثر في تفاصيل التركيبات الداخلية في هذا النموذج لأنها تقع خارج حدود هذا البحث[1].

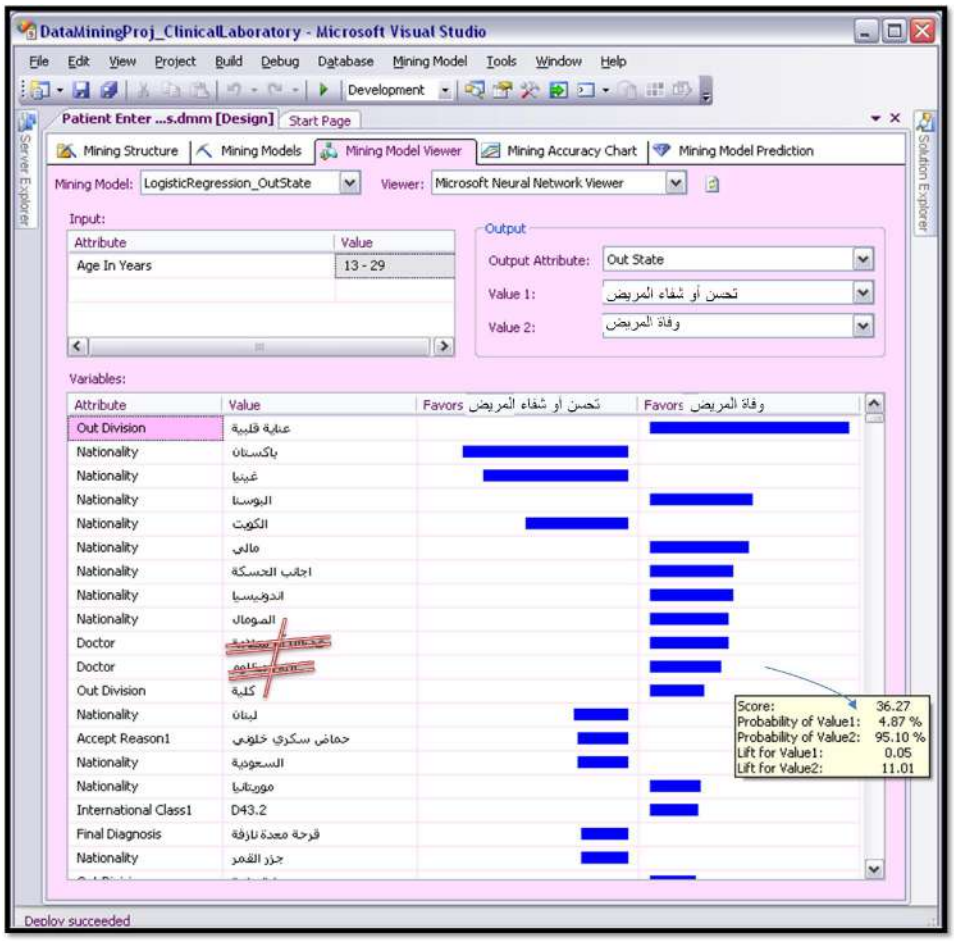
[1] للمزيد http://technet.microsoft.com/en-us/library/dd206985.aspx & [MacLennan, J.; Tang, Z. Crivat, B. 2009, **Data Mining with Microsoft SQL Server 2008**]



الشكل (5-4- 18): لقطات من مرأى شجرة المحتويات الشاملة (Generic Content Tree View) لنموذج الشبكات العصبونية (NeuralNetworks_OutState) المتعلق بتوقع حالة الخروج من المشفى

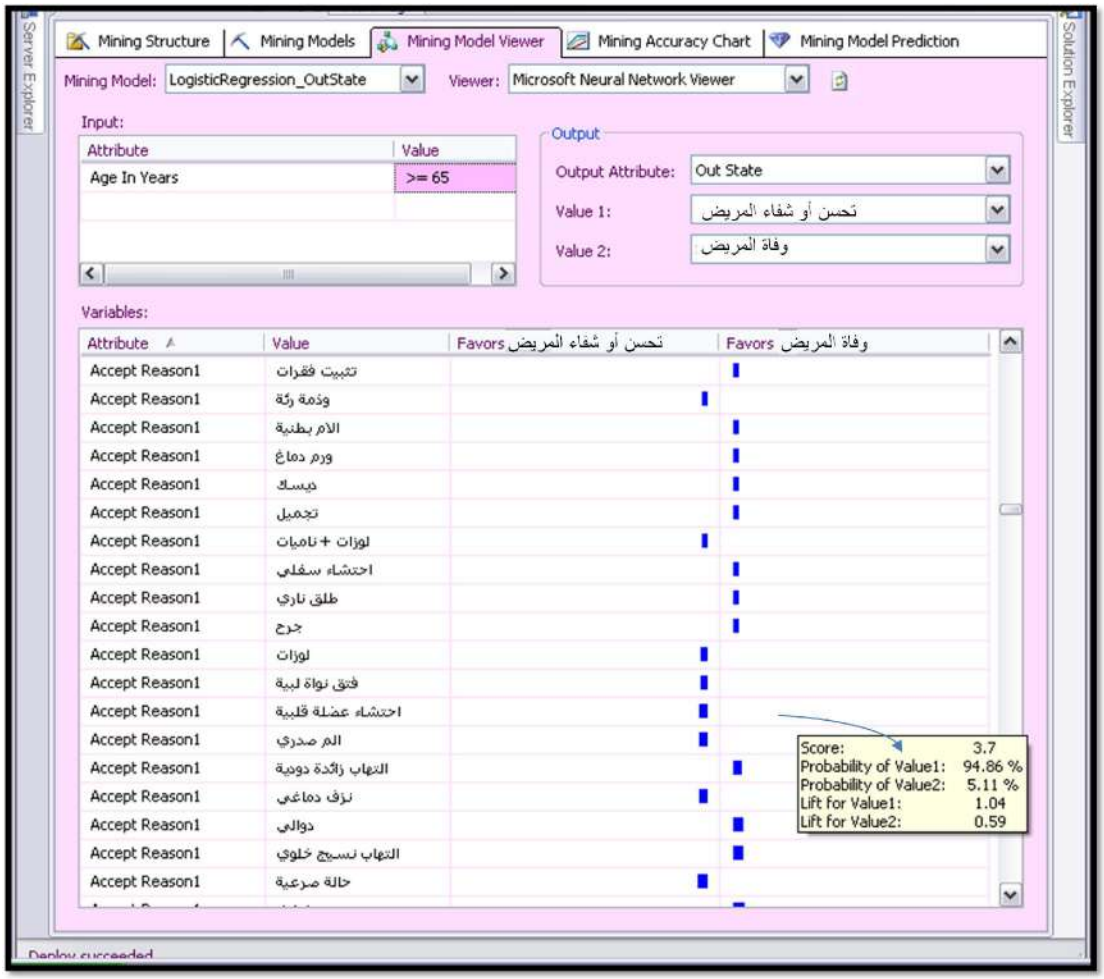
3•4•1•4•5 نموذج الانحدار المنطقي (LogisticRegresion_OutState)

يتشابه نموذج الانحدار المنطقي مع نموذج الشبكات العصبونية في الشكل العام لنافذة عرض النتائج، وقد لاحظنا من خلال استعراض النتائج أن هنالك بعض الاختلاف بين النموذجين في تقييم النتائج. فمثلاً مَنح هذا النموذج أعلى علامة وهي (100) لتأثير شعبة العناية القلبية على وفاة المرضى في الأعمار بين (13–29)، بينما أَعطى نموذج الشبكات العصبونية السابق علامة (28.98) لهذه الشعبة. كما وقدَّر نموذجُ الشبكات العصبونية السابق علامةً عاليةً (76.14) لتأثير اسم أحد الأطباء المشرفين في المشفى على ترجيح وفاة المرضى (الشكل (5–4–16))، بينما أعطى نموذج الانحدار المنطقي علامة أقل وهي (36.7) لتأثير هذا الطبيب كما في الشكل (5–4–19). ويعتبر اختلاف التقييم أمر طبيعي فهو يُعبر عن اختلاف وجهات نظر كل من النموذجين بسبب اختلاف الطرق الحسابية في الخوارزميات المستخدمة في كل منهما.



الشكل (5–4– 19): نافذة نتائج نموذج الانحدار المنطقي (LogisticRegresion_OutState) المتعلقة بتقييم تأثير كل من قيم خاصيات المرضى على حالة الخروج من المشفى للفئة العمرية 13–29 سنة، ومرتبة تنازلياً بحسب العلامات المحسوبة داخلياً لكل حالة.

وقد ظهرت أحياناً بعض الحالات التي تتعاكس فيها التقييمات بين النموذجين، فمثلاً أعطى نموذج الشبكات العصبونية علامة (5.33) لتأثير سبب احتشاء العضلة القلبية في توقع وفاة المريض الذي يزيد عمره عن 65 سنة (راجع الشكل (5-4-17))، بينما أعطى نموذج الانحدار المنطقي علامة (3.7) لتأثير هذا السبب في توقع شفاء المريض الذي يزيد عمره عن 65 سنة، كما في الشكل (5-4-20). أي أن النتيجتين متعاكستين تماماً...! وهنا تكمن خطورة الاعتماد على نموذج معين من نماذج التنقيب في تقييم الحالات الطبية أو اتخاذ أية قرارات طبية حساسة. وبناءً على ذلك لابد من تجريب هذه النماذج كثيراً قبل الاعتماد عليها فالبيانات التي تستند إليها هنا محدودة المتغيرات وليست بالغة الدقة، كما ويُستحسن أن لا يتم الاعتماد في الحقل الطبي على أية نتيجة ما لم تكن مدعمة من قبل عدة نماذج للتنقيب.



الشكل (5-4- 20): نافذة نتائج نموذج الانحدار المنطقي (LogisticRegresion_OutState) المتعلقة بتقييم تأثير كل من قيم خاصية سبب قبول المرضى على خاصية حالة الخروج من المشفى في الأعمار التي تزيد عن 65 سنة

2•4•5 مشروع التنقيب في بيانات نتائج التحاليل الطبية المخبرية

تنشأ بين نتائج التحاليل المخبرية المختلفة علاقات تأثير واقتران يُستفاد منها في تشخيص الأمراض، ولكن إلى أي مدى تنطبق الحقائق الطبية المعروفة على التحاليل الطبية التي أجريت فعلاً للمرضى؟. ألا يوجد خصوصية لتحاليل المرضى وللعلاقات بينها باختلاف صفات المرضى؟. وهل يوجد علاقات جديدة غير مكتشفة كامنة في هذه البيانات المخزنة عبر السنين؟؟. ألا يوجد أسئلة كثيرة يمكن أن يطرحها المهتمون بالقضايا الطبية لو أتيح لهم التعرف على إمكانيات أدوات التنقيب في البيانات؟؟. إن المشروع الحالي يهدف إلى التعريف بنمط المهام التي تستطيع أن تقوم بها أدوات التنقيب في البيانات عندما تُطبَّق على بيانات التحاليل الطبية المخبرية.

1•2•4•5 إنشاء المشروع وتحديد مصدر البيانات

حمل المشروع الجديد للتنقيب في البيانات اسم (DataMIning_ClinicalLaboratory)، وعُرِّف مستودع بيانات التحاليل الطبية (LocalHost.Clinical Laboratory Study) على أنه مصدر لبيانات هذا المشروع.

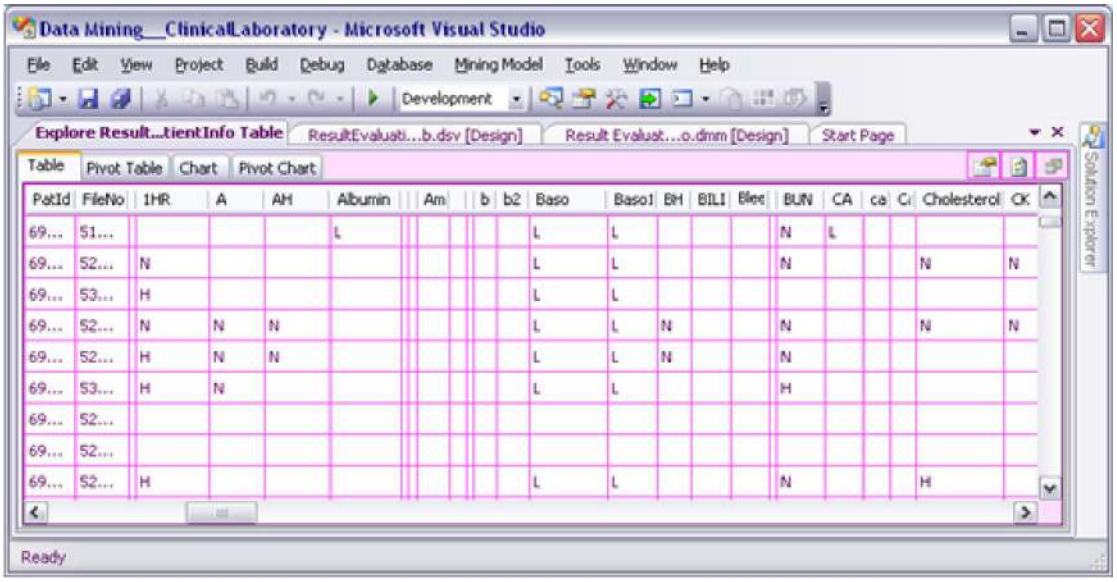
تَطلب تنفيذُ هذا المشروع الحصولَ على بيانات نتائج التحاليل الطبية بطريقة مناسبة لتطبيق التنقيب فيها، وقد جرت عدة تجارب للتنقيب في بيانات نتائج التحاليل الطبية قبل الوصول إلى الشكل النهائي لهذا المشروع. وكان الطريق إلى هذا المشروع طويلاً زمنياً ومحفوفاً بالصعوبات والقيود، وكثرت خلال هذه الفترة الحركةُ الرجعية على هذا الطريق إلى المراحل السابقة من مراحل اكتشاف المعرفة لتجريب بعض التعديلات وإحداث بعض التطويرات.. وعلى الرغم من أن بعض التطويرات كانت قد نتجت خلال فترات زمنية متباينة إلا أنها وردت في هذا البحث بحسب عنوان مرحلة اكتشاف المعرفة التي تنتمي إليها، وسنبين فيما يلي وباختصار مضمون تجارب التنقيب التي تأسس وتطور من خلالها هذا المشروع.

أ- نُفِّذت تجارب التنقيب الأولى على جدول نتائج التحاليل الطبية الفرعية (SubTestsDoing) الذي بُني في هذا البحث، والذي خُزِّنت فيه نتائج التحاليل الطبية الرقمية في حقل واحد، ولكن لم تحقق هذه المحاولات أهداف التنقيب المرجوة لأن بيانات نتائج التحاليل الطبية

موجودة في حقل واحد ولكل نتيجة منها المعنى الخاص بها والذي يختلف باختلاف اسم التحليل وعمر المريض وجنسه.

ب- اعتمدت تجارب التنقيب الثانية على فكرة استبدال نتائج التحاليل الطبية الرقمية برموز لفهم جميع التحاليل بطريقة موحدة، وهي الرموز التي تُعبِّر عن تقييم نتائج التحاليل الطبية (N للنتائج الطبيعية، H: للنتائج المرتفعة عن الحدود الطبيعية، L: للنتائج المنخفضة عن الحدود الطبيعية)، وقد ورد ذكر عملية التحويل هذه في أثناء الحديث عن تطوير مستودع البيانات في الفقرة (5•3•3•5-أ). وقد قدَّمت تجاربُ التنقيب المعتمدة على الرموز بعضَ النتائج المفيدة في مجال اكتشاف المعرفة المخبأة في البيانات، ولكن بقيت فكرة وجود نتائج التحاليل الطبية في حقل واحد تعيق دراسة العلاقة البينية بين نتائج التحاليل الطبية، وتَحُد من مخرجات عمليات التنقيب التي تزداد قوتها عادةً بتزايد عدد الخاصيات (المتغيرات، الحقول) التي تصف البيانات.

ج- اعتمدت تجارب التنقيب الثالثة على فكرة تخصيص حقل لكل نوع من أنواع التحاليل الطبية، وقد تم إنشاء جدول جديد انطلاقاً من هذه النقطة، وهو الجدول الذي تحدثنا عنه سابقاً في الفقرة (5•3•3•5-ب). وساعدت فكرة هذا الجدول على إبراز الدور الفعلي لنتائج التحاليل الطبية المخبرية في عمليات اكتشاف المعرفة من قواعد البيانات الطبية. وقد تم تصميم مرأى من هذا الجدول – في مستودع البيانات- يناسب عمليات التنقيب في هذا المشروع وأطلق عليه اسم (ResultEvaluation_Crosstab)، وتم اعتباره كمرأى لبيانات مشروع التنقيب الحالي ويحتوي هذا المرأى على 127 حقل (خاصية) لعرض نتائج التحاليل الطبية المختلفة، انظر الشكل (5-4-21)، حيث تم تصغير المسافات بين الحقول بهدف عرض كم أكبر من الخلايا التي تحتوي على قيم.



الشكل (5-4- 21): استعراض مرأى (ResultEvaluation_Crosstab) المستخدم في مشروع التنقيب في بيانات نتائج التحاليل الطبية

5•4•2•2 بناء هيكل التنقيب

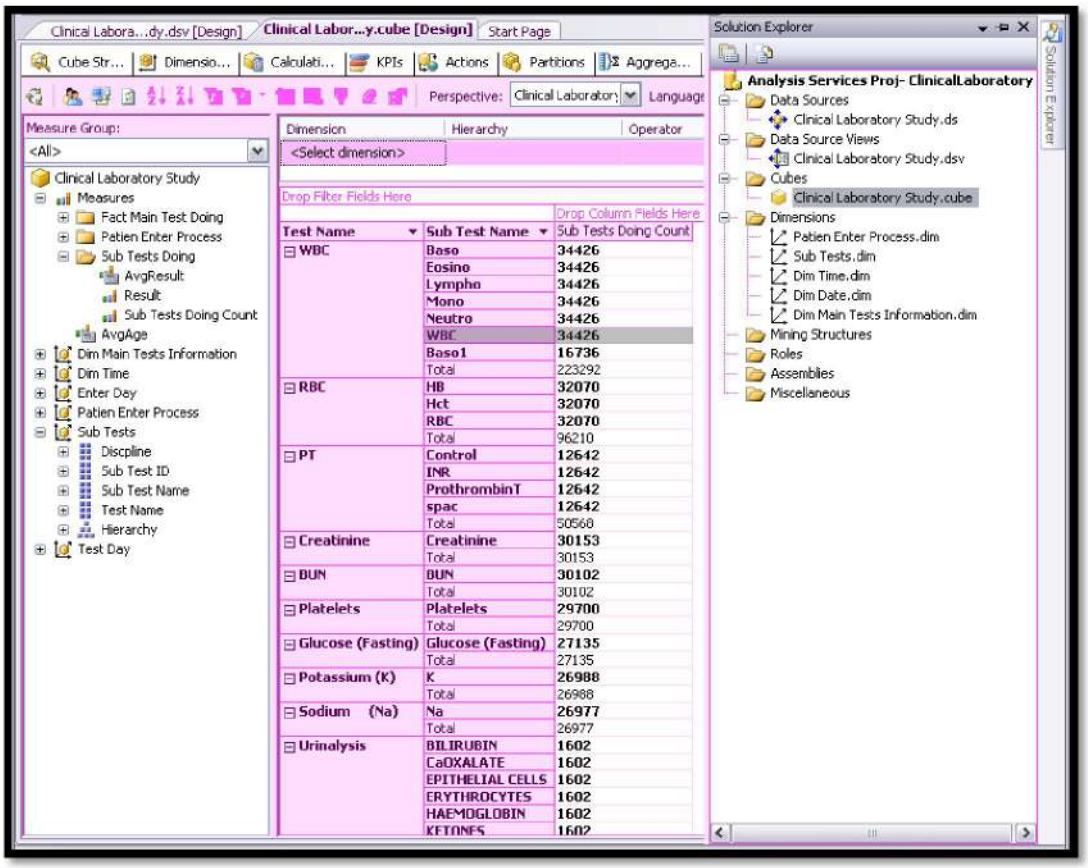
بُني هيكل التنقيب في هذا المشروع باستخدام إجراءات البناء في برنامج (Visual Studio) والتي ورد ذكرها في المشروع السابق، وتم انتقاء حقول التنقيب من الحقول الموجودة في مرأى البيانات (ResultEvaluation_Crosstab)، ويعتبر هذا المرأى كبيراً نسبياً مقارنةً بمرأى البيانات في المشروع السابق مما جعل عمليات تدقيق وبناء وتفعيل طلبات بناء هيكل ونماذج التنقيب أو أية تعديلات تطرأ عليها تأخذ الكثيرَ من الوقت مقارنة بالوقت المستغرق في المسألة السابقة حيث وصلت أحياناً إلى ساعة ونصف في هذا المشروع ، في حين لم تتجاوز المدد الزمنية في المشروع السابق العشر دقائق.

5•4•2•3 تصميم واستعراض نماذج التنقيب في بيانات نتائج التحاليل الطبية المخبرية

تتمحور الفكرة الأساسية في نماذج التنقيب المدروسة هنا حول محاولة التعرف على نتائج أي تحليل طبي اعتماداً على نتائج التحاليل الأخرى، وتم لذلك تصميم سبعة نماذج للتنقيب يستند كل نموذج إلى إحدى خوارزميات التنقيب، وقد اعتُبرت معظم حقول التحاليل المخبرية المتوفرة كمخرجات تنبؤية (Predict)[1] لنماذج التنقيب وأُعتبرت بقية الحقول كمدخلات

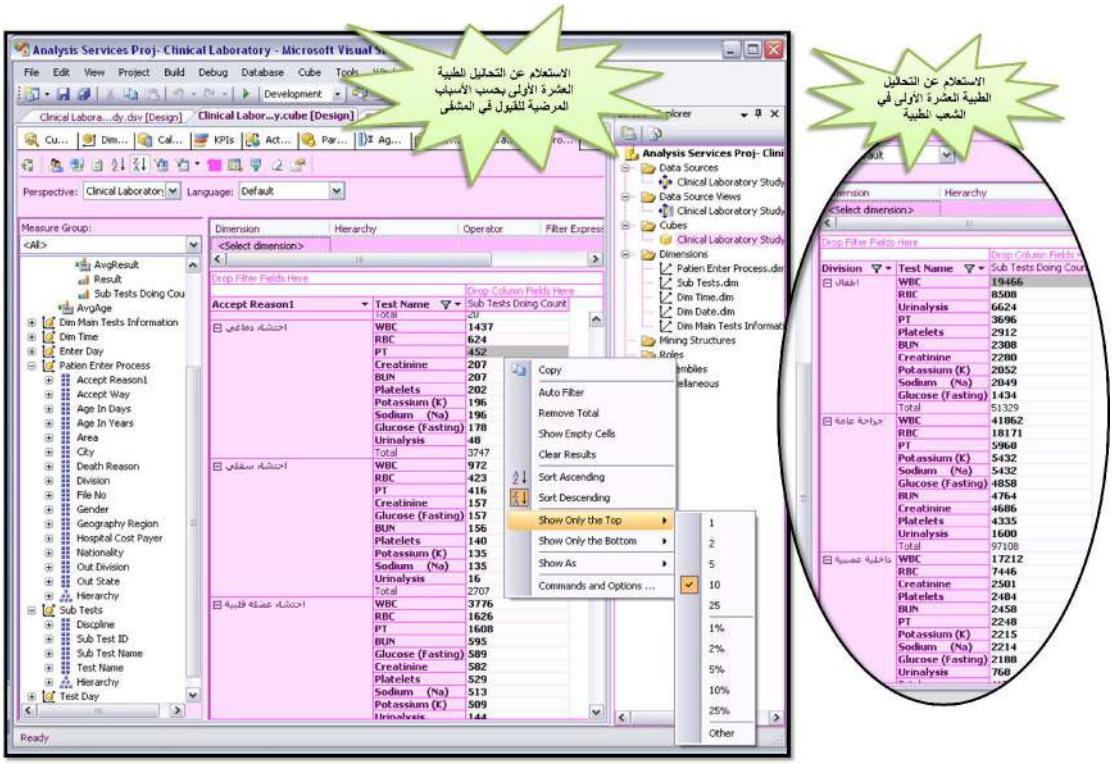
[1] تسمح مهمة التنبؤ (Predict) في حقل ما باستخدام نفس الحقل كمدخلات للتنبؤ بالحقول الأخرى التي تحمل مهمة التنبؤ، بينما تسمح مهمة (Predict Only) باستخدام الحقل للتنبؤ فقط، راجع الفصل الثالث.

(Input)، وقد ساهمت عدة أمور في اختيار مهام الحقول، وأهمها أمران الأول معطيات التغذية الراجعة من مرحلة تقييم النماذج حيث كانت تتم دراسة تأثير تغيير مهام الحقول على دقة نماذج التنقيب، والثاني الاستفادة من نتائج المعالجة التحليلية الآنية (OLAP)، حيث مُنِحت المهمة التنبؤية للتحاليل الأكثر تكراراً في البيانات. فمثلاً يُعتبر التحليل الطبي لتعداد كريات الدم البيضاء (WBC) وتعداد كريات الدم الحمراء (RBC) من أكثر التحاليل تكراراً في البيانات، ويترافقان مع الكثير من التحاليل الأخرى ومع الكثير من الأسباب المرضية المختلفة التي تحتاج إلى قبول في المشفى[1]، وهذا ما يظهر في الشكلين (5-4-22) و(5-4-23) المأخوذين من مشروع المعالجة التحليلية الآنية الخاص بهذه الدراسة.



الشكل (5-4- 22): استخدام مكعب البيانات الخاص بالدراسة في التعرف على أكثر التحاليل الطبية طلباً في مخبر التحاليل الطبية التابع للمشفى المدروس

[1] يعتبر التحليل الطبي المسمى تعداد الدم الكامل (Complete Blood Count(C.B.C)) بأنه التحليل الأول الذي يُطلب في المشفى للتعرف على حالة الدم ويشتمل على التحاليل الطبية التالية (WBC, RBC, Platelets, HB)



الشكل (5-4- 23): استخدام مكعب البيانات الخاص بالدراسة في التعرف على أكثر التحاليل الطبية طلباً في مخبر التحاليل الطبية التابع للمشفى المدروس مصنفة بحسب الأمراض وبحسب الشعب الطبية

ومن الجدير ذكره أن تركيز الاهتمام على التحاليل المتكررة لم يمنع من تجربة التنبؤ بالمتغيرات الأقل أهمية أو الأقل تكراراً، فهي أيضاً من المواضيع الهامة في الطب ولكن القلة الشديدة في هذه البيانات في قاعدة البيانات المدروسة حالت دون العثور على جوانب تدعم البحث الحالي.

تتصف النماذج في مشروع التنقيب في التحاليل الطبية المخبرية بتعدد المتغيرات المتنبأ بها، ولكن من الصعوبة بمكان أن يُعرَض الكم الهائل من النتائج بالنسبة لكل تحليل من التحاليل المتنبأ بها، وإنما تم اختيار التركيز على أبرز الأمثلة التي تساهم في التعرف على أداء كل من نماذج التنقيب، وفي تسليط نقاط ضوء على الإمكانيات المستقبلية الواسعة التي يمكن أن تقدمها نماذج التنقيب فيما لو استُثمرت بالشكل الصحيح في الحقل الطبي.

1•3•2•4•5 نموذج أشجار القرار (DecisionTrees_ PredictTestResult)

أظهر نموذج أشجار القرار الكثير من النتائج عن العلاقات التسلسلية بين نتائج التحاليل الطبية المخبرية، ويبين الشكل (5-4-24) شجرة القرار الخاصة بالتنبؤ بنتيجة التحليل الفرعي

لتعداد كريات الدم البيضاء[1]، وتُعبِّر درجة لون نهايات فروع الشجرة عن كثافة النتائج المرتفعة لتحليل (WBC)، فكلما ارتفعت الكثافة كلما كان اللون غامقاً. ويُقسم الشريط الملون في هذه النهايات بحسب النسب الاحتمالية لنتائج تحليل (WBC)، وتدل القيمة (Missing) لتحليل معين على عدم وجود أية نتيجة لهذا التحليل في هذا التفرع من الشجرة، ومعنى ذلك من ناحية طبيعة البيانات المدروسة أنه لم يتم إجراء هذا التحليل بالترافق مع بقية التحاليل في هذا التفرع. ويؤدي نقر زر الفأرة على إحدى نهايات فروع الشجرة إلى ظهور قائمة تفسيرية للتنبؤ باحتمالات تحقق كل من قيم المتغير المتنبأ به، إذا تحققت قيم المتغيرات الأخرى في هذا الفرع. وكمثال على ذلك نقرأ في شجرة القرار الخاصة بالتنبؤ بقيم تحليل تعداد كريات الدم البيضاء (WBC) بعض النتائج التي أُخذت من القوائم التفسيرية في الشكل (5-6-25) وذلك كما يلي:

1- إذا كانت نسبة العدلات مرتفعة ونسبة اللمفاويات منخفضة وعدد الصفائح ليس منخفضاً ونسبة الحمضات منخفضة فإن الاحتمال الأكبر (90.69%) هو أن يكون تعداد كريات الدم البيضاء مرتفعاً، ويُعبر عن هذه النتيجة في الشكل البياني على النحو التالي:

(Neutro = 'H' and Lympho = 'L' and Platelets not = 'L' and Eosino = 'L')→WBC='H'

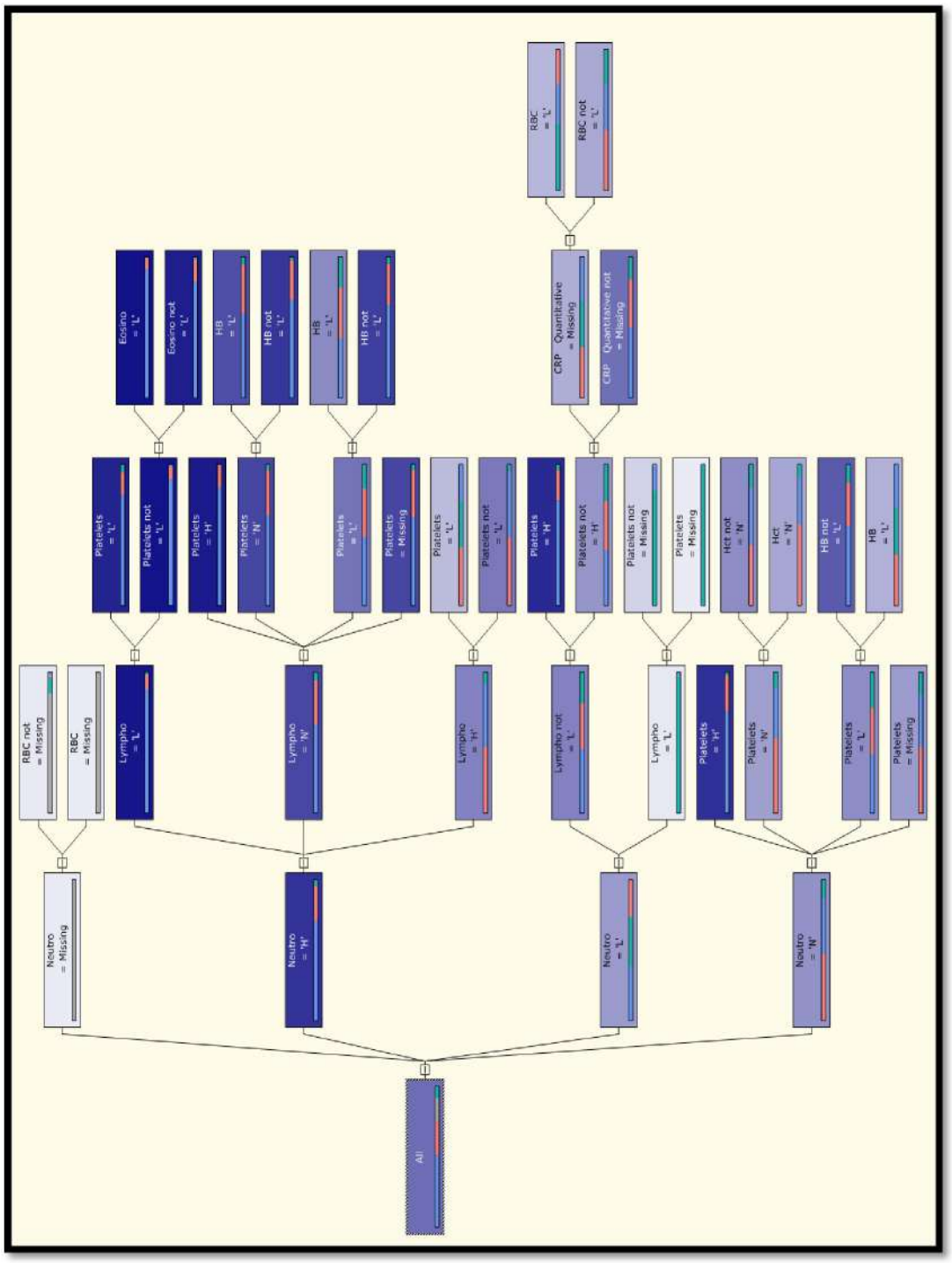
وقد تكون هذه النتيجة بديهية في تفسير النتائج المخبرية، ولكن ظهور مثل هذه النتيجة في البيانات المدروسة يعتبر من الاكتشافات التي تدعو إلى التفاؤل بقدرات تقانة التنقيب في البيانات على اكتشافات نتائج أو علاقات جديدة وغير مكتشفة سابقاً طالما أنها تدعم البديهيات. ومن جهة أخرى تدعم هذه النتيجة فكرة أن البيانات المدروسة صحيحة.

2- إذا كانت نسبة العدلات طبيعية وكان عدد الصفائح منخفضاً وكان الخضاب منخفضاً فإن احتمال أن يكون تعداد كريات الدم البيضاء مرتفعاً هو 30.56%، واحتمال أن يكون منخفضاً هو 34.33%.

3- إذا انخفضت نسبة العدلات وانخفضت نسبة اللمفاويات فإن 96.31% من الحالات سينخفض فيها تعداد كريات الدم البيضاء.

---

[1] يُعبِّر تعداد الكريات البيضاء (WBC) عن التعداد الإجمالي لكل أنواع الكريات البيضاء، ويقاس بعدد الخلايا في المليمتر المكعب من الدم، وهناك عدة أنواع لكريات الدم البيضاء هي العدلات (Neutro-phils) واللمفاويات(Lymph-cytes) ووحيدات النوى (Mono-cytes) والحمضات (Eosino-phils) والأساسات (Baso-phils)، ويقاس كل نوع من أنواع الكريات البيضاء بحسب نسبته إلى التعداد الإجمالي للكريات البيضاء.. المرجع [د. سيد الحديدي، 2006، **دلالات وتفسير النتائج المخبرية** ، ص141]



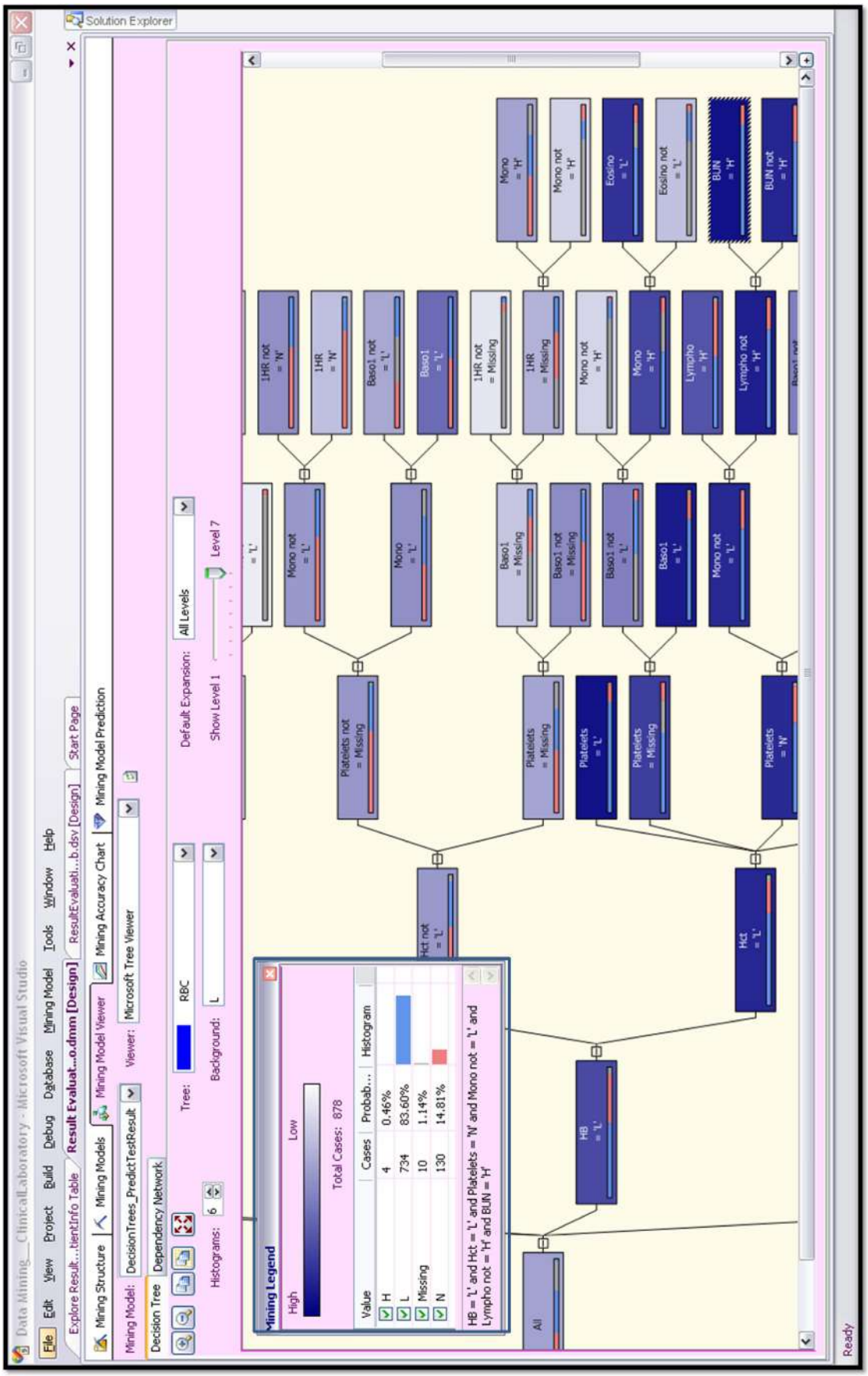
الشكل (5-4- 24): مخطط شجرة القرار الناتج عن نموذج التنقيب (DecisionTrees_PredictTestResult)، وقد استدعيت هذه الشجرة للتنبؤ أو استنتاج نتيجة تحليل تعداد كريات الدم البيضاء (WBC) استناداً إلى نتائج التحاليل الأخرى



الشكل (5-4- 25): أمثلة عن النتائج المتوقعة للتحليل الطبي (WBC) استنادا إلى نتائج التحاليل الطبية الأخرى والمستمدة من نموذج شجرة القرار (DecisionTrees_PredictTestResult)

ويعرض الشكل (5-4-26) مثالاً أخر عن شجرة القرار الخاصة بالتنبؤ بنتائج تحليل تعداد كريات الدم الحمراء (RBC). ونقرأ في إحدى تفرعات هذه الشجرة أنه عندما يكون الخضاب منخفضاً (HB='L') والهيماتوكريت منخفضاً (Hct='L') وعدد الصفائح طبيعياً (Platelets='N') وعدد وحيدات النوى ليس منخفضاً (Mono≠'L') وعدد اللمفاويات ليس مرتفعاً (Lemfo≠'H') وتحليل البولة النتروجينية[1] مرتفعاً (BUN='H') فإن احتمال أن يكون عدد كريات الدم الحمراء منخفضاً هو 83.6%، واحتمال أن يكون طبيعياً هو 14.8%، واحتمال أن يكون مرتفعاً 0.46%.

[1] تحليل مستوى البولة النتروجينية في الدم (BUN) هي اختصار (Blood Urea Nitrogen)، وهو أحد التحاليل الطبية التي تدل على وظائف الكلى، ويرتفع في حالات مثل الالتهاب الكلوي الحاد والمزمن، والفشل الكلوي.

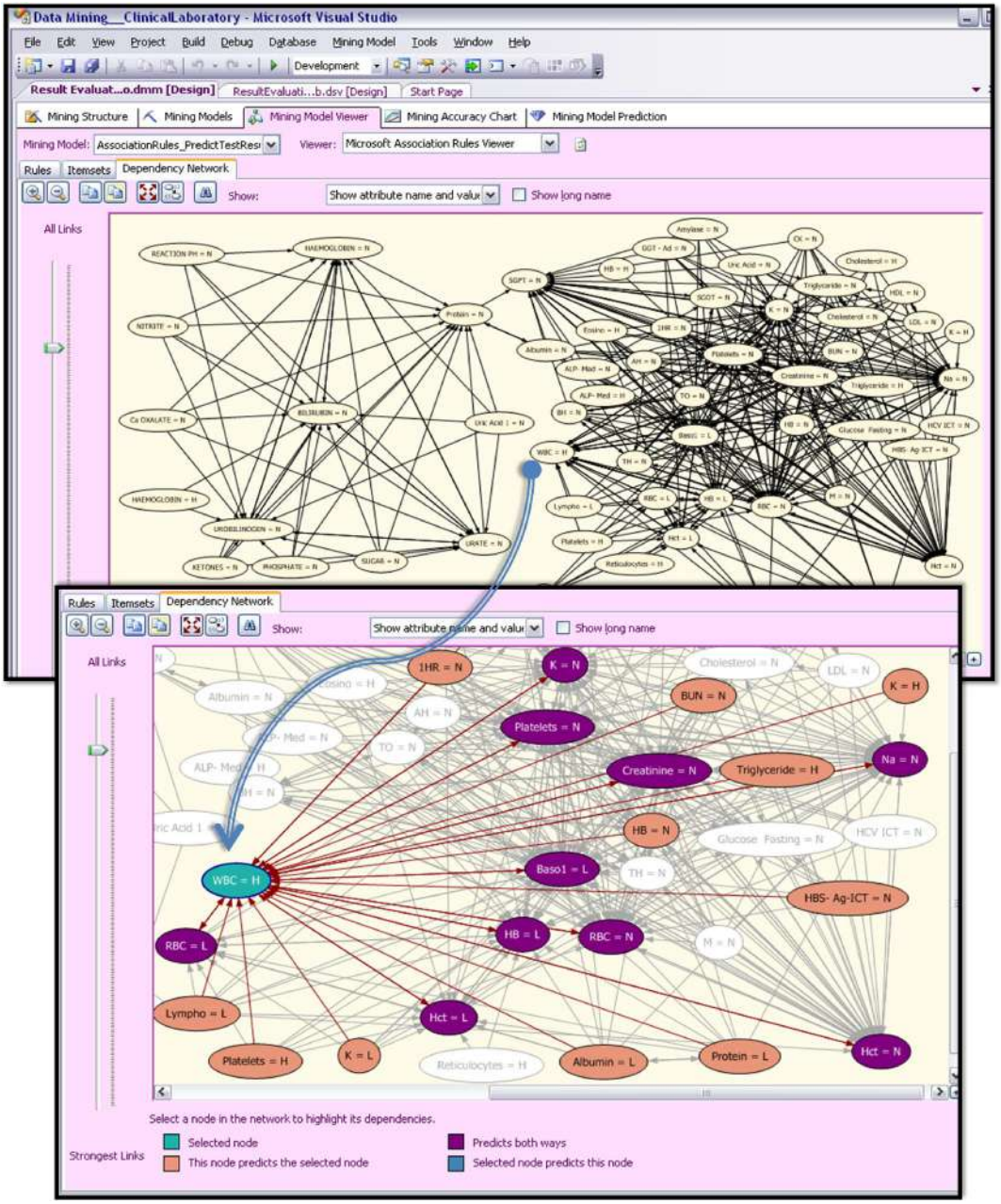


الشكل (5-4- 26): نافذة شجرة القرار في نموذج (DecisionTrees_PredictTestResult) الخاصة بالتنبؤ بنتائج التحليل الطبي لكريات الدم الحمراء (RBC)

2•3•2•4•5 نموذج قواعد الاقتران (AssociationRules_PredictTestResult)

أ- نافذة شبكة التبعية (Dependency Network)

أظهرت نافذة شبكة التبعية في نموذج التنقيب في التحاليل الطبية المعتمد على خوارزمية قواعد الاقتران وجودَ الكثير من التشابكات بين نتائج التحاليل الطبية، ويدل ذلك على وجود الكثير من قواعد الاقتران في البيانات المدروسة، انظر الشكل (5-4-27)



الشكل (5-4-27): في الأعلى منظر عام لنافذة شبكة التبعية في نموذج التنقيب المستند إلى خوارزمية قواعد الاقتران (AssociationRules_PredictTestResult)، وفي الأسفل جزء من هذه النافذة ويمثل علاقات اقتران نتائج التحاليل الطبية المتبادلة مع التحليل الطبي لتعداد كريات الدم البيضاء (WBC)

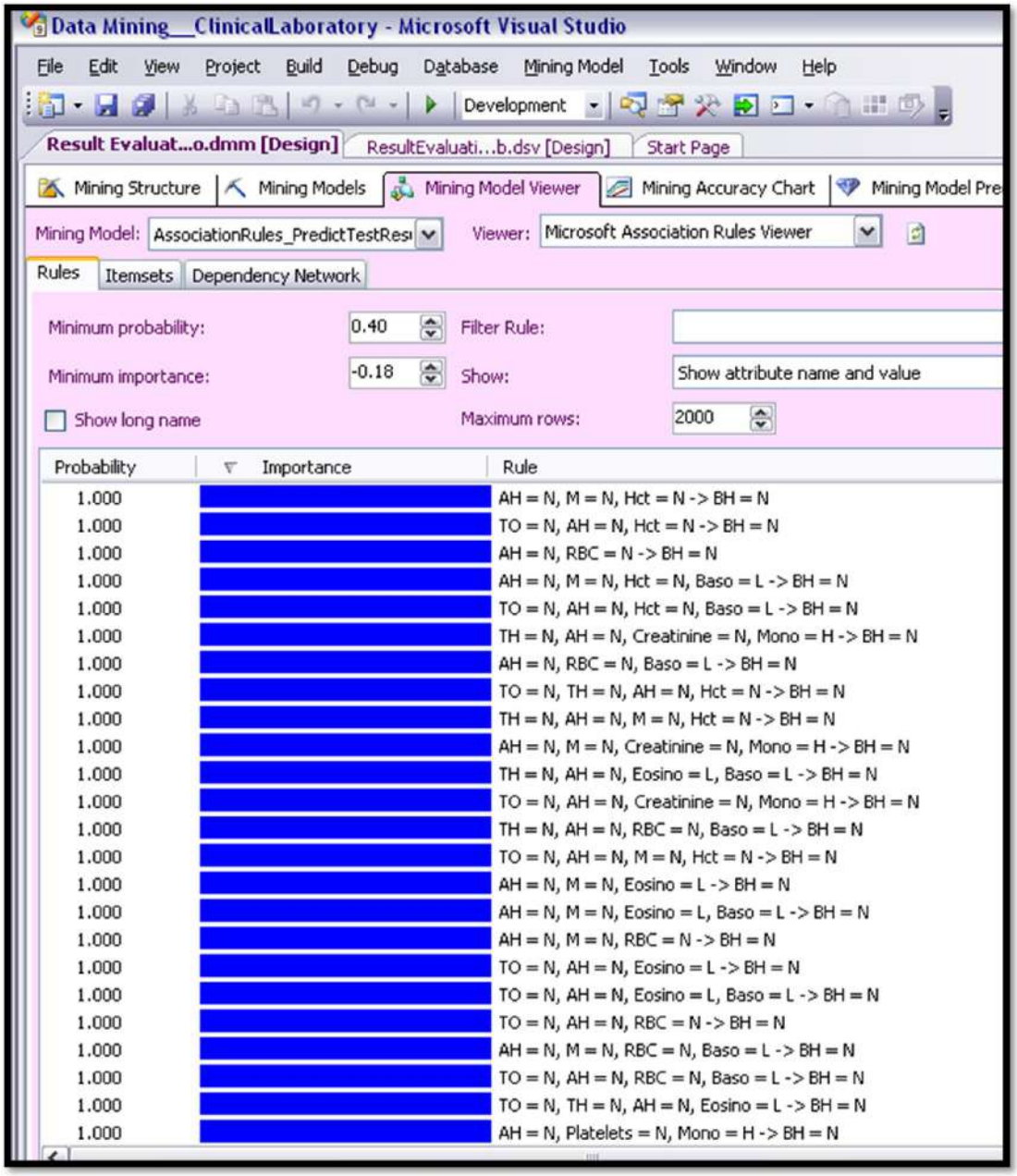
ب- نافذة قواعد الاقتران (Rules)

احتوت نافذة قواعد الاقتران على عدد كبير من القواعد، وقد تبيَّن وجودُ عدد كبير من القواعد التي كانت فيها معظم نتائج التحاليل الطبية طبيعية، ويعود ذلك -من ناحية قيم خاصيات البيانات- إلى التكرار الكبير للنتائج الطبيعية مقارنةً بالنتائج غير الطبيعية، أما من الناحية الطبية فيعود باختصار إلى وجود عدد كبير من التحاليل الطبية التي طُلِبت للتأكد من أن النتائج طبيعية أو لإزالة الشكوك الطبية بمرض معين، أو لأنها من التحاليل التي تُطلب لكل مريض يدخل المشفى مثل تحاليل تعداد الدم الكامل..الخ وهذا ما سمح بوجود عدد كبير من النتائج الطبيعية للتحاليل الطبية.

وكمثال على ذلك نجد في الشكل (5-4-28) أن قواعد الاقتران المتعلقة بالتحاليل الطبية الفرعية (TO, TH, AH, BH) - والمندرجة تحت التحليل الأساسي المسمى بتفاعل فيدال (WIDAL) الخاص بالتقصي عن الحمى التيفية[1]- قد ظهرت في بداية قائمة القواعد لأنها استحوذت على احتمالات التحقق الأكيد (1.00)، وعلى أهميات (Importances) كبيرة أيضاً، ولكن تبين بالتدقيق في هذه القواعد أن جميع نتائج هذه التحاليل الطبية هي نتائج طبيعية، وهذا يعود إلى وجود الكثير من الطلب على هذه التحاليل دون وجود إصابة بهذا المرض.

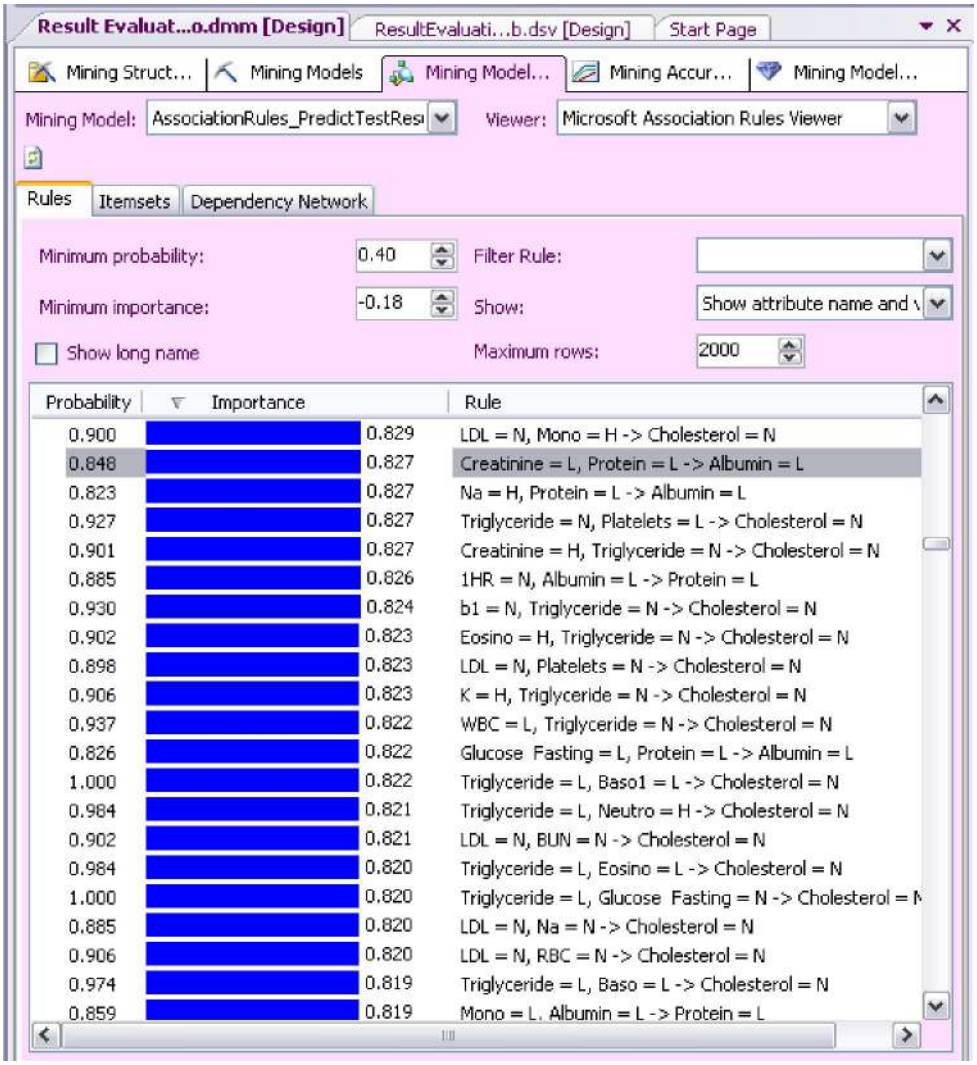
ومن ناحية أخرى فإن قلة النتائج غير الطبيعية بالنسبة لهذا المرض لم يسمح لها بتشكيل قواعد اقتران قوية. وهنا يمكن القول أنه لو كانت البيانات كافية ومتعددة المصادر لكان بالإمكان استنتاج الكثير من القواعد الهامة عن مثل هذه الأنواع من الأمراض وخصوصاً إذا رُبطت بيانات التحاليل الطبية مع البيانات الشخصية للمرضى ومنها بيانات مكان الإقامة، وذلك أن أدوت التنقيب في البيانات تستطيع أن تلعب دوراً هاماً في تصنيف الظواهر الصحية ووضع قواعد لها مثل الانتشار المكاني والزماني للأمراض والأوبئة.

[1] الحمى التيفية هي مرض ينشأ بسبب بكتريا السالمونيلا التي تنتقل إلى الإنسان من المشتقات الحيوانية وخصوصاً لحوم الدواجن والبيض. وتتلخص الأعراض بالارتفاع الشديد في درجة حرارة الجسم، والصداع والهزال العام والإسهال، وقد يظهر طفح جلدي علي شكل بقع وردية اللون.



الشكل (5-4-28): نافذة القواعد في نموذج التنقيب المستند إلى خوارزمية قواعد الاقتران (AssociationRules_PredictTestResult)، مرتبة تنازلياً بحسب الاحتمال، ويظهر في الشكل قواعد الاقتران التي تتواجد فيها التحاليل الطبية الفرعية الخاصة بالتقصي عن الحمى التيفية

إن سيطرة قواعد الاقتران التي تدعم النتائج الطبيعية للتحاليل الطبية لم يمنع من وجود بعض القواعد القوية الخاصة بالنتائج غير الطبيعية للتحاليل الطبية، وكمثال على ذلك نقرأ في القاعدة الثانية في الشكل (5-4-29) أنه " إذا انخفضت نتيجة كل من التحليلين الطبيين (Creatinine=L) و(Protein=L) فإنه باحتمال (0.848) وبأهمية (0.827) تكون نتيجة تحليل (Albumin) منخفضة "

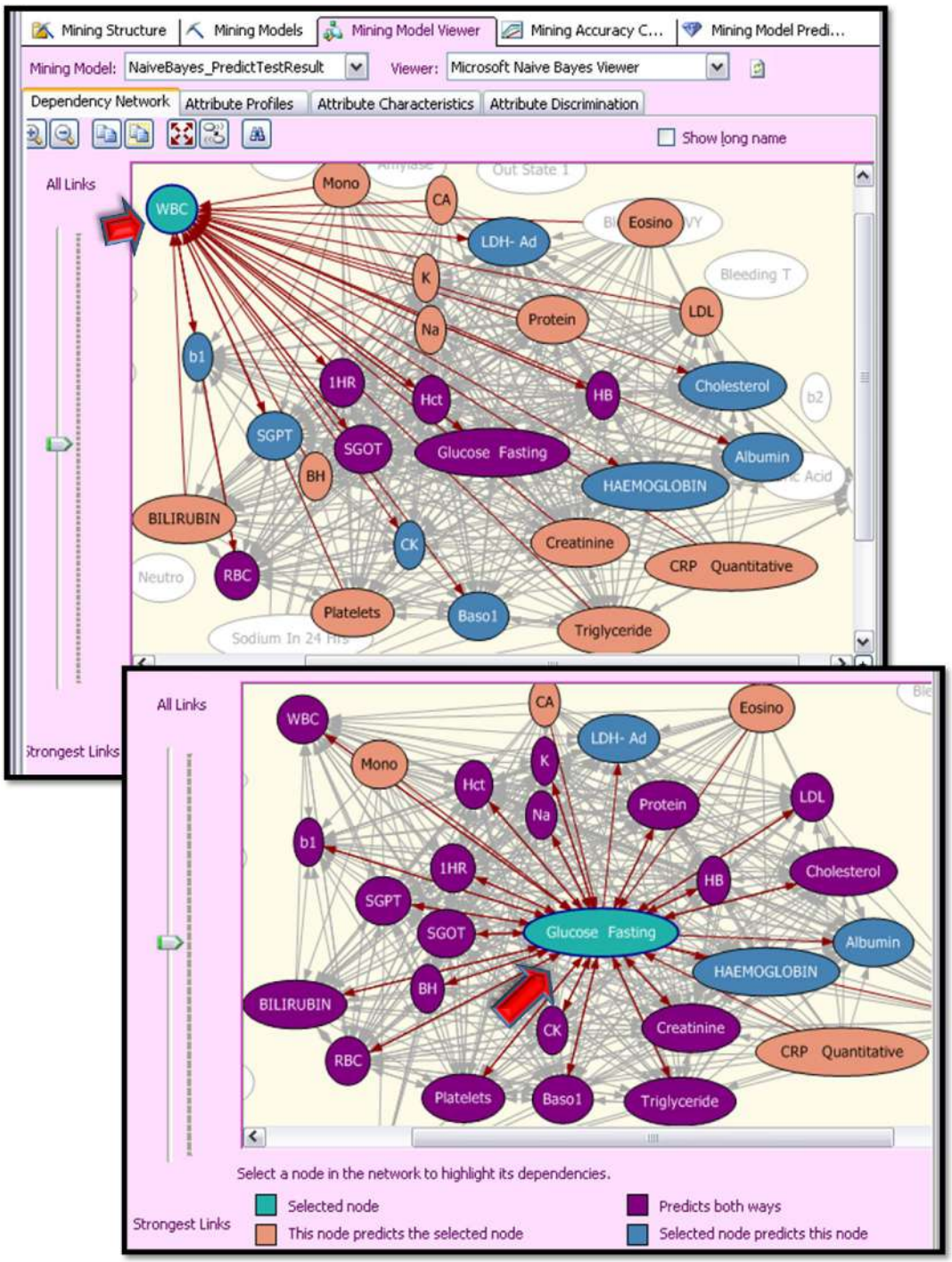


الشكل (5–4– 29): نافذة القواعد في نموذج التنقيب المستند إلى خوارزمية قواعد الاقتران
(AssociationRules_PredictTestResult)، مرتبة تنازلياً بحسب قوة القواعد ويظهر في الشكل بعض قواعد الاقتران الثلاثية القوية

3•3•2•4•5 نموذج نايف بايز (NaiveBayes_PredictTestResult)

أ- شبكة التبعية في نموذج (NaiveBayes_PredictTestResult)

أظهرت نافذة شبكة التبعية في هذا النموذج العلاقات الترافقية بين التحاليل الطبية، وإن الضغط بزر الفأرة على عقدة اسم أحد التحاليل الطبية في هذه الشبكة يساعد في التعرف على التحاليل الطبية المؤثرة والمتأثرة بهذا التحليل الطبي بغض النظر عن نتائج كل منهم، ويبين الشكل (5–4–30) أهم التحاليل الطبية المترافقة مع كل من التحليلين الطبيين تعداد الكريات البيضاء (WBC) والسكر الصيامي (Glucose-Fasting).

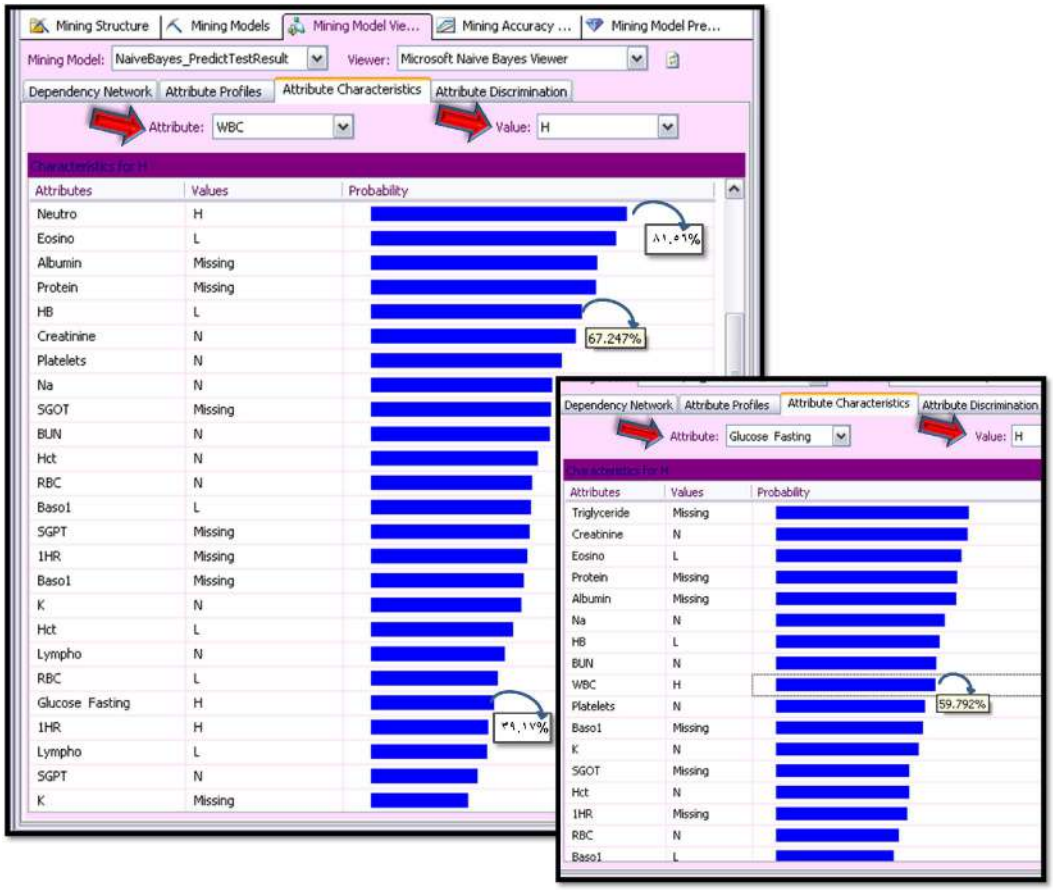


الشكل (5-4- 30): نافذة شبكة التبعية في نموذج التنقيب (NaiveBayes_PredictTestResult)، وتظهر في الشكل الأعلى العلاقات حول تحليل تعداد الكريات البيضاء (WBC)، بينما تظهر في الشكل الأسفل العلاقات حول تحليل السكر الصيامي (Glucose-Fasting)

ب- نافذة صفات الخاصيات (Attribute Characteristics)

يساعد التنقيب في بيانات التحاليل الطبية باستخدام نموذج نايف بايز في التعرف على احتمال ظهور النتائج المختلفة للتحاليل الطبية بالترافق مع قيمة محددة للتحليل الطبي المتنبأ به، وذلك في نافذة صفات الخاصيات. ونقترح مثالاً على ذلك استعراض احتمالات ترافق نتائج التحاليل

المختلفة مع ارتفاع تعداد كريات الدم البيضاء (WBC)، ويمكن التعبير عن مخرجات هذا النموذج كما في المثال التالي: إن ارتفاع نتيجة تحليل تعداد كريات الدم البيضاء (WBC=H) يترافق معه ارتفاع نتيجة تحليل العدلات (Neutro)[1] باحتمال قدره 81.56%، وانخفاض بنتيجة تحليل الخضاب (HB=L) باحتمال قدره 67.247%، وارتفاع في نتيجة تحليل السكر الصيامي (Glucose-Fasting=H) باحتمال قدره 39.17%..الخ، لاحظ النافذة اليسرى في الشكل (5-4-31). وبالمقابل إذا كان السكر الصيامي مرتفع عند مرضى المشفى فإن تعداد كريات الدم البيضاء يكون مرتفع باحتمال قدره 59.79%، انظر النافذة اليمنى في الشكل (5-4-31).



الشكل (5-4-31): أمثلة لنافذة صفات الخاصيات في نموذج (NaiveBayes_PredictTestResult)، ويظهر في النافذة اليسرى قيم التحاليل الطبية المرافقة لارتفاع تعداد الكريات البيضاء (WBC) واحتمال كل منها، ويظهر في النافذة اليمنى قيم التحاليل الطبية المرافقة لارتفاع تحليل السكر الصيامي (Glucose Fasting) واحتمال كل منها

[1] Neutro: العدلات (Neutro-phil) هي إحدى أنواع كريات الدم البيضاء وترتفع بشكل عام في حالات الالتهاب، وتنخفض في بعض حالات فقر الدم.

4•3•2•4•5 نموذج خوارزمية الشبكات العصبونية (NeuralNetwork_PredictTestResult)

يمتلك نموذجُ التنقيب المستند إلى خوارزمية الشبكات العصبونية نافذةً مرنةً لعرض النتائج، حيث يقوم فيها النموذج بإعطاء علامات (Score) بين [0–100] تعبِّر عن قوة تأثير نتائج التحاليل الطبية المختلفة على ترجيح إحدى نتيجتين للتحليل الطبي المتنبأ به المختار. ويبين الشكل (5–4–32) قوة تأثير قيم التحاليل الطبية المختلفة على ترجيح ارتفاع أو انخفاض تعداد الكريات البيضاء (WBC) لدى مرضى المشفى. ونقرأ في هذا الشكل النتائج التالية:

1- إن ارتفاع نتيجة التحليل الطبي لحمض البول (Uric Acid) يُرجِّح انخفاض تعداد كريات الدم البيضاء – وليس ارتفاعها– بقوة ترجيح 85.45.

2- إن انخفاض نتيجة التحليل الطبي للكولسترول من نوع (LDL) يُرجِّح ارتفاع تعداد كريات الدم البيضاء – وليس انخفاضها– بقوة ترجيح 60.48.

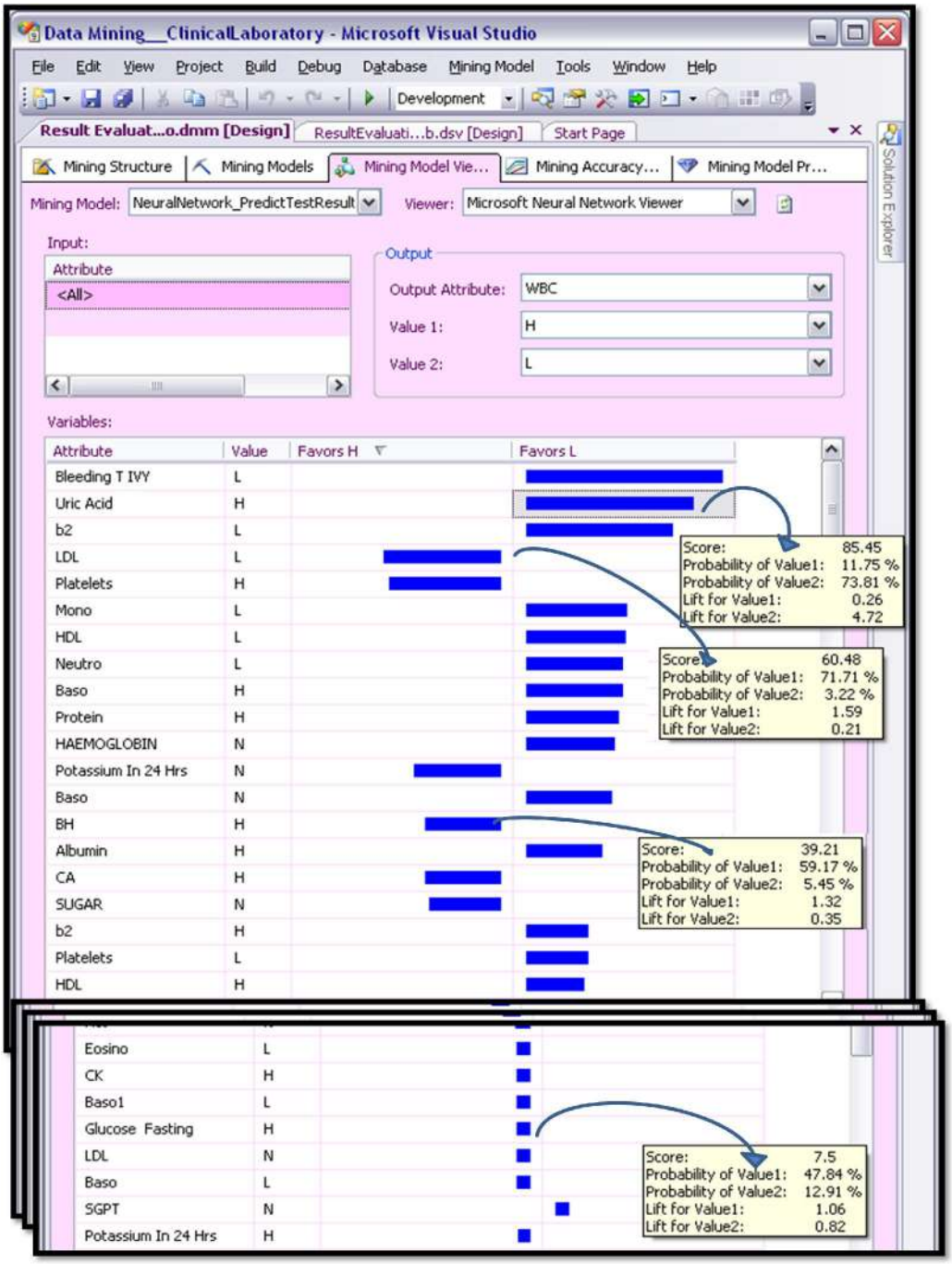
3- إن ارتفاع نتيجة تحليل السكر الصيامي يُرجِّح أن يكون تعداد الكريات البيضاء مرتفعاً وليس منخفضاً بعلامة ترجيحية قدرها 7.5. وبمقارنة هذه النتيجة مع نتيجة نموذج نايف بايز (راجع الشكل (5–4–31)) نجد أن ارتفاع السكر الصيامي يترافق معه ارتفاع تعداد الكريات البيضاء باحتمال قدره 59.79%، أي أن طريقة التقييم تختلف من نموذج إلى آخر.

4- إن ارتفاع نتيجة التحليل الطبي (BH) – وهو أحد تحاليل الحمى التيفية – يُرجِّح أن يكون تعداد الكريات البيضاء مرتفعاً وليس منخفضاً بعلامة ترجيحية قدرها 39.21. ونلاحظ أن القيم المرتفعة لتحليل (BH)[1] لم يكن لها وجود ضمن القواعد الهامة في نموذج قواعد الاقتران بل كان التواجد الغالب هو للقيم الطبيعية لهذا التحليل، ويدل ذلك على أن نموذج الشبكات العصبونية يُظهر تأثير القيم قليلة التكرار في التحاليل الطبية على نتائج التحليل المتنبأ به، ولكن بالمقابل يجب الحرص على تجنب تعميم مثل هذه النتائج غير المدعومة بعددٍ كافٍ من الحالات.

5- يدرس نموذج الشبكات العصبونية العلاقات البينية بين القيم غير المفقودة في جدول التحاليل الطبية، أي أنه يتجاهل تأثير القيم المفقودة في المدخلات (Inputs) على قيم التحاليل الطبية المتنبأ بها، وهذا ما يجعل التأثير الحقيقي لنتائج بعض التحاليل الطبية يظهر

[1] تكرر ارتفاع نتيجة تحليل (BH) مرتين فقط في البيانات المدروسة، بينما تكررت النتيجة الطبيعية (N) لهذا التحليل 2365 مرة وذلك بحسب نتائج الاستعلام من مستودع البيانات المدروس.

في هذا النموذج على خلاف غيره من النماذج التي سبق ذكرها مثل نموذج نايف بايز. ففي الواقع نجد أن التحاليل الطبية التي تُخزَّن في السجل الطبي للمريض هي فقط التحاليل التي تُجرى له والمتناسبة مع حالته المرضية مما يؤدي إلى الانتشار الكبير للقيم المفقودة في حقول التحاليل الطبية التي لم تُجرى له. وبناءً على ذلك نجد أن نموذج الشبكات العصبونية قد يكون أكثر توافقاً من النماذج السابقة مع هذه الطبيعة لبيانات التحاليل الطبية.



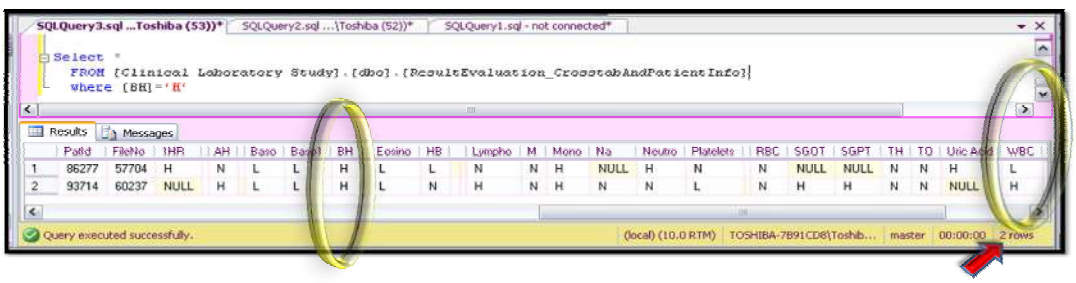
الشكل (5-4- 32): نافذة نتائج نموذج الشبكات العصبونية (NeuralNetwork_PredictTestResult) المتعلقة بتقييم تأثير كل من نتائج التحاليل الطبية على ارتفاع أو انخفاض تحليل تعداد كريات الدم البيضاء (WBC)

5•3•2•4•5 نموذج الانحدار المنطقي (LogisticRegrision_PredictTestResult)

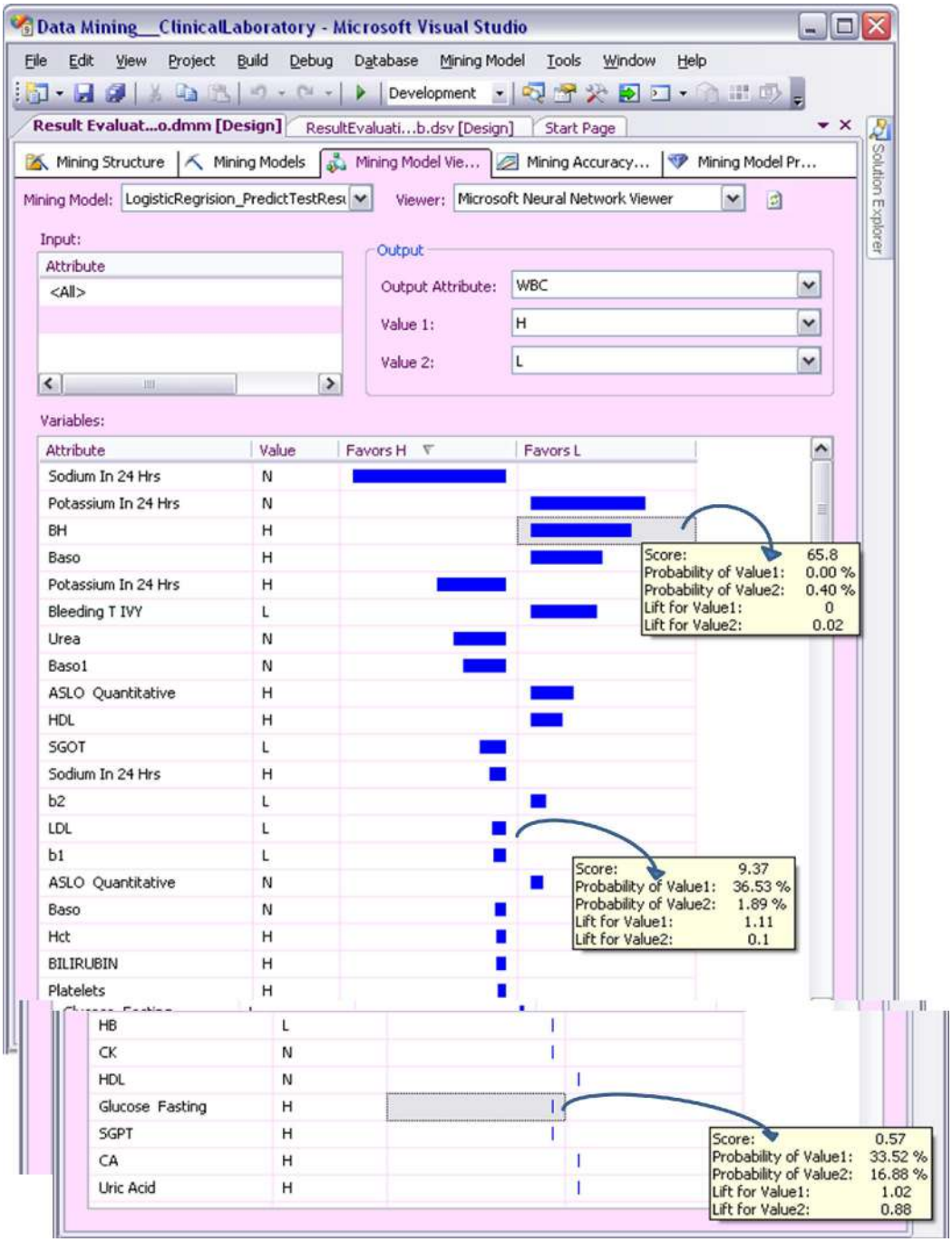
تتشابه نافذة نتائج نموذج الانحدار المنطقي مع الشكل العام لنافذة نموذج الشبكات العصبونية، ولكنها تختلف في طريقة تقييم النتائج، ويُظهر الشكل (5-4-34) قوة تأثير قيم التحاليل الطبية المختلفة على ترجيح ارتفاع أو انخفاض تعداد الكريات البيضاء (WBC) لدى مرضى المشفى باستخدام نموذج الانحدار المنطقي. ونعرض فيما يلي لبعض الاستنتاجات من هذا النموذج:

1- إن انخفاض نتيجة تحليل الكولسترول من نوع (LDL) يُرجِّح ارتفاع تحليل (WBC) بعلامة ترجيحية (60.48)، وبمقارنة هذه النتيجة مع النموذج السابق نجد أن نموذج الشبكات العصبونية أعطى نفس هذه النتيجة ولكن بعلامة ترجيحية أقل هي (9.37).

2- إن ارتفاع نتيجة تحليل (BH) يُرجِّح انخفاض نتيجة تحليل تعداد الكريات البيضاء بعلامة ترجيحية (65.8)، ولكن تتناقض هذه النتيجة كثيراً مع النتيجة التي صدرت عن نموذج الشبكات العصبونية والتي نصَّت على أن ارتفاع نتيجة تحليل (BH) يُرجِّح ارتفاع نتيجة تحليل تعداد الكريات البيضاء. وقد تبين من خلال الاستعلام عن سجلات المرضى التي تحتوي على قيم مرتفعة لتحليل (BH) – بالعودة إلى مستودع البيانات- أنه يوجد فقط سجلين موافقين، وأن أحدهما فيه قيمة منخفضة لتعداد الكريات البيضاء والآخر مرتفعة، انظر الشكل (5-4-33). ولذلك لا يمكن القول بأن أحد النموذجين أخطأ والثاني أصاب فلكل منهما وجهة نظر مختلفة في تقييم الأمور، وطالما أن تكرار البيانات التي نناقشها قليل جداً، فإنه لا يمكن الحكم بدقة من خلال هذه النتائج على جودة أي من هذين النموذجين.



الشكل (5-4- 33): الاستعلام من مستودع البيانات عن سجلات المرضى التي تحتوي على نتائج مرتفعة في تحليل (BH)، ونلاحظ أنه يوجد فقط سجلين من سجلات المرضى، وأحدهما فيه قيمة منخفضة لتعداد الكريات البيضاء والآخر مرتفعة



الشكل (5-4- 34): نافذة نموذج الانحدار المنطقي (LogisticRegrision_PredictTestResult) المتعلقة بتقييم تأثير كل من نتائج التحاليل الطبية على ارتفاع أو انخفاض تحليل تعداد كريات الدم البيضاء (WBC)

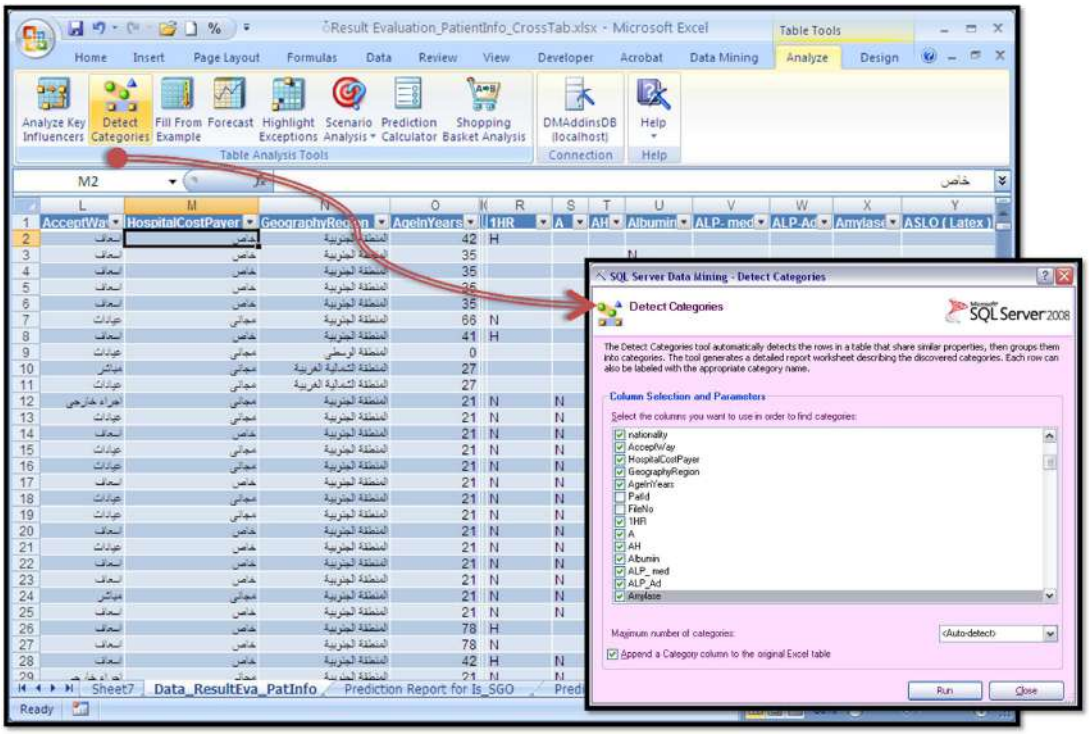
6•3•2•4•5 نموذج التجزئة العنقودية (Clustering Model) باستخدام أداة اكتشاف الفئات في برنامج اكسل

أُستخدِمت خوارزمية التجزئة العنقودية هنا لتعمل على البحث عن سجلات مرضى المشفى والتحاليل الطبية التي تحتوي على صفات مشتركة وتضعهم في فئات تحمل هذه الصفات. وقد تم اختيار استخدام برنامج اكسل لتطبيق هذه الخوارزمية لأن أداة اكتشاف الفئات التي تستخدم

هذه الخوارزمية والموجودة في برنامج اكسل تَعرض مخرجات التنقيب بطريقة متطورة ومرنة وأفضل من تلك التي يوفرها برنامج (Visual Studio)، راجع الفقرة (5•4•1•3•3).

تم إنشاء مرأى في مستودع بيانات التحاليل الطبية، واشتمل على البيانات الشخصية والمرضية للمرضى، وعلى تقييم نتائج التحاليل الطبية بشكلها النهائي الذي خصص فيه حقل لكل تحليل طبي، ثم استدعي إجراء استيراد هذا المرأى من برنامج اكسل، ونتجَ بالتالي جدول صالح لتطبيق أدوات التحليل والتنقيب عليه.

توجد أداة اكتشاف الفئات (Detect Categories) في اكسل ضمن قائمة التحليل (Analyze)، وتعمل هذه الأداة على إيجاد أسطر الجدول المتشابهة الخواص، ومن ثم تُجَزِّئُهم إلى فئات (Categories). وقد جرى استدعاء هذه الأداة واختيار معظم البيانات الشخصية والمرضية ونتائج التحاليل الطبية للمرضى كما هو مبين في الشكل (5–4–35). وتولد عن تنفيذ هذه الأداة أربعة أشكال من المخرجات كما هو آت:



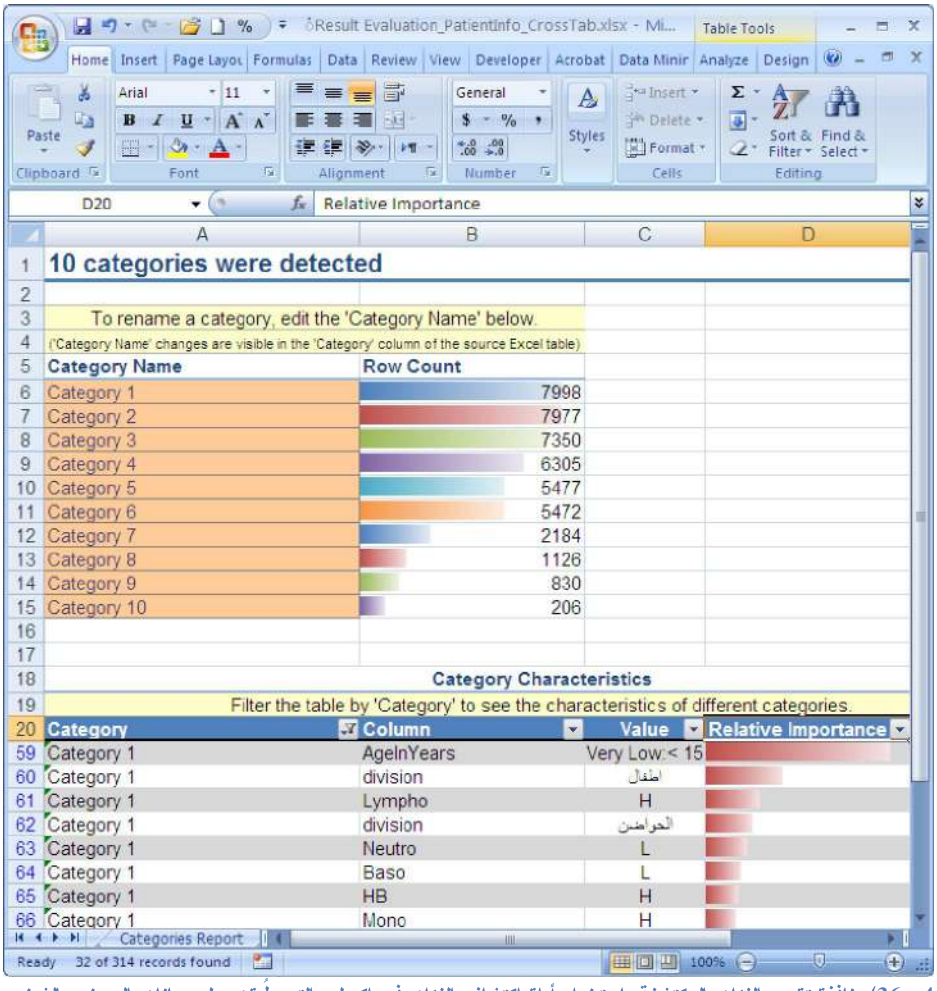
الشكل (5–4– 35): استدعاء أداة اكتشاف الفئات (Detect Categories) في اكسل واختيار الخاصيات المطلوبة لعملية التجزئة

أ– قائمة أسماء الفئات وأحجامها

تحتوي هذه القائمة على عشر فئات للبيانات مرتبة تنازلياً بحسب حجم كل منها، انظر في أعلى الشكل (5–4–36).

ب- جدول صفات الفئات (Category Characteristics)

يختص هذا الجدول بعرض صفات كل من الفئات الناتجة، حيث يمكن اختيار الفئة أو الفئات المطلوب استعراض خواصها من خلال ميزة الانتقاء (Filter) المرفقة مع هذا الجدول، ويعرض هذا الجدول قيم الخواص التي تتضمنها كل فئة ومرتبة أيضاً بحسب أهميتها النسبية أي بحسب مساهمتها في إضفاء الصفات الأساسية على الفئة، انظر القسم الأسفل من الشكل (5-4-36).



الشكل (5-4- 36): نافذة تقرير الفئات المكتشفة باستخدام أداة اكتشاف الفئات في اكسل والتي طُبقت على بيانات المرضى الشخصية والمرضية ونتائج التحاليل الطبية، وتظهر في الأعلى قائمة الفئات العشرة المكتشفة وحجم كل منها، ويظهر في الأسفل جدول مرن لعرض صفات كل فئة

وقد أُجريت دراسة صفات كل فئة من الفئات الناتجة بتمعن مع الأخذ بعين الاعتبار الأهمية النسبية المرفقة بكل صفة، وتم التوصل إلى تعيين أسماء لكل فئة من الفئات الناتجة بحسب خلاصة الصفات التي تتضمنها، ثم جرى استبدال الأسماء الافتراضية بالأسماء الجديدة في قائمة أسماء الفئات مما أدى إلى التجدد الآلي للأسماء الافتراضية أينما وردت في مخرجات أداة التنقيب هذه، انظر الشكل (5-4-37). ونبين فيما يلي التسميات المقترحة للفئات الناتجة:

1- فئة أمراض وتحاليل الأطفال والحواضن دون عمر 15 سنة.

2- فئة أمراض الاحتشاءات القلبية والدماغية للأعمار التي تزيد عن 51 سنة.

3- فئة الجراحات للأعمار فوق 32 سنة.

4- فئة أمراض الكلية وأمراض القلب في الأعمار التي تزيد عن 51 سنة.

5- فئة أمراض الدم للأعمار بين 15-32 سنة (راجع الملاحظة الهامشية في الصفحة 241).

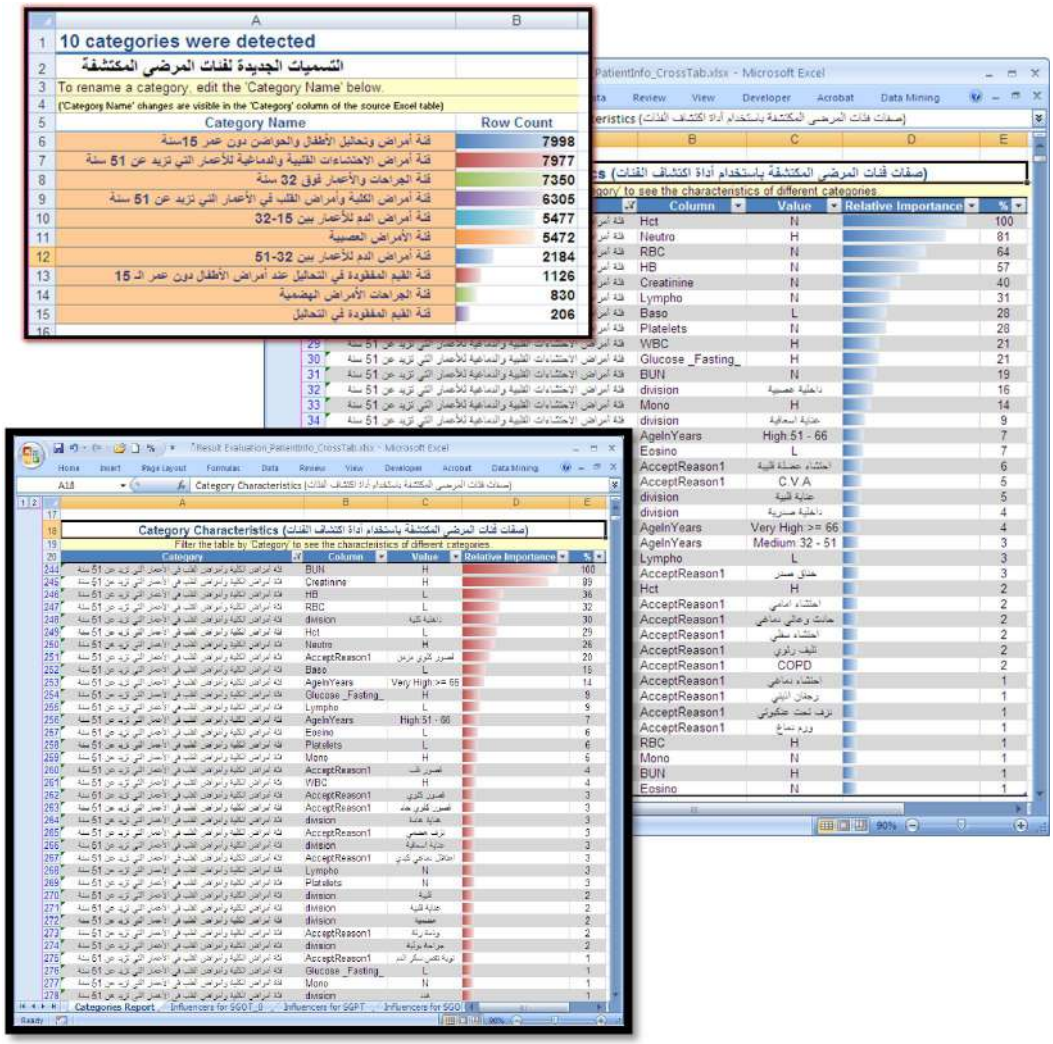
6- فئة الأمراض العصبية.

7- فئة أمراض الدم للأعمار بين 32-51 سنة.

8- فئة القيم المفقودة في التحاليل عند أمراض الأطفال دون عمر 15 سنة.

9- فئة الجراحات والأمراض الهضمية.

10- فئة القيم المفقودة في التحاليل الطبية.



الشكل (5-4- 37): مخرجات أداة اكتشاف الفئات بعد انتقاء التسميات المناسبة للفئات بحسب قيم الخصائص في كل منها، ويظهر في الأعلى قائمة الأسماء الجديدة، وفي الوسط جدول صفات الفئة المسماة (فئة أمراض الاحتشاءات القلبية والدماغية للأعمار التي تزيد عن 51 سنة)، وإلى الأسفل جدول صفات فئة أمراض الكلية وأمراض القلب في الأعمار التي تزيد عن 51 سنة

ويساعد الاطلاع على فئات المرضى في التعرف على الخصائص المميزة لكل فئة، فمثلاً نجد أن أبرز الصفات التي تنطبع بها "فئة أمراض الكلية وأمراض القلب في الأعمار التي تزيد عن 51 سنة" هي ما يلي:

1- تتركز أعمار هذه الفئة في الأعمار العالية التي تزيد عن 51 سنة، وإن تقسيم الأعمار إلى فئات يتم آلياً وداخلياً من قبل أداة التنقيب بما يتناسب مع البيانات.

2- أهم أسباب القبول في المشفى في هذه الفئة هي القصور الكلوي، وقصور القلب.

3- يحقق ارتفاع نتيجة كل من تحليل البولة النتروجينية (BUN) وتحليل الكرياتينين (Creatinine)[1] أهمية عالية بين أفراد هذه الفئة.

4- يحقق انخفاض نتيجة تحليل كل من الخضاب (HB)، وتعداد الكريات الحمراء (RBC)، والهيماتوكريت (Hct) أهمية متوسطة بين أفراد هذه الفئة.

ج- أدوات رسم بياني محوري (PivotChart Tools)

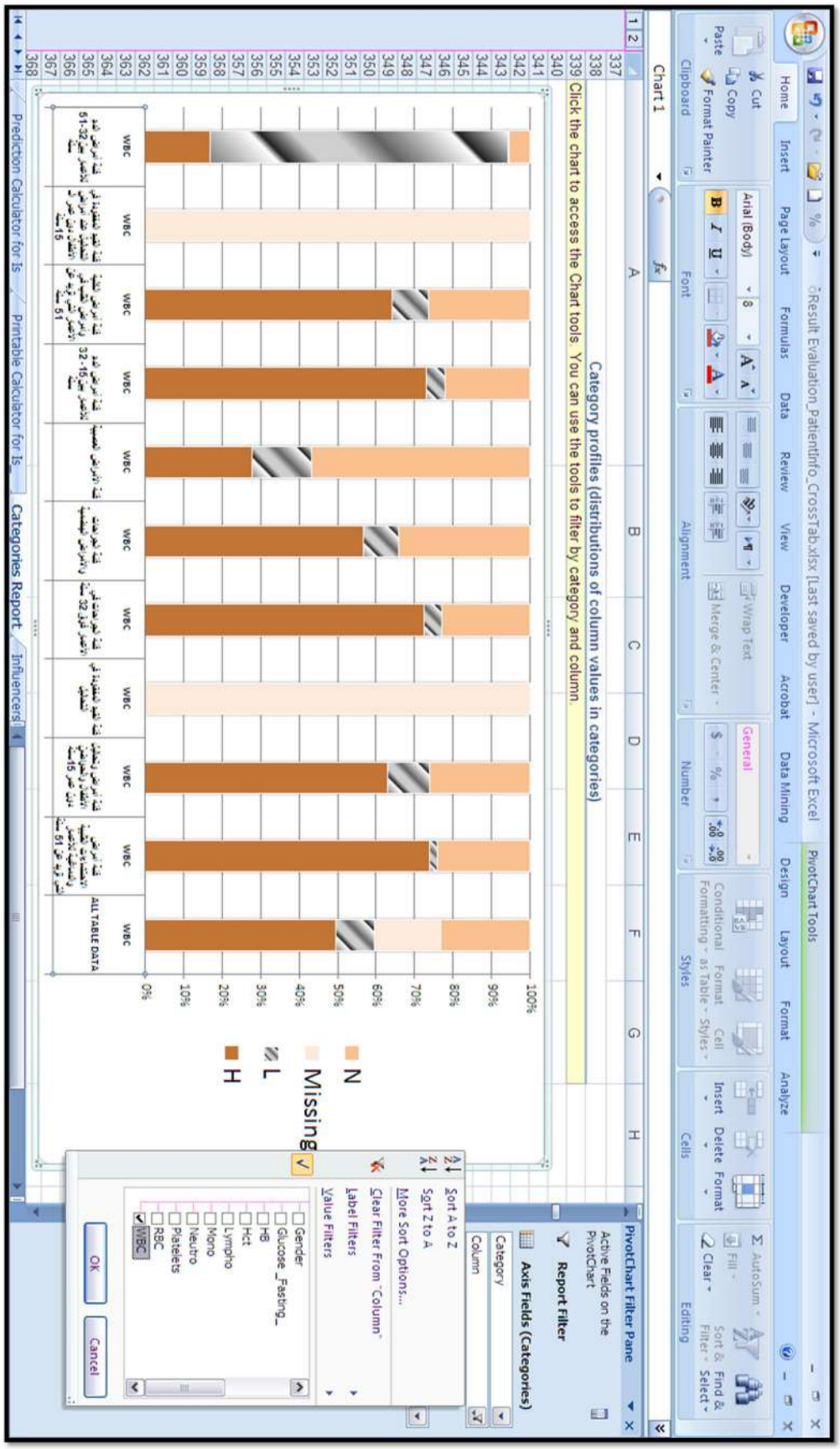
تُنتج أداة اكتشاف الفئات أدوات رسم بياني محوري مرفقة بقائمة تساعد على مشاهدة المخرجات من زوايا مختلفة، حيث يمكن مثلاً انتقاء مشاهدة الفئات بحسب واحدة من الخاصيات المتوفرة كالعمر أو مكان الإقامة أو نتائج أحد التحاليل. أي تُقدِّم هذه الأداة عدداً كبيراً من الاستعلامات بحسب الأفكار التي تخطر ببال المستخدم. ويبين الشكل (5-4-38) هذا الرسم البياني وذلك بعد تخصيصه بعرض نتائج تحليل كريات الدم البيضاء (WBC) في كل فئة. ونقرأ في هذا الرسم الملاحظات التالية:

1- نلاحظ أن هنالك فئتين من المرضى لم يُجرى لهم هذا التحليل إطلاقا (وهما الفئتان اللتان برزت فيهما صفات القيم المفقودة في التحاليل الطبية).

2- تتركز النتائج المنخفضة في هذا التحليل كثيراً في فئة أمراض الدم للأعمار بين 32-51.

3- تكثر النتائج الطبيعية (N) لتحليل كريات الدم البيضاء في مجموعة الأمراض العصبية.

[1] الكرياتينين هو مؤشر على سلامة وظائف الكلية، ويرتفع في حالات الفشل الكلوي الحاد والمزمن وحالات الانسداد البولي

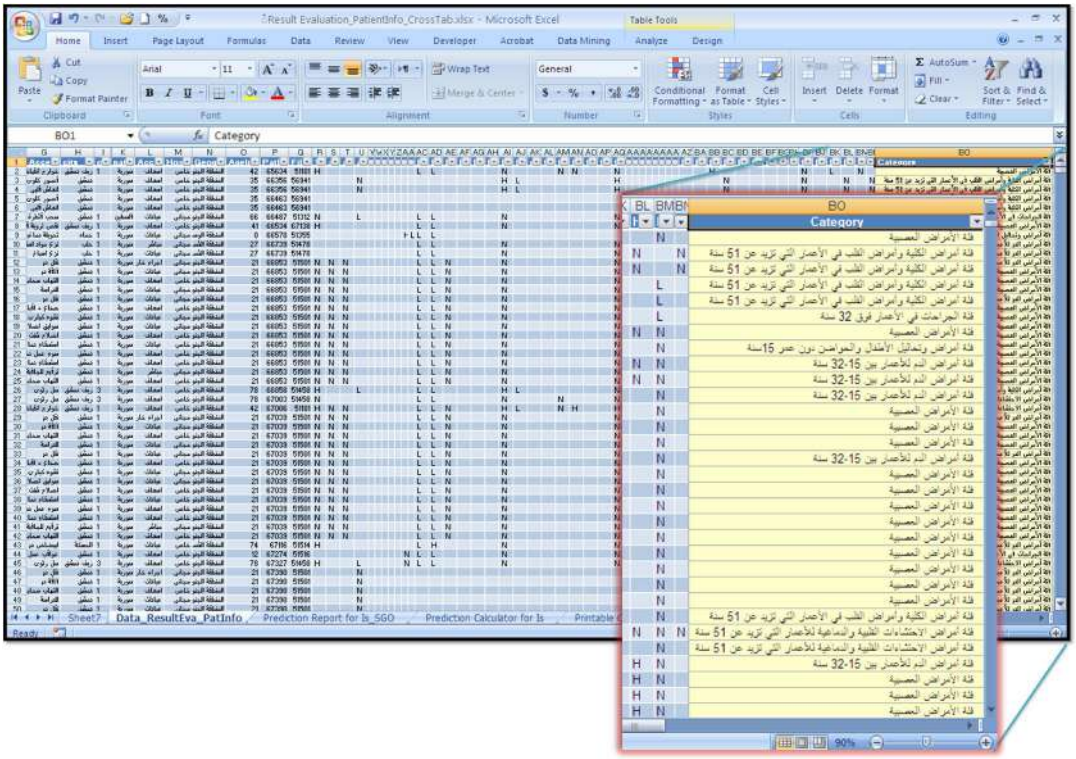


الشكل (5-4- 38): مخطط محوري مرفق مع تقرير فئات المرضى الناتج عن استخدام أداة التنقيب (Detect Category)، وقد جرى اختيار استعراض نتائج تحليل كريات الدم البيضاء (N, H ,L, Missing) في كل فئة من فئات المرضى

د- إنشاء حقل أسماء الفئات

تعمل أداة اكتشاف الفئات – اختيارياً- على إضافة عمود جديد إلى جدول البيانات الأساسي، ويتضمن هذا الحقل اسم الفئة التي ينتمي إليها كل سطر في الجدول، وتستجيب أسماء الفئات في هذا الحقل إلى التغيرات التي تُجرى في قائمة أسماء الفئات كما سبق وذُكر، ويبين الشكل (5-4-39) جدول البيانات الأساسي وقد ألحق به حقل أسماء الفئات بحلتها الجديدة.

يساعد استخدام أداة التنقيب هذه على التعرف على أسماء التحاليل الطبية في كل فئة من فئات المرضى ومن الممكن مستقبلاً وباستخدام بيانات أجود أن يتم تقسيم التحاليل إلى مجموعات وفق فئات المرضى، ويفيد ذلك في إيجاد بنى هرمية جديدة للتحاليل الطبية لاستخدامها في تطوير مكعب بيانات التحاليل الطبية كما سبق وأشرنا في الفقرة (**10•4•3•5**- ب).

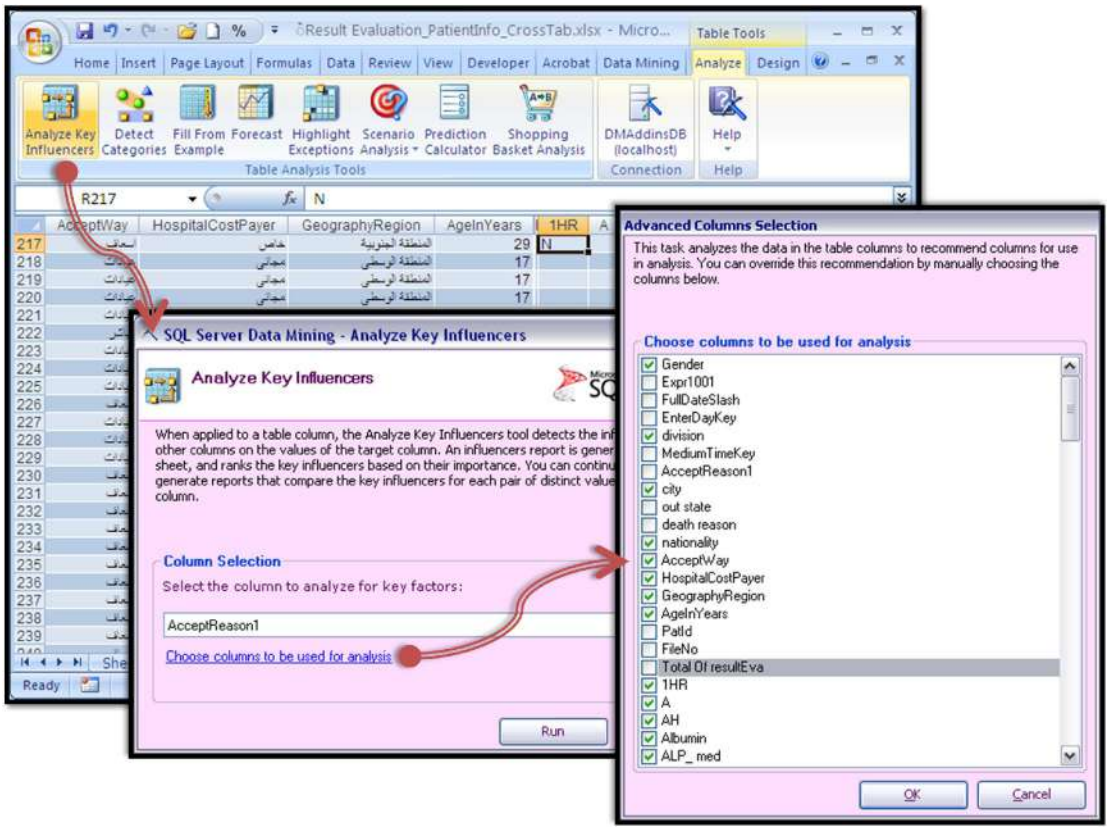


الشكل (5-4- 39): يظهر إلى اليمين الحقلُ الذي أضافته أداة التنقيب (Detect Category) إلى جدول البيانات المنقب فيها، ويَعرض هذا الحقل اسم الفئة التي ينتمي إليها كل مريض

7•4•2•4•5 نموذج تحليل عوامل التأثير الرئيسة (Analyze Key Influencers) في اكسل

تعمل أداة تحليل عوامل التأثير الرئيسة على استخدام خوارزمية التجزئة العنقودية لكي تُمكِّن المستخدم من اختيار عمود البيانات الذي يتضمن المخرجات المستهدفة أو القيمة المستهدفة، ومن ثم تقوم بتحليل الأنماط الموجودة في بيانات الجدول المدروس، وتحدد بالتالي أي العوامل يمتلك التأثير الأقوى على المخرجات المستهدفة.

وقد استُخدمت هذه الأداة هنا للتعرف على تأثير قيم الخاصيات المختلفة (البيانات الشخصية والمرضية ونتائج التحاليل الطبية للمرضى) في كل نوع من أنواع أسباب القبول. أي التعرف على الخاصيات الأبرز المرافقة لكل سبب من أسباب القبول، ولتحقيق ذلك تم اختيار حقل أسباب القبول ضمن خانة (Column Selection)، وتم اختيار الحقول الأخرى التي سيُدرس تأثيرها على سبب القبول من القائمة (Choose columns to be used for analysis) كما هو مبين في الشكل (5-4-40).



الشكل (5-4- 40): استدعاء أداة تحليل عوامل التأثير الرئيسة في برنامج اكسل (Analyze Key Influencers) للتنقيب في البيانات الشخصية والمرضية ونتائج التحاليل الطبية

وقد نتج عن استخدام هذه الأداة نوعين من التقارير، الأول تقرير عام عن أسباب القبول والثاني تقرير مقارنات ثنائية بين أسباب القبول، وذلك كما يلي:

أ- جدول تقرير العوامل المؤثرة على سبب القبول

قامت هذه الأداة بتقديم تقرير عن أهم العوامل المؤثرة على بعض من أسباب القبول التي أمكن التعرف على العوامل المؤثرة فيها، حيث لم يمتلك عدد كبير من أسباب القبول خاصيات مرافقة مميزة له. ويَعرض الشكل (5-4-41) جزءاً من هذا التقرير. وقد تم العثور في هذا التقرير على عددٍ من الاستنتاجات[1] نذكر منها:

1- تبين أن ورم الدماغ يتأثر بالمنطقة الجغرافية الشمالية الشرقية بأهمية نسبية 40%، أي من المرجح أن يكون هنالك تأثير لهذه المنطقة الجغرافية في التعرض لورم الدماغ، ولكن يتطلب الحكم الدقيق على هذا الحقيقة أن يتوفر للدراسة مزيدٌ من البيانات من المشافي المختلفة.

2- إن ارتفاع نتيجة تحليل (b1)[2] ترافق أو ساعد في تشخيص مرض التهاب الكبد بنسبة 100%، وترافق ارتفاع نتيجة تحليل (SGPT)[3] بنسبة73%. كما أن الفئة العمرية الأكثر تأثيراً أو ارتباطاً مع تشمع الكبد هي بين 51-66 سنة.

3- يتأثر حدوث النزف الدماغي بالعمر الذي يزيد عن 66 سنة بنسبة 100%.

4- إن ارتفاع نتيجة تحليل السكر في البول (SUGAR) كان مؤثراً بنسبة 96%، في تشخيص الحماض السكري الخلوني، في حين كان تأثير العمر بين 15-32 على هذا المرض نسبته 88%، وكانت نسبة تأثير أن يكون المريض أنثى 52%... الخ

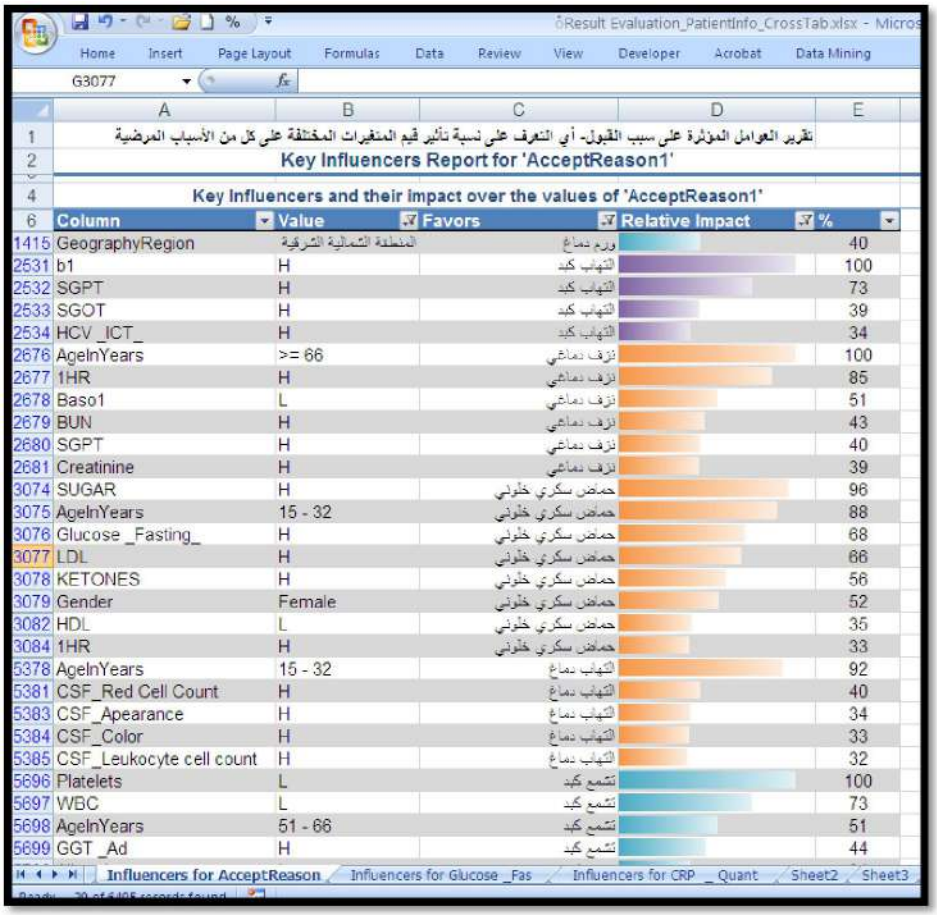
5- أُعطيت للعمر بين 15-32 أهمية نسبية 92% في حدوث التهاب الدماغ.

---

[1] تعتبر هذه النتائج خاصة بالبيانات المدروسة وربما تكون هنالك بيانات مفقودة أو مشوهة خارجة عن سيطرة البحث وقد أدت إلى الوصول إلى هذه الاستنتاجات، ولكن المهم هنا هو بيان قدرات أدوات التنقيب في التوصل إلى مثل هذا النوع من التقارير.

[2] وهو تحليل البيليروبين ويرتفع في عدة حالات مثل أمراض الكبد الناتجة عن تأذي الخلايا الكبدية.

[3] SGPT أنظيم ألانين ترانسفراز (Alanine Transferase) يقيس الوظيفة الكبدية، يزداد في عدة حالات مثل التهاب الكبد الانتاني الحاد. [د. سيد الحديدي، 2006، **دلالات وتفسير النتائج المخبرية**، ص 31، 118]



	A	B	C	D	E
1	تقرير العوامل المؤثرة على سبب القبول- أي التعرف على نسبة تأثير قيم المتغيرات المختلفة على كل من الأسباب المرضية				
2	Key Influencers Report for 'AcceptReason1'				
4	Key Influencers and their impact over the values of 'AcceptReason1'				
6	Column	Value	Favors	Relative Impact	%
1415	GeographyRegion	المنطقة الشمالية الشرقية	ورم دماغ		40
2531	b1	H	التهاب كبد		100
2532	SGPT	H	التهاب كبد		73
2533	SGOT	H	التهاب كبد		39
2534	HCV _ICT_	H	التهاب كبد		34
2676	AgeInYears	>= 66	نزف دماغي		100
2677	1HR	H	نزف دماغي		85
2678	Baso1	L	نزف دماغي		51
2679	BUN	H	نزف دماغي		43
2680	SGPT	H	نزف دماغي		40
2681	Creatinine	H	نزف دماغي		39
3074	SUGAR	H	حماض سكري خلوني		96
3075	AgeInYears	15 - 32	حماض سكري خلوني		88
3076	Glucose _Fasting_	H	حماض سكري خلوني		68
3077	LDL	H	حماض سكري خلوني		66
3078	KETONES	H	حماض سكري خلوني		56
3079	Gender	Female	حماض سكري خلوني		52
3082	HDL	L	حماض سكري خلوني		35
3084	1HR	H	حماض سكري خلوني		33
5378	AgeInYears	15 - 32	التهاب دماغ		92
5381	CSF_Red Cell Count	H	التهاب دماغ		40
5383	CSF_Apearance	H	التهاب دماغ		34
5384	CSF_Color	H	التهاب دماغ		33
5385	CSF_Leukocyte cell count	H	التهاب دماغ		32
5696	Platelets	L	تشمع كبد		100
5697	WBC	L	تشمع كبد		73
5698	AgeInYears	51 - 66	تشمع كبد		51
5699	GGT _Ad	H	تشمع كبد		44

الشكل (5-4- 41): تقرير العوامل المؤثرة على سبب القبول في المشفى باستخدام أداة تحليل عوامل التأثير الرئيسة في برنامج اكسل (Analyze Key Influencers)

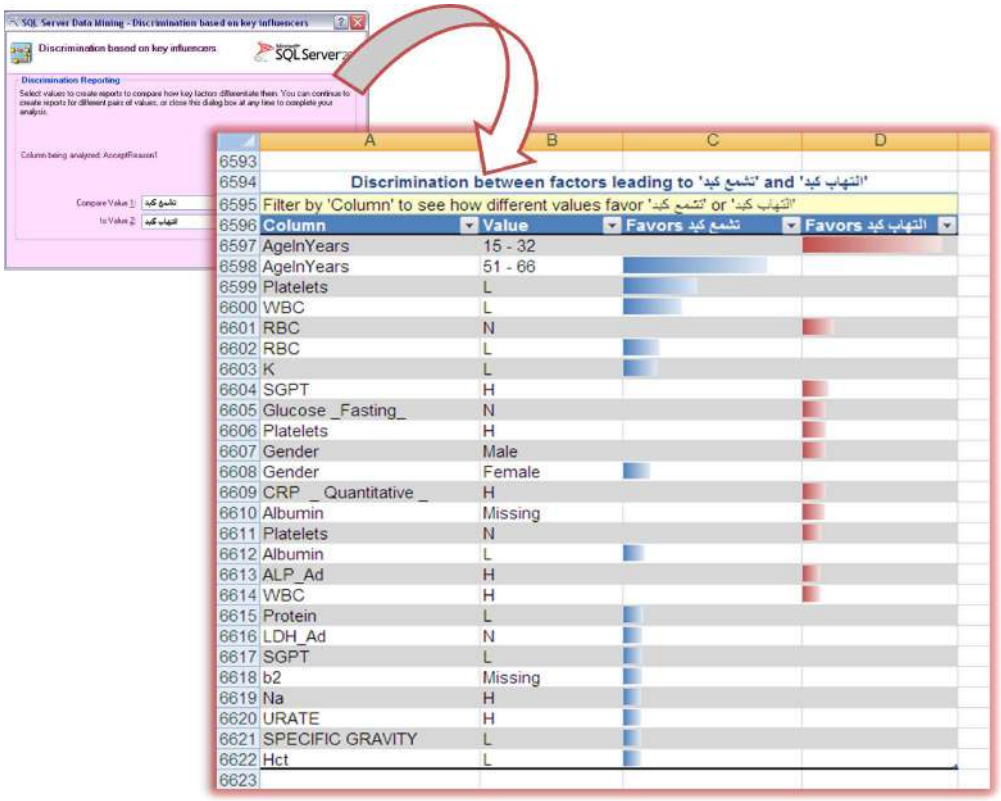
ب- أداة توليد تقارير المقارنة

وتظهر هذه الأداة على شكل نافذة صغيرة لاستدعاء التقارير، وتُستخدم في مقارنة تأثير المتغيرات المختلفة على قيمتين من قيم المتغير المدروس، وتقدم هذه المقارنة معلومات أوسع من تلك التي يقدمها التقرير السابق لأنها تقارن معظم العوامل المؤثرة على كل من القيمتين. وتفيد هذه الأداة هنا في التعرف على الاختلافات بين الأمراض التي تجمعها عناصر مشتركة، فمثلاً نجمت عن مقارنة كل من التهاب الكبد وتشمع الكبد من ناحية العوامل المترافقة مع كل منهما النتائج التالية (انظر الشكل (5-4-42)):

1- يعتبر العمر بين 15-32 سنة على أنه العمر الأكثر ترافقاً مع التهاب الكبد، بينما يعتبر العمر بين 51-66 سنة أكثر ترافقاً مع تشمع الكبد.

2- يتأثر التهاب الكبد بجنس الذكور بينما يتأثر تشمع الكبد بجنس الإناث.

3- تُظهر نتائج تحاليل (Platelets, WBC, RBC,K) انخفاضاً عند المصابين بتشمع الكبد ولا يظهر لها نتائج واضحة عند المصابين بالتهاب الكبد.



الشكل (5-4- 42): أداة توليد تقارير المقارنة في نموذج تحليل عوامل التأثير الرئيسة (Discrimination based on Key Influencers)، ويظهر في الشكل تقرير المقارنة بين تشمع الكبد والتهاب الكبد من ناحية العوامل المؤثرة على كل منهما

ويمكن عمل المزيد من المقارنات بحسب وجهات نظر المستخدم أو المستفيد من هذا النوع من المقارنات.

وهكذا نكون قد تعرفنا على مخرجات مختلفة لأدوات التنقيب في البيانات الطبية، وقد تكون بعض الاستنتاجات الناتجة عن نماذج التنقيب في هذا المبحث بديهية بالنسبة للأخصائيين الطبيين، وإن هذا يعتبر نقطة ايجابية في إثبات قدرات أدوات التنقيب في البيانات في الوقوف على حقيقة بعض الأمور، وبالتالي تزداد الثقة بإمكانيات هذه الأدوات على اكتشاف الحقائق المخفية في البيانات مستقبلاً. ومن ناحيةٍ أخرى إذا وجد الأخصائيون الطبيون أن بعض النتائج تتنافى مع الحقائق الطبية المرتبطة بمرض معين وتتناسب أكثر مع مرض آخر... فإن هذا يستدعي أن تتم إعادة تقييم الأساليب الحالية التي يتبعها بعض الأطباء في الكشف عن الأمراض، أو التحقق من جودة توثيق البيانات الطبية في الأنظمة الحاسوبية.

الفصل الخامس

المبحث الخامس

مرحلة تقييم المعرفة المكتشفة من بيانات مخبر التحاليل الطبية

مقدمة

تكوَّنت مرحلة التنقيب في بيانات مخبر التحاليل الطبية من مشروعي تنقيب، واختص الأول بدراسة البيانات الشخصية والمرضية للمرضى، واشتمل بدوره على مجموعتين من النماذج، واختص المشروع الثاني بدراسة بيانات التحاليل الطبية لمرضى المشفى.

تتوجه بالتالي المرحلة الحالية نحو تقييم تلك النماذج بحسب المشروع (أو مجموعة النماذج) الذي تنتمي إليه، وذلك باستخدام أداتين من أدوات التقييم هما مخطط الارتقاء ومصفوفة التصنيف.

1•5•5 تقييم قدرات نماذج مشروع التنقيب في البيانات الشخصية والمرضية للمرضى

توجَّهت المجموعة الأولى من نماذج التنقيب في هذا المشروع نحو دراسة بيانات المرضى المتوفين في المشفى، وذلك بهدف اكتشاف أسباب الوفاة الأكثر تأثراً بالخاصيات المختلفة للمرضى، وأيضاً بهدف التعرف على قدرة هذه البيانات على تحديد أو تصنيف المرضى المتوفيين ضمن مجموعات. ويتطلب التطبيق الواقعي الفعال لمثل هذه المجموعة من نماذج التنقيب إلى المزيد من خاصيات البيانات عن المرضى، لكي يُستفاد منها في التوصل إلى اكتشافات مفيدة حول توزع الأمراض أو توزع العوامل المؤهِّبة للوفاة، فمثلاً قد يكون لطبيعة عمل المريض (الأمراض المهنية) والوضع العائلي والوزن والأمراض السابقة له، ومكان

الإقامة ومكان العمل..الخ تأثير كبير على التعرض لأسباب معينة للوفاة. وقد اتضح في المبحث السابق أن البيانات المتاحة للدراسة تمتلك فقط جزءاً بسيطاً من هذه القدرات، لأنها تحتوي على عدد محدود جداً من الخاصيات وكانت مخرجاتها أيضاً قليلة، وبالتالي من غير المتوقع أن يَنتج عنها نماذج عالية الأداء، وهذا ما أثبتته نتائج تقييم النماذج في هذه المجموعة، والتي سنأتي على ذكرها في هذا المبحث.

توجَّهت المجموعة الثانية من نماذج التنقيب في هذا المشروع نحو دراسة بيانات المرضى الذين تخرجوا من المشفى بحالة شفاء أو بحالة وفاة، وقد توجهت النماذج المصممة هنا نحو توقع حدوث الوفاة انطلاقاً من البيانات الشخصية والمرضية للمرضى، وعلى اعتبار أن هذه النماذج تسعى إلى التنبؤ بإحدى قيمتين (الشفاء أو الوفاة) فقد كان تقييم أداء هذا النماذج أفضل وأوضح من تقييم أداء نماذج المجموعة الأولى. ولذلك سيتم أولاً استعراض تفاصيل نتائج تقييم المجموعة الثانية من النماذج، ومن ثم ستُعرض نتائج تقييم المجموعة الأولى باختصار.

يتيح برنامج (Visual Studio 2008) إمكانية تقييم قدرات كل من نماذج التنقيب على توقع المتغير المتنبأ به، وذلك باستخدام نافذة مخطط دقة التنقيب (Mining Accuracy Chart) المرفقة مع مشروع التنقيب الذي قمنا ببنائه. وتتم عملية اختبار دقة التنقيب باستخدام ثلاثة نوافذ فرعية هي نافذة الإدخال (Input Selection)، ونافذة مخطط التقييم (Lift Chart)، ونافذة مصفوفة التصنيف (Classification Matrix)، ونبين فيما يلي مكونات عملية تقييم النماذج.

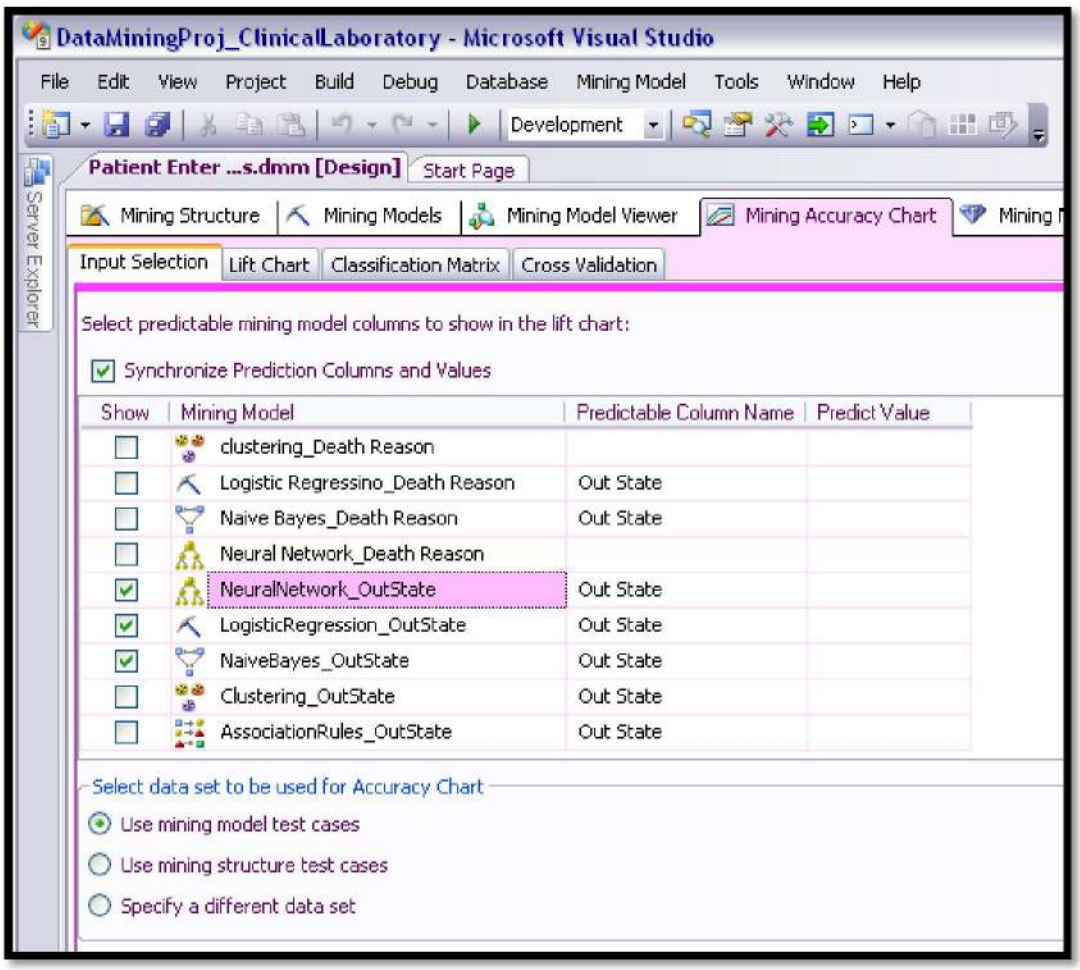
1•1•5•5 اختيار مفردات التقييم

تُعتبر نافذة الإدخال (Input Selection) النافذة الأساسية لاختيار عناصر التقييم، ويبين الشكل (5-5-1) العناصر المختارة في هذا التقييم وهي كما يلي:

أ- تحتوي الخانات الموجودة في حقل اسم عمود التنبؤ في هذا المشروع على اسمين الأول هو سبب الوفاة (DeathReason) ويترافق مع نماذج المجموعة الأولى، والاسم الثاني هو حالة الخروج من المشفى (OutState) ويترافق مع نماذج المجموعة الثانية. وقد تم اختيار أولاً تقييم نماذج التنبؤ الخاصة بحالة الخروج من المشفى (OutState).

ب- جرى انتقاء نماذج التنقيب الخاصة بحالة الخروج والمرغوب مقارنتها، وهي نماذج المجموعة الثانية (NaïveBayes_OutState)، و(NeuralNetwork_OutState)، و(LogisticRegressino_OutState).

ج- تم اختيار أن تكون البيانات التي ستُستخدم في تقييم النماذج هي بيانات الاختبار التي احتجزها البرنامج جانباً عند بناء كل نموذج تنقيب (Mining Model Test Cases) وتعادل نسبتها الافتراضية 30% من البيانات المتوفرة للنموذج. وكان السبب وراء اختيار هذه البيانات لأنها تُماثل البيانات المنتقاة التي عمل عليها النموذج وهي حالات الشفاء وحالات الوفاة، بينما تتضمن بيانات الاختبار في هيكل التنقيب (Mining Structure Test Cases) جميع حالات خروج المرضى أي الشفاء والوفاة والتحويل إلى مشفى آخر والقيم المفقودة.



الشكل (5-5- 1): نافذة اختيار العناصر المرغوبة لاستدعاء أدوات اختبار دقة التنقيب

2•1•5•5 مخطط الارتقاء (Lift Chart) لنماذج توقع حالة خروج المرضى من المشفى

1•2•1•5•5 مخطط الارتقاء (Lift Chart) العام

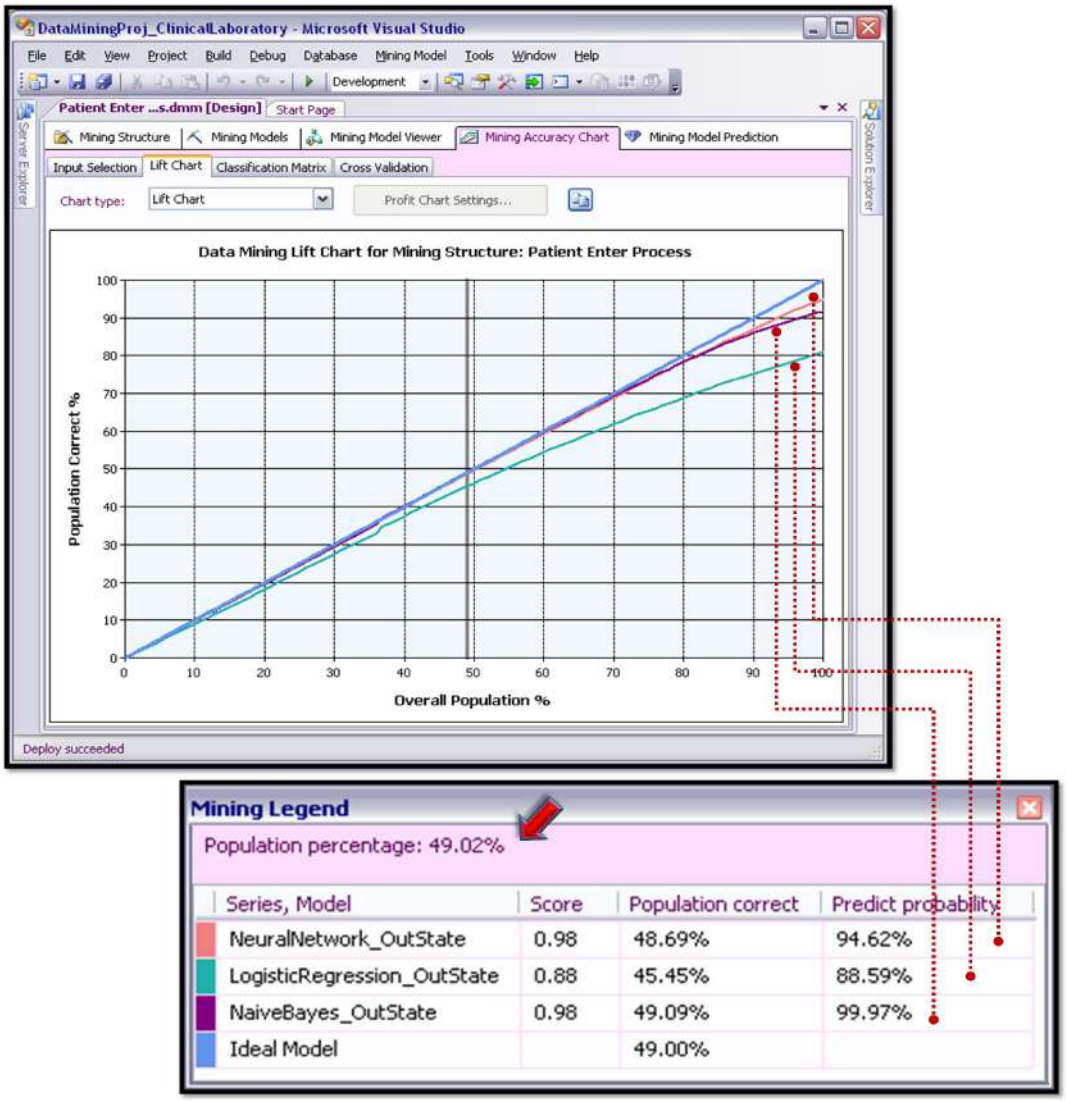
يَعرض مخططُ الارتقاء الذي تم استدعاؤه بناءً على معطيات الإدخال التمثيلَ البياني لقدرات كل من نماذج التنقيب الثلاثة على تحسين مستوى التنبؤ بقيم المتغير المدروس (حالة الشفاء وحالة الوفاة)، ونقرأ في هذا المخطط الموجود في الشكل (5-5-2) ما يلي:

أ- يُستخدم محوري المخطط الأفقي (Overall Population%) والمحور الشاقولي (Population Correct) للتعرف على النسبة الصحيحة من بيانات (الشفاء والوفاة) التي يستطيع أن يتنبأ بها كل من نماذج التنقيب باستخدام نسبة أو جزء من بيانات الاختبار.

ب- يُلاحظ في هذا المخطط أن جميع النماذج تقترب من خط النموذج المثالي (Ideal Model) مما يدل على أن هذه النماذج تتمتع بقدرات تنبؤ عالية.

ج- تَبين في قائمة التفسير المرفقة مع مخطط الارتقاء أن كل من نموذجي التنقيب (NeuralNetworks_OutState) و(NaïveBayes_OutState) قد حصلا على أعلى علامة تقييم (Score) في القدرة على توقع حالة خروج المرضى من المشفى وهي (0.98).

د- تُشير القائمة التفسيرية في هذا المخطط أن نسبة البيانات التي تتناسب مع أفضل الكفاءات لدى جميع نماذج التنقيب هي 49.02% (Population Percentage)، وتدُل هذه النسبة على أنه لو تم استخدام 49.02% من البيانات المتاحة (بيانات الاختبار) فإنه يمكن التنبؤ – بشكل صحيح – بحالة خروج 49.9% من المرضى باستخدام نموذج التنقيب (NaïveBayes_OutState)، بينما يمكن التنبؤ بحالة خروج 48.69% من المرضى باستخدام نموذج التنقيب (NeuralNetworks_OutState)، والتنبؤ بحالة خروج 45.45% باستخدام نموذج التنقيب (LogisticRegression_OutState). أي كلاً من النماذج الثلاثة تصلح بشكل جيد للتنبؤ ولكن أفضل نموذج للتنقيب من ناحيتي علامة التقييم ونسبة البيانات الصحيحة المتنبأ بها هو نموذج (NaïveBayes_OutState).



الشكل (5-5- 2): مخطط الارتقاء (Lift Chart) العام لتقييم قدرة بعض نماذج التنقيب المنشأة على توقع حالتي خروج المريض من المشفى (تحسن أو وفاة)

2•2•1•5•5 مخطط الارتقاء المخصص لتقييم قدرة النماذج على التنبؤ بقيمة محددة

يمكن تخصيص مخطط الارتقاء لاختبار قدرة نماذج التنقيب على التنبؤ بحالة واحدة من حالات الخروج من المشفى وهي حالة الوفاة، وذلك بإضافة خيار إلى نافذة الإدخال وهو انتقاء (حالة الوفاة) في عمود (Predict Value)، راجع الشكل (5-5-1)، ويختلف بالتالي شكل ومعنى مخطط الارتقاء الناتج هنا عن مخطط الارتقاء العام السابق، وهو يُستخدم لحل نمط معين من المسائل، فمثلاً لدينا المسألة التالية:

"يُقدِّم المشفى المدروس خدماتٍ مأجورةٍ للمرضى بأسعار معتدلة، ولكن قد يعجز بعض المرضى عن دفع كامل التكاليف الضرورية مما يقلل من فرصهم في نيل الرعاية الصحية

الكافية ويزيد بالتالي من احتمالات تعرضهم للوفاة، وقد أرادت إحدى الجهات المهتمة بالأمور الطبية أن تدعم علاج الأشخاص المعرضين للوفاة للتقليل من نسب الوفيات وذلك من خلال تقديم معونات مجانية لتكثيف الاهتمام بهم ومتابعتهم بدقة."

إذاً تحتاج هذه الجهة إلى التعرف على أكبر عدد من الأشخاص المستهدفين دون أن تدفع الكثير من التكاليف. ويساعد مخطط الارتقاء المخصص للتنبؤ بحالة الوفاة على حل مثل هذه المشكلة، فهو يعمل على تحديد نسبة المرضى التي يمكن اختيارها لتوجيه العناية الطبية الإضافية لها من أصل جميع المرضى المدروسين، ويبين أيضاً عند كل نسبةٍ الحدَّ الأقصى من المرضى المستهدفين الممكن التعرف عليهم، ونقرأ في هذا المخطط المبين في الشكل (3-5-5) ما يلي:

أ- يمثل المحور الأفقي نسبة المرضى التي يمكن اختيارها لتوجيه العناية الطبية الإضافية لها، ويمثل المحور الشاقولي نسبة المرضى المعرضين للوفاة والمتنبأ بهم في كل نموذج.

ب- يحتوي التمثيل البياني لهذا المخطط على نموذج تخمين عشوائي (Random Guess Model) يأخذ شكل محور قطري في الشكل البياني، ويمثل الخط الأساسي الذي ترتفع فوقه خطوط نماذج التنقيب المختارة.

ج- يحتوي المخطط أيضاً على نموذج مثالي (Ideal Model) يُمثل جميع نماذج التنقيب في هذا المخطط ويأخذ شكل زاوية تعلو التمثيل البياني، ويُعبِّر عن النتائج المثالية لمجموعة بيانات الاختبار فيما لو تم إنشاء نموذج ذو كفاءة عالية في التنبؤ، ويدل نموذج التنقيب المثالي في هذا المخطط على أنه يمكن التوصل مثالياً إلى جميع المرضى المتوقع وفاتهم باستخدام ما يقارب 3.92% من البيانات، وبمعنى أوضح تُعبِّر هذه النسبة تقريباً عن نسبة المرضى المتوفين في بيانات الاختبار وهي (3.7%)[1]. وكلما اقتربت الخطوط البيانية لنماذج التنقيب من هذا الخط عند نقطة (3.92% في المحور الأفقي) كلما ارتقت (زادت) كفاءة نموذج التنقيب.

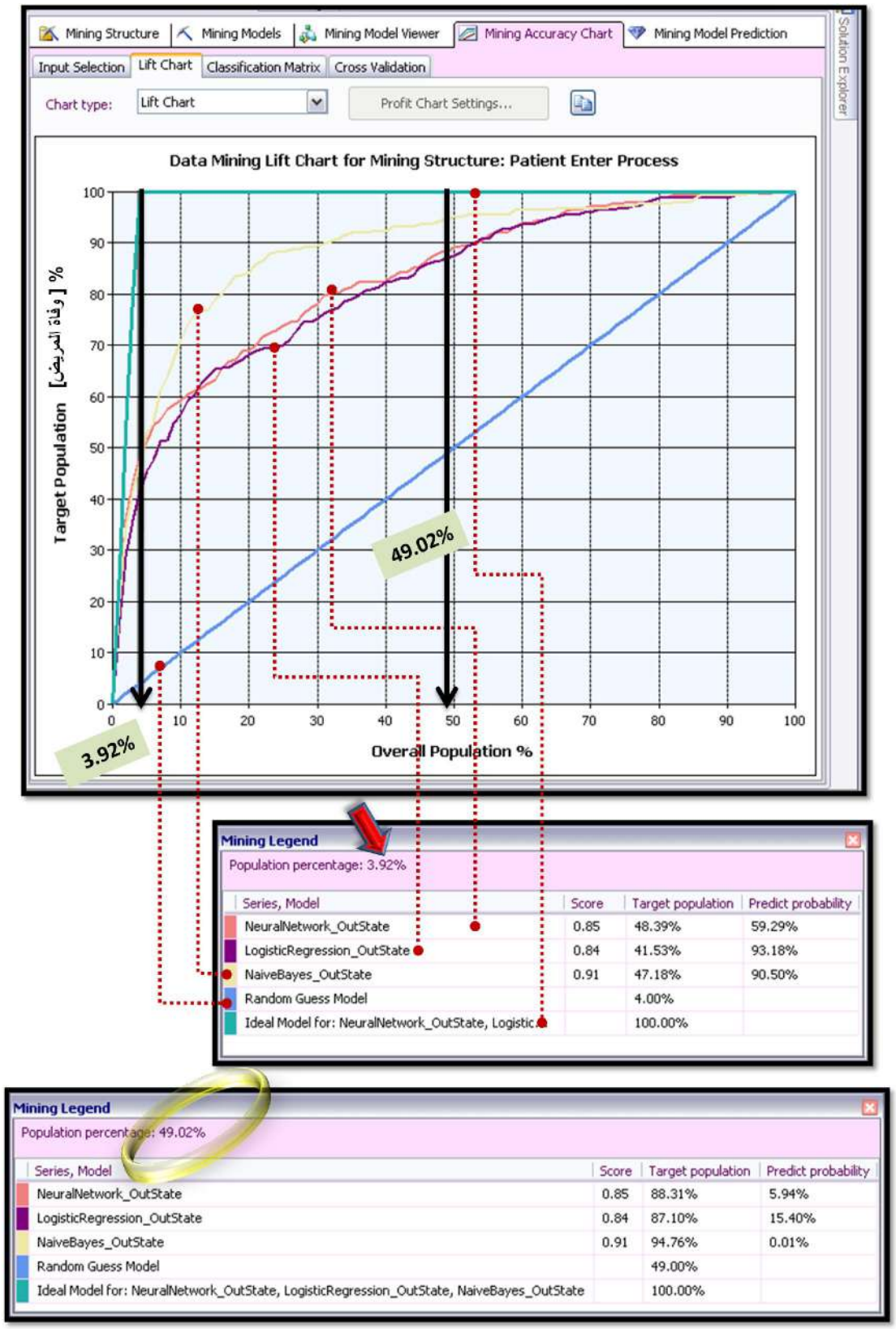
[1] نسبة المرضى المتوفين في بيانات الاختبار في نماذج التنقيب المدروسة (566÷15430×100= 3.7)

د- أظهرت القائمة التفسيرية المرفقة مع المخطط عند نقطة (3.92%) من المحور الأفقي أن نموذج (NaïveBayes_OutState) يستطيع أن يتنبأ بـــ 47.18% فقط من حالات الوفاة لأنه حَصَل على أعلى علامة تقييم لقدرته على التنبؤ وهي (0.91)، يليه نموذج (NeuralNetworks_OutState) الذي حاز على علامة (0.85) ويستطيع أن يتنبأ بـــ 48.39% فقط من حالات الوفاة عند نفس النقطة، أي أن كل من النموذجين يستطيع أن يتنبأ بأقل من نصف المرضى المعرضين للوفاة بينما يستطيع النموذج المثالي أن يتنبأ عند هذه النقطة بكامل المعرضين للوفاة.

ه- قمنا بانتقاء نقاط جديدة على المحور الأفقي للتعرف على نسبة المرضى المعرضين للوفاة الذين يمكن اكتشافهم باستخدام نماذج التنقيب السابقة، ولإظهار دقة تلك النماذج. وتبين على سبيل المثال أنه إذا تم توجيه العناية إلى 49.02% من المرضى في مجموعة الاختبار (الذين ينتقيهم النموذج) فإن العناية ستطال 94.76% من المعرضين للوفاة باستخدام نموذج التنقيب (NaïveBayes_OutState) الذي حصل على أعلى علامة في دقة التنبؤ وهي (0.91)، وهو أفضل نموذج للتنبؤ مقارنة مع النموذجين الآخرين. أي أن الجهة المهتمة بدعم موضوع الصحة عليها أن تُقدم العناية الإضافية إلى نصف المرضى إذا أرادت أن تصل إلى معظم المعرضين للوفاة.

و- إن تحديد النسبة المناسبة من المرضى – لتوجيه الاهتمام المجاني لها- استناداً إلى نتائج مخطط الارتقاء يعود إلى رأي الجهة الطبية التي ستتحمل التكاليف حيث أنه كلما زادت نسبة المرضى المختارين كلما ارتفعت التكاليف[2]. وإن التطبيق الواقعي لهذه الفكرة يحتاج إلى معرفة المزيد من الخاصيات الشخصية والمرضية عن المرضى كالحالة الاجتماعية ونتائج الفحوص الطبية السريرية وقياس الضغط وتخطيط القلب ونتائج الصور الشعاعية..الخ، وعندها ستكون بالتأكيد قدرات تنبؤ نماذج التنقيب أكثر واقعيةً وأقوى من القدرات الناتجة عن البيانات الحالية المتاحة في هذا البحث.

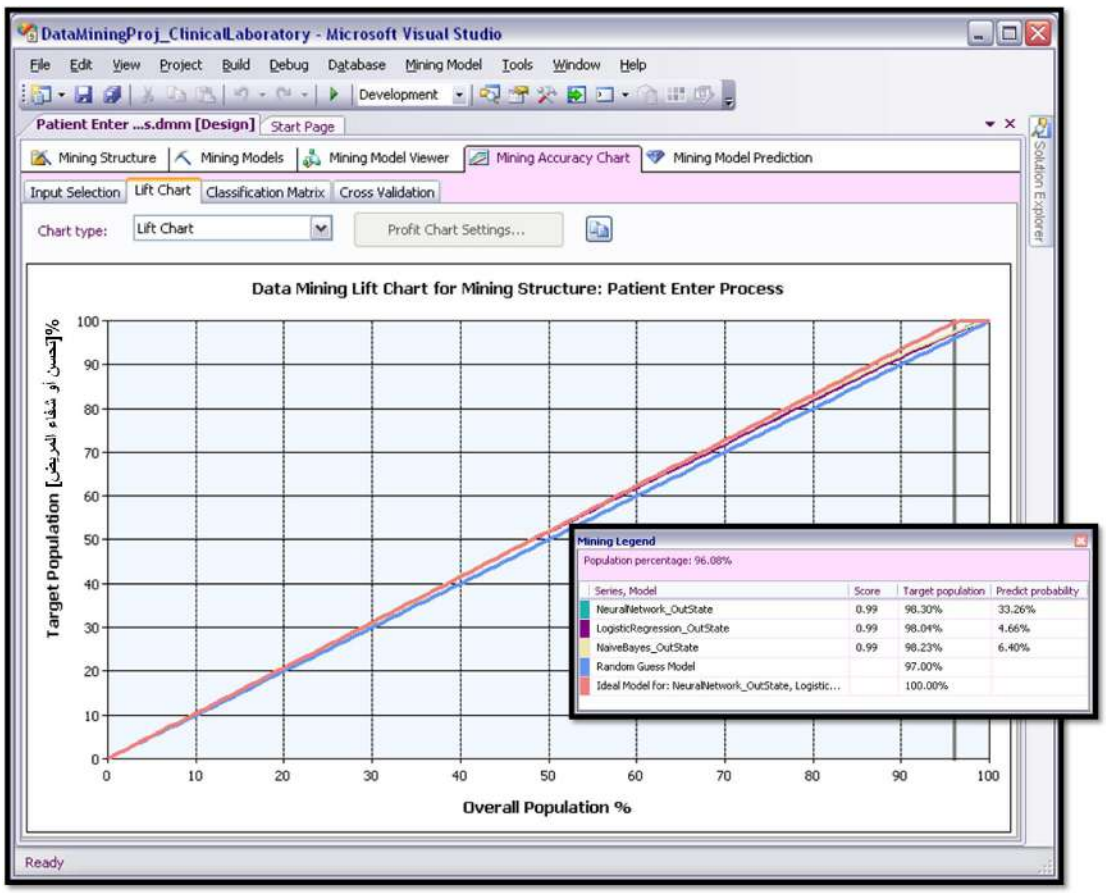
[2] تساعد أداة تنبؤ نموذج التنقيب التي سيرد ذكرها في المبحث التالي على التعرف على المرضى المستهدفين.



الشكل (5–5– 3): مخطط الارتقاء المخصص لاختبار قدرة نماذج التنقيب على التنبؤ بخروج المرضى بحالة وفاة، مرفق بقائمتي تفسير مخطط الارتقاء عند نسبة 3.92%، وعند نسبة 49.02% من المرضى في مجموعة الاختبار

قدَّم مخطط الارتقاء المخصص السابق فائدة كبيرة لعملية تقييم قدرات نماذج التنبؤ وذلك لأن فئة المرضى المتوفين صغيرة مقارنةً بفئة المرضى المتعافين، مما جعل علامات تقييم النماذج في مخطط الارتقاء العام تتأثر بتقييم النتائج التي تشكل الأكثرية على حساب الأقلية. وقد لاحظنا الانخفاض القليل الذي حصل في علامات تقييم النماذج في المخطط المخصص، والانخفاض الكبير الذي ظهر في نسب التنبؤ (Target Population) عما كان عليه الوضع في المخطط العام. ولتوضيح هذه الفكرة لاحظ الشكل البياني (5-5-4) الذي يُمثل مخطط الارتقاء المخصص لتقييم قدرات نماذج التنقيب على التنبؤ بخروج المرضى بحالة تحسن وليس بحالة وفاة، حيث تظهر العلامات المرتفعة التي أعطيت لكل النماذج والتي تفوق العلامات التي نالتها هذه النماذج في مخطط الارتقاء العام.

ونتيجةً لذلك نجد أنه من المهم أن تؤخذ المسألة المدروسة بعين الاعتبار عند تقييم النماذج المكتشفة وذلك لأن قراءة المعارف الناتجة يحتاج إلى معرفة جيدة بالمجال المدروس.



الشكل (5-5- 4): مخطط الارتقاء المخصص لاختبار قدرة نماذج التنقيب على التنبؤ بخروج المرضى بحالة تحسن، وذلك عند نسبة 96.08% من البيانات

3•1•5•5 مصفوفة التصنيف (Classification Matrix) لنماذج التنبؤ بحالة الخروج من المشفى

قُسِّمت بيانات المرضى (في حالتي الشفاء والوفاة) في مشروع التنقيب إلى مجموعتين الأولى مجموعة التجريب (Training Cases) وتشكل 70%، وهي التي اُستخدمت في بناء نماذج التنقيب وتتكون من 566 حالة وفاة ومن 14864 حالة شفاء عند الخروج من المشفى، وتسمى المجموعة الثانية بمجموعة الاختبار (Test Cases) ونسبتها 30% من هؤلاء المرضى، وهي التي تم استخدامها هنا في اختبار نتائج نماذج التنقيب الناتجة، وتتألف هذه المجموعة من 6590 حالة، منها 248 حالة وفاة و6342 حالة تحسن أو شفاء.

استمدَّت مصفوفة التصنيف بياناتها من مجموعة الاختبار التي طُبقت عليها نماذج التنقيب الواردة في الشكل (5-5-1)، وقد قام البرنامج الحاسوبي بتصنيف حالات خروج المرضى في كل نموذج وفق بعدين الأول للقيم الحقيقية (Actual) والثاني للقيم المتنبأ بها (Predicted)، ونتج عن ذلك خلايا تتضمن أعداد الحالات التي أخطأ فيها التنبؤ أو أصاب، انظر الشكل (5-5-5). ونقرأ في هذه المصفوفة ما يلي:

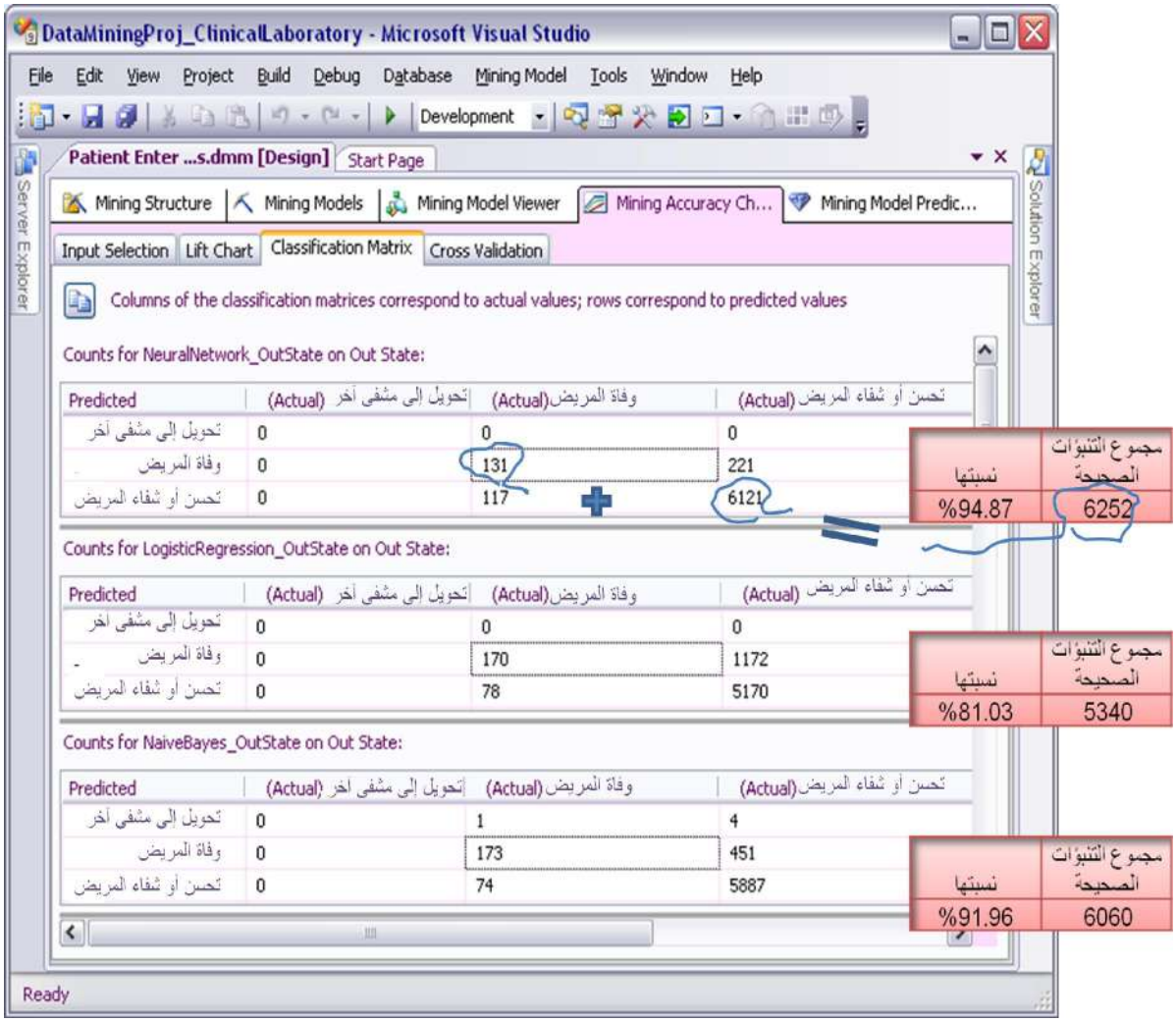
أ- نلاحظ وجود عناوين في المصفوفات لحالات تحويل المرضى إلى مشفى آخر، وإن هذا لا يتعارض مع فكرة عدم وجود هذه الحالات في بيانات التنقيب، فالقيم الحقيقية لها في المصفوفات تساوي الصفر، ولكن يوجد فقط أربع تنبؤات خاطئة بهذه الحالة في نموذج (NaïveBayes_OutState)، وقد تنتج هذه التنبؤات عن ظهور حالات في بيانات الاختبار لم يتمكن النموذج من اختيار الحالة المناسبة لها (شفاء أو وفاة) فتوقع أن تكون منتمية إلى حالة التحويل إلى مشفى آخر.

ب- تمَّكن نموذج التنقيب (NeuralNetworks_OutState) من التنبؤ بشكل صحيح بحالة شفاء 6121 مريض وبحالة وفاة 131 مريض ولكنه تنبأ بشكل خاطئ بحالة شفاء 117 مريض وبحالة وفاة 221 مريض. ونستنتج أن عدد الحالات المتنبأ بها بشكل صحيح في هذا النموذج هو (6121+131 = 6252) أي بنسبة 94.87% من حالات مجموعة الاختبار.

ج- استطاع نموذج التنقيب (LogisticRegresion_OutState) التنبؤ بشكل صحيح بما نسبته 81.03% من حالات مجموعة الاختبار.

د- حقق نموذج التنقيب (NaïveBayes_OutState) نجاحاً في التنبؤ بنسبة 92.03% من حالات مجموعة الاختبار.

ه- قدم نموذج الشبكات العصبونية (NeuralNetworks_OutState) أعلى نسبة تنبؤ مقارنة بالنموذجين الآخرين، وهي 94.87%.



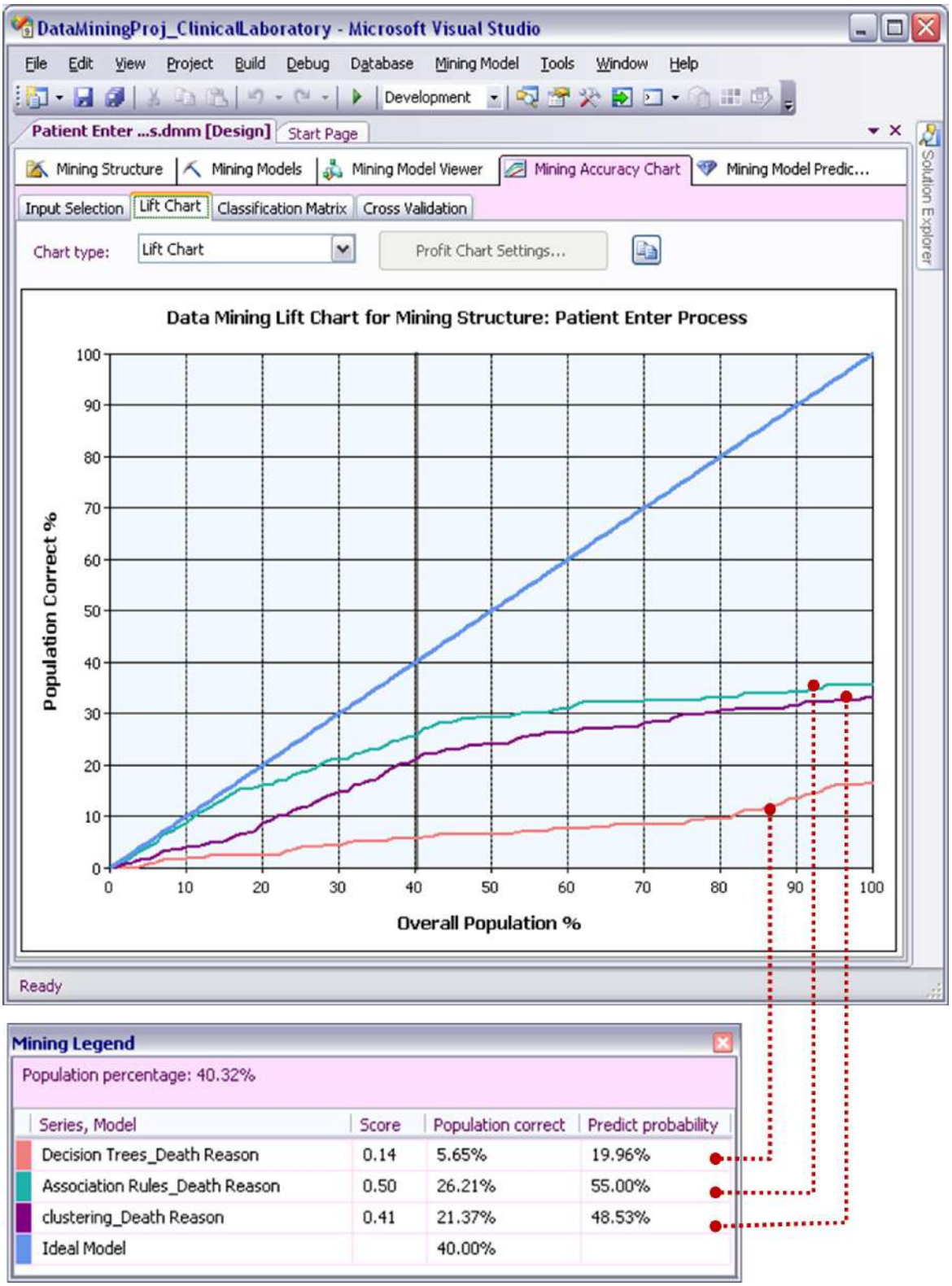
الشكل (5-5- 5): نافذة مصفوفة التصنيف (Classification Matrix)، وتُبين مقارنة بين أعداد القيم المتنبأ بها الصحيحة منها والخاطئة في مجموعة بيانات الاختبار لكل نموذج من نماذج التنقيب الخاصة بالتنبؤ بحالة الخروج من المشفى (الشفاء أو الوفاة)

إذاً ما هو النموذج الأفضل علماً أنه تبين في مخططي الارتقاء العام والمخصص أن أفضل نموذج للتنبؤ هو نموذج (NaïveBayes_OutState)، بينما أظهرت مصفوفة التصنيف أن نموذج الشبكات العصبونية (NeuralNetworks_OutState) قدم أعلى نسبة تنبؤات؟.

إن قرار الاختيار يجب أن يأخذ بعين الاعتبار معنى نتائج التنقيب ووجهة نظر الجهة المستفيدة من النماذج. وبناءً على ذلك وعلى اعتبار أن حالة الوفاة هي حالة حرجة فمن المفترض أن نحصل على نموذج تقل فيه أخطاء التنبؤ بشفاء الأشخاص المعرضين فعلاً للوفاة.. وبالتعمق أكثر في بيانات مصفوفة التصنيف يتبين أن نموذج الشبكات العصبونية قد تنبأ خطأً بشفاء 117 مريض (ولكنهم توفوا)، أي بنسبة 47.17% (117÷248%)، بينما تنبأ نموذج نايف بايز خطأً بشفاء 74 مريض أي بنسبة 29.83% (74÷248%).. وبذلك نجد أن نموذج التنقيب المخصص للتنبؤ بحالة خروج المريض من المشفى والمعتمد على خوارزمية نايف بايز (NaïveBayes_OutState) هو النموذج الذي تقل فيه الأخطاء الخطيرة، ويكون بالتالي هو النموذج الأفضل (الأقل مخاطرةً) نسبياً بين النماذج المدروسة للتنبؤ بحالة خروج المرضى من المشفى، إضافةً إلى أنه كان النموذج الأفضل في مخطط الارتقاء.

4•1•5•5 مخطط الارتقاء (Lift Chart) للنماذج المتعلقة بدراسة أسباب وفاة المرضى

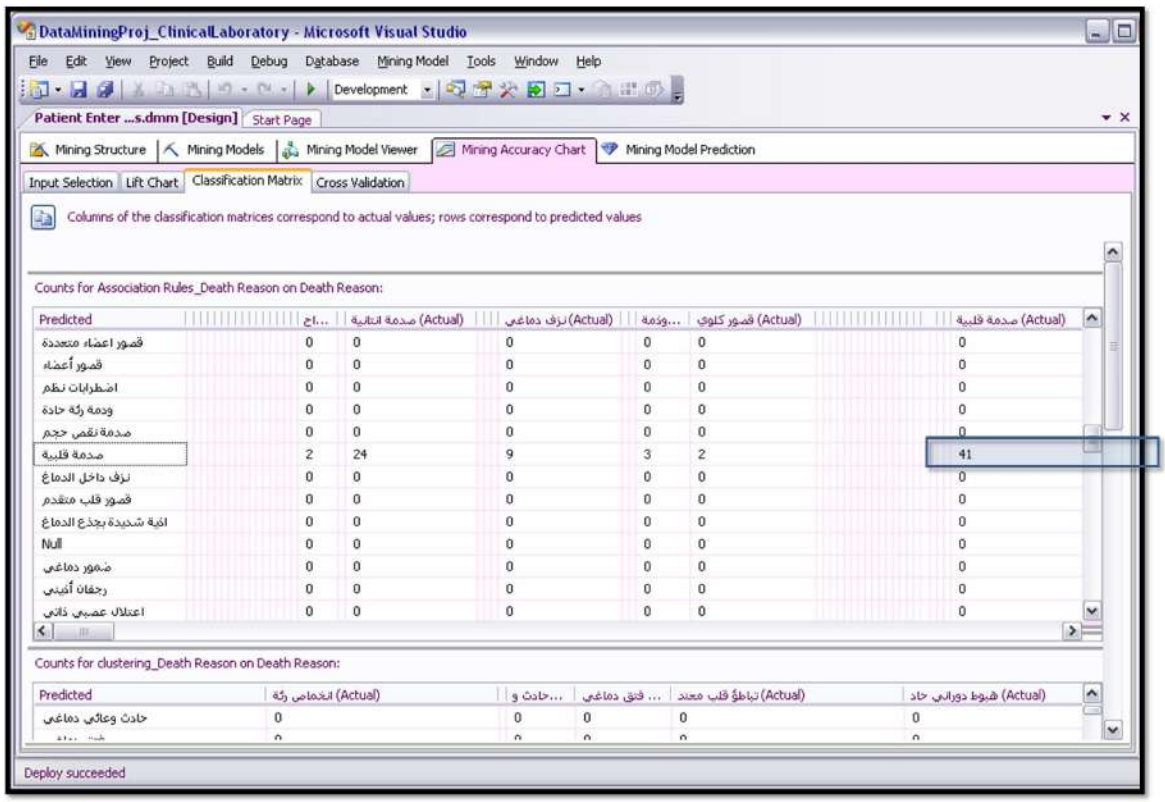
يحتاج استدعاء مخطط الارتقاء لتقييم النماذج المتعلقة بأسباب الوفاة (DeathReasons) إلى العودة إلى نافذة الإدخال الموجودة في الشكل (5-5-1) واختيار سبب الوفاة في حقل اسم عمود التنبؤ واختيار نماذج التنقيب المناسبة وهي (DecisionTrees_DeathReason)، (AssociationRules_DeathReason)، (Clustering_DeathReason). وقد أظهر مخطط الارتقاء الذي تم استدعاؤه أن هذه النماذج أقل قدرةً على التنبؤ من تلك النماذج الخاصة بالتنبؤ بحالة خروج المرضى من المشفى، وقد تبين باختصار أن نموذج التنقيب المستند إلى خوارزمية قواعد الاقتران (AssociationRules_DeathReason) هو النموذج الأقوى في توقع أسباب الوفاة لأنه حظي بأعلى علامة للدقة وهي 0.50، ولكنها علامة قليلة نسبياً حيث أنه إذا استُخدم 40.32% من البيانات المتاحة فإنه يمكن التنبؤ على الأكثر بسبب وفاة 26.21% من المرضى باستخدام هذا النموذج، انظر الشكل (5-4-6).



الشكل (5–5– 6): مخطط الارتقاء (Lift chart) لنماذج التنقيب الثلاثة الخاصة بالتنبؤ بأسباب الوفاة لدى مرضى المشفى

5•1•5•5 مصفوفة التصنيف لنماذج التنبؤ بأسباب الوفاة

تحتوي مصفوفة التصنيف في كل نموذج من نماذج التنبؤ بأسباب الوفاة على عددٍ كبيرٍ من الأسطر والأعمدة بحسب عدد أسباب الوفاة الموجودة والتي تزيد عن المائة، ولذلك لا يمكن مشاهدة كافة حقول المصفوفات الناتجة في صفحة واحدة.. وكمثال على نتائج هذه المصفوفات جرى اختيار الاطلاع على حقول مصفوفة نموذج قواعد الاقتران المتعلقة بسبب الوفاة بالصدمة القلبية وهو واحد من أسباب الوفاة الثلاثة الأولى في المشفى، وقد قمنا – لضرورة العرض– بتصغيرُ المسافات بين أعمدة مصفوفة التصنيف من أجل إظهار بعض الأعمدة التي تحتوي قيم تنبؤية صحيحة أو خاطئة تتعلق بالصدمة القلبية، لاحظ الشكل (5–5–7). ويتضح في هذا الشكل أن النموذج قد تنبأ بشكل صحيح بجميع وفيات الصدمة القلبية وعددهم 41 في مجموعة الاختبار، ولكنه تنبأ خطأً بوفاة ما يزيد عن 116 مريض بالصدمة القلبية ولكنهم توفوا بأسباب أخرى، وهذا ما يعكس ضعف القدرة على التنبؤ في هذا النموذج والذي أثبته مخطط الارتقاء السابق.



الشكل (5–5– 7): نافذة مصفوفة التصنيف لنماذج التنقيب الخاصة بالتنبؤ بأسباب الوفاة انطلاقاً من البيانات الشخصية والمرضية لمرضى المشفى، ويظهر في الشكل جزء من مصفوفة التصنيف لنموذج قواعد الاقتران

2•5•5 تقييم قدرات نماذج مشروع التنقيب في نتائج التحاليل الطبية المخبرية

توجهت نماذج التنقيب الخمسة الأولى[3] في هذا المشروع نحو دراسة العلاقات البينية بين نتائج التحاليل الطبية، وقد استُخدمت بعض التحاليل الطبية كمتغيرات تنبؤية في هذه النماذج، وذلك بهدف التعرف على مدى تأثر نتائج هذه التحاليل بنتائج التحاليل الأخرى. وينتج عن هذا المشروع مخططات ومصفوفات تقييم بقدر عدد التحاليل الطبية المتنبأ بها، وتختلف قدرات نماذج التنقيب باختلاف التحليل الطبي المتنبأ به. ويتطلب استخدام أحد النماذج للتنبؤ بنتائج أحد التحاليل الطبية الموجودة أن يتم تقييم النماذج بالنسبة إلى هذا التحليل تحديداً، حيث من الممكن أن يكون أحد النماذج هو الأفضل في التنبؤ بنتائج تحليل ما، ويكون نموذج آخر هو الأفضل في التنبؤ بنتائج تحليل آخر، وقد تكون جميع النماذج ضعيفة الأداء في التنبؤ بقيم بعض التحاليل الطبية.

1•2•5•5 اختيار مفردات التقييم في نافذة المخطط البياني لدقة التنقيب

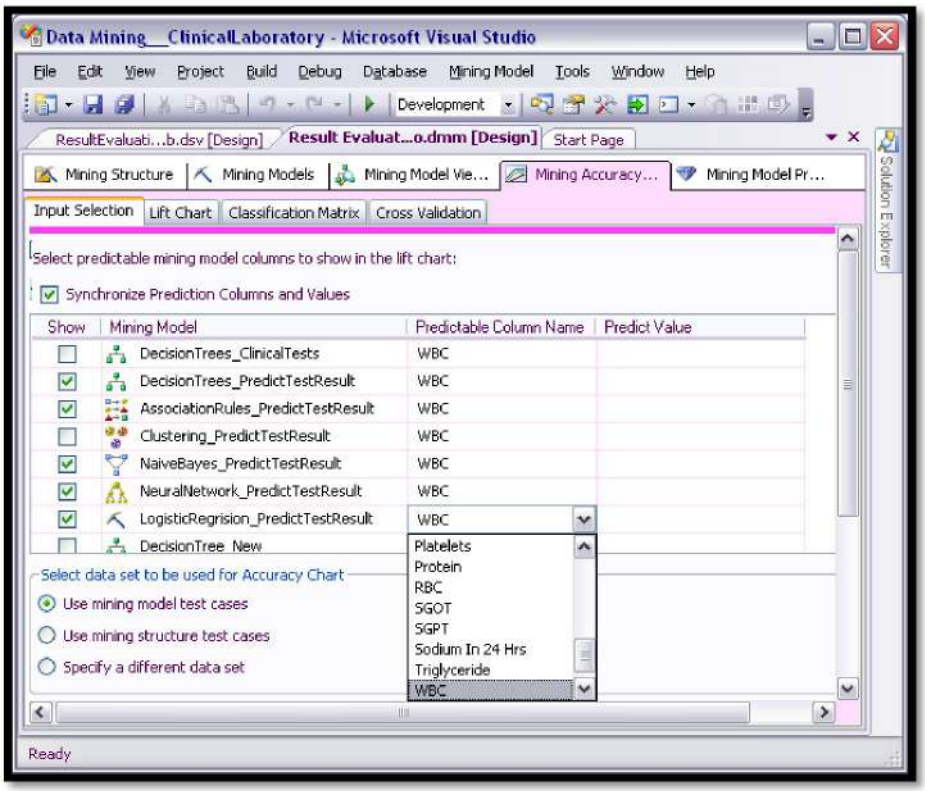
جرى استعراض مخططات التقييم بالنسبة للعديد من التحاليل الطبية، ولكن يضيق المجال هنا لتقييم جميع هذه التحاليل الطبية، ولذلك تم اختيار تحليل تعداد كريات الدم البيضاء لتقييم النماذج التي تتنبأ به خصوصاً أنه استحوذ على قسط كبير من الاهتمام في مرحلة التنقيب في البيانات، ونبين فيما يلي عناصر التقييم التي تم اختيارها في نافذة الإدخال:

أ- اختيار تحليل تعداد كريات الدم البيضاء (WBC) في حقول العمود المتنبأ به.

ب- انتقاء نماذج التنقيب الموجودة في هذا المشروع، انظر الشكل (5-5-8).

ج- استخدام بيانات مجموعة اختبار نماذج التنقيب للتحقق من أداء نماذج التنقيب المستندة إلى بيانات التجريب.

[3] لن يتم التطرق إلى تقييم النموذجين الأخيرين في هذا المشروع، لأنهما أولاً بنيا بشكل منفصل عن النماذج الأخرى، وثانياً لأن الهدف منهما هو دمج خاصيات المرضى مع نتائج التحاليل الطبية والحصول على نماذج تنقيب ديناميكية باستخدام برنامج اكسل.



الشكل (5-5- 8): نافذة انتقاء عناصر تقييم نماذج التنقيب الخاصة بمشروع التنقيب في نتائج التحاليل الطبية، وقد تم انتقاء تقييم قدرات النماذج في التنبؤ بنتائج تحليل تعداد كريات الدم البيضاء

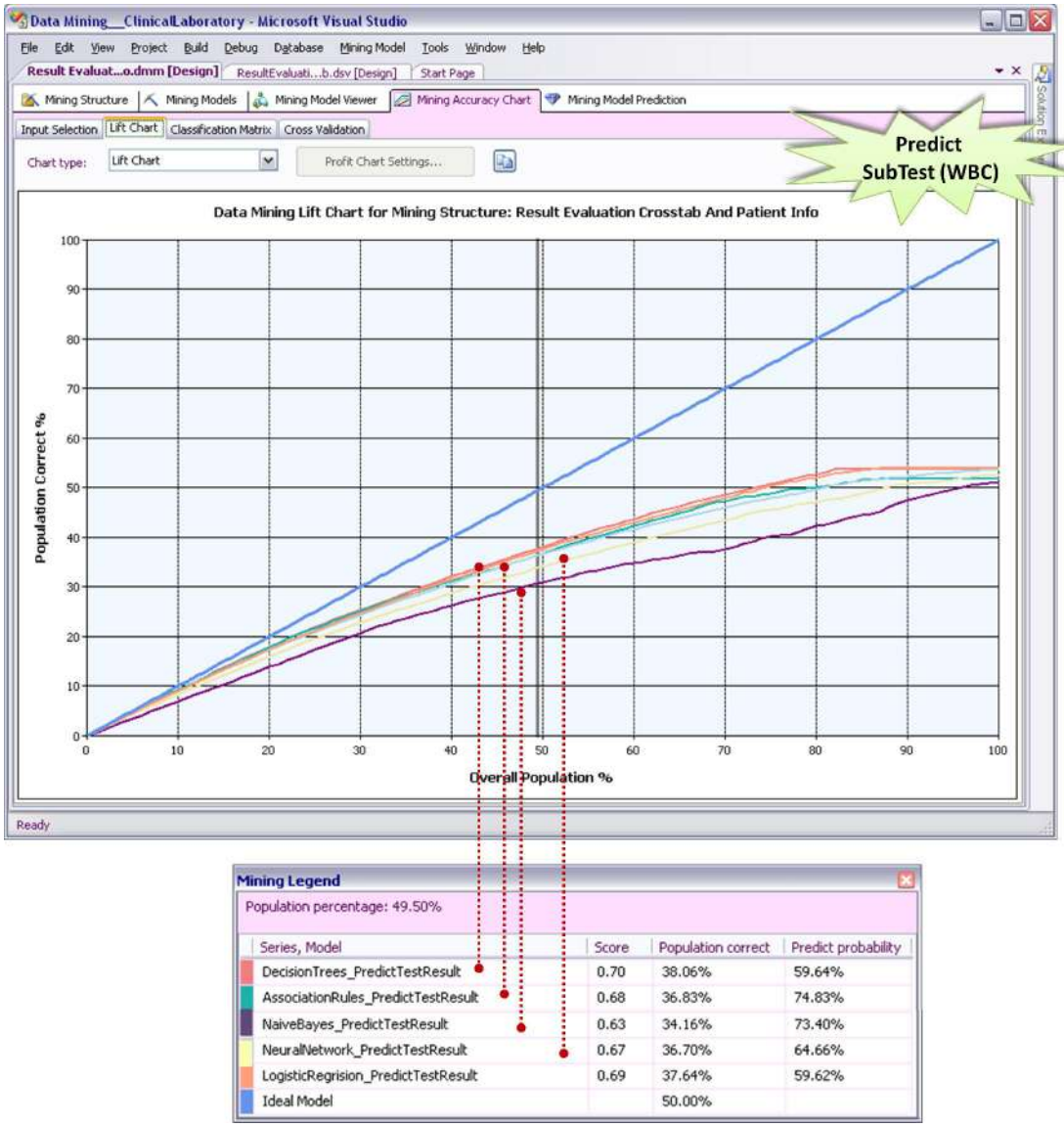
5•5•2•2 مخطط الارتقاء (Lift Chart) لنماذج توقع نتيجة تحليل تعداد كريات الدم البيضاء

5•5•2•2•1 مخطط الارتقاء (Lift Chart) العام

أظهر مخطط الارتقاء العام الذي يُقيِّم أداء نماذج التنقيب التي تتنبأ بنتائج التحليل الطبي الفرعي لتعداد كريات الدم البيضاء النتائج التالية:

أ- إن أداء جميع النماذج متقارب، ولكن الأفضل بينها هو النموذج المعتمد على خوارزمية أشجار القرار (DecisionTrees_PredictTestResult)، حيث أنه حقق أعلى علامة تقييم وهي (0.70)، يليه نموذج الانحدار المنطقي (LogisticRegrision_PredictTestResult) الذي نال علامة (0.69)، انظر الشكل (5-5-9).

ب- يستطيع نموذج أشجار القرار أن يتنبأ بـــ 38.06% من قيم نتائج التحليل الطبي (WBC) باستخدام 49% من البيانات، وهي نسبة جيدة مقارنة بالنسبة المثالية التي يحققها النموذج المثالي وهي 50%.



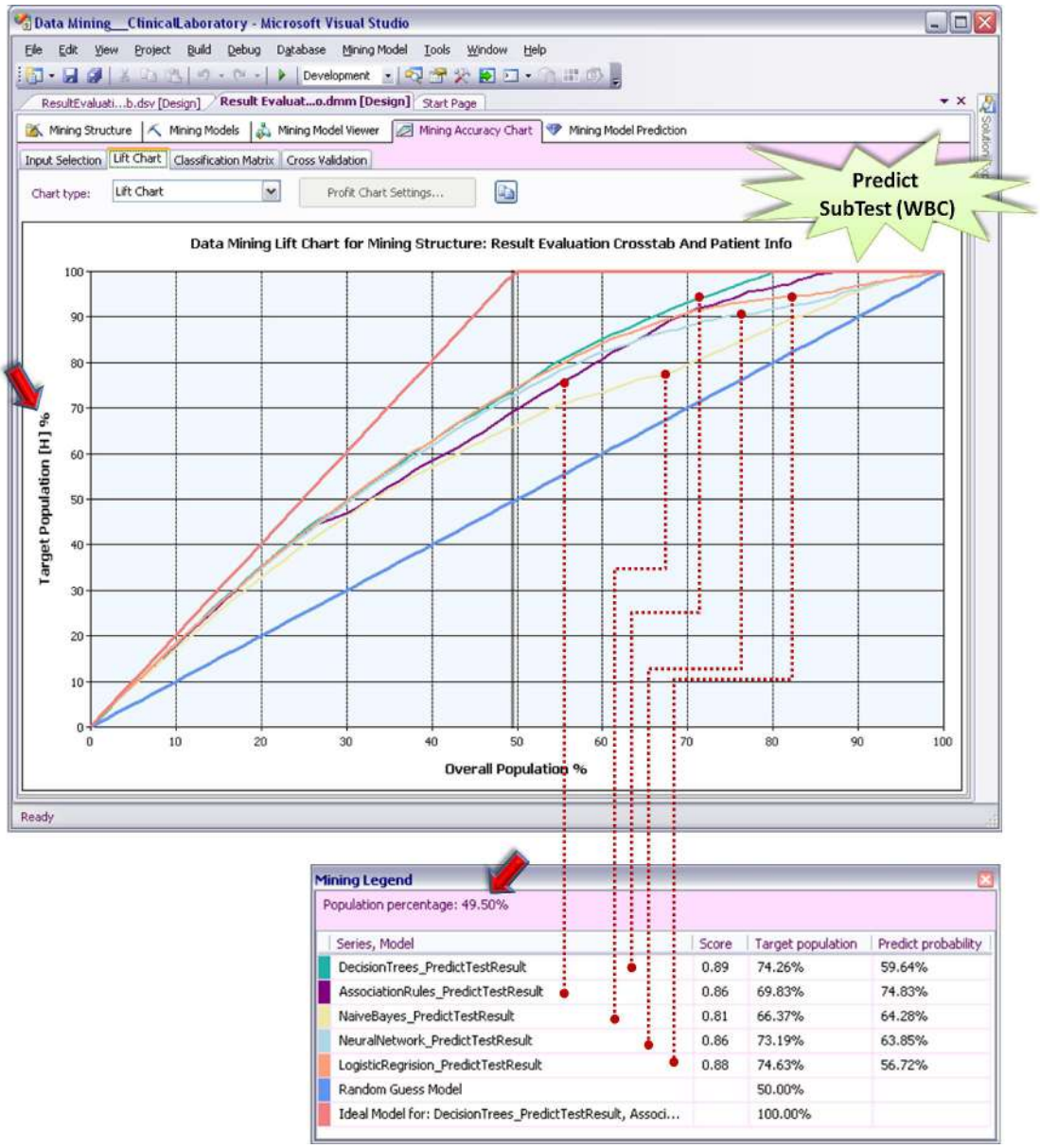
الشكل (5-5- 9): مخطط الارتقاء العام لتقييم قدرة نماذج التنقيب على التنبؤ بنتائج التحليل الطبي (WBC)

2•2•2•5•5 مخطط الارتقاء المخصص لتقييم قدرة النماذج على التنبؤ بالقيمة المرتفعة لتحليل تعداد كريات الدم البيضاء

أدى تخصيص مخطط الارتقاء بتقييم قدرة النماذج على التنبؤ بالقيم المرتفعة لتحليل (WBC) إلى الحصول على علامات تقييم أعلى من العلامات في مخطط الارتقاء العام السابق، ونقرأ في المخطط المخصص الناتج ما يلي:

أ- أحرز نموذج أشجار القرار أعلى علامة أداء في التنبؤ وهي (0.89)، وتلاه مباشرةً نموذج الانحدار المنطقي الذي نال علامة (0.88).

ب- تستطيع نماذج التنقيب أن تتنبأ كحد أقصى بما لا يزيد عن ثلاثة أرباع حالات ارتفاع تعداد كريات الدم البيضاء. حيث حقق نموذج الانحدار المنطقي أعلى نسبة تنبؤ صحيح وهي (74.63%)، يليه مباشرةً نموذج أشجار القرار الذي قُدِّرت تنبؤاته الصحيحة بنسبة (74.26%).

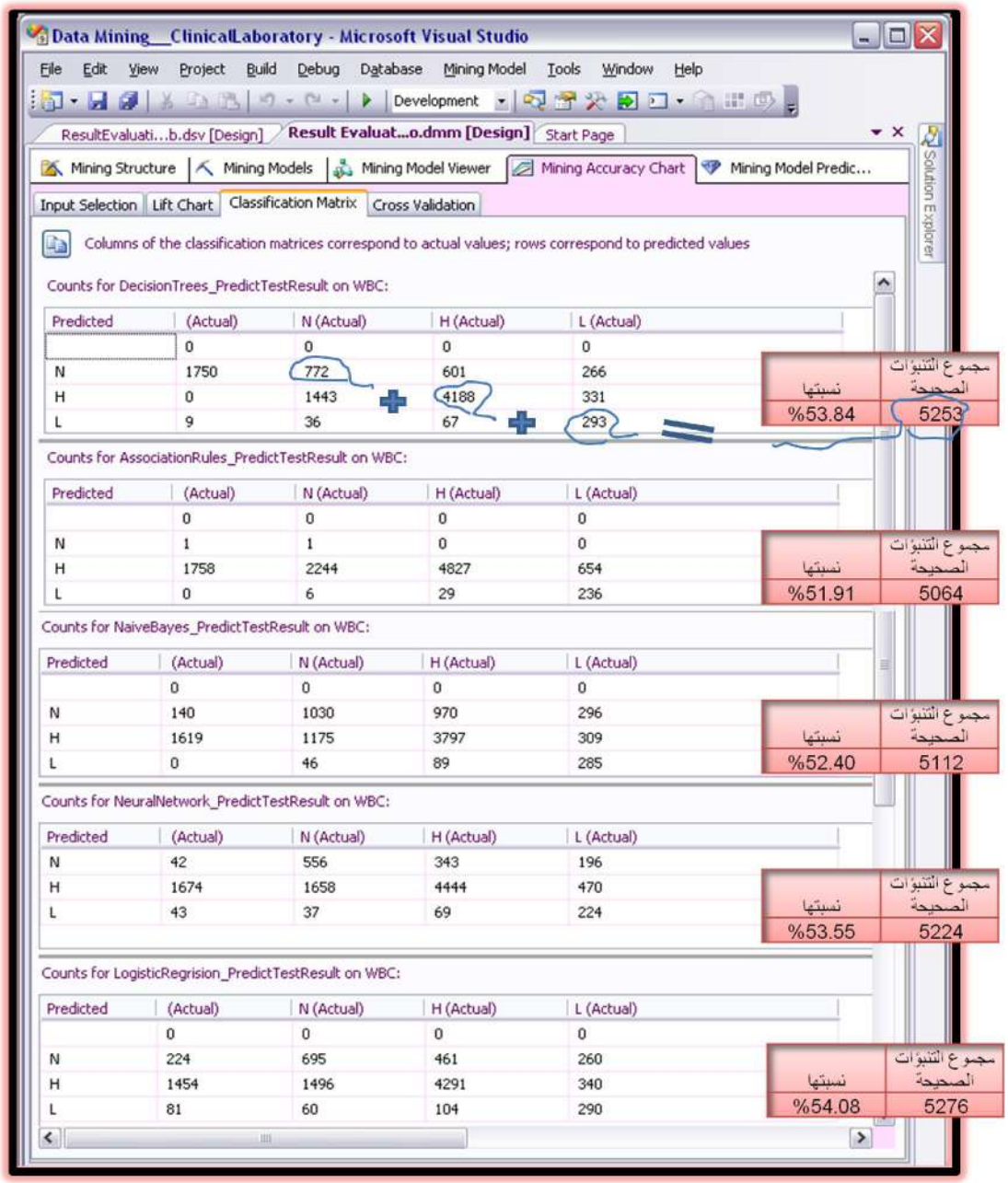


الشكل (5-5- 10): مخطط الارتقاء المخصص لتقييم قدرة النماذج على التنبؤ بالقيمة المرتفعة لتحليل تعداد الكريات البيضاء

3•2•5•5 مصفوفة التصنيف (Classification Matrix)

تتضمن مصفوفة التصنيف نتائج التنبؤات التي أعطتها كل من نماذج التنقيب في بيانات مجموعة الاختبار مقارنةً مع القيم الحقيقية، وأظهرت مصفوفة التصنيف أن جميع نماذج التنقيب الخاصة بالتنبؤ بنتائج التحليل الطبي (WBC) قادرة على التنبؤ الصحيح بما يقارب

نصف الحالات المدروسة، أي أنها ذات قدرات تنبؤية متوسطة، وأن أعلى نسبة للتنبؤات الصحيحة في هذه النماذج تقدر بـــ 54.08% في نموذج الانحدار المنطقي (LogisticRegrision_PredictTestResult)، أما نموذج أشجار القرار الذي حاز على أعلى علامة تقييم فقُدِّرت نسبة تنبؤاته الصحيحة بـــ 53.84%، ولكن تبقى النسبتين متقاربتين كثيراً، انظر الشكل (5-5-11).



الشكل (5-5- 11): نافذة مصفوفة التصنيف الخاصة بعرض نتائج التنبؤات التي قامت بها كل من نماذج التنقيب الخاصة بالتنبؤ بنتائج التحليل الطبي لتعداد كريات الدم البيضاء (WBC) استناداً إلى نتائج التحاليل الطبية الأخرى المقترنة معه. وقد أضيفت إليها مستطيلات تتضمن مجموع التنبؤات الصحيحة ونسبتها إلى مجموع البيانات المتنبأ بها في كل نموذج (أي نسبة إلى المجموع 9756)

وبمقارنة نتائج التقييم في هذا المشروع نجد أن هناك اختلافات بسيطة بين قدرات كل من نموذجي التنقيب أشجار القرار (DecisionTrees_PredictTestResult) والانحدار المنطقي (LogisticRegrision_PredictTestResult)، وأن أي منهما يصلح للاستخدام في التنبؤ بنتائج تحليل تعداد الكريات البيضاء (WBC)، ولم يختلف القرار في حالة التنبؤ بالقيم المرتفعة لهذا التحليل فكلا النموذجين يمتلك ذات القدرة تقريباً.

إذاً اختصت المرحلة الحالية من مراحل عملية اكتشاف المعرفة بتقييم نماذج التنقيب بطريقة حسابية وبغض النظر عن مضمون البيانات وعن معنى المعرفة الناتجة. ولكن من الناحية التطبيقية يُفضل أن يُبنى قرار اختيار نموذج التنقيب الأنسب بالترافق مع فهم مجال التطبيق ومع تفسير المعرفة المكتشفة، ومع اختبار قدرة النماذج المكتشفة على إثبات المعارف والحقائق الحالية قبل البحث عن معارف جديدة. وقد تمت مراعاة هذه الجوانب في أثناء استعراض نماذج التنقيب في المبحث الرابع من الفصل الخامس، وكان من ضمن النتائج الملفتة للانتباه أن نموذجي التنقيب (LogisticRegrision_PredictTestResult) و(NeuralNetwork_PredictTestResult) كانا يتجاهلان تأثير القيم المفقودة في المدخلات (Inputs) على قيم التحاليل الطبية المتنبأ بها، وهذا ما يجعل التأثير الحقيقي لنتائج بعض التحاليل الطبية يظهر في كل من هذين النموذجين على خلاف غيرهم من النماذج.
وبالتقاطع بين معطيات التقييم المختلفة يكون النموذج المستند إلى الانحدار المنطقي (LogisticRegrision_PredictTestResult) هو النموذج الذي تم ترجيحه من قبل أكثرية المصادر. وإلى جانب ذلك فقد يلعب الهدف من عملية التنقيب دوراً أساسياً في اختيار نموذج التنقيب المناسب، وربما يكون من الأفضل اعتماد أكثر من نموذج بحسب احتياجات الجهة المستفيدة من المعلومات، حيث تُحقق بعض النماذج فائدة تصنيف البيانات ضمن مجموعات مثل النماذج المستندة إلى التجزئة العنقودية، بينما تُقدم نماذج أخرى قواعد هامة لتتابع وترافق قيم المتغيرات مثل النماذج المستندة إلى قواعد الاقتران.

الفصل الخامس

المبحث السادس

مرحلة استخدام المعرفة المكتشفة من بيانات مخبر التحاليل الطبية

مقدمة

تتواصل هذه المرحلة مع مرحلة التنقيب في البيانات، وقد كان من الصعب فصل بعض المعارف الأولية المكتشفة من عمليات التنقيب عن النماذج التي أنتجتها، ولذلك عُرضت تلك المعارف مباشرةً ضمن مبحث مرحلة التنقيب في البيانات. أما هذه المرحلة فتتضمن تقريراً سريعاً لأبرز ما تم التوصل إليه من نتائج في مرحلة التنقيب، وتتضمن عدداً من الأمثلة عن الاستخدامات الممكنة للمعرفة المكتشفة من قواعد بيانات التحاليل الطبية المدروسة.

ويعتبر برنامج اكسل من أهم البرامج الحاسوبية التي تُستخدم في مرحلة استخدام المعرفة المكتشفة حيث أنه يتلقى المعارف المكتشفة ويحولها إلى أشكال عرض مختلفة مفيدة ومفهومة للمستخدمين، كالتنبؤ بالدخل الشهري للمنشأة الطبية، وتتبع تتالي التعرض للأمراض المختلفة، وتصميم أداتين إحداهما الكترونية والأخرى ورقية للتنبؤ بالأشخاص المعرَّضين للوفاة، ودراسة المتغيرات الطبية المختلفة المرافقة لتغير نتائج أحد التحاليل الطبية، والتعرف على الأسباب الكامنة وراء تحويل المرضى إلى المشافي الأخرى.

1•6•5 تقرير نتائج عملية التنقيب

أنتج كلٌّ من نماذج التنقيب في مرحلة التنقيب في البيانات مجموعة من المعارف، وكان لكل نموذج منهم طريقته المختلفة في استنباط المعارف من قواعد البيانات. وبالرغم من أن أداء بعض النماذج كان أفضل من غيره، فإن الأداء كان متقارباً كثيراً، ويمكن بالتالي الاعتماد على مخرجات أكثر من نموذج، لأن كلاً منها يتعرَّف أو يكتشف حقائق مختلفة عن غيره. وسنقدم فيما يلي تقريراً عن أبرز النتائج التي قدَّمها كل نموذج، وذلك في الجدول التالي.

الجدول (5-6- 1): أبرز النتائج التي قدمتها نماذج التنقيب في مشاريع التنقيب التي بنيت في مرحلة التنقيب في البيانات

المشروع	مجموعة النماذج	اسم النموذج (علامة التقييم في مخطط الارتقاء)	أهم النتائج
مشروع التنقيب في البيانات الشخصية والمرضية لمرضى المشفى	نماذج التنقيب التي تهدف إلى اكتشاف أسباب الوفاة الأكثر تأثراً بالخاصيات المختلفة للمرضى	نموذج (DecisionTrees_DeathReason) (Score=0.14)	• لا يمكن الحصول إلا على عقدة وحيدة في الشجرة تتضمن احتمال كل سبب من أسباب الوفاة.
		نموذج (AssociationRules_DeathReason) (Score=0.50)	• الأسباب الرئيسة الأولى في الوفاة هي القصور التنفسي الحاد والصدمة الإنتانية والصدمة القلبية. • مشكلة داء الأغشية الهلامية ذات التصنيف الدولي الأساسي (P22.0) والتصنيف الدولي للوفاة (029-4) والتي تحدث للمرضى في شعبة الحواضن والقادمين من المنطقة الجنوبية تؤدي إلى الوفاة بالقصور التنفسي الحاد بالتأكيد (الاحتمال=1)، وتقدر أهمية هذه القواعد بـ (0.878). • إذا كان سبب القبول في المشفى هو الصدمة الإنتانية، وكان جنس المريض ذكر وتوفي هذا المريض فإن سبب الوفاة هو الصدمة الإنتانية باحتمال قدره 1، وتقدر أهمية هذه القاعدة بـ 0.781. • عندما يكون جنس المريض أنثى، ويزيد عمرها عن السبعين عاماً وقُبلت في المشفى في شعبة الداخلية الصدرية فإن احتمال أن يكون سبب وفاتها هو القصور التنفسي الحاد يتراوح بين (0.941-0.944). وتتراوح أهمية هذه النتيجة بين (0.925- 0.931). • سبب الوفاة بالصدمة الإنتانية يتأثر بسبب القبول بالحروق وعند المرضى القادمين من مدينة دير الزور.
		نموذج (Clustering_DeathReason) (Score=0.41)	• تم توزيع البيانات بحسب خاصياتها في عشرة عقد، حيث تتشابه الخواص بين بيانات كل عقدة، فمثلاً تتصف البيانات في العقدة الأولى بـأن 53% من أسباب الوفاة فيها تعود إلى الصدمة القلبية، ويشكل الذكور 70% من المرضى فيها، وأن 96% منهم تزيد أعمارهم عن 46 سنة، وأن (95%) منهم أتوا بحالة إسعافية إلى المشفى من مناطق قريبة كدمشق (53%) وريف دمشق (35%).
	نماذج التنقيب الخاصة بالتنبؤ بحالة الخروج من المشفى (الشفاء أو الوفاة)	نموذج (NaiveBayes_OutStae) (Score=0.98)	• تعتمد حالة خروج المرضى من المشفى على أربعة خاصيات فقط من خاصيات المدخلات وهي عمر المريض وطريقة قبوله في المشفى والشعبة التي قُبل فيها وشعبة التخرج. • 80% من الذين توفوا كانوا قد أتوا بحالة إسعاف، وأن 29.5% من الذين توفوا تقل أعمارهم عن 13 سنة وأن 29.5% منهم تزيد أعمارهم عن 65%، و23.7% منهم تتراوح أعمارهم بين [29-46]، وأن 25% من الذين توفوا كانوا في شعبة العناية الإسعافية.
		نموذج (NeuralNetworks_OutState) (Score=0.98)	• إن سبب المرارة المحصاة يأخذ علامة 5.09 في التأثير على خروج المريض بحالة تحسن، أما سبب احتشاء العضلة القلبية فيأخذ علامة 5.33 في التأثير على خروج المريض بحالة وفاة. ونلاحظ أن 97.81% من المرضى الذين عانوا من المرارة المحصاة قد خرجوا بحالة تحسن، وأن 95.06% من المرضى الذين تعرضوا لاحتشاء العضلة القلبية قد خرجوا بحالة تحسن وليس وفاة.
		نموذج (LogisticRegresion_OutState) (Score=0.88)	• أعطى نموذج الانحدار المنطقي علامة (3.7) لتأثير سبب احتشاء العضلة القلبية في توقع شفاء المريض الذي يزيد عمره عن 65 سنة، وليس توقع الوفاة.

مشروع التنقيب في بيانات نتائج التحاليل الطبية المخبرية	نماذج التنبؤ بنتائج التحاليل الطبية اعتمادا على نتائج التحاليل الطبية الأخرى	**نموذج أشجار القرار (DecisionTrees_PredictTestResult) ((WBC) Score=0.70)**	• إذا كانت نسبة العدلات (Neutro) مرتفعة ونسبة اللمفاويات (Lympho) منخفضة وعدد الصفائح (Platelets) ليس منخفضاً ونسبة الحمضات (Eosino) منخفضة فإن احتمال أن يكون تعداد كريات الدم البيضاء مرتفعاً هو (90.69%). • إذا كانت نسبة العدلات (Neutro) طبيعية وكان عدد الصفائح (Platelets) منخفضاً وكان الخضاب (HB) منخفضاً فإن احتمال أن يكون تعداد كريات الدم البيضاء مرتفعاً هو 30.56%، واحتمال أن يكون منخفضاً هو 34.33%. • إذا انخفضت نسبة العدلات (Neutro) وانخفضت نسبة اللمفاويات (Lympho) فإن 96.31% من الحالات سينخفض فيها تعداد كريات الدم البيضاء. • عندما يكون الخضاب منخفضاً (HB) والهيماتوكريت منخفضاً (Hct) وعدد الصفائح طبيعياً (Platelets) وعدد وحيدات النوى ليس منخفضاً (Mono≠'L') وعدد اللمفاويات ليس مرتفعاً (Lemfo≠'H') وتحليل البولة النتروجينية مرتفعاً (Bun) فإن احتمال أن يكون عدد كريات الدم الحمراء منخفضاً هو 83.6%، واحتمال أن يكون طبيعياً هو 14.8%، واحتمال أن يكون مرتفعاً 0.46%.
		نموذج قواعد الاقتران (AssociationRules_PredictTestResult) ((WBC) Score=0.68)	• استحوذت قواعد الاقتران التي تحمل فيها التحاليل الطبية نتائج طبيعية على أقوى الاحتمالات وأكبر الأهميات. • إذا انخفضت نتيجة كل من التحليلين الطبيين (Creatinine=L) و(Protein=L) فإنه باحتمال (0.848) وبأهمية (0.827) تكون نتيجة تحليل (Albumin) منخفضة.
		نموذج نايف بايز (NaiveBayes_PredictTestResult) ((WBC) Score=0.63)	• إن ارتفاع نتيجة تحليل تعداد كريات الدم البيضاء (WBC=H) يترافق معه ارتفاع نتيجة تحليل العدلات (Neutro) باحتمال قدره 81.56%، وانخفاض بنتيجة تحليل الخضاب (HB=L) باحتمال قدره 67.247%، وارتفاع في نتيجة تحليل السكر الصيامي (Glucose Fasting=H) باحتمال قدره 39.17%. • إذا كان السكر الصيامي مرتفع عند مرضى المشفى فإن تعداد كريات الدم البيضاء يكون مرتفع باحتمال قدره 59.79%.
		نموذج خوارزمية الشبكات العصبونية (NeuralNetwork_PredictTestResult) ((WBC) Score=0.67)	• إن ارتفاع نتيجة التحليل الطبي لحمض البول (Uric Acid) يُرجح انخفاض تعداد كريات الدم البيضاء –وليس ارتفاعها- بقوة ترجيح 85.45. • إن انخفاض نتيجة التحليل الطبي للكولسترول من نوع (LDL) يُرجح ارتفاع تعداد كريات الدم البيضاء -وليس انخفاضها- بقوة ترجيح 60.48. • إن ارتفاع نتيجة تحليل السكر الصيامي يُرجح أن يكون تعداد الكريات البيضاء مرتفعاً وليس منخفضاً بعلامة ترجيحية قدرها 7.5. • إن ارتفاع نتيجة التحليل الطبي (BH) – وهو أحد تحاليل الحمى التيفية- يُرجح أن يكون تعداد الكريات البيضاء مرتفعاً وليس منخفضاً بعلامة ترجيحية قدرها 39.21.
		نموذج الانحدار المنطقي (LogisticRegrision_PredictTestResult) ((WBC) Score=0.69)	• إن انخفاض نتيجة تحليل الكولسترول من نوع (LDL) يُرجح ارتفاع تحليل (WBC) بعلامة ترجيحية (60.48). • إن ارتفاع نتيجة تحليل (BH) يُرجِّح انخفاض نتيجة تحليل تعداد الكريات البيضاء بعلامة ترجيحية (65.8).

	نماذج التعرف على فئات البيانات الشخصية والمرضية ونتائج التحاليل الطبية	**نموذج التجزئة العنقودية (Clustering Model) باستخدام أداة اكتشاف الفئات في برنامج اكسل**	تمت تجزئة البيانات إلى 10 فئات، حيث تتشارك عناصر كل فئة بقيم الخاصيات المتشابهة وهي: 1- فئة أمراض وتحاليل الأطفال والحواضن دون عمر 15 سنة. 2- فئة أمراض الاحتشاءات القلبية والدماغية للأعمار التي تزيد عن 51 سنة. 3- فئة الجراحات للأعمار فوق 32 سنة. 4- فئة أمراض الكلية وأمراض القلب في الأعمار التي تزيد عن 51 سنة. 5- فئة أمراض الدم للأعمار بين 15-32 سنة. 6- فئة الأمراض العصبية. 7- فئة أمراض الدم للأعمار بين 32-51 سنة. 8- فئة القيم المفقودة في التحاليل عند أمراض الأطفال دون عمر 15 سنة. 9- فئة الجراحات والأمراض الهضمية. 10- فئة القيم المفقودة في التحاليل الطبية.
		نموذج تحليل عوامل التأثير الرئيسة (Analyze Key Influencers) في اكسل	• يتأثر ورم الدماغ بالمنطقة الجغرافية الشمالية الشرقية بأهمية نسبية 40%. • إن ارتفاع نتيجة تحليل (b1) ترافق أو ساعد في تشخيص مرض التهاب الكبد بنسبة 100%، وترافق ارتفاع نتيجة تحليل (SGPT) مع التهاب الكبد بنسبة73% من الحالات. كما أن الفئة العمرية الأكثر تعرضاً لتشمع الكبد هي بين 51-66 سنة. • يتأثر حدوث النزف الدماغي بالعمر الذي يزيد عن 66 سنة بنسبة 100%. • إن ارتفاع نتيجة تحليل السكر في البول (SUGAR) كان مؤثراً بنسبة 96%، في تشخيص الحماض السكري الخلوني، في حين كانت نسبة تأثير العمر بين 15-32 سنة على هذا المرض 88%، وكانت نسبة تأثير أن يكون المريض أنثى 52%. • تعطى للعمر بين 15-32 أهمية نسبية 92% في حدوث التهاب الدماغ. • يعتبر العمر بين 15-32 سنة على أنه العمر الأكثر ترافقاً مع التهاب الكبد، بينما يعتبر العمر بين 51-66 سنة أكثر ترافقاً مع تشمع الكبد. • تُظهر نتائج تحاليل (Platelets, WBC, RBC,K) انخفاضاً عند المصابين بتشمع الكبد ولا يظهر لها نتائج واضحة عند المصابين بالتهاب الكبد.

يتبين من الجدول السابق أنه قد تختلف أو تتعاكس النتائج أحياناً بين نماذج التنقيب المختلفة بسبب اختلاف الأساليب الحسابية المتبعة في كل منهم، وهذا ما تمت الإشارة إليه في مرحلة التنقيب في البيانات. ويَظهر في الجدول أيضاً عدد من النتائج البديهية بالنسبة للمجال المدروس، كما يوجد بعض النتائج الجديدة، وعندما تسترعي بعض النتائج اهتمام المختصين بالمجال المدروس فإنها لابد وأن تحرضهم على التعمق في دراسة بعض الحالات وبالتالي إجراء أبحاث طبية موجهة نحو هدف محدد وقد تكون بإتباع أساليب البحث الطبية التقليدية.

2•6•5 الاستعلام المباشر من نماذج التنقيب وتصدير النتائج

1•2•6•5 بناء الاستعلام من نماذج التنقيب

تتيح أداة تنبؤ نموذج التنقيب (Data Mining Prediction) في برنامج (Visual Studio) إمكانية تطبيق نماذج التنقيب المكتشفة على أية سجلات للمرضى تحتوي على بيانات مشابهة للبيانات المستخدمة في بناء نماذج التنقيب، ويمكن بالتالي الاستعلام بالتفصيل عن النتائج الحقيقية والمقدرة لكل مريض.

لنفترض أن الجهة المهتمة بالأمور الطبية أرادت الحصول على جدول بأسماء المرضى المتوفين وبعض خاصياتهم والتنبؤات بحالة خروجهم من المشفى التي قدمها نموذج التنقيب الأجود. ولتحقيق ذلك يتم إتباع الخطوات التالية المتمثلة في الشكل (5-6-1):

أ- استدعاء نافذة أداة تنبؤ النموذج (Data Mining Prediction).

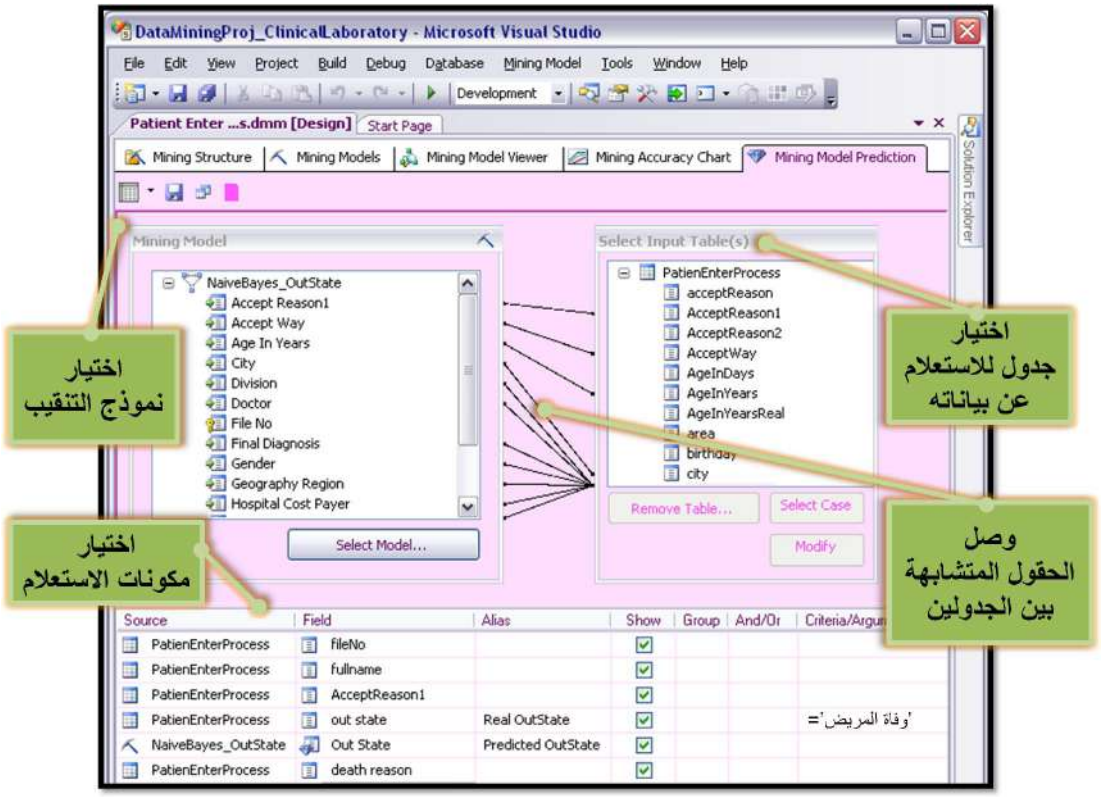
ب- اختيار نموذج التنقيب (NaïveBayes_OutState) لأنه حقق أفضل النتائج بالنسبة للقدرة على التنبؤ بحالة خروج المرضى من المشفى باستخدام البيانات الشخصية والمرضية للمرضى كما سبق وورد في المبحث السابق.

ج- اختيار جدول بيانات المرضى الشخصية والمرضية (PatientEnterProcess).

د- وصل كل من الحقول المتشابهة بين جدول المرضى وجدول نموذج التنقيب.

ه- تحديد مكونات الاستعلام من حيث الحقول ومن حيث الشروط، وقد تم اشتراط انتقاء سجلات المرضى المتوفيين.

ويستفاد من هذه الأداة في التطبيق المستقبلي على أي جدول حديث التاريخ لبيانات المرضى أو أي جدول للمرضى من مشفى آخر، شريطة أن يحتوي الجدول الجديد على نفس الخاصيات التي استُخدمت في عمليات التنقيب، وأن تكون البيانات الجديدة نظيفة، ويمكن بالتالي إجراء الاستعلام التفصيلي على التوقعات التي يقدمها نموذج التنقيب المختار لكل مريض من مرضى الجدول الجديد.

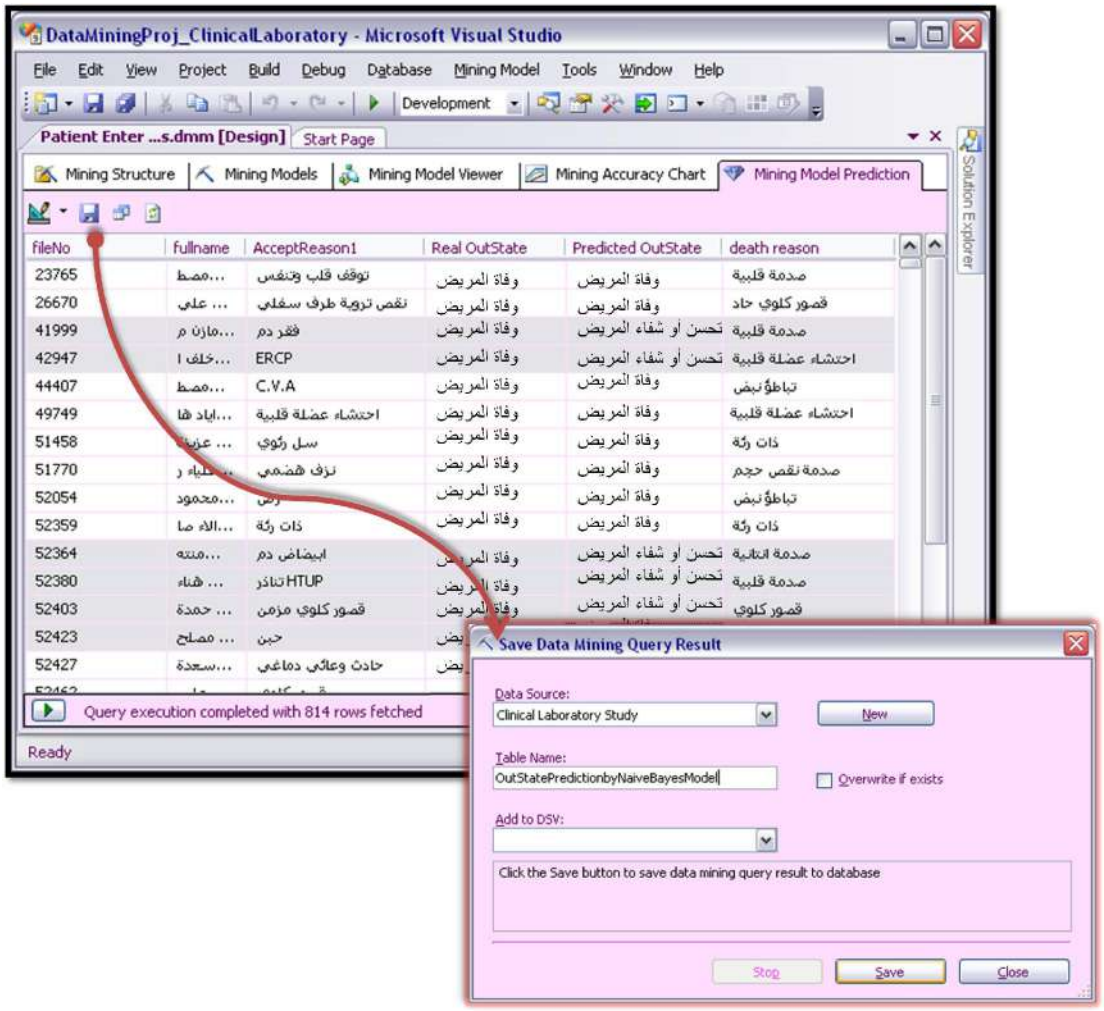


الشكل (5–6– 1): نافذة تصميم الاستعلامات عن تنبؤات نماذج التنقيب (Mining Model Prediction)، وتتضمن استدعاء جدول ما لتطبيق نتائج أحد نماذج التنقيب على بياناته، واختيار مكونات الاستعلام، وقد تم اختيار تطبيق الاستعلام على بيانات جدول بيانات المرضى الذي بُنيت عليه نماذج التنقيب في هذه الدراسة، وتحديداً المرضى الذين توفوا في المشفى

2•2•6•5 استعراض الاستعلام وتخزينه

تظهر نتائج الاستعلام عن توقعات نموذج التنقيب (NaiveBayes_OutState) التي طُلبت في الفقرة السابقة على شكل جدول كما في الشكل (5–6–2). ويُعرض في الجدول حالة خروج المرضى (Real OutStat) المستمدة من جدول المرضى والحالة المقابلة المتوقعة من خلال نموذج التنقيب. ويلاحظ وجود بعض الاختلافات أحياناً بين القيم الحقيقية والقيم المتنبأ بها، وهو أمر عادي حيث أن نسبة التنبؤات الخاطئة كانت في مجموعة الاختبار في هذا النموذج 29.83%، راجع فقرة **(3•1•5•5)**.

كما تتيح هذه الأداة إمكانية تخزين نتائج هذا الاستعلام في أي مخزن للبيانات، وقد تم هنا اختيار تخزين نتائج هذا الاستعلام في جدول جديد في مستودع البيانات الخاص بالدراسة (Clinical Laboratory Study)، انظر الشكل (5–6–2).

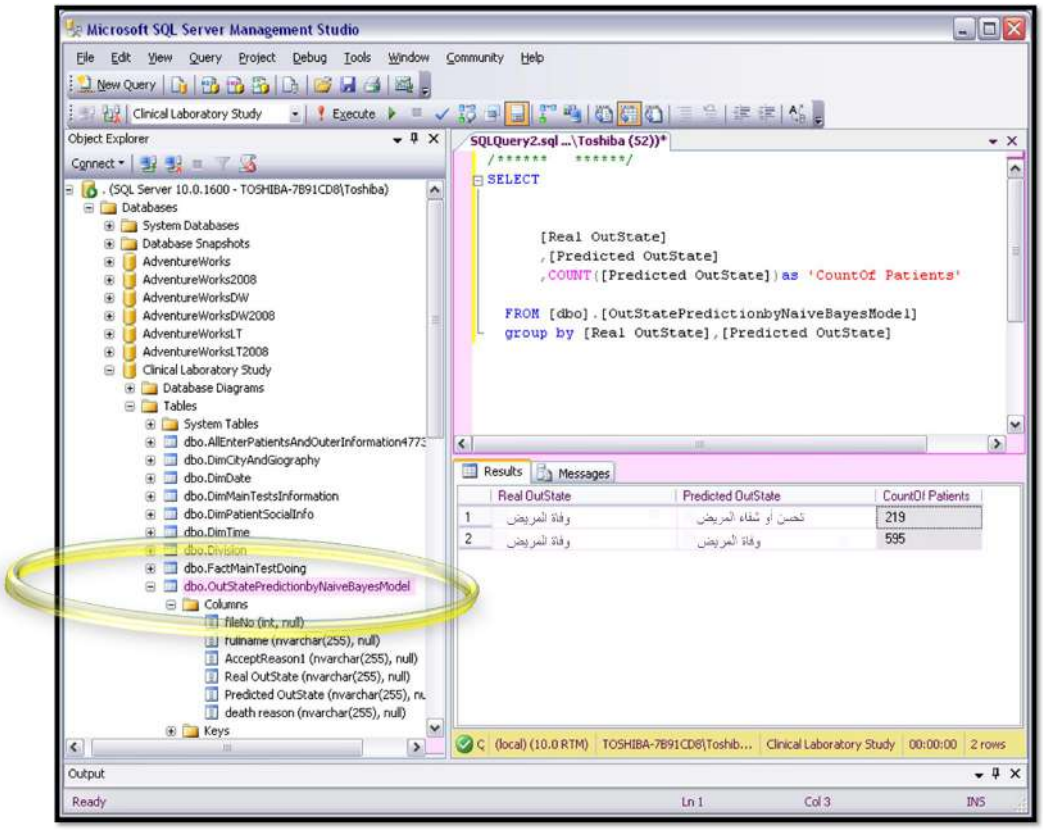


الشكل (5–6– 2): نافذة نتائج الاستعلام عن توقعات نموذج (NaïveBayes_OutState) حول بيانات المرضى المتوفين في جدول بيانات المرضى (PatientEnterProcess)، واستدعاء نافذة الإجراء الخاص بتخزين جدول الاستعلام في مستودع بيانات التحاليل الطبية

3•2•6•5 استخدام الاستعلام الناتج المخزن في مستودع بيانات التحاليل الطبية

انضم الاستعلام الناتج عن عملية التخزين السابقة إلى مجموعة جداول مستودع بيانات مخبر التحاليل الطبية، وتفتح بذلك نتائج التنقيب في البيانات أبواباً جديدة لأبحاث جديدة، حيث يمكن استثمار نتائج التنقيب في البيانات في بناء استعلامات جديدة، وفي إجراء دراسات إحصائية جديدة، وفي إنشاء تقارير تفيد في بناء القرارات. ويبين الشكل (5–6–3) نتائج الاستعلام التلخيصي من الجدول الجديد عن أعداد التوقعات الصحيحة والتوقعات الخاطئة لحالة وفاة

المرضى مقارنة بالقيم الحقيقية، حيث تقدر التوقعات الصحيحة بنسبة 73.09% (595÷814%).



الشكل (5-6- 3): نافذة برنامج (SQL Server Management Studio) ويظهر فيها انضمام جدول الاستعلام عن توقعات نموذج التنقيب (OutStatePredictionByNaiveByesModel) إلى مجموعة جداول مستودع بيانات التحاليل المخبرية، كما يظهر في هذه النافذة الاستعلام التلخيصي عن أعداد التوقعات الصحيحة والتوقعات الخاطئة لحالة وفاة المرضى

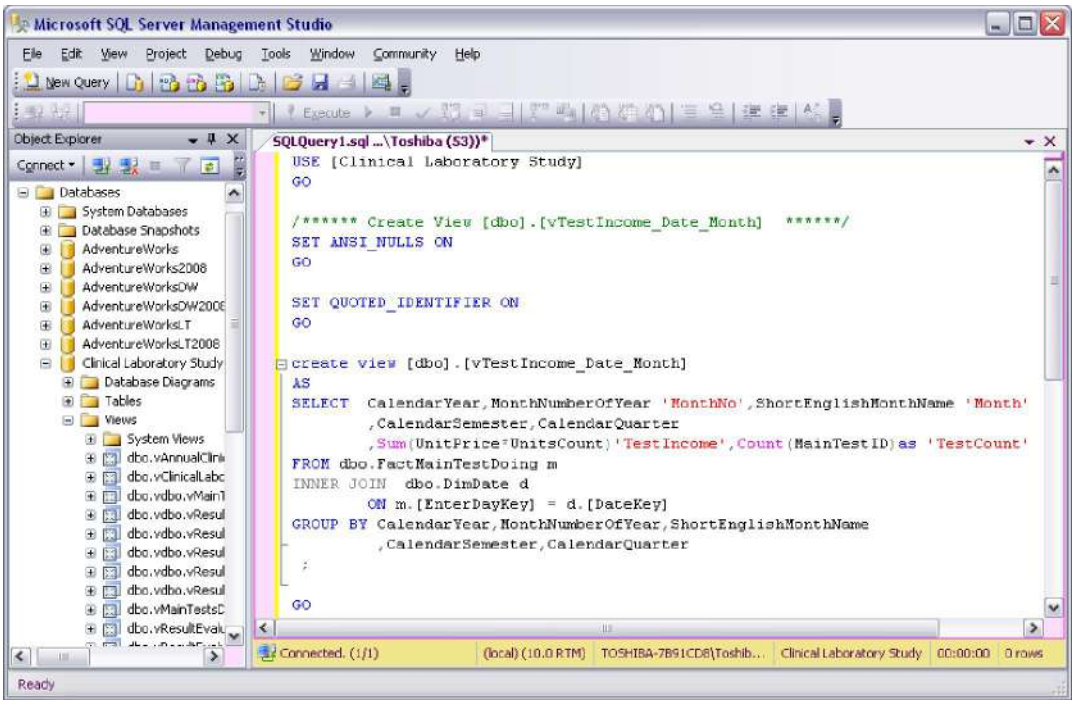
3•6•5 التنبؤ بالإيراد الشهري لمخبر التحاليل الطبية لعام 2009

قدمت أدوات المعالجة التحليلية الآنية (OLAP) في مرحلة تحضير البيانات الكثير من المعلومات المفيدة الناتجة عن استعراض وتلخيص وترتيب البيانات وفق الأبعاد أو الخاصيات المختلفة. فمثلاً تبين أن الإيراد المادي للمخبر قد بلغ ذروة الارتفاع خلال الربع الثالث لكل من العامين المدروسين، وهي معلومة مفيدة وتساعد المخبر في تكوين الملاحظة الأولية عن التغيرات التي تحدث بشكل دوري في فترات زمنية معينة، ولكنها تخلو من الأرقام المستقبلية

وتحتاج إلى تدعيمها ومعالجتها باستخدام الأدوات الأخرى مثل أدوات التنقيب في البيانات التي تعمل في كثير من الأحيان على متابعة المهام التي بدأت بها أدوات المعالجة التحليلية الآنية.

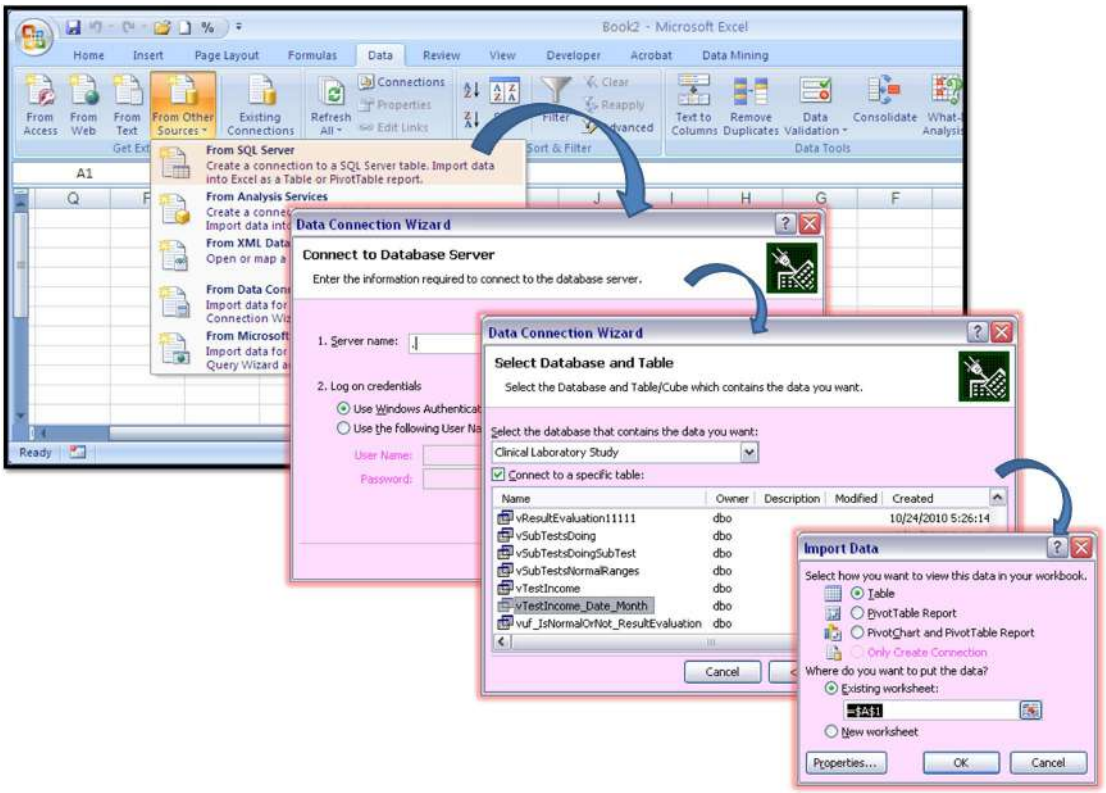
ونبين فيما يلي الخطوات التي تم إتباعها للتنبؤ بالإيراد الشهري للمخبر خلال عام 2009، باستخدام أدوات التنقيب التنبؤي في البيانات.

أ- جرى إنشاء مرأى في مستودع بيانات التحاليل الطبية باستخدام لغة الاستعلام (SQL) في برنامج (SQL Server Management Studio) وذلك لحساب الأعداد والإيرادات الشهرية للتحاليل الطبية في المخبر المدروس خلال الفترة 2007-2008، انظر الشكل (5-6-4). وتعتبر هذه العملية كمثال عن قدرات لغة الاستعلام (SQL) في تلخيص البيانات والتي تختلف عن عملية (OLAP) في أنها تُنفذ طلبات محددة وغير متعددة الأبعاد، ويمكن اعتبار بعض المهام التي تقوم بها هذه اللغة في تلخيص البيانات وفق خاصيات معينة على أنها جزء من المهام التي تستطيع أن تقوم بها (OLAP).



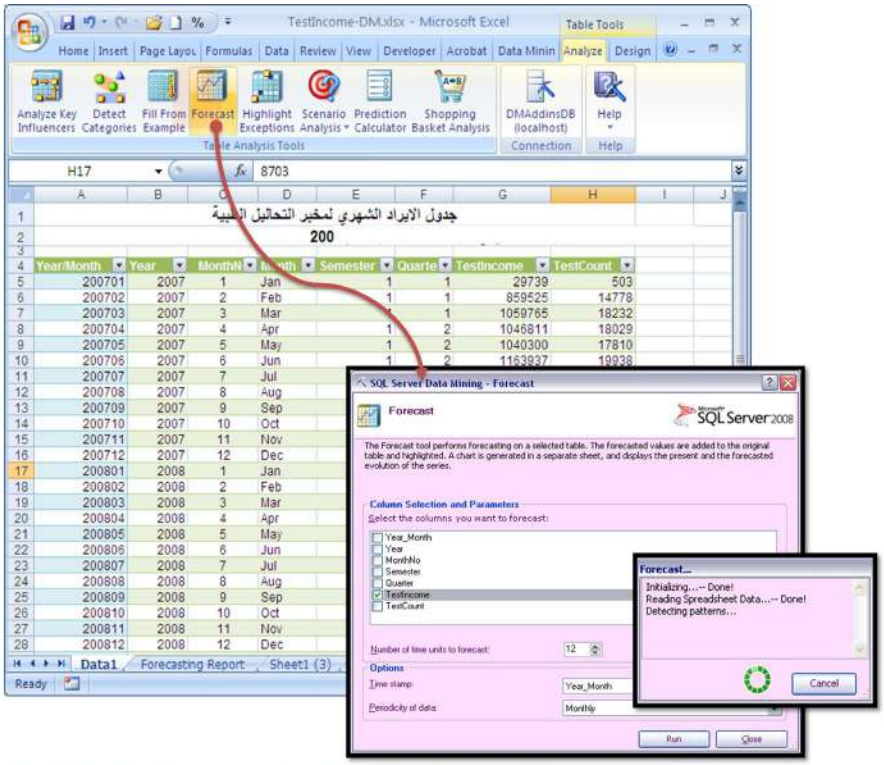
الشكل (5-6- 4): إنشاء مرأى في برنامج (SQL Server Management Studio) لحساب الإيرادات الشهرية للتحاليل الطبية

ب- تم الاستيراد في برنامج اكسل للمرأى السابق وذلك كما هو مبين في الشكل (5-6-5)، وينتج بالتالي جدول قابل للتنقيب فيه باستخدام أدوات التنقيب المتوفرة في برنامج اكسل.

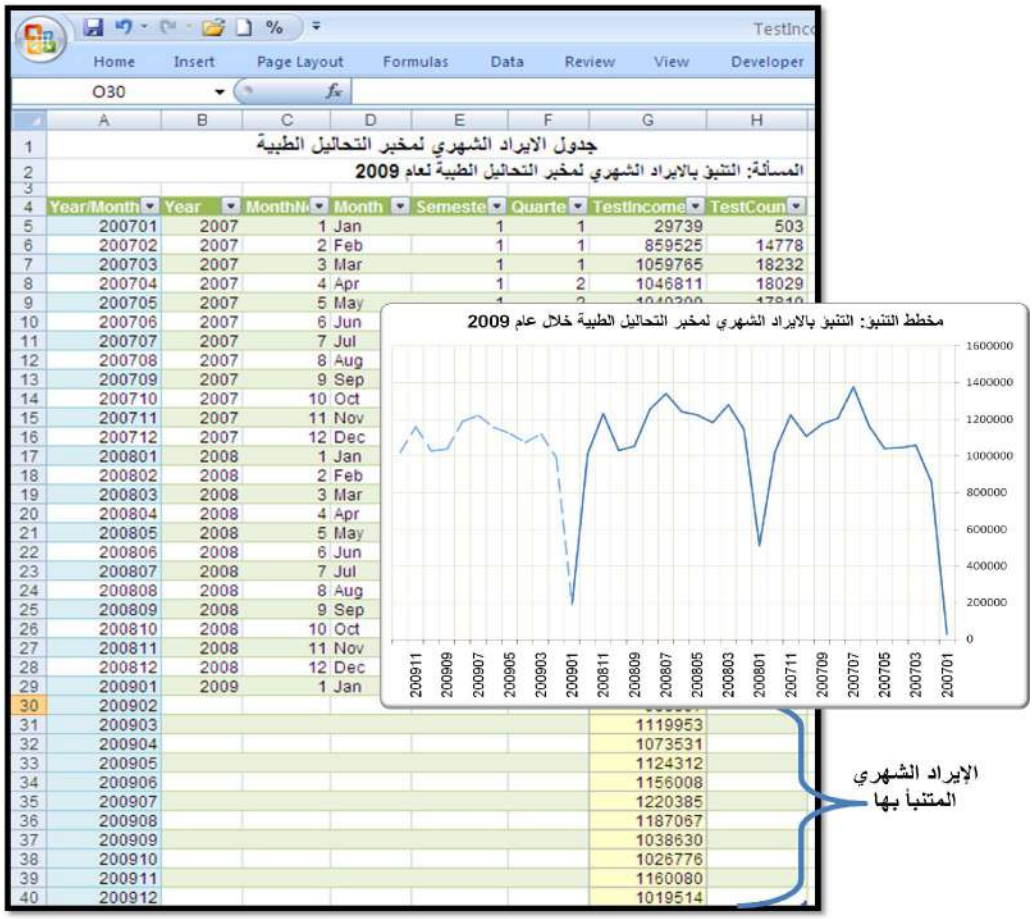


الشكل (5–6– 5): الاستيراد في برنامج اكسل لمرأى الإيرادات الشهرية للتحاليل الطبية الموجود في مستودع بيانات التحاليل الطبية

ج- تم استدعاء إجراء التنبؤ (Forecast) الموجود في قائمة التحليل (Analyze)، واختيار التنبؤ بالإيراد الشهري لمخبر التحاليل الطبية لمدة لا تقل عن 12 شهر اعتماداً على الإيراد الشهري خلال عامي 2007–2008، انظر الشكل (5–6–6). ويقوم بالتالي هذا الإجراء بإضافة البيانات المتوقعة إلى حقل الإيراد الشهري في الجدول الذي طُبِّق عليه هذا الإجراء، ويقوم أيضاً بإنتاج مخطط بياني عن الإيراد الشهري الحالي والمتوقع لمخبر التحاليل الطبية، لاحظ الشكل (5–6–7).



الشكل (– –): استدعاء إجراء التنبؤ ( ) في برنامج اكسل للتنبؤ بالإيراد الشهري لمخبر التحاليل الطبية خلال عام 2



الشكل (5–6– 7): مخطط التنبؤ بالدخل الشهري لمخبر التحاليل الطبية خلال عام 2009

أظهر المخطط البياني للإيرادات الشهرية الحقيقية والمتوقعة لمخبر التحاليل الطبية أنه يوجد انخفاض في الإيراد الشهري للتحاليل الطبية خلال الأشهر الباردة (كانون الأول، كانون الثاني، شباط)، وأن هنالك تزايداً في هذا الإيراد خلال الأشهر الدافئة وخصوصاً شهر تموز. وتبعاً لذلك فمن المتوقع أن ترتفع إيرادات التحاليل الطبية (وأعداد التحاليل الطبية أيضاً) خلال فصل الصيف من عام 2009 وستصل إلى ذروتها في شهر تموز.

4•6•5 استعمال المعرفة المكتشفة في تتبع تتالي الأسباب المرضية المؤدية إلى تكرار دخول المرضى إلى المشفى

تُعتبر دراسة تتالي الأمراض التي يتعرض لها المرضى وتسبب دخولهم المتكرر إلى المشفى على درجة من الأهمية في اكتشاف العلاقات الخفية بين الأمراض التي يتعرض لها الإنسان عبر الزمن. وعالمياً تُجرى مثل هذه الدراسات على سجل الصحة الالكتروني[1] حيث تتاح متابعة حياة السكان الصحية ودراسة الأجيال واستنتاج بعض العلاقات الترافقية واكتشاف المعارف التي تحتويها قواعد البيانات الطبية الضخمة والمتعددة المصادر، وذلك بالاستعانة بأدوات اكتشاف المعرفة من قواعد البيانات.

وتشكل القاعدة التالية مثالاً افتراضياً عن نمط القواعد التي قد تنتج من دراسة تتالي الأمراض

" تبين أن 30% من الأشخاص الذكور الذين حصل معهم التهاب سحايا بعمر 5–9 سنوات قد ظهر معهم ورم في الدماغ بعمر 50–60 "

يُلاحظ من هذا المثال أنه قد لا تكون هنالك علاقة طبية معروفة بين هذه الأمراض، ولكن دراسة البيانات الهائلة المتراكمة عبر الزمن الطويل تكشف خفايا الأمور، وتطرح العديد من القضايا البحثية الهامة التي قد تسترعي اهتمام الباحثين الطبيين، وتدفعهم إلى الخوض أكثر في تفاصيل الأحداث، والبحث عن أسبابها وإجراء التجارب الطبية المناسبة. ومن الأمثلة الأخرى الجيدة التي تدل على أهمية دراسة البيانات الطبية عبر الزمن تلك الدراسات التي تكتشف ترافق تناول بعض أنواع الأدوية (أو الأغذية) مع ظهور أمراض مستقبلية لدى متناولي هذه

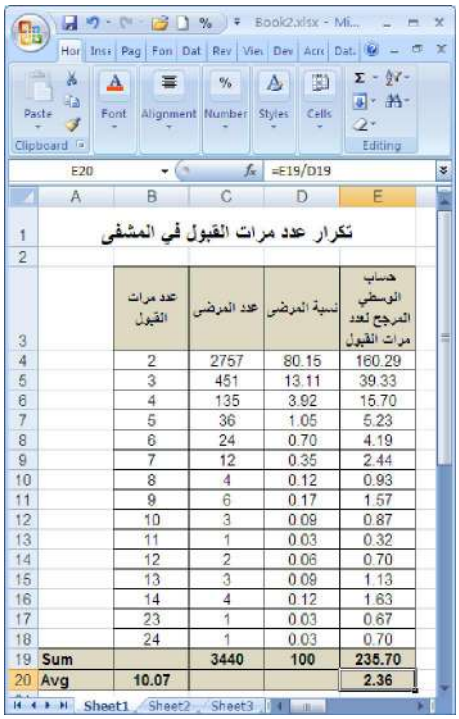
[1] كما سبق وورد في الفصل الرابع

الأدوية لفترات طويلة من الزمن، وهذا ما يدفع الجهات الصحية لإيقاف إنتاج وبيع أصناف من الأدوية كانت تعتبر آمنة لحقبة من الزمن ثم ثبت أن لها أضرار مخفية ومخيفة تفوق الفوائد المرجوة منها.

تتصف البيانات الطبية المتاحة لهذه الدراسة بقصر الفترة الزمنية التي تغطيها، وبأنها تمثل بيانات مشفى واحد فقط، ولكن يمكن استخدامها في تقديم نموذج جيد عن دراسة تتالي الأسباب المرضية المؤدية إلى تكرار دخول بعض المرضى إلى المشفى.

بلغ عدد سجلات قبول المرضى في المشفى والمتاحة للدراسة 47731 سجل، وتمثل 43063 مريض. أي تكرر دخول 3440 مريض إلى المشفى، ويوجد بعض السجلات للمرضى في عام 2005 كما سبق وورد في مرحلة استيعاب البيانات، وقد احتُفِظ بهذه البيانات لاستخدامها في حل مثل هذه المسألة.

أظهر الاستعلام التلخيصي لبيانات المرضى الذين تكرر قبولهم في المشفى أن عدد مرات القبول في المشفى يتراوح بين [2–24] قبول، ويُقدَّر وسطي عدد مرات القبول المرجح بأعداد المرضى بـــ 2.36 مرة، انظر الشكل (5–6–8).



تكرار عدد مرات القبول في المشفى

	عدد مرات القبول	عدد المرضى	نسبة المرضى	حساب الوسطي المرجح لعدد مرات القبول
4	2	2757	80.15	160.29
5	3	451	13.11	39.33
6	4	135	3.92	15.70
7	5	36	1.05	5.23
8	6	24	0.70	4.19
9	7	12	0.35	2.44
10	8	4	0.12	0.93
11	9	6	0.17	1.57
12	10	3	0.09	0.87
13	11	1	0.03	0.32
14	12	2	0.06	0.70
15	13	3	0.09	1.13
16	14	4	0.12	1.63
17	23	1	0.03	0.67
18	24	1	0.03	0.70
19	Sum		3440	100	235.70
20	Avg	10.07			2.36

الشكل (5–6– 8): جدول تلخيصي لإظهار تكرار قبول المرضى في المشفى

يُنظر إلى تتالي الأسباب المرضية المؤدية إلى تكرار القبول في المشفى وكأنه سلة الأمراض التي يتعرض لها المريض خلال الزمن. وبناءً على ذلك فقد تمت دراسة هذه المسألة بالاستعانة بأداة التنقيب التي تعتمد على خوارزمية قواعد الاقتران وتُستخدم عادةً في تحليل سلة المشتريات في برنامج اكسل، وذلك وفق الخطوات التالية المبينة في الشكل (5-6-9):

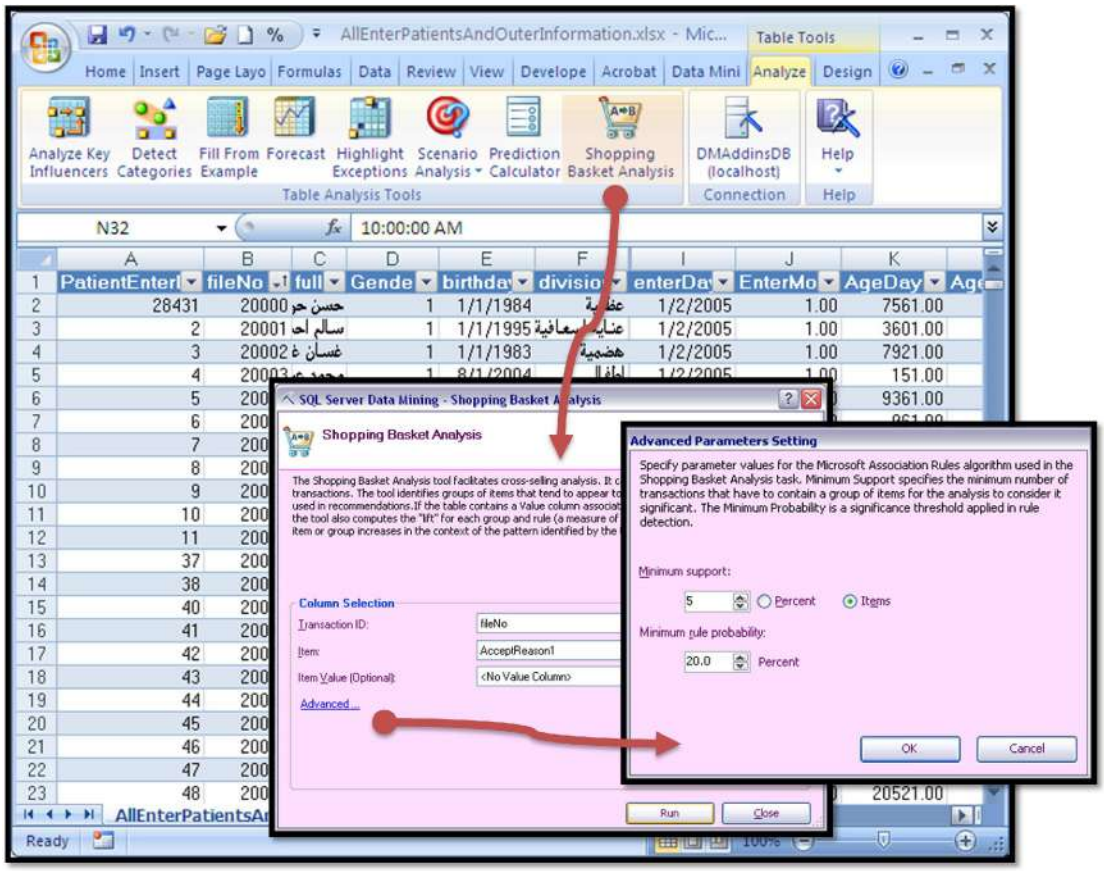
أ- الاستيراد في برنامج اكسل للجدول الذي يحتوي على بيانات المرضى الذين تكرر قبولهم في المشفى.

ب- استدعاء إجراء تحليل سلة المشتريات (Shopping Basket Analysis) الموجود في قائمة التحليل (Analyze).

ج- اعتبار رقم اضبارة المريض على أنه الخاصة المميزة لمجموعة الأسباب المرضية لكل مريض.

د- اختيار خاصية الأسباب المرضية (Accept Reason1) في خلية البند المراد دراسة (Item) الترافق بين مكوناته.

ه- تتحدد المخرجات الناتجة عن تحليل سلة الأسباب المرضية بحسب قيم المحددات (Parameters) التي تستخدمها خوارزمية قواعد الاقتران في عملية التنقيب هذه، وتقترح أداة التنقيب عادةً قيماً افتراضية لهذه المحددات، ولكن بدايةً أدى اعتماد هذه المحددات في هذه المسألة إلى قلة القواعد الناتجة بسبب صغر حجم البيانات التي يتم التنقيب فيها، ولذلك تم تخفيض الحد الافتراضي الأدنى لتكرار الترافق بين الأسباب المرضية (Minimum Support) من 10 إلى 5 أي لجعل إجراء التنقيب يكتشف قواعد الاقتران التي يكون الحد الأدنى لتكرارها هو خمسة حالات، كما وتم تخفيض قيمة الاحتمال الأدنى (Minimum Rule Probability) لتحقق هذه القواعد في البيانات من 40% إلى 20%، وأدت هذه التخفيضات إلى ظهور قواعد اقتران إضافية في النتائج.



الشكل (5–6– 9): خطوات تحليل سلة الأسباب المرضية التي تؤدي إلى تكرار قبول المرضى في المشفى باستخدام أداة التنقيب (Shopping Basket Analysis)

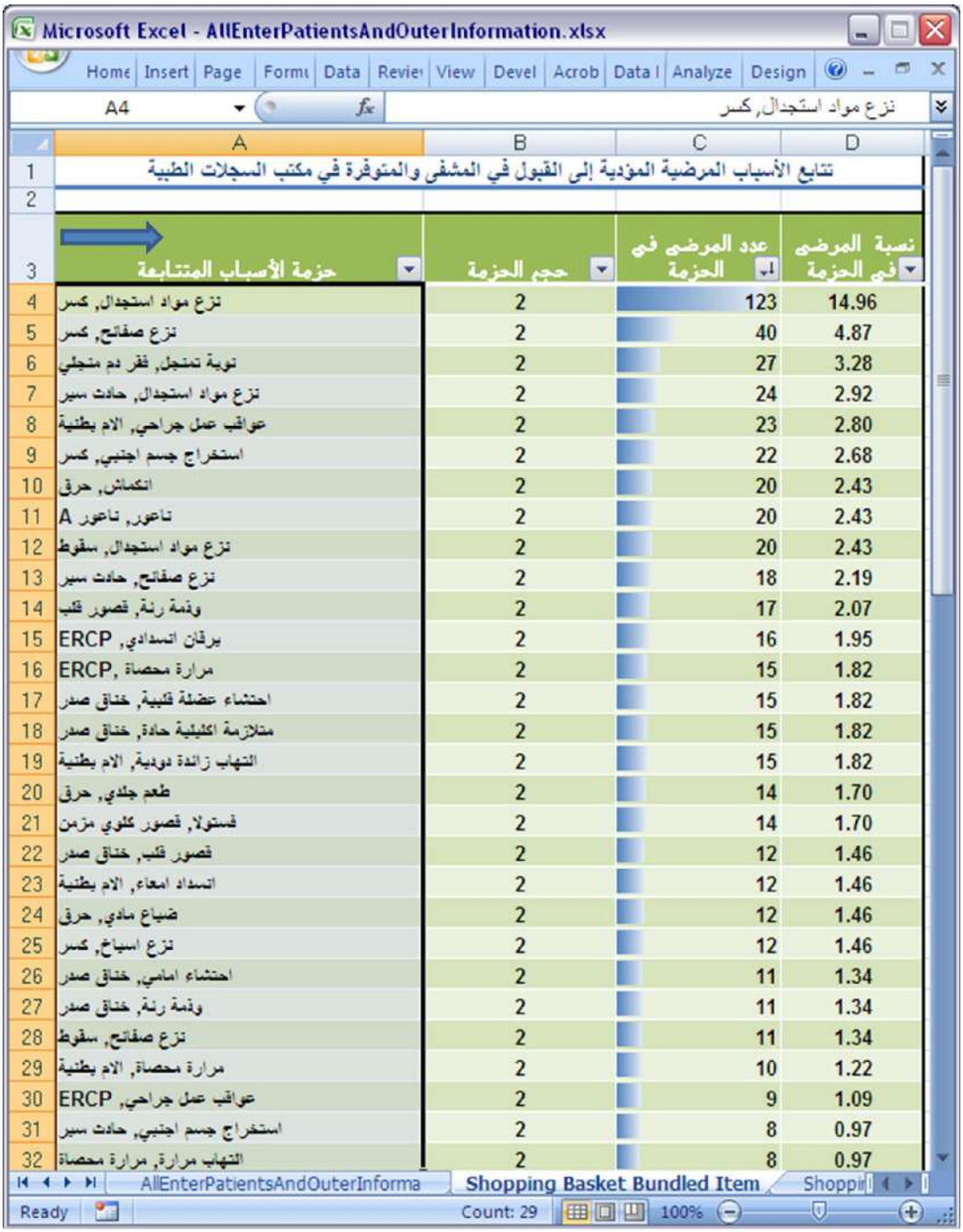
نتج عن تطبيق الإجراء السابق جدولُ بمكونات حزم الأسباب المرضية التي أدت إلى القبول المتتالي للمرضى في المشفى، مصنفة تنازلياً بحسب حجم الحزمة (عدد المرضى في الحزمة)، انظر الشكل (5–6–10)، ونقرأ في هذا الشكل الاستنتاجات التالية:

أ- تُشكل الأسباب المرضية المرتبطة بالكسور النسبة الأكبر من الأسباب التي يتكرر القبول في المشفى لأجلها. فمثلاً نجد أن 123 مريض دخلوا إلى المشفى بحالة كسر، عادوا ودخلوا إلى المشفى لنزع مواد استجدال، و40 عادوا لنزع الصفائح.

ب- 17 مريض دخلوا إلى المشفى بسبب قصور القلب، عادوا إلى المشفى بسبب وذمة رئة.

ج- تظهر في البيانات الناتجة بعض الترادفات في معنى المصطلحات الطبية والتي تعتبر من العيوب التي تُشيب قواعد البيانات الطبية المدروسة، والتي تحتاج إلى خبرة طبية للتعرف عليها، وتتطلب مستقبلاً إيجاد حلول لها قبل إدخال البيانات إلى الأنظمة الحاسوبية، وليس عند وصول البيانات إلى مرحلة التنقيب، وذلك علماً أنه بُذل الكثير من الجهد في مرحلة

تحضير البيانات للتقليل من تشوه بيانات الأسباب المرضية. ويُمكن أن نكون ايجابيين تجاه هذه النتيجة ونقول أن اكتشاف العيوب في البيانات يُعتبر أيضاً بأنه اكتشاف معرفة.



Microsoft Excel - AllEnterPatientsAndOuterInformation.xlsx

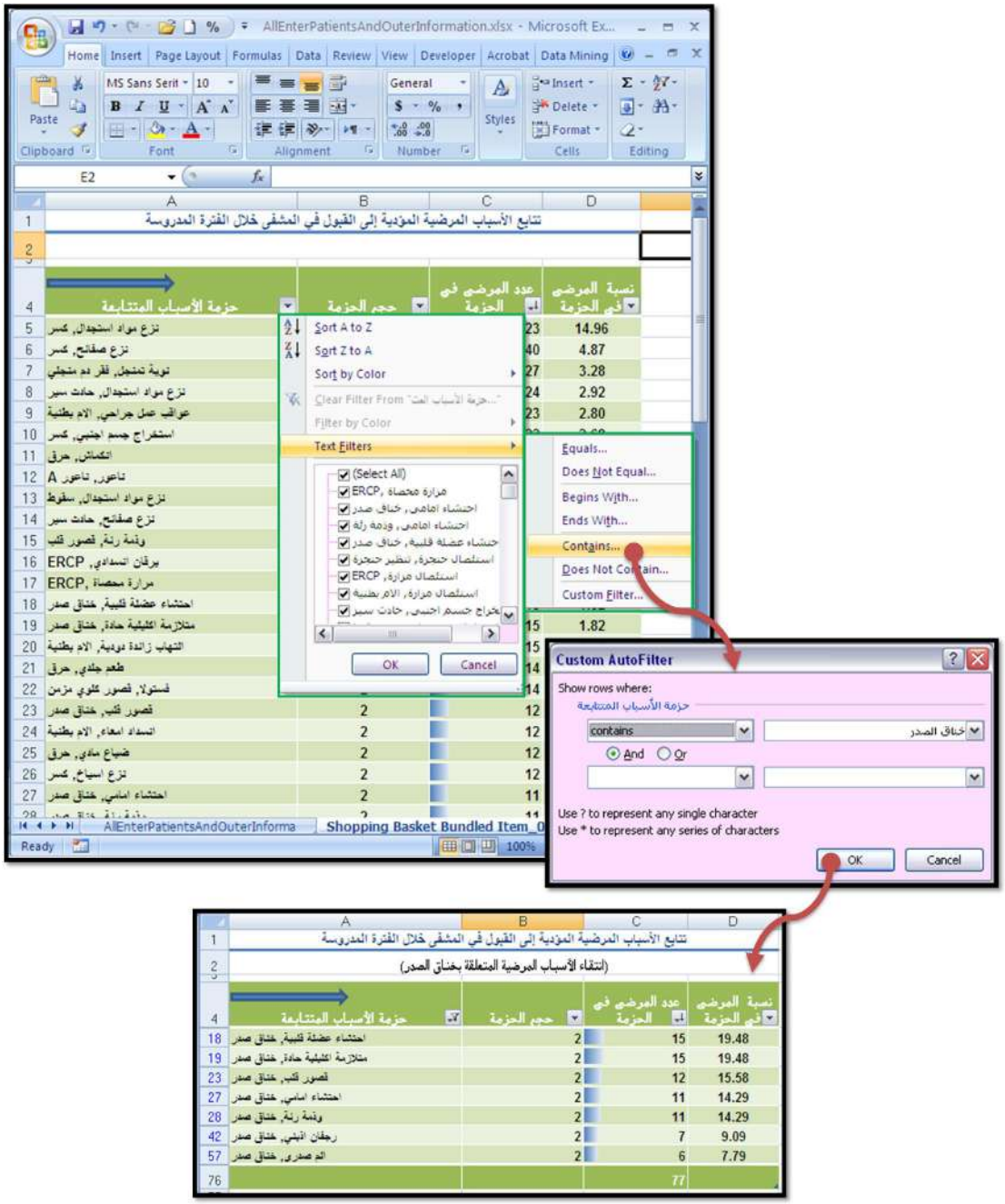
	A	B	C	D
1	تتابع الأسباب المرضية المؤدية إلى القبول في المشفى والمتوفرة في مكتب السجلات الطبية			
2				
3	حزمة الأسباب المتتابعة	حجم الحزمة	عدد المرضى في الحزمة	نسبة المرضى في الحزمة
4	نزع مواد استجدال, كسر	2	123	14.96
5	نزع صفائح, كسر	2	40	4.87
6	ثوبة تمتجل, قفر دم متجلي	2	27	3.28
7	نزع مواد استجدال, حادث سير	2	24	2.92
8	عواقب عمل جراحي, الام بطنية	2	23	2.80
9	استخراج جسم اجنبي, كسر	2	22	2.68
10	انكماش, حرق	2	20	2.43
11	A تاعور, تاعور	2	20	2.43
12	نزع مواد استجدال, سقوط	2	20	2.43
13	نزع صفائح, حادث سير	2	18	2.19
14	وذمة رئة, قصور قلب	2	17	2.07
15	ERCP, برقان انسدادي	2	16	1.95
16	ERCP, مرارة محصاة	2	15	1.82
17	احتشاء عضلة قلبية, خناق صدر	2	15	1.82
18	متلازمة اكليلية حادة, خناق صدر	2	15	1.82
19	التهاب زائدة دودية, الام بطنية	2	15	1.82
20	طعم جلدي, حرق	2	14	1.70
21	فستولا, قصور كلوي مزمن	2	14	1.70
22	قصور قلب, خناق صدر	2	12	1.46
23	انسداد امعاء, الام بطنية	2	12	1.46
24	ضياع مائي, حرق	2	12	1.46
25	نزع اسياخ, كسر	2	12	1.46
26	احتشاء امامي, خناق صدر	2	11	1.34
27	وذمة رئة, خناق صدر	2	11	1.34
28	نزع صفائح, سقوط	2	11	1.34
29	مرارة محصاة, الام بطنية	2	10	1.22
30	ERCP, عواقب عمل جراحي	2	9	1.09
31	استخراج جسم اجنبي, حادث سير	2	8	0.97
32	التهاب مرارة, مرارة محصاة	2	8	0.97

AllEnterPatientsAndOuterInforma | Shopping Basket Bundled Item | Shoppi

Ready | Count: 29 | 100%

الشكل (5-6- 10): تتابع الأسباب المرضية المؤدية إلى الدخول إلى المشفى خلال الفترة المدروسة (2007-2008)، ويوجد بعض البيانات عن المرضى في عام 2005

د- يترافق "خناق الصدر" مع عدة أسباب مرضية أخرى في الحزم المختلفة، وللتعرف بالتفصيل على هذه الحزم يمكن انتقاء الحزم التي تحتوي على عبارة "خناق صدر" ودراسة مكوناتها بصورة مستقلة، وذلك كما هو مبين في الشكل (5-6-11).



الشكل (5-6- 11): استخدام أدوات الانتقاء في اكسل لاختيار جزء من الحزم الموجودة في جدول تتابع الأسباب المرضية والمتعلقة بمرض "خناق الصدر"

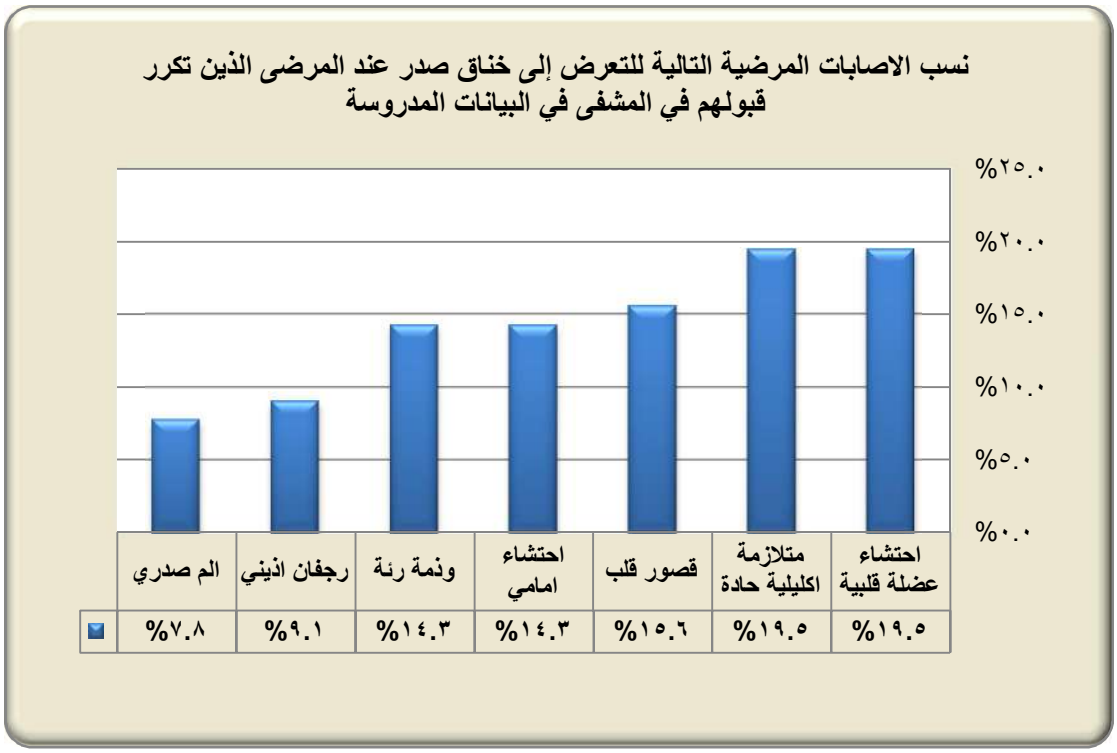
وقد نَتجتْ عن عملية الانتقاء الحقيقةُ التالية:

"إن 77 مريض عانوا من خناق صدر، ثم حدثت معهم الإصابات المرضية التالية التي قُبلوا بسببها في المشفى ثانيةً:

- 15 منهم حصل معهم احتشاء عضلة قلبية.
- 15 مريض منهم حدث معهم متلازمة إكليلية حادة.
- 12 مريض منهم حصل معهم قصور قلب.

- 11 منهم حصل معهم احتشاء أمامي.
- 11 منهم حصل معهم وذمة رئة.
- 7 منهم حصل معهم رجفان أذيني.
- 6 منهم حصل معهم ألم صدري."

ويمكن تحويل الحقيقة السابقة[1] إلى نسب مئوية تُعبر عن احتمالات الإصابة بالمشاكل الصحية المختلفة عند مجموعة المرضى المدروسة وذلك حسب ما هو مبين في الشكل البياني (5-6-12).



الشكل (5-6- 12): نسب الإصابات المرضية التي تلت التعرض إلى خناق الصدر عند المرضى الذين تكرر قبولهم في المشفى المدروس

وتُعبِّر النسب السابقة فقط عن الفئة التي استخلصت عنها قواعد الاقتران في هذه المسألة، ولا يمكن تعميم هذه النتيجة على كل المرضى الذين أصيبوا بخناق الصدر، وذلك أنه يوجد في بيانات سجلات المرضى 978 سجل تحتوي على سبب خناق الصدر[2]، ومنها

[1] تحتوي الحقيقة السابقة على بعض التقاطعات في المعاني بسبب عدم إتباع مصطلحات موحدة في الإشارة إلى الأمراض، فمثلاً يُعتبر الاحتشاء الأمامي أحد تفرعات احتشاء العضلة القلبية..، وبغض النظر عن المعاني الطبية فإن هذه الحقيقة هي مثال نموذج عن الحقائق التي تقدمها أداة التنقيب في البيانات. وهنا تبرز ضرورة تقيد المشافي بالتوثيق الالكتروني وفقاً للتصنيف الدولي للأمراض، مما يقلل من التكرار والتداخل بين الأسباب المرضية.

[2] بحسب الاستعلام من جدول سجلات المرضى في قاعدة بيانات السجلات الطبية.

181 سجل تكرر دخول أصحابهم إلى المشفى، ولكن يعود سبب ظهور فقط 77 مريض منهم في الحزم (قواعد الاقتران الناتجة) إلى أنهم اجتازوا شروط خوارزمية التنقيب هنا وهي أن يكون الحد الأدنى لتكرار الحالة في البيانات هو 5 حالات، وأن يكون الحد الأدنى لاحتمال أن تنطبق عليهم القاعدة المكتشفة هو 20%.

ه- أظهرت البيانات تكرار وجود السبب المرضي المعبر عنه بالآلام البطنية في الحزم المختلفة، وتم لذلك انتقاء الحزم التي تحتوي على هذا السبب فنتج الجدول المعروض في الشكل (5-6-13)، ونتجت الحقيقة التالية:

"إن 73 مريض عانوا من آلام بطنية لم يتم تبريرها مباشرة عند القبول في المشفى حصلت معهم الحالات التالية التي قبلوا بسببها في المشفى لاحقاً:

- 23 منهم أجريت لهم جراحات مختلفة (غير موثقة في البيانات الأصلية)، ثم عادوا إلى المشفى بشكوى عواقب عمل جراحي.
- 15 مريض منهم عادوا إلى المشفى وشُخِّص لهم التهاب زائدة دودية.
- 12 منهم ظهر معهم فيما بعد انسداد أمعاء.
- 10 منهم عادوا إلى المشفى وتبين وجود مرارة محصاة.
- 8 أشخاص عادوا إلى المشفى بسبب السقوط، وتعتبر هذه الفكرة غير مبررة طبياً، وهنا نجد أن أدوات التنقيب في البيانات تساعد في اكتشاف علاقات غريبة بين البيانات وقد لا يمكن تفسيرها بصورة منطقية أحياناً.
- 5 أشخاص عادوا إلى المشفى وتم استئصال المرارة لهم."

يَحملُ هذا النوع من التنقيب في تتبع تتالي الأمراض الكثيرَ من الفوائد بالنسبة لشركات التأمين، حيث تُساعد النتائج المكتشفة في تقدير التكاليف الطبية المترتبة على الأمراض الحالية وعلى الإصابات المستقبلية المتتالية المتوقعة. كما أن الاحتمالات المتوقعة للإصابة بأمراض معينة بعد التعرض لمرض ما، تساعد الأطباء على اتخاذ الإجراءات الوقائية أو حتى الاستعداد المسبق ومعرفة كيفية مواجهة تطورات المرض إن حدثت. ويمكن مستقبلاً تطبيق هذا النوع من التنقيب على الأعراض المرضية التي تظهر تباعاً على المرضى، وذلك للتعرف على الخطورة المترافقة مع كل قيمة من قيم الأعراض.

	A	B	C	D
1	تتابع الأسباب المرضية المؤدية إلى القبول في المشفى خلال الفترة المدروسة			
2	(انتقاء الأسباب المرضية المتعلقة بالآلام البطنية)			
4	حزمة الأسباب المتتابعة	حجم الحزمة	عدد المرضى في الحزمة	نسبة المرضى في الحزمة
9	عواقب عمل جراحي, الام بطنية	2	23	31.51
20	التهاب زائدة دودية, الام بطنية	2	15	20.55
24	انسداد امعاء, الام بطنية	2	12	16.44
30	مرارة محصاة, الام بطنية	2	10	13.70
35	سقوط, الام بطنية	2	8	10.96
58	استئصال مرارة, الام بطنية	2	5	6.85
76			73	

الشكل (5-6- 13): استعراض حزم الأسباب المتتابعة التي تحتوي على مشكلة الآلام البطنية

5•6•5 تصميم أداة في اكسل للتنبؤ بالأشخاص المعرضين للوفاة باستخدام الصفات الشخصية والأسباب المرضية ونتائج التحاليل الطبية

يدعمُ برنامجُ اكسل بالتعاون مع مخدم SQL Server تصميمَ أداتي تنبؤ الأولى تفاعلية (Interactive)، والثانية قابلة للطباعة حيث تعمل كل منهما على تقييم البيانات الجديدة بالنسبة لدورها في التنبؤ بالمخرجات، أي يمكن استخدامها مباشرة للتنبؤ بنتائج الحالة المدروسة حالما توفرت عنها بيانات المدخلات، وتستند هاتين الأداتين إلى خوارزمية الانحدار المنطقي. وقد تم القيام بالخطوات التالية للحصول عليهما:

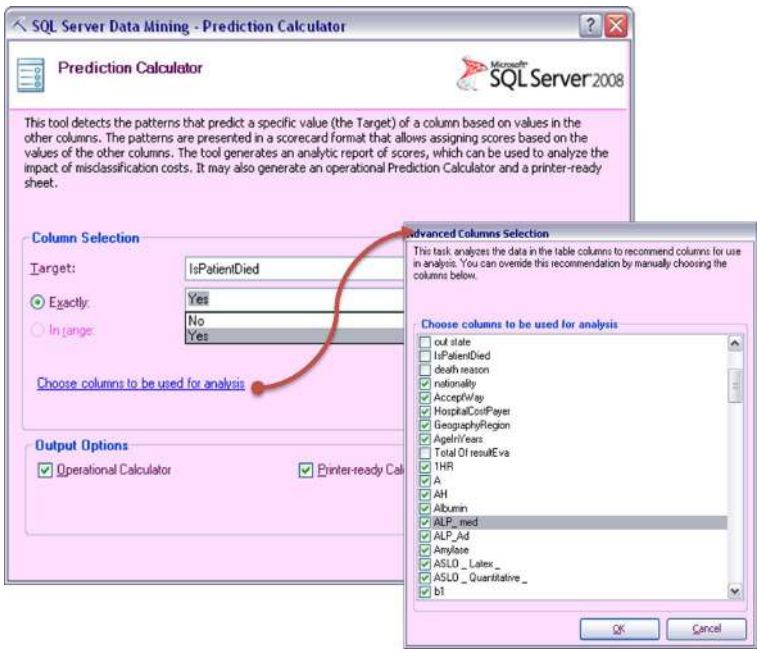
أ- تم في البداية القيام بالاستيراد في اكسل لمرأى البيانات الشخصية والمرضية وخاصيات تقييم نتائج التحاليل الطبية من مستودع بيانات التحاليل الطبية.

ب- جرت إضافة حقل جديد (IsPatientDied) إلى جدول البيانات الناتج بحيث يأخذ النتيجة "Yes" إذا كانت حالة الخروج عند المريض هي "وفاة المريض"، ويأخذ النتيجة "No" في بقية الحالات.

ج- تم استدعاء إجراء حساب التنبؤ (Prediction Calculator) الموجود ضمن قائمة التحليل (Analyze)، واختيار أن يكون الهدف هو حقل (IsPatientDied)، وتحديداً القيمة الايجابية (Yes) التي تدل على وفاة المريض، كما جرى انتقاء معظم خاصيات المرضى الشخصية والطبية المتوفرة، انظر الشكل (5-6-14).

د- عَمِل هذا الإجراء داخلياً على دراسة تأثير كل من قيم الخاصيات المختلفة على المتغير المتنبأ به، واقترح بالتالي نقاطاً تقييميةً مقابلةً لكل قيمة من قيم خاصيات المدخلات وتُعبِّر عن هذا التأثير. فمثلاً قَدَّر هذا الإجراء أن المرضى المقبولين بذات الرئة تُعطى لهم 55 نقطة، وتعطى للمرضى الذين تزيد أعمارهم عن 65 نقطة واحدة، وتُعطى للقادمين من ريف دمشق نقطة واحدة...الخ. كما توصَّل هذا الإجراء أيضاً إلى قيمة تدعى عتبة التقييم وتُمثل الحد الأدنى لمجموع النقاط التي تُميِّز المعرضين للوفاة عن غير المعرضين للوفاة.

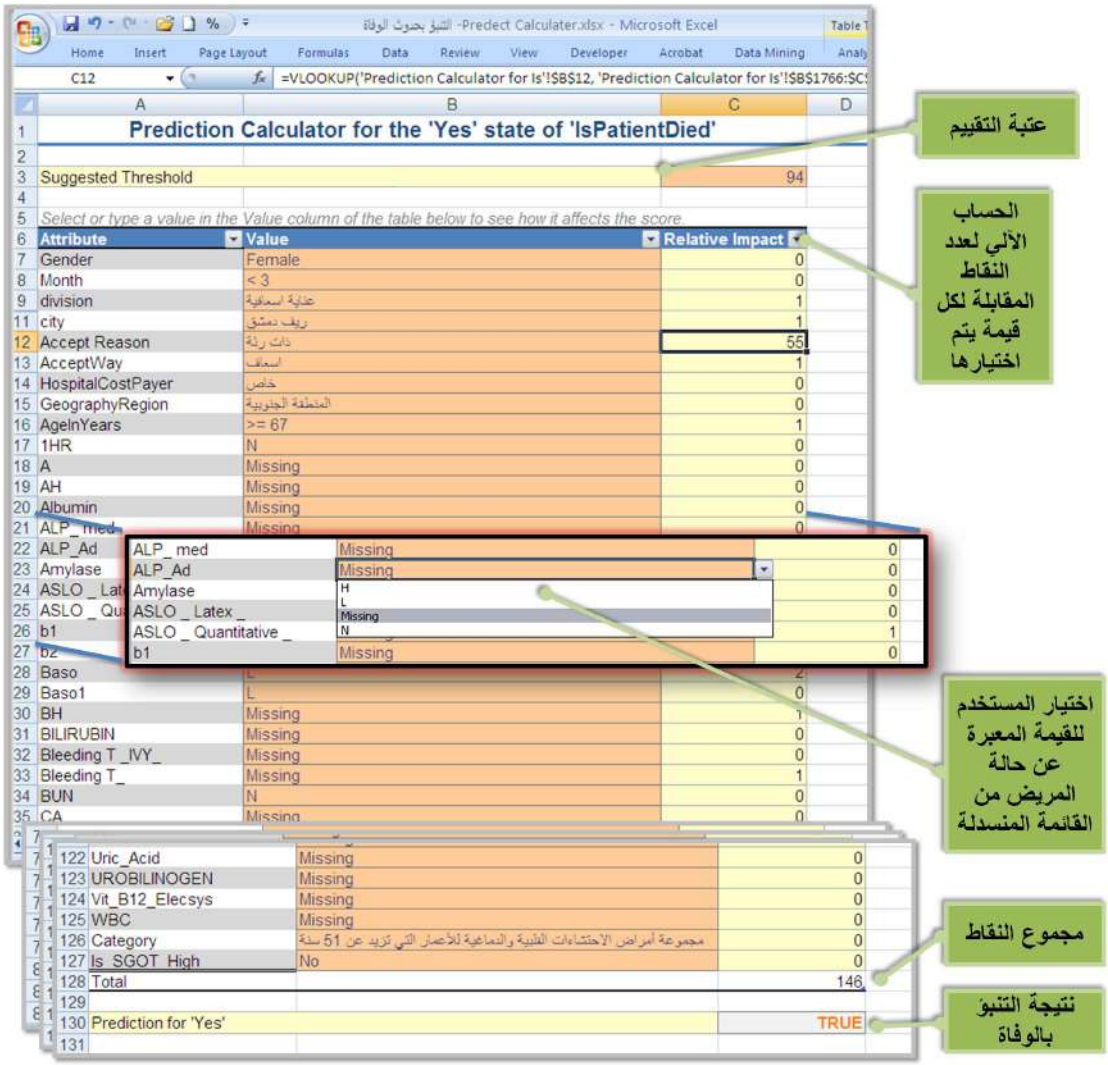
ه- وقد نتج عن هذا الإجراء ورقتي عمل في اكسل تحتويان على أداتين للتنبؤ الأولى آلية تفاعلية والثانية يدوية قابلة للطباعة.



الشكل (5-6- 14): استدعاء إجراء تصميم أداتي تنبؤ بالوفاة الأولى آلية تفاعلية والثانية ورقية

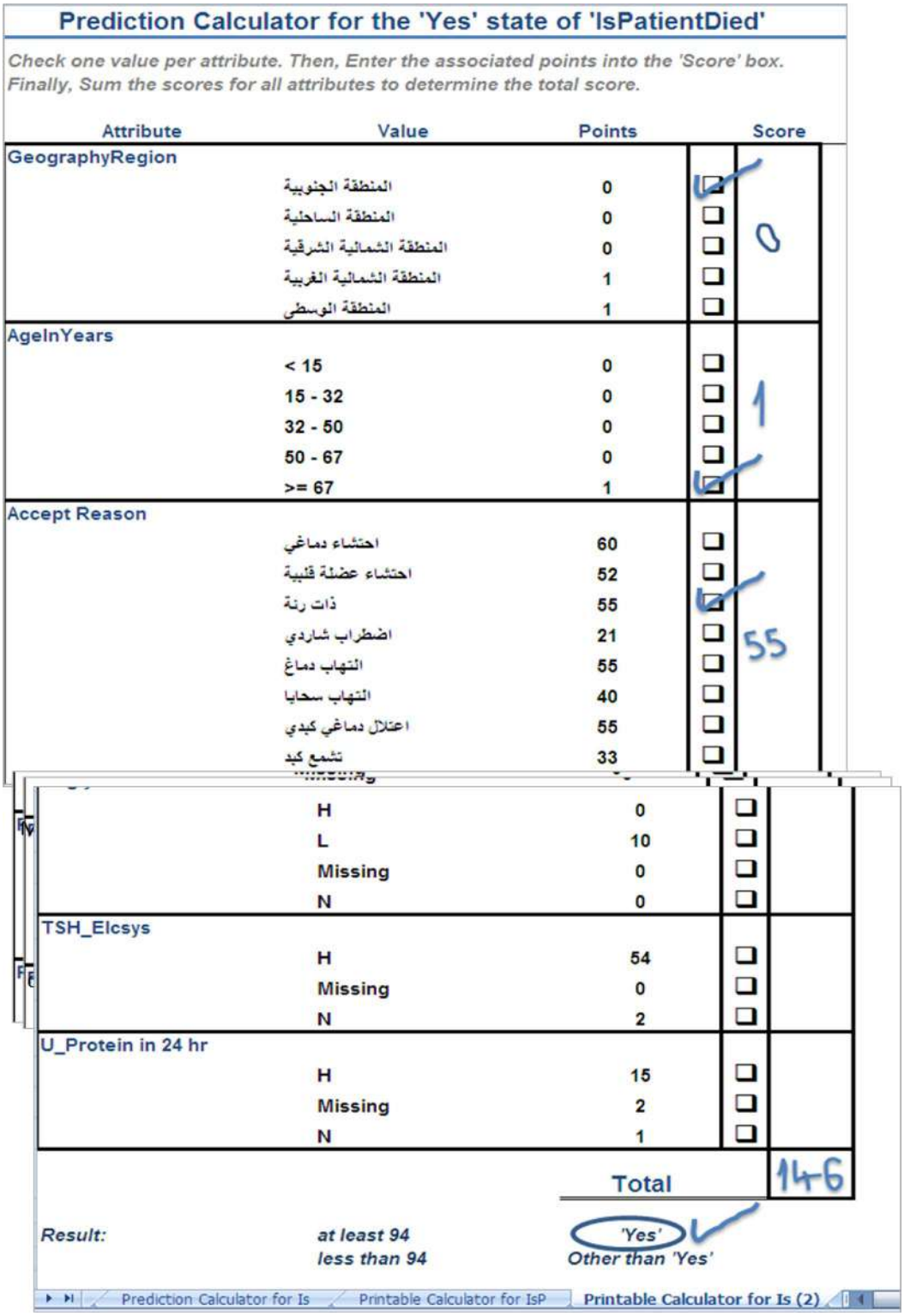
تأخذ الأداة الأولى شكل جدول تفاعلي كما في الشكل (5-6-15) بحيث يُمكن أن يقوم المستخدم إما بكتابة قيمة الخلية أو باختيار قيمة الخلية من القائمة المنسدلة منها، وبالتالي تقوم هذه الأداة باحتساب آلي لمجموع النقاط المقابلة لقيم خاصيات هذا المريض وتقارنه مع العتبة الخاصة بالبيانات، فإذا تجاوز مجموعُ النقاط العتبةَ دل ذلك على أن حالة المريض تحتمل الوفاة، وكلما ارتفع عدد النقاط كلما كان المريض أكثر عرضة للوفاة وبالتالي يجب تكثيف العناية به والتركيز على التخفيف من عوامل الخطورة لديه. أما إذا نقص مجموع النقاط عن العتبة دل ذلك على أن حالة المريض جيدة وقل توقع الوفاة.

ويمكن مستقبلاً ربط أداة التنبؤ بالوفاة بقواعد بيانات المرضى بحيث يتمكن مستخدمو هذه القواعد في الحقل الطبي من التنبؤ بمصير أي مريض من خلال البيانات المدخلة إلى أنظمة قواعد البيانات والمنتقلة آلياً إلى أداة التنبؤ بالوفاة.



الشكل (5-6- 15): أداة التنبؤ الآلي بحدوث الوفاة استناداً إلى الصفات الشخصية والمرضية ونتائج التحاليل الطبية للمرضى، وقد تم اختيار مثال واقعي من البيانات المدروسة، وكان مجموع النقاط يساوي 146 أكبر من العتبة 94، مما دل على توقع الوفاة لهذه المريضة (TRUE)، ومن الملاحظ أنه لا تأثير للجنس في توقع الوفاة

تأخذ أداة التنبؤ الثانية في اكسل شكل استمارة ورقية، وهي متطابقة تماماً في مضمونها مع الأداة السابقة، ولكن يمكن طبع هذه الاستمارة على أوراق واستخدامها في تعبئة بيانات المريض يدوياً ومن ثم يتم جمع النقاط الناتجة ومقارنتها بالعتبة المقترحة من أجل توقع حالة المريض، انظر الشكل (5-6-16).



Prediction Calculator for the 'Yes' state of 'IsPatientDied'

Check one value per attribute. Then, Enter the associated points into the 'Score' box. Finally, Sum the scores for all attributes to determine the total score.

Attribute	Value	Points		Score
GeographyRegion				
	المنطقة الجنوبية	0	☑	0
	المنطقة الساحلية	0	☐	
	المنطقة الشمالية الشرقية	0	☐	
	المنطقة الشمالية الغربية	1	☐	
	المنطقة الوسطى	1	☐	
AgeInYears				
	< 15	0	☐	1
	15 - 32	0	☐	
	32 - 50	0	☐	
	50 - 67	0	☐	
	>= 67	1	☑	
Accept Reason				
	احتشاء دماغي	60	☐	55
	احتشاء عضلة قلبية	52	☐	
	ذات رئة	55	☑	
	اضطراب شاردي	21	☐	
	التهاب دماغ	55	☐	
	التهاب سحايا	40	☐	
	اعتلال دماغي كبدي	55	☐	
	تشمع كبد	33	☐	
	H	0	☐	
	L	10	☐	
	Missing	0	☐	
	N	0	☐	
TSH_Elcsys				
	H	54	☐	
	Missing	0	☐	
	N	2	☐	
U_Protein in 24 hr				
	H	15	☐	
	Missing	2	☐	
	N	1	☐	
		Total		146
Result:	*at least 94*	*'Yes'*	☑	
	less than 94	*Other than 'Yes'*		

Prediction Calculator for Is | Printable Calculator for IsP | Printable Calculator for Is (2)

الشكل (5-6- 16): مقاطع من الاستمارة الورقية لأداة التنبؤ بالوفاة الناتجة عن تطبيق إجراء التنقيب في البيانات (Predict Calculation)

لا يمكن في الواقع استخدام هذه الأداة بوضعها الحالي في توقع حالة المرضى، وذلك لأنها بنيت باستخدام بيانات قليلة جداً عن المرضى سواء من ناحية صغر حجم البيانات أو من ناحية قلة المتغيرات المستخدمة في بناء هذه الأداة، ومن الناحية الطبية يحتاج توقع مثل هذه

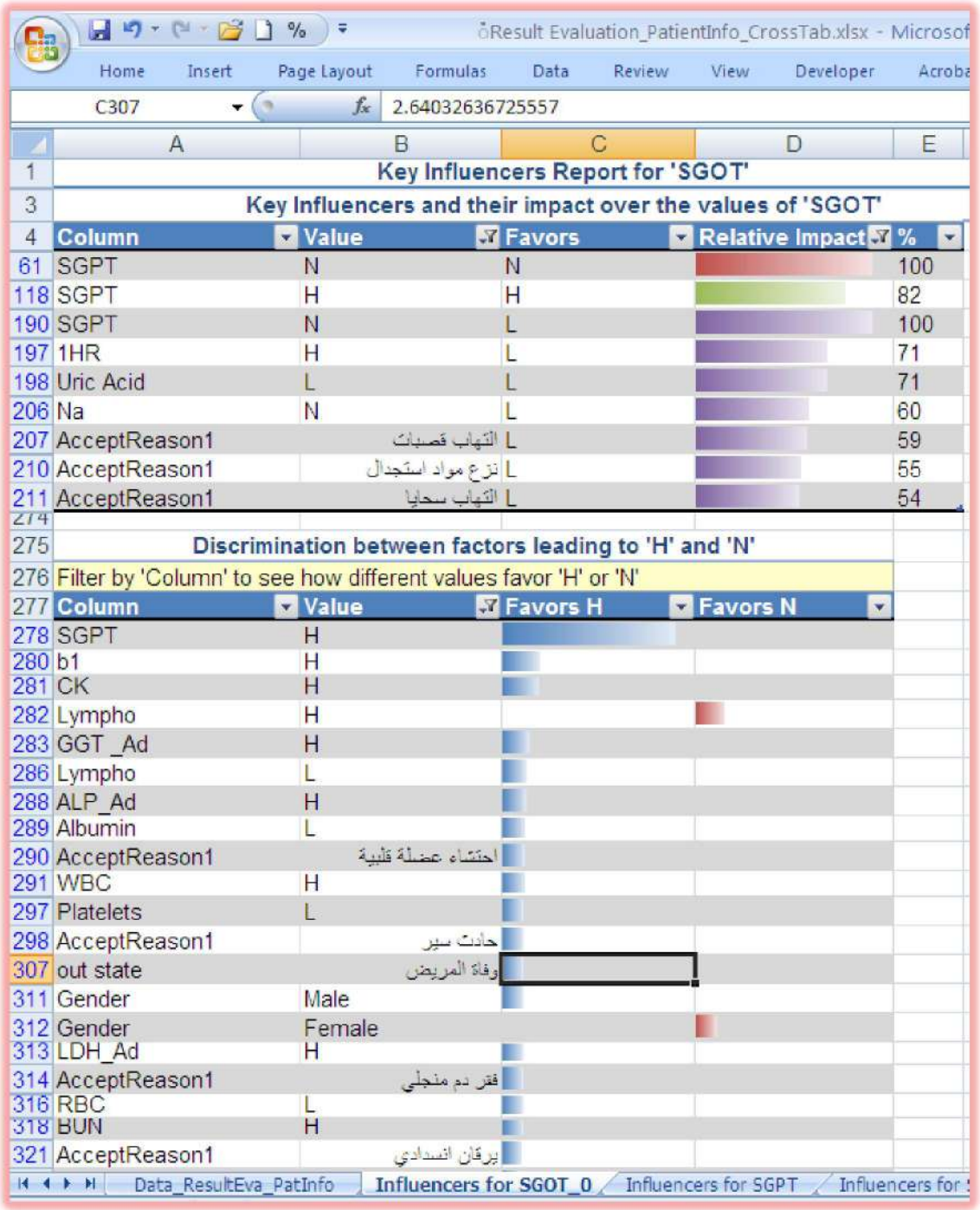
الأمور إلى المزيد من البيانات عن المرضى مثل الوزن والطول وضغط الدم ونتائج الصور الشعاعية والحالة السريرية والأمراض والعمليات السابقة..الخ. وبناءً على ذلك تعتبر أداة التنبؤ بالوفاة هذه كتصميم أولي أو نموذجي لأداة يمكن تطويرها مستقبلاً بإضافة خاصيات طبية وشخصية أكثر كماً ونوعاً إليها، بحيث يتمكن الأطباء من استخدامها في التعرف على مقدار مساهمة العوامل المختلفة في توقع شفاء أو وفاة المريض عند خروجه من المشفى، ويمكن بالتالي اطلاع المريض أو ذويه بمدى الخطورة المتوقعة له، كما ويمكن أن يُستخدم مثل هذا النوع من الأدوات في اتخاذ بعض القرارات مثل اتخاذ قرار إجراء عمل جراحي معين أو تطبيق أسلوب علاجي محدد في ضوء التوقعات المحسوبة لكل مريض بحسب خاصياته.

6•6•5 اكتشاف المتغيرات الطبية الأكثر تأثيراً على تغيُّر نتائج أحد التحاليل الطبية

يُمكن تطبيق هذه المسألة على أي تحليل طبي، وكمثال على ذلك جرى تطبيق هذه الفكرة على التحليل الطبي لأنظيم (SGOT)[1]، حيث تقول الدراسات الطبية بأن هذا الأنظيم يزداد في عدد من الحالات مثل احتشاء العضلة القلبية الحاد والرضوض العضلية والتهاب الكبد الانتاني والتهاب البنكرياس، ويترافق ارتفاع هذا الأنظيم مع ارتفاع في أنظيم آخر هو (SGPT) ما عدا بعض الحالات مثل احتشاء العضلة القلبية. وقد اخترنا أن ندرس العوامل المختلفة المؤثرة في نتائج هذا التحليل باستخدام أداة التنقيب (Analyze Key Influence)، فتبين أن ارتفاع نتيجة أنظيم (SGOT) تأثر بنسبة 82% بارتفاع نتيجة تحليل أنظيم (SGPT)، وأن اعتدال أو انخفاض نتيجة أنظيم (SGOT) تحقق معه اعتدال نتيجة (SGPT) بنسبة 100%. كما وأظهر تقرير المقارنة بين المؤثرات في النتيجتين المعتدلة والمرتفعة لتحليل (SGOT) أن احتشاء العضلة القلبية يؤثر بنسبة 4% تقريباً في ارتفاع نتيجة هذا التحليل، وأن وفاة المرضى يؤثر بنسبة 2.6% في ارتفاع نتيجة التحليل (أي ارتبطت بعض حالات الوفاة مع ارتفاع هذا الأنظيم وليس انخفاضه)، وأن جنس المريض الذكر يؤثر في ارتفاع نتيجة التحليل بنسبة

[1] أنظيم أسبرتات ترانسفراز (Aspartate Trasferase) يدل على الوظيفة الكبدية [د. سيد الحديدي، 2006، **دلالات وتفسير النتائج المخبرية**، ص 31، 118]

2.5%، بينما يؤثر جنس المريض الأنثى في اعتدال النتيجة بنسبة 2.5% أيضاً، انظر الشكل (5-6-17).



Result Evaluation_PatientInfo_CrossTab.xlsx - Microsof

Home | Insert | Page Layout | Formulas | Data | Review | View | Developer | Acroba

C307 | 2.64032636725557

Key Influencers Report for 'SGOT'

Key Influencers and their impact over the values of 'SGOT'

	Column	Value	Favors	Relative Impact	%
61	SGPT	N	N		100
118	SGPT	H	H		82
190	SGPT	N	L		100
197	1HR	H	L		71
198	Uric Acid	L	L		71
206	Na	N	L		60
207	AcceptReason1	التهاب قصبات	L		59
210	AcceptReason1	نزع مواد استجدال	L		55
211	AcceptReason1	التهاب سحايا	L		54

Discrimination between factors leading to 'H' and 'N'

Filter by 'Column' to see how different values favor 'H' or 'N'

	Column	Value	Favors H	Favors N
278	SGPT	H		
280	b1	H		
281	CK	H		
282	Lympho	H		
283	GGT _Ad	H		
286	Lympho	L		
288	ALP_Ad	H		
289	Albumin	L		
290	AcceptReason1	احتشاء عضلة قلبية		
291	WBC	H		
297	Platelets	L		
298	AcceptReason1	حادث سير		
307	out state	وفاة المريض		
311	Gender	Male		
312	Gender	Female		
313	LDH_Ad	H		
314	AcceptReason1	فقر دم منجلي		
316	RBC	L		
318	BUN	H		
321	AcceptReason1	برقان انسدادي		

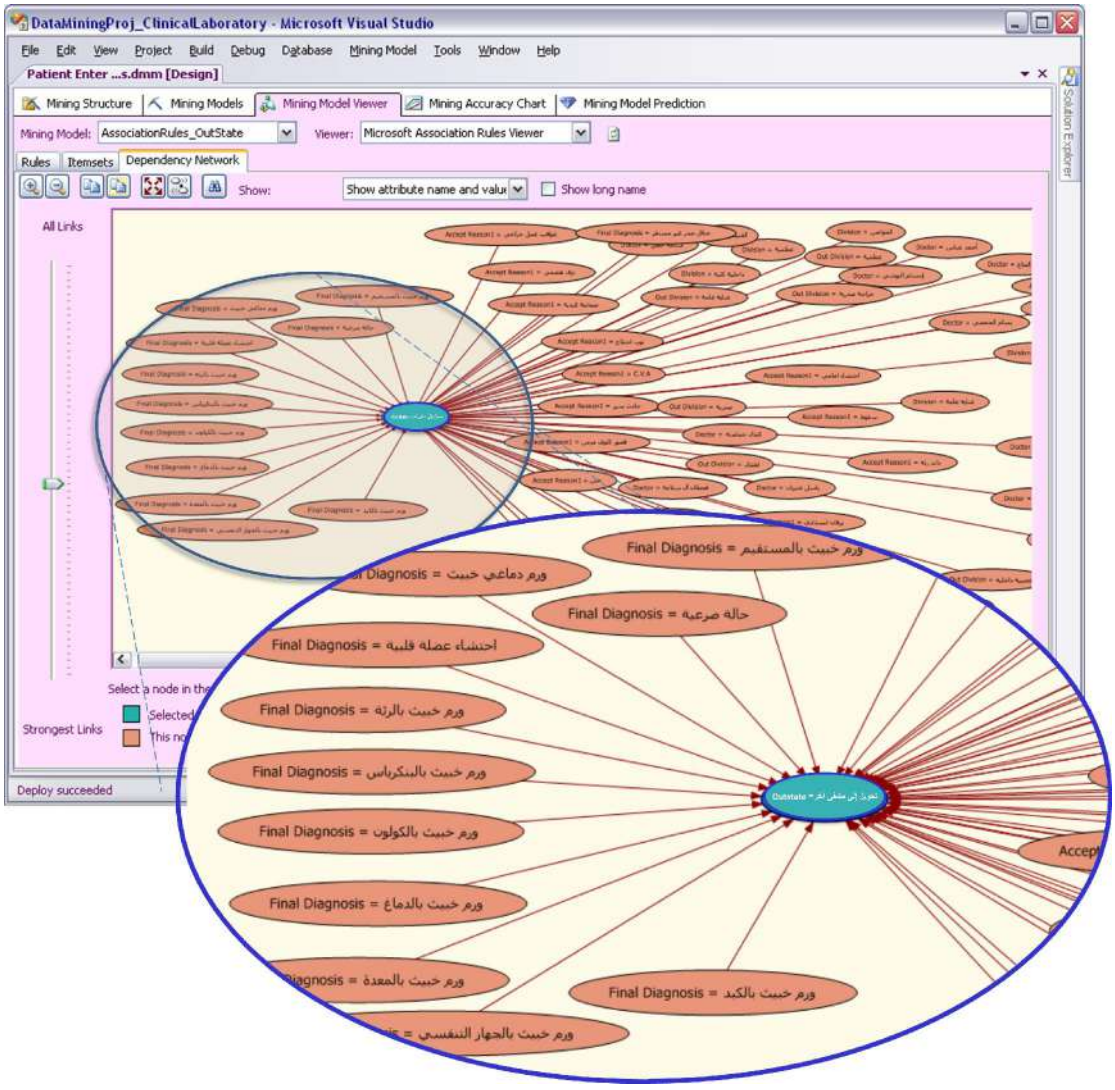
Data_ResultEva_PatInfo | Influencers for SGOT_0 | Influencers for SGPT | Influencers for

الشكل (5-6- 17): تحليل العوامل المؤثرة على ارتفاع وانخفاض نتيجة التحليل الطبي (SGOT)
باستخدام أداة (Analyze Key Influence)

7•6•5 استخدام تقانة التنقيب في البيانات في البحث عن الأسباب الكامنة وراء تحول المرضى إلى مشفى آخر

يُعتبر التعرف على صفات الزبائن الذين يتحولون إلى المنافسين من أهم استخدام تقانة اكتشاف المعرفة في عالم الأعمال، ويعتبر المريض من وجهة نظر إدارة المشفى على أنه زبون يتلقى خدمات من منظمتهم، وتهتم إدارة المشفى عادةً بالحفاظ على زبائنها وتجنب تحولهم إلى مشفى آخر منافس بسبب النقص أو التقصير في خدماتها المقدمة. وربما يصعب على تلك الإدارة وخصوصاً في المشافي الكبيرة أن تقيِّم جودة الخدمات التي تقدم في أقسامها، وتتعرف على أماكن الخلل والتقصير التي تجعل المرضى ينتقلون إلى المشافي الأخرى... وبالتالي يمكن مساعدتهم من خلال تقانة اكتشاف المعرفة، ولكن يحتاج تطبيق هذه التقانة على أرض الواقع إلى الكثير من البيانات عن المرضى وعن المشفى. وسيتم فيما يلي تقديم نموذج مبسط عن هذه المسألة وباستخدام البيانات المتاحة للدراسة.

اعتمدت عملية اكتشاف أسباب تحويل المرضى إلى المشافي الأخرى على البيانات الشخصية والمرضية للمرضى، حيث تم استخدام نماذج التنقيب الموجودة في مشروع التنقيب الأول التي تتنبأ بحالة خروج المرضى من المشفى. ولتحقيق هدف المسألة الحالية فقد جرى التركيز على انتقاء ومقارنة حالة شفاء أو تحسن المرضى مع حالة تحويل المرضى إلى المشافي الأخرى. وباختصار تبين من خلال استعراض مصفوفة التصنيف أن نموذج التنقيب المستند إلى خوارزمية قواعد الاقتران هو أفضل نموذج في وصف حالة التحويل إلى مشفى آخر. وتم لذلك الاعتماد على مخرجات هذا النموذج، والتي قَدَّمت بالفعل أوضح وأفضل النتائج، حيث أظهرت شبكة التبعية الخاصة بهذا النموذج أن معظم الأسباب الكامنة وراء تحويل المرضى إلى المشافي الأخرى تقترن مع التشخيص النهائي للمرضى بوجود أورام خبيثة بأنحاء متفرقة من الجسم. وبناءً على ذلك يمكن الاستنتاج بأن التحويل كان معظمه إلى مشفى مختص بعلاج الأورام الخبيثة، وهو سبب منطقي وطبيعي، انظر الشكل (5–6–18). وبالتالي فإن الأسباب الأساسية للتحول إلى المشافي الأخرى لم تكن أسباباً تنافسية بالنسبة للمشفى المدروس.



الشكل (5-6- 18): شبكة التبعية في نموذج قواعد الاقتران (AssociationRules_OutState) الخاص بالتنبؤ بحالة خروج المرضى من المشفى مع التأكيد على حالة تحويل المرضى إلى مشفى آخر

عرضنا فيما سبق لبعض الاستخدامات الممكنة للمعارف المكتشفة من قواعد البيانات المدروسة، وإن المجال واسعٌ جداً للمزيد من الاستخدامات. ويمكن أن تتسع مستقبلاً دائرة البيانات إلى الخارج لتشمل مزيداً من الجهات التي تقدم خدمات طبية للمرضى، وأن تتسع داخلياً بزيادة عدد الخاصيات أو بالتركيز على مسائل معينة أو فئات معينة من المرضى.

الخاتمة

النتـــائـج

والمقترحات والأعمال المستقبلية

الخاتمة

قدَّم هذا البحث تجربةً عمليةً في تطبيق تقانة اكتشاف المعرفة من قواعد البيانات على بيانات حقيقية مستمدة من الواقع المحلي. وقد تم بداية التعرف على المفاهيم المحيطة بهذه التقانة، وعلى أبرز مجالات تطبيقها في العالم، ثم جرى عرض الإطار النظري العام لمراحل عملية اكتشاف المعرفة الذي تم الاسترشاد به أثناء تنفيذ هذه العملية على بيانات مخبر التحاليل الطبية المدروس، كما وتم التعرف على الحقل المدروس كونه مصدراً للبيانات المدروسة ومجالاً علمياً يُستعان به في انجاز الأبحاث الطبية المتنوعة.

النتـــائج

سعى البحث الحالي إلى تحقيق عدد من الأهداف كما سبق وذكرنا في الفصل التمهيدي، وانطلاقاً من هذه الأهداف ومما تم تحقيقه في هذا البحث سنقدم فيما يلي عرضاً تلخيصياً لأبرز الأفكار والأعمال والنتائج التي أتى بها هذا البحث والتي تعكس الدور الذي تستطيع أن تؤديه تقانة اكتشاف المعرفة في مجال دراسة البيانات بشكل عام، والدور الخاص الذي أدته في مجال دراسة بيانات مخابر التحاليل الطبية المتاحة.

1- قدَّم هذا البحث تعريفاً بالمكونات النظرية والتطبيقية لتقانة اكتشاف المعرفة من قواعد البيانات، وهي تُعتبر من الحقول البحثية الحديثة التي قلَّت الكتابة عنها باللغة العربية.

2- تم التعريف بالأعمال المختلفة التي تؤديها تقنية التنقيب في مجالات متنوعة من الحياة، حيث حققت هذه التقنية تطويراً كبيراً لأساليب العمل، وخصوصاً في المنظمات التي تُعنى بالزبائن وتحرص على استمرار رضاهم وولاءهم لها. كما قدَّمت العديد من الخدمات المفيدة للإنسانية والعلم، وخصوصاً عند استخدامها في البحوث الطبية والجينية. وقد تعدَّت أيضاً خدماتُ هذه التقنية الأرضَ التي نعيش عليها، ووصلت إلى اكتشاف أقاصي الفضاء.

3- تأكدت أهمية وجود نظام معلومات طبي يحتوي بيانات مخبر التحاليل الطبية والبيانات الأخرى للمرضى، عندما احتاجت هذه الدراسة إلى بيانات شخصية ومرضية للمرضى تتوفر حقولها في البرنامج الحاسوبي المُستخدم في المخبر، ولكن جرى إهمال توثيقها أثناء الاستخدام الفعلي للبرنامج بسبب أن الهدف الأساسي من استخدام هذا البرنامج في المخبر المدروس هو لطباعة نُسخ ورقية عن نتائج التحاليل الطبية للمرضى. وتم لذلك تسليط الضوء في الفصل الرابع من الدراسة على المفاهيم المتعلقة بنظم المعلومات والتقانات الحديثة التي ترعى التخزين والتبادل الالكتروني للبيانات الطبية.

4- اهتم البحث بعرض فكرة عامة عن الأبحاث العلمية المنجزة في حقل التحاليل الطبية، وذلك لبيان موقع البحث الحالي منها. وقد تم الاقتراح في هذا البحث لتقسيم تلك الأبحاث إلى مجموعتين أساسيتين الأولى هي الأبحاث المخبرية الطبية التقليدية التي تُنجز بالاعتماد على سحب عينات إحصائية من الأشخاص أو المرضى، وتُجمع عنهم بيانات موافقة لفرضيات البحث وبعيداً عن البيانات المخزنة مسبقاً في الحواسيب. أما المجموعة الثانية من الأبحاث فتترافق مع التطورات التقنية المحققة في مجال مخابر التحاليل الطبية، حيث تؤدي التطورات في الأجهزة المستخدمة في المخابر إلى ظهور أنواع جديدة من الأبحاث تسعى لاختبار طرق جديدة لإجراء التحاليل أو تختبر أنواعاً جديدة من التحاليل.. كما أدى الاعتماد على تخزين بيانات مخابر التحاليل الطبية في الحواسيب إلى ظهور أبحاث طبية مخبرية تعتمد على وجود فرضيات ولكنها تُختبر بعد مرور فترة طويلة من الزمن. وأخيراً ساعد وجود تقانة اكتشاف المعرفة على إجراء نوع جديد من الأبحاث الطبية المخبرية تكشف المجهول في البيانات الحقيقية المخزنة في الحواسيب دون وجود فرضيات مسبقة للبحث ولكن بوجود مسائل مطروحة للتنقيب.

5- أثبت هذا البحث أن بيانات مخبر التحاليل الطبية المتاحة (البيانات الشخصية والمرضية ونتائج التحاليل الطبية للمرضى) قابلة للتحول إلى معارف مفيدة مكتشفة، وقد كانت هذه هي الفكرة الأساسية التي تمحورت حولها مشكلة البحث التي طُرحت في بداية العمل. وقد بقيت هذه النتيجة مغمضة طويلاً إلى أن وصلت رحلة البيانات المحفوفة بالمشقات والصعوبات إلى محطة استعراض نماذج التنقيب.

6- جرى تسليط الضوء على أبرز مشاكل البيانات المحلية، حيث ذخرت قواعد البيانات المدروسة بالأخطاء التي يعود معظمها إلى التصاميم المرنة لتلك القواعد، فهي تعتمد على تعميم نمط الحقل النصي على معظم حقولها مهما تكن طبيعة البيانات، ولا تحتوي على قيود كافية للمستخدمين عند إدخال البيانات إليها، ولا يوجد فيها مخططات كيان-علاقة جيدة..الخ. ولا نقصد توجيه اللوم فقط على مصممي البرامج الحاسوبية التي ترفد هذه القواعد بالبيانات، وإنما قد يضطر المصمم في كثير من الأحيان نزولاً عند رغبة المستخدمين إلى بناء تصاميم مرنة تنال رضاهم وراحتهم، وتسمح لهم بالكتابة بأية طريقة يختارونها، وخصوصاً أن الهدف الأساسي لمثل هذه الأنظمة هو التوثيق الالكتروني والطباعة الورقية لنتائج البيانات المخزنة.

7- احتوت قاعدة البيانات مكتب القبول المدروسة على عدد من الترادفات في المصطلحات الطبية المعبرة عن الأمراض، وأدى ذلك إلى زيادة وهمية في تنوع الأسباب الطبية المؤدية إلى دخول المشفى، وقد بُذل الكثير من الجهد لتوحيد طرق التعبير عن الأمراض أثناء مرحلة تنظيف البيانات، وتم تخفيض أعداد هذه الأسباب إلى الثلث تقريباً.

8- احتاجت قواعد البيانات المدروسة إلى الكثير من الجهد والوقت لإصلاحها و تحويلها إلى قواعد بيانات صالحة لإجراء عمليات التنقيب فيها، وقد اتجهت عمليات الإصلاح نحو تحقيق ثلاثة غايات على التوازي، هي تنظيف البيانات، وإعادة تنسيقها، واشتقاق حقول وجداول بيانات جديدة اعتماداً على محتويات الجداول الأساسية.

9- تم دمج قاعدة بيانات مخبر التحاليل الطبية وقاعدة بيانات السجلات الطبية وقاعدة بيانات مكتب القبول في قاعدة بيانات جديدة نظيفة مكونة من الجداول المفيدة في عمليات التحليل والتنقيب، وقد جرى تصميم مخطط كيان-علاقة لتمثيل العلاقات المنطقية بين تلك الجداول.

10- تم التوصل في هذا البحث إلى عدد من الحلول للمشاكل التي عانت منها البيانات المدروسة، والتي ظهر معظمها أثناء التطبيق العملي، ونذكر منها:

أ- تحويل طريقة عرض نتائج التحاليل الطبية الرئيسة إلى نتائج فرعية مقابلة لكل تحليل طبي فرعي.

ب- تحويل جميع نتائج التحاليل الطبية إلى قيم رقمية، بعد أن تنوعت طرق التعبير عن هذه النتائج، وكانت تُخزَّن في حقل من نمط نصي.

ج- تعريف القيم الطبيعية لجميع التحاليل الطبية الفرعية، وكذلك التحاليل الطبية التي كان يُعبَّر عن نتائجها بالرموز أو العبارات النصية.

د- توحيد طريقة تقييم جميع نتائج التحاليل الطبية مهما اختلفت من خلال استبدال جميع النتائج الرقمية إلى ما يناسبها من تقييمات (مرتفع عن الحد الطبيعي، طبيعي، أقل من الحد الطبيعي) مقارنةً بالقيم الطبيعية التي تختلف باختلاف اسم التحليل الطبي وجنس وعمر المريض.

ه- استخدام رقم اضبارة المرضى لربط جداول قواعد البيانات الثلاثة المستخدمة في هذا البحث، حيث أن كلاً منها كانت تعمل بصورة مستقلة تماماً عن غيرها، ولا يوجد أي رابط يجمعها معاً.

و- تحويل أسطر نتائج التحاليل الطبية إلى أعمدة، بسبب ظهور الحاجة أثناء تنفيذ بعض عمليات التنقيب في البيانات إلى تخصيص حقل لكل تحليل من التحاليل الطبية بدلاً من الحقل الذي يجمع بداخله كافة نتائج التحاليل الطبية، وذلك من أجل التعرف على التأثير المتبادل لنتائج كل من التحاليل الطبية مع بعضها ومع قيم الخاصيات المختلفة للمرضى، وتوقع نتائج بعض التحاليل الطبية اعتماداً على نتائج التحاليل الأخرى، وعلى الصفات المختلفة للمرضى.

11- تم بناء مستودع جديد للبيانات مشتق من بيانات قواعد البيانات الثلاثة المتاحة للدراسة، وقد جرى تطويره عدة مرات بنتيجة ظهور احتياجات جديدة من البيانات لخدمة عمليات التنقيب.

12- جرى تصميم مكعب للبيانات له بنية مخطوطة مجموعة النجوم، أي تتعدد فيه وتتداخل جداول الحقيقة والأبعاد، وتم تزويده بالعديد من المقاييس الجديدة التي ظهرت الحاجة إليها أثناء تطبيق عمليات المعالجة التحليلية الآنية على هذا المكعب.

13- ساعد استخدام تقنية المعالجة التحليلية الآنية على التعرف بعمق على البيانات المدروسة، والتوصل إلى الكثير من المعلومات التي يمكن أن يُستفاد منها مباشرةً أو بعد

إخضاعها للمعالجات الإحصائية. وقد تم باختصار التعرف من خلال هذه التقنية على التقلبات الزمنية في إيرادات (أعداد) التحاليل الطبية، وجرى استعراض بيانات المكعب بحسب البُنى الهرمية للأبعاد التي يتكون منها مثل البُنى الجغرافية لمكان إقامة المرضى وبُنى تصنيفات التحاليل الطبية، وبحسب الخاصيات المختلفة للمرضى (العمر، الجنس، الأسباب المرضية للقبول في المشفى).. وكان لكل من تلك الاستعلامات فوائده التي ظهرت في متن البحث.

14- ساعدت تقنية المعالجة التحليلية الآنية على التعرف على بعض المسائل التي يمكن حلها باستخدام تقانة التنقيب في البيانات مثل التنبؤ بالإيراد الشهري للتحاليل الطبية في العام الذي تلى الأعوام المدروسة.

15- جرت الاستفادة في هذا البحث من الاستخدامات العظيمة الكامنة في برنامج اكسل، وذلك في إجراء عدد من عمليات المعالجة التحليلية الآنية، وفي تنفيذ بعض عمليات التنقيب في البيانات.

16- لم يكن سهلاً وسريعاً وضع مسائل اكتشاف معرفة في ضوء البيانات المتاحة للدراسة، وإنما تم بدايةً وضع خطوط عريضة للمسائل الممكن طرحها وحلها، ثم تمت صياغة المسائل توازياً مع التطورات المحققة في كل مرحلة من مراحل اكتشاف المعرفة.

17- أنجزت تقنية التنقيب في البيانات الطبية المدروسة النوع الأول من مهام التنقيب، وهو التنقيب الوصفي، حيث قامت بعض نماذج التنقيب التي اعتمدت على خوارزمية التجزئة العنقودية بتقسيم بيانات المرضى إلى فئات تحمل عناصر كل منها صفات متشابهة (شخصية ومرضية ونتائج التحاليل الطبية). وتعتبر هذه التقسيمات معرفة مكتشفة من قواعد البيانات، وتفيد في التعرف على الصفات المشتركة التي تجمع المرضى المنتمين إلى كل فئة من الفئات، فمثلاً تتصف فئة المرضى الذين يُصابون بأمراض الكلية (وخصوصاً القصور الكلوي) أو القلب (خصوصاً قصور القلب) بأن أعمارهم تزيد غالباً عن 51 سنة، وغالباً ما ترتفع لديهم نتائج بعض التحاليل الطبية مثل البولة النتروجينية والكرياتينين، وتنخفض أخرى مثل الخضاب وتعداد كريات الدم الحمراء والهيماتوكريت. كما قامت نماذج تنقيب أخرى تعتمد أيضاً على خوارزمية التجزئة

العنقودية باستخراج العوامل المؤثرة على التعرض لبعض الأمراض، وأظهرت المعرفة المكتشفة منها أن بعض الأمراض تتأثر بالمكان الجغرافي لإقامة المريض، فمثلاً تبيَّن أن الأورام التي تحصل للدماغ تتأثر بالمنطقة الجغرافية الشمالية الشرقية، وأن بعض الأمراض تحدث في أعمار معينة مثل التهاب الكبد الذي ترافق بشدة مع الفئة العمرية بين 15–32 سنة، بينما ترافق تشمع الكبد مع الأعمار بين 51–66 سنة. كما تبين أن بعض الأمراض تتأثر بالجنس مثل التهاب الكبد الذي يتأثر بجنس الذكور بينما يتأثر تشمع الكبد بجنس الإناث.

18- أنجزت تقنية التنقيب في البيانات الطبية المدروسة النوع الثاني من مهام التنقيب، وهو التنقيب التنبؤي، حيث استُخدمت أداة اكسل (Forecast) للتنبؤ بالإيراد الشهري (أعداد التحاليل الطبية) لمخبر التحاليل الطبية خلال عام 2009، وتبين أنه سيبلغ ذروته في شهر تموز، وتساعد هذه المعرفة في عمليات التخطيط المالي والإداري للمخبر. كما جرى العثور على نماذج جيدة للتنبؤ بنتائج أحد التحاليل الطبية اعتماداً على نتائج التحاليل الأخرى، مثل النموذج المعتمد على خوارزمية أشجار القرار. وتم أيضاً اقتراح أداة للتنبؤ بالأشخاص المعرضين للوفاة استناداً إلى قيم خاصياتهم الشخصية والمرضية ونتائج التحاليل الطبية، وتعتمد على خوارزمية الانحدار المنطقي، حيث تساعد هذه الأداة في التعرف على عوامل الخطورة عند كل مريض، وتفيد في اطلاعه أو اطلاع ذويه على ذلك، وأيضاً اتخاذ التدابير الطبية اللازمة لإنقاذ حياته بالوقت المناسب، واتخاذ القرار الطبي الحكيم والأنسب لكل مريض مثل إجراء أو عدم إجراء عمل جراحي معين له.

19- إن القرارات الطبية حساسة ويترتب عليها المصير الصحي للمريض، ولذلك يجب توخي الحذر عند اختيار نموذج التنقيب الأنسب في دراسة خاصيات المريض، ولابد من تجريب النماذج كثيراً قبل الاعتماد على أي منها. كما ويمكن اعتماد مخرجات عدة نماذج، فإذا كانت جميعها تدعم نتيجة ما، فعندها تتأكد صحة هذه النتيجة.

20- ساعدت نتائج نماذج التنقيب أحياناً في هذا البحث في التحقق من جودة البيانات المدخلة إلى أنظمة عمل قواعد البيانات، والتي لم يتم التمكن من تنظيفها أثناء مرحلة تهيئة

البيانات، فمثلاً لم يكن بالإمكان تنظيف حقل التصنيف الدولي لأسباب الوفاة بسبب عدم وجود أي مرجعيات في قاعدة بيانات السجلات الطبية عن مضمون هذا الحقل، وقد كانت بيانات هذا الحقل ثمرة الجهد الفردي للعاملين في هذا القسم، فاتضح أثناء تصفح نماذج التنقيب في المشروع الأول أن التصنيف الدولي لبعض أسباب الوفاة كان يتوافق مع الخاصيات الأخرى للمرضى مثل شعبة القبول وسبب القبول وسبب الوفاة.. مما يدل على صحة تلك البيانات التي جرى استعراضها.

21- كشفت بعض أدوات التنقيب وجود عيوب في البيانات، على الرغم من خضوعها لعمليات التنظيف، فمثلاً تبين أثناء تتبع تتالي الأسباب المرضية المؤدية إلى دخول المرضى إلى المشفى أنه بالرغم من الانجازات الكبيرة المحققة في توحيد العبارات المتشابهة في حقل الأسباب المرضية، فلا يزال هنالك بعض التعبيرات المترادفة أو المتداخلة عن الأسباب المرضية التي يصعب التخلص منها جميعاً في مرحلة التنظيف مثل (نزع مواد استجدال، نزع صفائح، نزع أسياخ، استخراج جسم أجنبي..)، وبذلك يُعتبر دور أدوات التنقيب ايجابياً في التعرف على مثل تلك العيوب والأخطاء في البيانات والتمكن مستقبلاً من التخلص منها.

22- تمكَّن نموذج الشبكات العصبونية من دراسة العلاقات البينية بين القيم غير المفقودة في جدول التحاليل الطبية، حيث أنه تجاهل تأثير القيم المفقودة في المدخلات (Inputs) على قيم التحاليل الطبية المتنبأ بها، مما أدى إلى إظهار التأثير الحقيقي لنتائج بعض التحاليل الطبية في هذا النموذج على خلاف غيره من نماذج مشروع التنقيب الثاني مثل نموذج نايف بايز. كما أن هذا النموذج يُظهر تأثير القيم قليلة التكرار في التحاليل الطبية على نتائج التحليل الطبي المتنبأ به. وبناءً على ذلك نجد من الناحية التطبيقية أن نموذج الشبكات العصبونية قد يكون أكثر توافقاً من النماذج الأخرى مع طبيعة بيانات التحاليل الطبية التي تمتلئ طبيعياً بالقيم المفقودة المقابلة للتحاليل الطبية التي لم تُجرى للمريض.

23- قدَّم مخططُ الارتقاء المخصص فوائدَ كبيرةً لدراسة القدرة الحقيقية لنماذج التنقيب على التنبؤ بمتغير محدد، وقد برزت فائدته الكبيرة عندما استُخدم لاختبار قدرة النماذج على التنبؤ بالفئات ذات الأحجام الصغيرة، فمثلاً انخفضت دقة التنقيب عندما تم تخصيص

مخطط الارتقاء للتنبؤ بحالة الوفاة بدلاً من التنبؤ بجميع حالات الخروج من المشفى، وذلك أن المخطط العام كان يتحيز للفئة الأكبر وهي فئة الخروج بحالة شفاء، وكانت بالتالي دقة التنقيب فيه مرتفعة بصورة متوهة، ولكنها اتضحت بدقة عندما جرى تخصيصه. كما يُساعد مخطط الارتقاء المخصص للتنبؤ بحالة الوفاة في التعرف على الأشخاص المعرضين لخطر الوفاة انطلاقاً من بياناتهم الشخصية والمرضية، وبالتالي يمكن اختيار نسبة من هؤلاء وتوجيه العناية المركزة المناسبة المجانية لهم وإبعادهم (إن شاءت الأقدار) عن خطر الموت.

24- تم في هذه الدراسة اتخاذ قرار اختيار نموذج التنقيب الأنسب للتنبؤ بحالة خروج المرضى من المشفى بعد الأخذ بعين الاعتبار أن حالة الوفاة هي حالة حرجة، وقد تم اختيار النموذج المستند إلى خوارزمية نايف بايز لأنه النموذج الذي قلت فيه أخطاء التنبؤ بشفاء الأشخاص المعرضين فعلاً للوفاة، إضافة إلى أنه كان النموذج الأفضل بحسب تقييمات مخطط الارتقاء. وبهذا نستنتج أن قرار اختيار نموذج التنقيب الأفضل يجب أن يُبنى بالترافق بين الاستيعاب العميق لأهداف مسألة اكتشاف المعرفة المطروحة، وبين معنى النتائج المتولدة عن النماذج، وبين نتائج أدوات التقييم الحسابي لقدرات النماذج.

25- يَحملُ التنقيبُ في تتبع تتالي الأمراض الكثيرَ من الفوائد بالنسبة للأطباء، حيث تساعدُ معرفةُ الاحتمالات المتوقعة للإصابة بأمراض معينة -بعد التعرض لمرض ما- الأطباءَ على اتخاذ الإجراءات الوقائية أو حتى الاستعداد المسبق ومعرفة كيفية مواجهة تطورات المرض إن حدثت. كما تُساعد النتائج المكتشفة أيضاً شركات التأمين في تقدير التكاليف الطبية التي ستترتب مستقبلاً على الأمراض الحالية أو القديمة للمريض الراغب بالحصول على تأمين صحي.

26- ظهر في هذه الدراسة العديد من النتائج البديهية مثل علاقات التأثير المتبادل بين نتائج التحاليل الطبية، ويُعتبر ظهور مثل هذه النتائج في البيانات المدروسة من الاكتشافات التي تدعو إلى التفاؤل بقدرات تقانة التنقيب في البيانات المدروسة على اكتشاف نتائج أو علاقات جديدة وغير مكتشفة سابقاً طالما أنها تدعم البديهيات. وقد ساعدت هذه

الاكتشافات على زيادة الثقة بالمعارف الجديدة التي تم اكتشافها في البيانات المدروسة على الرغم من قلة عدد الخاصيات التي تمتلكها، وعلى الرغم من العيوب التي كانت تشيب البيانات الأساسية وجرى التخلص منها.

27- لعبت تقانة اكتشاف المعرفة من قواعد بيانات مخبر التحاليل الطبية دوراً جديداً وفعالاً في دراسة هذه البيانات، وقد استطاعت أن تحقق عدداً من الانجازات التي لا تستطيع أن تقوم بها أدوات الاستعلام من قواعد البيانات التي ورد ذكرها نظرياً في الفصل الأول وظهرت عملياً في أثناء عرض واجهات البرنامج الحاسوبي المستخدم في المخبر المدروس، وأثناء استيعاب البيانات المدروسة وتهيئتها. كما استطاعت هذه التقانة أن تقوم بإظهار جوانب في البيانات يصعب على الأساليب الإحصائية التقليدية القيام بها، وأهمها اكتشاف الصفات والعلاقات الخفية بين البيانات دون وجود فرضيات أو معرفة مسبقة عنها. كما أنها استطاعت أن تكشف عن وجود عدد من القضايا البحثية أو الفرضيات التي يُمكن اختبارها والتأكد منها باستخدام الأساليب الإحصائية. وتدَعَّمت بذلك الفكرة الأساسية التي انطلقنا منها عندما تحدثنا عن أهمية هذا البحث في الفصل التمهيدي.

المقتـــرحـــات والأعمال المستقبلية

نورد فيما يلي مجموعة من المقترحات والأعمال المستقبلية التي تكمِّل الخطوة التي بدأها هذا البحث في مجال اكتشاف المعرفة من قواعد البيانات المحلية:

1- قدم البحث الحالي نموذجاً عن استخدام تقانة اكتشاف المعرفة في الحقل الطبي، ويتطلب التطبيق المستقبلي لهذا النموذج على أرض الواقع أن تتوفر خاصيات أخرى كثيرة عن المرضى (شخصية ومرضية)، وأن تتمتع بجودة أعلى من تلك التي أتيحت لهذه الدراسة.

2- لن نطالب بتوجيه العناية لعمليات إدخال البيانات إلى الأنظمة الحاسوبية، ولن نقترح إدخال المزيد من الخاصيات مستقبلاً على قواعد البيانات التي أتيحت لهذه الدراسة. فهذه المطالب ضعيفة ومحدودة الطموح في العصر الحالي، وإنما هنالك حاجة قوية لبناء نظم جديدة للمعلومات في المراكز الصحية تنطلق من المعايير والتجارب العالمية

الناجحة، وتتمحور حول الحصول على سجلات طبية أو صحية كفوءة.

3- يجب الاهتمام على المستوى الوطني بتوحيد المصطلحات المطبقة في البرامج الحاسوبية، مثل استخدام طريقة موحدة في الترميز أو التعبير عن أسماء التحاليل الطبية، والالتزام بالتصنيف الدولي للأمراض وأسباب الوفاة..الخ. ويساهم ذلك في تحسين جودة البيانات والمعلومات والمعارف الناتجة عنها، وتسهيل الانتقال نحو تطبيق نظم المعلومات الطبية بدل الأنظمة الضعيفة الموجودة في المراكز الطبية الحالية، وتيسير الربط المستقبلي لقواعد البيانات الطبية الموجودة في الحواسيب معاً تحت جناح تقانة المعلومات الصحية.

4- من الضروري الاهتمام ببناء نظم قواعد بيانات محلية متينة، وغير ضعيفة، وأن تكون مرنة ولكن بطريقة منطقية وهادفة وليس بطريقة عشوائية.

5- ترافق انجاز هذا البحث مع توجه الاهتمام في سورية نحو بناء نظم معلومات طبية، ولذلك فإننا انطلاقاً من تجربتنا في التعامل مع البيانات الطبية في هذا البحث نتمنى أن تنجح هذه النظم، ونؤكد أيضاً على ضرورة أن يكون لتقانة اكتشاف المعرفة دورٌ فعال فيها منذ تأسيسها وليس بعد أن تمتلئ بالبيانات.

6- إن تطبيق تقانة اكتشاف المعرفة لا يزال محدوداً في بلادنا، ولكن بدأت هذه التقانة تلقى اهتماماً محلياً، حيث عُقدت في الشهر العاشر من عام 2010 الندوة الوطنية الأولى حول تطبيقات ذكاء الأعمال في سورية. وقد جرى في هذه الندوة طرح بعض التجارب المحلية لتطبيق تقانة التنقيب في البيانات، ومنها في مجال رفع جودة التعليم العالي في إحدى الجامعات الخاصة. كما جرى التأكيد في توصيات هذه الندوة على ضرورة استخدام ودعم مستخدمي هذه التقانة في عالم الأعمال وفي البحث العلمي المحلي.. ويشارك البحث الحالي الرأي والاهتمام الذي تبديه بلادنا بهذه التقانة، فهي فعلاً من التقانات التي من يتعرف عليها ويستخدمها يزداد قناعةً بأهميتها، ولا يستطيع أن يتخلى عنها.

7- قدمت تقانة اكتشاف المعرفة من قواعد البيانات المدروسة أداةً للتنبؤ بالأشخاص المعرضين للوفاة ولذلك يُقترح أن يتم إلحاق مثل هذه الأداة بنظم قواعد البيانات الطبية

في المشافي، بهدف انجاز التحليل الآلي المباشر لبيانات أي مريض وتوقع مدى الخطورة المحيطة به.

8- قد تستدعي بعض الاكتشافات التي تُنتجها تقانة اكتشاف المعرفة في مثل هذا البحث أن يتم إجراء المزيد من الدراسات للتحقق من صحتها وأن يُجمع عنها بيانات من عدة مصادر. فمثلاً إن اكتشاف وجود تأثير للمنطقة الجغرافية الشمالية الشرقية من سورية على التعرض لأورام الدماغ قد ينال الاهتمام من قبل المعنيين بهذا الأمر لكي يقوموا بالتحري عن العوامل الكامنة وراء انتشار مثل هذا المرض في تلك المنطقة، واتخاذ التدابير اللازمة للحد منه.

9- تُعتبر بعض المعارف المكتشفة على أنها فرضيات يمكن التحقق منها في دراسات مستقبلية باستخدام الاستدلال الإحصائي، مثل التأكد من صحة الفرضية التي تقول بأنه يوجد فارق بين الذكور والإناث في التعرض لبعض الأمراض مثل التهاب الكبد وتشمع الكبد؟.

10- يمكن مستقبلاً إجراء دراسات إحصائية على بيانات نتائج التحاليل الطبية لإيجاد علاقات رياضية بينها مثل العلاقة بين ارتفاع السكر الصيامي والكوليسترول، لاستخدامها أثناء مرحلة تهيئة البيانات في تعبئة القيم المفقودة عند الضرورة.

11- يمكن مستقبلاً تشكيل بنى هرمية جديدة للتحاليل الطبية، بحيث ترتبط التحاليل بالأمراض التي تحتاج إليها، وبالخاصيات المختلفة للمرضى. ويكون بالتالي إجراء التحاليل الطبية منطقياً ويحقق تخفيضاً في التكاليف الطبية، وذلك أن تقليص هذه التكاليف أصبح من المطالب التي تلقى اهتماماً عالمياً كبيراً كما سبق وذكرنا في الفصل الثاني.

12- اهتم البحث الحالي بتسليط الضوء على الإمكانيات التي تستطيع أن تقدمها تقانة اكتشاف المعرفة في دراسة البيانات الطبية، ولكن يحتاج التطبيق الحقيقي المستقبلي لهذه التقانة على بيانات طبية تتمتع بالكفاءة إلى الاستعانة برأي طبي (خارج إطار هذا البحث) لتوجيه هذه العمليات، واكتشاف المعارف الكامنة فيها من وجهة نظر طبية تعمل إلى جانب وجهات النظر الإحصائية والمعلوماتية.

قائمة المصطلحات

نبين فيما يلي قائمة بالمصطلحات المستخدمة في هذا البحث مرتبة أبجدياً بحسب التسمية الانكليزية

المصطلح الانكليزي	المصطلح العربي
Academic Research Models	نماذج البحث الأكاديمي
Aggregate	اندماج، تكامل
Algorithm Parameters	محددات الخوارزمية
Analyses Services	خدمات التحليل
Analyze Key Influencers Tool	أداة تحليل عوامل التأثير الرئيسة
Association Algorithms	خوارزميات الاقتران
Association Rules Algorithm	خوارزمية قواعد الاقتران
Attribute	خاصية
Attribute Profile	اللمحة المختصرة للخاصية
Barcode	رِمَازٌ قُضْباني
Business Intelligence (BI)	ذكاء الأعمال
Classification algorithms	خوارزميات التصنيف
Classification Matrix	مصفوفة التصنيف
Cluster Characteristics	صفات العقدة
Cluster Diagram	الرسم البياني العنقودي
Clustering Algorithm	خوارزمية التجزئة العنقودية
Clusters Discrimination	تمييز العقد
Confidence	الثقة وتساوي الاحتمال
Constellation Schema	مخطوطة مجموعة النجوم
Constructed data	بيانات تركيب
Cross Query	استعلام تصالبي
Cross-Industry Standard Process for Data Mining (CRISP-DM)	العملية القياسية للتنقيب في البيانات في الصناعة
Cube Structure	بنية المكعب
Data Cleaning	تنظيف البيانات
Data Cube	مكعب البيانات
Data Formatting	تنسيق البيانات
Data Mining	التنقيب في البيانات
Data Mining Algorithms	خوارزميات التنقيب في البيانات

المصطلح الانكليزي	المصطلح العربي
Data Mining Model	نموذج التنقيب في البيانات
Data Mining Prediction Tool	أداة تنبؤ نموذج التنقيب
Data Mining Structure	هيكل التنقيب في البيانات
Data Source	مصدر البيانات
Data Source View	مرأى مصدر البيانات
Data Warehouse	مستودع البيانات
Debugging	عملية التدقيق والبناء والتفعيل
Decision Trees Algorithm	خوارزمية أشجار القرار
Dependency Network	شبكة التبعية
Descriptive Data Mining	التنقيب الوصفي في البيانات
Detect Categories Tool	أداة اكتشاف الفئات
Diagram	رسم بياني
Dice	تقطيع إلى مكعبات أصغر
Dimension Table	جدول البعد
Drill-Down	نثر للأسفل في مكعب البيانات
Electronic Health Record (HER)	سجل الصحة الالكتروني
Electronic Medical Record (EMR)	السجل الطبي الالكتروني
Entity-Relationship Diagram (ERD)	نموذج الكيان–العلاقة
Facts	حقائق
Feed Back	التغذية الراجعة
Filter	انتقاء
Generic Content Tree View	مرأى شجرة المحتويات الشاملة
Health Information Technology (HIT)	تقانة المعلومات الصحية
Hierarchy	بنية هرمية
Hospital Information System (HIS)	نظام المعلومات الطبية في المشفى
Ideal Model	نموذج مثالي
Image Mining	التنقيب في الصور
Integration Services	خدمات التكامل
Knowledge Discovery in Databases	اكتشاف المعرفة من قواعد البيانات
Laboratory Information System (LIS)	نظام معلومات مخبر التحاليل الطبية
Lift Chart	مخطط الارتقاء

المصطلح الانكليزي	المصطلح العربي
Linear Regression Algorithm	خوارزمية الانحدار الخطي
Local Host	مخدم محلي
Logistic Regression Algorithm	خوارزمية الانحدار المنطقي
Machine Learning	تعليم الآلة
Market Basket Analysis	تحليل سلة التسوق
Measures	مقاييس
Metadata	البيانات المفسِّرة
Mining Accuracy Chart	مخطط دقة التنقيب
Mining Models	نماذج التنقيب
Missing Values	قيم مفقودة
Monitoring Mechanisms	آليات مراقبة
Naïve Bayes Algorithm	خوارزمية نايف بايز (الاحتمالات)
Natural Language Processing (NLP)	معالجة اللغة الطبيعية
Neural Networks Algorithm	خوارزمية الشبكات العصبونية
Neurons	خلايا عصبونية
Noisy Data	بيانات مشوهة
Online Analytical Processing (OLAP)	المعالجة التحليلية الآنية
On-Line Transaction Processing (OLTP)	معالجة المبادلات الآنية
Pattern	نمط
Pivot Table	جدول محوري
Pivot Table	مخطط محوري
Predictive Data Mining	التنقيب التنبؤي في البيانات
Profit Chart	مخطط المنفعة أو الربح
Random Guess Model	نموذج تخمين عشوائي
Regression Algorithms	خوارزميات الانحدار
Roll-Up	طي للأعلى في مكعب البيانات
Segmentation Algorithms	خوارزميات التقطيع
Sequence Analysis Algorithms	خوارزميات التحليل التتابعي
Sequence Clustering Algorithm	خوارزمية التجزيء العنقودي التتابعي
Sequential Pattern Mining	تنقيب الأنماط التتابعية
Six-Step Knowledge Discovery Process	طريقة المراحل الستة لاكتشاف المعرفة

المصطلح الانكليزي	المصطلح العربي
Slice	تقطيع شرائح في مكعب البيانات
Snowflake Schema	مخطوطة رقاقة الثلج
SQL Server 2008 Data Mining Add-ins	الوظائف المضافة للتنقيب في البيانات عبر مخدم لغة الاستفسار المهيكلة
Star Schema	المخطوطة النجمية
Structural Query Language (SQL)	لغة الاستعلام المهيكلة
Supermarket Loyalty Card	بطاقة الولاء المستخدمة في السوق المركزية
Support	الدعم وتساوي التكرار
Testing Data	بيانات الاختبار
Training Data	بيانات التجريب
Transaction	مناقلة، مبادلة، واقعة، معاملة
View	مرأى وجمعها مرائي
Visual Clustering	التجزئة العنقودية البصرية
Visualization	العرض المرئي
Web Usage Mining (WUM)	التنقيب في استخدام شبكة الانترنت



قائمة المراجع

أولاً. المراجع العربية

[1] إلياس، محمود. (2008)، **"أتمتة مخابر التحاليل الطبية في سورية بين الواقع والمأمول"**، مجلة المعلوماتية – العدد (28) – شهر حزيران 2008، الجمعية العلمية السورية للمعلوماتية، سوريا.

[2] بن سالم، إيمان & خلادي، محمد. (2008)، **"التنقيب في البيانات وعلاقته بالإحصاء: نظرة شاملة"**، جامعة منتوري قسنطينة، كلية علوم الهندسة، الجزائر.

[3] د. أبو سعدي، داود؛ د. أبو عسلي، عماد وآخرون. (2008)، الإحصاء الطبي، مركز تعريب العلوم الصحية، دولة الكويت.

[4] د. الحديدي، سيد. (2006)، **دلالات وتفسير النتائج المخبرية**، شعاع للنشر والعلوم، حلب، سوريا.

[5] د. الساري، اسماعيل. (2010)، **"استخدام اختبارات الرقاقات الحيوية النانوية المعتمدة على اللعاب في تشخيص احتشاء عضل القلب الحاد في وحدة العناية: دراسة الجدوى"**، مجلة التشخيص المخبري، المجلد 5، العدد 7 – نيسان (2010)، بحث مترجم من مجلة (Clinical Chemistry55(8): 1530-1538, 2009)، هيئة مخابر التحاليل الطبية في سوريا، سوريا.

[6] د. الكفري، عبير & أ.د. إسماعيل، محمد طاهر. (2008)، **"التراجع الهائل للطفيليات المعوية البشرية وبخاصة الديدان في دمشق وريفها– دراسة تشخيصية في الفترة بين عامي 1998–2007"**، مجلة التشخيص المخبري– المجلد5– العدد1– تشرين الأول 2008، هيئة مخابر التحاليل الطبية في سورية.

[7] د. بشور، نهيد. (2010)، **"اعتلال الكلية المزمن"**، مجلة التشخيص المخبري المجلد 5 ، العدد 7، ربيع الثاني 1431 – نيسان 2010، هيئة مخابر التحاليل الطبية في سوريا، سوريا.

[8] د. مقداد، ماهر. (2010)، **"اللعاب كسائل حيوي للتطبيقات التشخيصية للأمراض"**، كلية الصيدلة – جامعة حلب– حلب– سورية، المجلد 5 ، العدد 6 كانون الثاني 2010.

[9] د. مورا، سامية؛ نادر، رفاعي وآخرون. (2009)، "مقارنة تراكيز كوليستيرول الـ LDL المحسوبة رياضياً وفق معادلة فريدوالد Friedewald؛ مع المُقاسة مباشرة وارتباطها بحوادث القلب الوعائية CVD في 27331 امرأة" ترجمة د. جالا قاسم العتمة، مجلة التشخيص المخبري، المجلد 5 , العدد 5 – تشرين أول 2009، هيئة مخابر التحاليل الطبية في سوريا، سوريا.

[10] الرزو، حسن مظفر. (2008)، **"تطبيقات التنقيب المعلوماتي على موارد المعرفة الإسلامية"**، /http://www.alukah.net/Culture/10336/3123

[11] شاينين، مارتين. شباط (2009)، **"تعزيز وحماية جميع حقوق الإنسان، المدنية والسياسية والاقتصادية والاجتماعية والثقافية، بما في ذلك الحق في التنمية"** (تقرير المقرر الخاص المعني بتعزيز وحماية حقوق الإنسان والحريات الأساسية في سياق مكافحة الإرهاب) مجلس حقوق الإنسان، الجمعية العامة– الأمم المتحدة.

[12] فهمي، محمد شامل. (2005)، **الإحصاء بلا معاناة: المفاهيم مع التطبيقات باستخدام برنامج SPSS الجزء 1–2**، الرياض، السعودية.

[13] النقري، مهيب. (2007)، **"التنقيب في السجلات الطبية للتشخيص باستخدام الحاسوب"**، العدد (20) – شهر تشرين الأول 2007، مجلة المعلوماتية، الجمعية العلمية السورية للمعلوماتية، سوريا.

ثانياً. المراجع الأجنبية

[14] Adamo, J. (2001)**, Data Mining for Association Rules and Sequential Patterns**, Springer- Verlag New York, Inc.

[15] Baylis, P. (2009)**, "Better Healthcare with Data Mining",** Technical report, SPSS and IPM Company, Shared Medical Systems Limited, UK.

[16] Beastall, GH. (2008), **The Modernization of Pathology and Laboratory Medicine in The UK: Networking Into The Future**. Clin Biochem Rev 2008; 29: 3-10.

[17] Bossuyt, X. & Verweire, K. (2007), **"Laboratory Medicine: Challenges and Opportunities"**, American Association for Clinical Chemistry, Bossuyt et al. 53 (10): 1730. 2007 http://www.clinchem.org/cgi/reprint/53/10/1730.

[18] Bossuyt, X. & Verweire, K. (2007), **Laboratory Medicine: Challenges and Opportunities**, American Association for Clinical Chemistry, Inc.

[19] Bresfelean, Vasile P. (2008), **Data Mining Applications in Higher Education and Academic Intelligence Management**, Babes-Bolyai University, Faculty of Economics and Business Administration. Romania.

[20] Canadian Institutes of Health Research. (2002), **Secondary use of personal information in Health Research**, Ethics Office Publications, Canada.

[21] Chapman, P.; Khabaza, Th. & Shearer, C. (2000), **CRISP-DM: Step-by-Step Data Mining Guide, SPSS Inc, USA**.

[22] Cios, K.J.; Pedrycz, W.; Swiniarski, R. W. (2007), **Data Mining: A Knowledge Discovery Approach**, Springer Science & Business Media, USA.

[23] Ciriani, T. & Fasano, G. & Haydock, M. (2003), **Operations Research in Space and Air**, Kluwer Academic Publishers, Dordrecht, The Netherlands.

[24] Debuse, J. & Iglesia, B. (1999), **"A Methodology for Knowledge Discovery: The KDD Roadmap"**, University of East Anglia, School of information Systems Technical Report SYS-C99-01.

[25] Eapen, A. G. (2004), **Application of Data mining in Medical Applications**. Waterloo, Ontario, Canada. (thesis).

[26] Fayyad, U.; Piatetsky, G. & Smyth P. (1996), **"From data mining to knowledge discovery in databases"**, AI Magazine;17(3):37-54.

[27] Glymour, C. & Madigan, D. (1997), **Statistical Themes and Lessons for Data Mining**, Kluwer Academic Publisher, Netherlands.

[28] Goulbourne, G. (1998), **"Knowledge Discovery in Database",** University of Liverpool, Liverpool.

[29] Han, J. & KAmber, M. (2001), **Data Mining Concepts and Techniques,** Academic Press, USA.

[30] Hanna, M. (2004), **Data-mining algorithms in Oracle9i and Microsoft SQL Server**, Liverpool, UK.

[31] Harrisson, J. (2008), **Clinics in Laboratory Medicine**, Saunders, USA.

[32] Harrisson, J. (2008), **"Introduction to Mining of Clinical Data"**, University of Virginia, Elsevier Inc, USA.

[33] Hirano, Sh. & Tsumoto, Sh. (2007), **"Temporal Data Mining in Hospital Information Systems Analysis of Clinical Courses of Chronic Hepatitis",** Department of Medical Informatics, Shimane University, School of Medicine, 89-1 Enya-cho, Izumo, Shimane 693-8501, Japan. http://wacong.org/icmed/Volume1.Number1.files/2.pdf

[34] Hosking, J. & Pednault, E.& Sudan, M. (1997), **A Statistical Perspective on Data Mining**, IBM Research Division, Yorktown Heights.

[35] Inmon, B. (2010), **"Including Unstructured Text in the Data Warehouse"**, www.b-eye-network.com.

[36] Jacobson, R. (2006), **SQL Server (2005) Analysis Services**, Microsoft Press, USA.

[37] **Kazmierczak**, S. & Leen, T. (2007), **"Reduction of Multi-Dimensional Laboratory Data to a two-dimensional plot a novel technique for the identification of laboratory error"**, Clin Chem Lab Med (2007);45(6):749–752, New York.

[38] Le Duff, F. & Muntean, C. (2004), **Predicting Survival Causes After Out of Hospital Cardiac Arrest Using Data Mining Method**, MEDINFO, IOS Press, Amsterdam.

[39] Lehmann, Th.; Guld, M. & Others. (2005), **Automatic categorization of medical images for content-based retrieval and data mining**, Department of Medical Informatics, Medical Faculty, Aachen University of Technology, Germany, ELSEVIER.

[40] Leung, K. & Eddie, Y. (2004), **Data Mining on DNA Sequences of Hepatitis B Virus by Nonlinear Integrals**, The Chinese University of Hong Kong, China.

[41] Luan, J. (2004), **"Data Mining Applications in Higher Education"** , Cabrillo College, SPSS Inc.

[42] Lushbough, C. & Anderson, B. (2007), **Field Stream Database System-Data mining Storage for Biological Data,** University of South Dakota.

[43] MacLennan, J.; Tang, Z. Crivat, B. (2009), **Data Mining with Microsoft SQL Server (2008)**, Wiley Publishing, Inc., Indianapolis, Indiana.

[44] Maglothin, R. (2002), **"Data Mining in DNA: Using the SUBDUE Knowledge Discovery System to Find Potential Gene Regulatory Sequences"**.

[45] Mayatt, G.J. (2007), **Making Sense of Data: A Practical Guide to Exploratory and Data Mining Data Analysis**, John Wiley & Sons Inc, Hoboken, New Jersey.

[46] Michael J. A. & Berry**,** (2004)**, Data Mining Techniques For Marketing, Sales, and Customer Relationship Management,** Wiley Publishing, Inc., Indianapolis, Indiana.

[47] **Microsoft SQL Server 2008 Data Mining Add-Ins for Office 2007 Help**.

[48] Mirela, D. & Tiberiu, S. & Adina, B. (2009), **"Toward A Distributed Data Mining System For Tourism Industry"** , Stefan cel Mare University of Suceava Faculty of Economic Science and Public Administration.

[49] Mullins, I. & Siadaty, M. & Others. (2005), **Data mining and clinical data repositories: Insights from a 667,000 patient data set**, Computers in Biology and Medicine 36 2006, 1351–1377, www.intl.elsevierhealth.com/journals/cobm.

[50] Ohsaki, M. (2003), **A Rule Discovery Support System for Sequential Medical Data,** Shizuoka University, Japan.

[51] Perez, A. & Seijas, A. (1997), **Approximation to Generic Knowledge Discovery in Database Systems,** Madrid, Spain.

[52] **Dunham** , **M. H.** (2002), **DATA MINING Introductory and Advanced Topics.ppt**. Prentice Hall

[53] Ranjan, J. & Ranjan, R. (2010), **" Application Of Data Mining Techniques In Higher Education In India" ,** Institute of Management Technology**,** Ideal Institute of Technology, India, Journal of Knowledge Management Practice, Vol. 11, Special Issue 1, January (http://www.tlainc.com/articlsi10.htm).

[54] Rogers, G. & Joyner, E. (1997), **"Mining Your Data For Health Care Quality Improvement",** SAS Institute, USA.

[55] Santangelo, J.; Rogers, P. & Others. (2007), **"Using Data Mining Tools to Discover Novel Clinical Laboratory Test Batteries"**, The Ohio Stat. University Medical Center Columbus. AMIA (2007) Symposium Proceeding.

[56] Schulz, H. (2007), **"Visual Graph Mining for Regenerative System"**, wwwmosi.informatik.uni-rostock.de/diemosiris.

[57] Shah, Sh. & Kusiak, A. (2004), **"Data mining and genetic algorithm based gene/SNP selection"**, Artificial Intelligence in Medicine, USA, ELSEVIER.

[58] **SQL Server Tutorials (2008).**

[59] Adriaans, P. (1999), **Introduction to Data Mining and Knowledge Discovery**, Two Crows Corporation, third edition, Potomac, USA.

[60] Wang, H.; Wang, Ch. & Zeng, Z. (2006)**, Design and Implementation of a Web Usage Mining Model Based On Fpgrowth and Prefixspan,** Business School, University of Shanghai for Science and Technology.

[61] Weiss, G. M. (2003), **"Data Mining In Telecommunications"** , Department of Computer and Information Science, Fordham University.

[62] Witten, I. & Frank, E. (2005), **Data Mining: Practical Machine Learning Tools and Techniques**, second edition, Morgan Kaufmann Publishers, USA.

[63] Yang, Q. (2005), **"10 Challenging Problems in Data Mining Research"** prepared for ICDM, Hong Kong University of scientific & Technology.

[64] Zheng, Y.; Zhao, M. & Others, (2008), **Tour the World: building a web-scale landmark recognition engine**, National University of Singapore, Singapore & Google Inc. U.S.A.

ثالثاً. بعض مواقع الانترنت

[65] http://damascushospital.org/

[66] http://en.wikipedia.org

[67] http://healthit.hhs.gov/ (By The Office Of The National Coordinator (ONC))

[68] http://informatics.suite101.com/article.cfm/text_analytics_in_data_mining_software

[69] http://informationweek.com/story/showArticle.jhtml?articleID=165702390

[70] http://it.toolbox.com/blogs/opensource-analytics/database-vs-data-warehouse-8286

[71] http://marketingblog.topsourcellc.com/page/2

[72] http://msdn.microsoft.com/en-us/library/ms175595.aspx

[73] http://scla.org.sy/

[74] http://technet.microsoft.com/en-us/library/dd206985.aspx

[75] http://thawra.alwehda.gov.sy

[76] http://www.alriyadh.com/2007/05/01/article246118.html

[77] http://www.america.gov/....

[78] http://www.b-eye-network.com/view/12870

[79] http://www.bnikhaled.com/vb/t3464.html

[80] http://www.cs.ou.edu/academics/syllabi_08/spring09/CS_5083_McGovern.pdf

[81] http://www.healthdatamanagement.com/news/health_reform_obama_database_claims_fraud....

[82] http://www.micsymposium.org/mics_2004/Vaughn.pdf

[83] http://www.sas.com/offices/europe/norway/nyheter/university-of-louisville-analytics.html

[84] http://www.scientificblogging.com/news_releases/ngc_6397_shows_the_natal_kick_of_white_dwarf_stars

[85] http://www.scientificblogging.com/news_releases/ngc_6397_shows_the_natal_kick_of_white_dwarf_stars

[86] http://www.shvoong.com/internet-and-technologies/blog/1827096-database-vs-data-warehouse/

[87] http://www.slideshare.net/dataminingcontent/data-mining-the-sky-3191852

[88] http://www.tech4t.co.uk/pages/t4t-viscovery-applications.htm

[89] http://zimmer.csufresno.edu/~fringwal/sro.html

[90] people.dsv.su.se/~benkt/2I1224/oh2.ppt

[91] www.acm.org

[92] www.data-mining-guide.net/

[93] www.datamininglab.blogspot.com

[94] www.dmreview.com

[95] www.kdnuggets.com

[96] www.mohp.gov.eg (ICD-10)

[97] www.statoo.com/dmdd/

[98] www.the-data-mining.com

[99] www.worldscientific.com