



الجمهورية العربية السورية

جامعة دمشق

كلية الهندسة الميكانيكية والكهربائية

قسم الهندسة الطبية الحيوية

نظام تشخيص آلي لمرض السكري باستخدام خوارزميات التصنيف

رسالة مقدمة لنيل درجة الماجستير في الهندسة الطبية الحيوية

إعداد

م. منى صالح

إشراف

أ. د. م. رشا مسعود

2025-2024



الجمهورية العربية السورية

جامعة دمشق

كلية الهندسة الميكانيكية والكهربائية

قسم الهندسة الطبية الحيوية

قرار لجنة الحكم

قدم هذا البحث لمتطلبات الحصول على درجة الماجستير في الهندسة الطبية في قسم الهندسة الطبية – كلية الهندسة الميكانيكية والكهربائية – جامعة دمشق.

لجنة الحكم:

الدكتور	الاختصاص	التوقيع
د. رشا مسعود.	تحكم طبي حيوي	
د. وائل الإمام.	إحصاء حيوي	
د. فايزة القبيلي.	الكيمياء الحيوية السريرية	

تاريخ المناقشة: في تمام الساعة (9 صباحاً) يوم 1 / 6 / 2025 في قاعة السيمينار في قسم الهندسة الطبية.

تصريح خطي:

أنا الموقع أدناه أصرح بعلمي الكامل وقبولي:

1. أن كل ما ينتج عن البحث والأطروحة هو ملكية فكرية ومالية بالتشارك مع جامعة دمشق، وأني ألتزم

بأخذ موافقة الجامعة في حال رغبتني بنشر البحث أو الأطروحة أو جزء منها نصاً أو مضموناً خارج

إطار الجامعة (من دور نشر أو مكنتبات أو مواقع إلكترونية وغيرها من وسائل النشر).

2. أنه لا يوجد جزء من هذه الأطروحة تم اقتباسه من عمل آخر أو أنجز للحصول على شهادة أخرى

في جامعة دمشق أو أية جامعة أو معهد تعليمي داخل أو خارج الجمهورية العربية السورية.

الاسم والتوقيع:

المهندسة منى صالح

شكر وتقدير

أشكر الله وأحمده حمداً كثيراً طيباً مباركاً على نعمه التي لا تُعدّ ولا تحصى، إذ منّ عليّ بإتمام هذا العمل المتواضع، ويسرّ لي إنجازَه على الوجه الذي أرجو أن يرضى به عنيّ.

ولا يسعني إلا أن أتقدّم بخالص الشكر والامتنان لجامعة دمشق العريقة وكلية الهندسة الميكانيكية والكهربائية التي أدين لها بالكثير الذي لا يمكنني الوفاء به، وأخص بالشكر أعضاء الهيئة التدريسية في قسم الهندسة الطبية من أساتذة ودكاترة ومهندسين لما قدموه من علم ومعرفة.

أتقدم بجزيل الشكر والامتنان إلى الدكتورة رشا مسعود التي لم تبخل عليّ بعلمها وتوجيهاتها السديدة، فكانت نعم الداعمة والموجهة في كل مراحل البحث والارتقاء به إلى المستوى المطلوب.

وأتقدّم بشكري الجزيل إلى أساتذتي الموقّرين (الدكتور وائل الإمام، الدكتورة فايزة القبيلي، الدكتورة رشا مسعود) في لجنة المناقشة، لتفضّلهم بقراءة العمل وتقديم ملاحظاتهم التي حسّنته وسدّت مكان النقص التي اعترته.

ولا يفوتني أن أعبر عن بالغ امتناني لعائلتي العزيزة، التي كانت السند الدائم لي، ووفرت لي الدعم النفسي والمعنوي الذي لا يقدر بثمن، وتحملت انشغالي وأولتني الاهتمام في كل حين (أبي، أمي، أولادي، أخوتي)

وكل الحب والتقدير إلى رفيق دربي وسندي الداعم وإليه يرجع الفضل في نجاحي وتميزي خلال سنوات دراستي (زوجي)

وأخيراً، كل الشكر لكل من مدّ لي يد العون، ولو بكلمة أو نصيحة، فجزاكم الله عني خير الجزاء.

المهندسة منى صالح

الملخص:

يعد السكري من الأمراض المزمنة التي تؤدي إلى ارتفاع مستوى السكر في الدم، وعدم السيطرة على ضبط هذا المستوى يؤدي إلى أضرار في العديد من أجهزة الجسم ولا سيما القلب والأوعية الدموية والأعصاب، وانطلاقاً من أهمية الكشف المبكر عن هذا المرض كمحاولة للتقليل من مخاطره أو تأخير الإصابة به، وتحسين الحالة الصحية العامة للناس، والتخطيط لحياتهم مستقبلاً، عمل هذا البحث على استخدام خوارزميتي الانحدار اللوجستي والغابة العشوائية لبناء ثمانية نماذج، باستخدام قاعدة البيانات الهندية (PID)، التي تمت معالجة القيم المفقودة والمتطرفة فيها، وتوحيد نطاق البيانات، وموازنة عدد العينات باستخدام تقنيات التوزين: الزيادة العشوائية للعينات (ROS)، الانقاص العشوائي للعينات (RUS)، توليد عينات اصطناعية جديدة (SMOTE)، ثم جرى استخدام طريقة البحث الشبكي مع التحقق المتقاطع لضبط المعاملات الفائقة لنماذج التصنيف ثم تدريبها واختبارها، وتقييم أدائها لتحديد النموذج الأفضل.

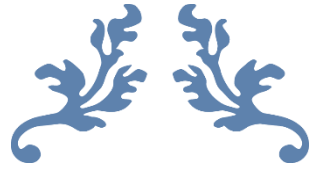
كما تم بناء قاعدة بيانات محلية اعتماداً على نتائج الدراسة المقطعية، وتم استخلاص سمات التنبؤ بمرض السكري من النوع الثاني من خلال تطبيق الاختبارات الإحصائية:

اختبار (Manwhitney, T-test) على السمات الرقمية، واختبار (Chi-Squares) على السمات الفئوية، كما تم الحصول على أفضل عشر سمات للتنبؤ بالسكري باستخدام خوارزمية تقليل السمات العودية (RFE)، وتم إعادة تدريب النموذج على البيانات المحلية بعد ضبط معاملاته الفائقة باستخدام طريقة البحث الشبكي والتحقق المتقاطع، ثم تقييم أداء هذا النموذج.

تفوق نموذج الغابة العشوائية على الانحدار اللوجستي عند التنبؤ بالسكري باستخدام قاعدة البيانات PID عند عدم موازنة البيانات، حيث كانت الدقة 87%، ودرجة F1 تساوي 84%، أما بالنسبة للمساحة تحت منحنى الخواص فقد بلغت 0.81.

كان أداء نموذج الغابة العشوائية على البيانات المحلية أفضل، حيث حقق أفضل معايير أداء عند أفضل ست وعشر سمات للتنبؤ بالسكري، حيث كانت الدقة 97%، ودرجة F1 تساوي 97%، بينما المساحة تحت منحنى الخواص كانت تساوي 0.96.

الكلمات المفتاحية: السكري من النوع الثاني، الكشف المبكر عن السكري، تعلم الآلة، المعاملات الفائقة، الانحدار اللوجستي، الغابة العشوائية.



الفصل الأول

الإطار العام للبحث



1.1 المقدمة:

شهدت تقنيات تعلم الآلة في الآونة الأخيرة تطوراً كبيراً، وأصبحت قادرة على تحليل وفهم البيانات الضخمة بطريقة فعالة ودقيقة، وأصبحت قادرة على اكتشاف الأمراض بشكل مبكر، الأمر الذي من شأنه أن يحسن فرص العلاج والإدارة الفعالة للمرضى، لذلك جرى العمل على تسخير هذه التقنيات المختلفة لتحليل البيانات المتاحة واختيار الأفضل منها للوصول إلى النماذج المناسبة للتنبؤ والكشف المبكر عن السكري من النوع الثاني بدقة وكفاءة عالية، حيث تم بناء عدة نماذج تصنيف باستخدام تعلم الآلة للكشف عن الحالات السليمة والحالات المصابة بالمرض.

2.1 مشكلة البحث:

يوجد حوالي 425 مليون مصاباً بالسكري حول العالم، وتشير التوقعات إلى إصابة ما يقدر بنحو 578 مليون فرد سيصاب بالسكري بحلول عام 2030، كما يعد هذا المرض سبب رئيسي لوفاة حوالي 2-5 مليون شخص كل عام، وما زال هناك 230 مليون شخص مصاب بالسكري دون تشخيص غير مدركين لحالتهم بسبب محدودية موارد الرعاية الصحية، كما أن تشخيص هذا المرض ليس بالأمر السهل، فأعراضه كثيرة و يظهر بعضاً منها في أمراض أخرى، أي أن هنالك أعراض مشتركة في كثير من الأمراض، لذلك كان لابد من العمل على الكشف المبكر عن هذا المرض، أو التنبؤ بخطر حدوثه.

3.1 تساؤلات البحث:

من التساؤلات التي تم طرحها:

1. هل يمكن الاستفادة من تعلم الآلة في بناء نموذج للتنبؤ والكشف المبكر عن السكري؟

2. ماهي عوامل الخطر التي يمكن أن تتنبأ بهذا المرض؟

4.1 فرضيات البحث:

لا بد من وجود عدة عوامل وسمات مرتبطة بشكل وثيق مع خطر الإصابة بمرض السكري من النوع الثاني، وتساعد في بناء نماذج للتنبؤ والكشف المبكر عن هذا المرض، مما يتيح للأفراد الحصول على الرعاية الصحية المطلوبة، والإدارة العامة للمرض.

5.1 مسلمات البحث:

من الأمور التي يمكن اعتبارها مسلمات لهذا البحث: الكشف عن السكري من النوع الثاني في وقت مبكر أمر مهم جداً، وقد أكد عليه جميع الباحثين والأطباء كمفتاح للعلاج الناجح والتخفيف من مضاعفات هذا المرض.

6.1 أهداف البحث:

1. إنشاء نموذج يتنبأ بدقة عن السكري باستخدام خوارزميات تعلم الآلة اعتماداً على بيانات ال PID والتأكد من مدى صلاحيتها في التنبؤ.
2. بناء نموذج تصنيف باستخدام خوارزميات تعلم الآلة قادر على تشخيص السكري اعتماداً على بيانات يتم جمعها محلياً.
3. تقييم أداء النموذج المقترح مع دراسات بحثية أخرى مشابهة لدراستنا.

7.1 أهمية البحث:

تكمن أهمية البحث في:

1. استخدام النموذج الناتج عن البحث كأداة مساعدة وداعمة لمقدمي الرعاية الصحية.
2. الكشف المبكر عن السكري، مما يسمح بالتدخل في الوقت المناسب، ومنع أو تأخير الإصابة به، وتحسين النتائج الصحية العامة للناس.

3. يمكن للمهنيين تنفيذ التدابير الوقائية وتقديم خطط رعاية شخصية للأفراد المعرضين للخطر.

8.1 مبررات البحث:

من المبررات ازدياد حالات الإصابة بالسكري من النوع الثاني بشكل كبير، وظهور بعض أعراض هذا المرض في أمراض أخرى مما يؤدي إلى التشخيص غير الدقيق، وتأكيد الباحثين على أهمية التنبؤ والكشف المبكر عن الإصابة بهذا المرض لتحقيق أفضل النتائج وإيجاد العلاج الفعال لجميع المرضى، لذلك كان لا بد من إيجاد أداة فعالة لتحديد عوامل التنبؤ بهذا المرض، وبالتالي الكشف المبكر لحالات الإصابة به.

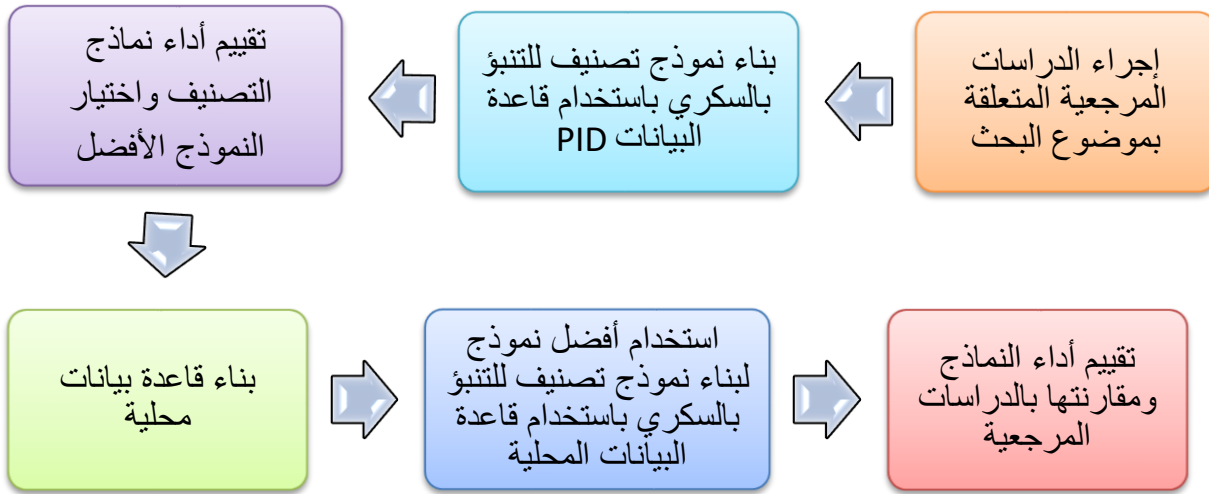
9.1 محدوديات البحث:

سيتم العمل مع البيانات الطبية المتوفرة في قاعدة البيانات التي يمكن تأمينها، بالإضافة إلى عمل دراسة مقطعية يتم من خلالها ملء استبيان مسحي لبناء قاعدة بيانات محلية.

10.1 مخطط البحث:

يوضح الشكل (1) المخطط العام الذي تم العمل وفقه في هذا البحث.

بعد دراسة وتحليل المراجع العلمية والدراسات المرجعية التي تناولت السكري والتنبؤ والكشف المبكر عن الأشخاص المعرضين لخطر الإصابة بالسكري، جرى العمل على تأمين قاعدة البيانات المناسبة لهذا البحث وتم استخدامها.



الشكل (1) المخطط الصندوقي لمراحل إنجاز البحث

ثم باستخدام لغة Python، تم العمل ضمن بيئة Google Colab التي تتوفر فيها كل المكتبات المطلوبة لهذا البحث لتحليل ومعالجة وتهيئة البيانات لبناء نماذج تعلم الآلة وضبط معاملاتها الفائقة (Hyper parameters)، حيث جرى في هذا البحث استخدام خوارزميتي الانحدار اللوجستي والغابة العشوائية لبناء ثمانية نماذج، ثم تدريبها واختبارها بعد ضبط معاملاتها الفائقة وتقييم أدائها، ومقارنة النتائج للوصول إلى النموذج الأفضل. كما جرى في هذا البحث بناء قاعدة بيانات محلية اعتماداً على نتائج الدراسة المقطعية (cross sectional study)، وتم استخلاص سمات التنبؤ بمرض السكري من النوع الثاني باستخدام الاختبارات الإحصائية المناسبة، ثم ضبطت المعاملات الفائقة لأفضل نموذج وتم تدريبه واختباره على أفضل السمات التي تم استخراجها باستخدام خوارزمية تقليل السمات اعتماداً على قاعدة البيانات المحلية، وبعد ذلك قيم أداء نموذج التصنيف، وقورنت النتائج مع الدراسات المرجعية، ثم كتبت الرسالة.

11.1 منهج البحث:

المنهج المتبع في هذا البحث هو المنهج المتبع في البحوث العلمية الهندسية، حيث يعد هذا البحث بمثابة دراسة تجريبية وتحليلية وبرمجية، فبعد دراسة وتحليل المعلومات في المراجع العلمية التي تناولت السكري، والتنبؤ بالأفراد المعرضين لخطر الإصابة بهذا المرض، أو الكشف المبكر عنه، والمتغيرات الحيوية والفيزيولوجية المرتبطة به، تم جمع وتأمين قاعدة البيانات التي استخدمت في هذا البحث، ثم دراسة وتحليل هذه البيانات، ثم بناء نماذج التنبؤ والكشف المبكر عن السكري وذلك بتجريب عدة خوارزميات تعلم آلة وتقييم أدائها لتحديد النموذج الأفضل لهذا البحث، إضافة إلى بناء قاعدة بيانات محلية اعتماداً على دراسة مقطعية مسحية يتم من خلالها جمع بيانات من أشخاص سليمين ومرضى بعد التوجه إلى عدة مراكز للرعاية الصحية المتواجدة ضمن محافظة دمشق وريفها، وتحليل واستخلاص السمات التي لها علاقة بالمرض باستخدام الطرق الإحصائية، ثم تدريب النموذج الأفضل واختباره على البيانات الجديدة وتقييم أدائه مع أفضل السمات، وفي النهاية تحليل جميع النتائج، ووضع التوصيات المستقبلية، وكتابة الرسالة.

12.1 أقسام الرسالة:

تقسم هذه الرسالة إلى سبعة فصول رئيسية، وهي:

1. الفصل الأول: الإطار العام وفيه فكرة عامة عن البحث، وتناول مشكلة البحث وهدفه وأهميته، ومبرراته بالإضافة إلى منهج البحث والمخطط العام لخطوات العمل.
2. الفصل الثاني: الأسس النظرية للبحث ويستعرض معلومات عن السكري، أنواعه، ومضاعفاته، وعوامل الخطورة للإصابة به.

3. الفصل الثالث: تعلم الآلة لتشخيص السكري ويستعرض معلومات عن تعلم الآلة والتعريف بأنواعها،

والأسس النظرية لخوارزميات التصنيف التي استخدمت في بناء نماذج البحث، وطرق تقييم الأداء، بالإضافة إلى شرح طريقة تقليل السمات المستخدمة.

4. الفصل الرابع: أدبيات البحث ويشمل شرح للدراسات المرجعية وما قام به الباحثين الآخرين في هذه الدراسات ذات الصلة بهذا البحث.

5. الفصل الخامس: منهج البحث وإجراءاته ويتضمن شرح مفصل لجميع الخطوات المتبعة في بناء النماذج المطلوبة ابتداء من تأمين قاعدتي البيانات ومعالجة البيانات انتهاء بضبط المعاملات الفائقة لبناء النماذج وتقييم أدائها بالإضافة إلى الطريقة المتبعة لتحديد أفضل سمات التنبؤ بالسكري اعتماداً على قاعدة البيانات المحلية.

6. الفصل السادس: النتائج والمناقشة ويستعرض ويناقش نتائج ضبط المعاملات الفائقة، وعرض

نتائج تقييم النماذج ومناقشتها من مصفوفة الارتباك (confusion matrix) ومقاييس الأداء

(Performance measures)، كما يعرض نتائج الاختبارات الإحصائية المطبقة على قاعدة

البيانات المحلية، ويعرض ويناقش نتائج طريقة اختيار أفضل السمات للتنبؤ بالسكري، إضافة إلى

مناقشة نتائج مقاييس الأداء لهذه النماذج، وفي النهاية مقارنة نتائج أفضل نموذج مع الدراسات

المرجعية ذات الصلة من حيث الدقة (Accuracy)، ودرجة F1، والمساحة (AUC) تحت

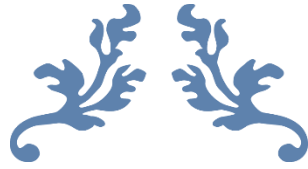
منحنى الخواص.

7. الفصل السابع: الخاتمة التي تتضمن ملخص البحث، وأهم الاستنتاجات والمميزات التي توصل

إليها هذا البحث، بالإضافة إلى التوصيات والآفاق المستقبلية.

13.1 خاتمة:

عمل هذا الفصل على إعطاء فكرة عامة عن البحث، حيث تم العمل على تسخير تقنيات تعلم الآلة من أجل بناء نموذج للتنبؤ بالسكري من النوع الثاني، وقد كانت نتائج هذا البحث واعدة ومشجعة.



الفصل الثاني
الأسس النظرية للبحث
السكري



1.2 مقدمة

يقدم هذا الفصل الأسس النظرية للبحث، حيث يستعرض بعض المعلومات الهامة عن السكري، وأنواعه، ومضاعفاته، وعوامل الخطورة التي تتعلق بالسكري، بالإضافة إلى التعريف بتعلم الآلة والأسس النظرية لخوارزميات التصنيف التي تم استخدامها في هذه الدراسة لبناء نماذج التصنيف، وكيفية إيجاد مقاييس تقييم الأداء لهذه النماذج وكفاءتها.

2.2 لمحة عن السكري

السكري مرض مزمن يحدث عندما يعجز البنكرياس عن إنتاج الأنسولين بكمية كافية، فقد تكون كمية الأنسولين التي يتم إفرازها أقل من المطلوب أو يكون هناك توقف تام عن إنتاجه، ويطلق على هذه الحالة "قصور الأنسولين"، أو أن الكمية المفرزة منه كبيرة في بعض الحالات كالأفراد المصابين بالسمنة، ولكن هناك مقاومة من قبل أنسجة وخلايا الجسم تعوق وظيفة الأنسولين، ويطلق على هذه الحالة "مقاومة الأنسولين"، وفي كلتا الحالتين يكون الغلوكوز غير قادر على دخول الخلايا، مما يؤدي إلى تراكمه في الدم وإمكانية ظهوره في البول، مع ازدياد تراكم السكر في الدم بدلا من دخوله خلايا الجسم، قد يؤدي إلى حدوث أضرار وخيمة على العديد من أجهزة الجسم، ولاسيما الأوعية الدموية الدقيقة في شبكية العين، و حوصلات الكلى، وتلك التي تغذي الأعصاب [1] .

3.2 أنواع السكري:

1.3.2 السكري من النوع الأول

يسمى هذا النوع بداء السكري المعتمد على لأنسولين، غالبا ما يستهدف الأطفال الصغار أو المراهقين الشباب، ونادرا ما يظهر بعد سن الأربعين، ينتج عن شذوذ في إفراز الأنسولين من طرف خلايا بيتا لجزر لانغرهانس المتواجدة في البنكرياس، أي إفراز الأنسولين بكميات ضئيلة جدا تكاد تكون منعدمة

أو عدم إفرازه مطلقا، وفي هذه الحالة يجب حقن الأنسولين يوميا بالكمية اللازمة وإلا فقد يتعرض المصاب إلى الغيبوبة، وقد تصل أحيانا إلى الموت، ومن أبرز أعراضه، عطش غير عادي مع كثرة شرب الماء، كثرة التبول، إنهاك وتعب شديد (جسدي ونفسي وجنسي)، الإحساس بالجوع وفقدان الوزن.

[2]

2.3.2 السكري من النوع الثاني

يسمى هذا النوع بالسكري غير المعتمد على الأنسولين، يظهر نتيجة نقص جزئي في هرمون الأنسولين، مما يجعله غير كاف للسيطرة على المستوى السوي للغلوكوز، يمثل هذا النوع ما يقارب 95 % من حالات الإصابة بالسكري وذلك حسب ما جاء في المجلة الطبية لمستشفى جينيف HUG، وهو غالبا ما يستهدف البالغين، يعتمد في علاجه على الأدوية الخافضة لسكر الدم أو على المواد التي تحسن من إنتاج هرمون الأنسولين التي تساهم في المساعدة على الاستعمال الأمثل للسكر في الجسم [2]

3.3.2 سكري الحمل:

هو من أنواع السكري الأقل انتشارا ويظهر أثناء الحمل فقط في النساء اللواتي لم يصبن بمرض السكري في السابق، وغالبا يعود سكر الدم إلى معدلاته السوية بعد الولادة ليعاود بالظهور في الحمل التالي، لذلك يجب على الأم الحامل إجراء فحص لسكر الحمل بين الأسبوع الرابع والعشرين والأسبوع الثامن والعشرين من الحمل خاصة إذا كان [3]:

1. عمرها 25 عاما وأكثر
2. أحد أفراد عائلتها مصابا بالسكري
3. ولادة سابقة لطفل كبير الوزن (أكبر من 4.5 كيلو غرام) أو ولادة طفل ميت.

نظام تشخيص آلي لمرض السكري باستخدام خوارزميات التصنيف، صالح.

4. السمنة المتكررة أثناء الحمل. [3]

4.3.2 مرض السكر الثانوي:

يحدث لوجود علة مرضية تؤثر على الخلايا المفرزة للأنسولين في البنكرياس وأهم هذه العلة:

1. أمراض البنكرياس المزمنة (الالتهاب المزمن).
2. أورام الغدة فوق الكلوية.
3. استئصال البنكرياس في حال وجود أورام سرطانية مثلاً.
4. اضطرابات في الغدد الصماء مثل مرض العملاقة بسبب زيادة إنتاج هرمون النمو وفرط إفراز الغدة الدرقية، ومرض كوشينغ الذي يؤدي إلى زيادة معدلات هرمون الكورتيزول.
5. استعمال بعض الأدوية مثل: الكورتيزون، وهرمون الغدة الدرقية. [3]

4.2 طرق تشخيص السكري سريريا:

هناك اختبارات سريرية عديدة للكشف عن السكري وهي:

1.4.2 اختبار السكر الصيامي:

وهو قياس نسبة السكر في الدم بعد مرور حوالي ثمان إلى عشر ساعات من الصيام المتصل عن

الطعام والشراب ماعدا (الماء)، فإذا كانت قيمة السكر في الدم:

1. يساوي (126 mg/dl) أو أكثر، فالشخص مصاب بالسكري مع وجوب التأكد من هذه النتيجة بإعادتها بعد يومين.

2. أقل من (100 mg/dl)، فالشخص سوي.

3. يساوي (100 mg/dl) حتى أقل من (126 mg/dl)، فالشخص في مرحلة ما قبل السكري. [4]

2.4.2 اختبار سكر الدم العشوائي:

عند وجود أعراض السكري مثل (كثرة العطش وشرب الماء وكثرة مرات التبول) مع مستوى سكر الدم

العشوائي:

1. يساوي (200 mg/dl) أو أكثر، فالشخص مصاب بالسكري.
2. أقل من (140 mg/dl)، فالشخص سوي.
3. يساوي (140 mg/dl) إلى (200 mg/dl)، فالشخص في مرحلة ما قبل السكري. [4]

3.4.2 اختبار تحمل الجلوكوز الفموي (OGTT):

وذلك بتحليل سكر الدم بعد مرور ساعتين من شرب محلول سكري يحتوي (75) غرام من الجلوكوز،

مع ملاحظة أن الشخص يكون صائماً قبل إجراء هذا الاختبار، فإذا كانت قيمة السكر في الدم:

1. يساوي (200 mg/dl) أو أكثر، فالشخص مصاب بالسكري.
2. أقل من (140 mg/dl)، فالشخص سوي.
3. يساوي (140 mg/dl) إلى (200 mg/dl)، فالشخص في مرحلة ما قبل السكري. [4]

4.4.2 اختبار معدل السكر التراكمي بالدم (الهيموغلوبين السكري) (HbA1c):

تؤخذ عينة دم وريدية في أي وقت من أوقات اليوم، ولا علاقة لنتيجة التحليل بحالة الصيام

فإذا كانت قيمة السكر التراكمي (الهيموغلوبين السكري):

1. أكثر من (6.4%) فالشخص مصاب بالسكري.
2. أقل من (5.7%) فالشخص سوي.

3. بين (5.7-6.4%) فالشخص في مرحلة ما قبل السكري [4]

5.2 أعراض السكري:

1. شدة العطش والاكثار من شرب المياه:

ويحدث العطش بسبب التأثير الاسمولي، حيث أن الزيادة الكبيرة في مستوى السكر في الدم يجري إفرازها عن طريق الكلى، ولكن هذا يحتاج إلى الماء لحمله، وبالتالي يؤدي إلى فقدان كبير للسوائل من الجسم، وهذا بدوره يؤدي إلى الجفاف، وفي النوع الثاني من السكري تكون زيادة التبول والعطش خفيفة في البداية، ولكنها تزداد سوءاً بالتدرج خلال أسابيع أو أشهر، وبعد ذلك يشعر المريض بالتعب الحاد، وقد يتطور ذلك إلى اضطراب في الرؤيا، ويمكن أن يتعرض الشخص إلى التجفاف. [3]

2. زيادة عدد مرات وكمية التبول:

هي إحدى الأعراض الأخرى لارتفاع السكر في الدم، لأن الكميات الزائدة من الجلوكوز في الدم تتم تصفيتها بواسطة الكليتين اللتين تحاولان التخلص منها بإفراز مزيد من الملح والماء. وتدعى هذه العملية بالإنتاج الزائد للبول، وتُعد غالباً المؤشر الأول للإصابة بالسكري. وإن لم يُفعل شيء لوقف هذه العملية، فإن الشخص سيصاب بسرعة بالجفاف والعطش. وكما ذكر سابقاً فإن الأنسولين يعمل إضافة إلى تنظيم الجلوكوز في الجسم، على منع خسارة الوزن والمساعدة على بناء أنسجة الجسم، وبالتالي فإن الشخص الذي يفتقر للأنسولين أو لا يعمل هذا الهرمون كما ينبغي لديه، سيخسر حتماً بعض الوزن، بشكل عام من المفترض ألا تزيد كمية الماء المفقودة عن طريق البول عن (1 - 2) لتر يومياً، وهذا يعني الذهاب للحمام حوالي (6-7) مرات يومياً، لكن مريض السكر

قد يفقد كميات أكبر من الماء عن طريق البول وقد تصل إلى 3 لتر أو أكثر يومياً، أي ما يعادل الذهاب

للحمام بمقدار عشر مرات يومياً. [5]

3. نقص الوزن:

يحدث نقص الوزن لدى مرضى النوع الأول من السكري دون وجود سبب واضح، على الرغم من الأكل الطبيعي أو حتى زيادة الأكل لدى الشخص، ويعزى ذلك إلى إفراز هرمون الجلوكاجون (المضاد لعمل الأنسولين) والذي يفرز من البنكرياس أيضا ويعمل على استقلاب البروتينات والدهون وتحويلها إلى سكر وبالتالي يحدث نقص في الوزن، كما يرجع نقص في الوزن أيضا إلى فقدان كمية كبيرة من الماء بسبب كثرة التبول. [3]

4. عدم وضوح الرؤية والدوار أو الصداع وعدم التركيز.

5. الاضطرابات النفسية:

إن الانفعالات النفسية الحادة كالقلق وعدم الرغبة في العمل تساهم في الإصابة بالسكري، لكنها ليست المسبب المباشر كون أغلب الأشخاص لديهم استعداد وراثي للسكري، وله تأثير نفسي على علاقات المصاب الشخصية وعلى عاداته الغذائية، من خوف وقلق وإحباط وعدم استقرار أو التأثير على عاداته الغذائية وفقدان الشهية، وعدم التعاون والخوف ويؤدي لعزوفه عن الطعام كون ظروف الإصابة تحتم تناول غذاء مخصص، مما يجعل الطعام غير مقبول من الناحية النفسية [6].

كما أظهرت الدراسات أن مرضى السكري أكثر عرضة للإصابة بالاضطرابات النفسية مقارنة مع الأشخاص الأصحاء، وقد يعود ذلك لارتفاع هرمونات التوتر كالكورتيزول التي تفرزها الغدد الصم بسبب الاضطراب النفسي، وقد أظهرت الدراسات أن معدل ارتفاع هذا الهرمون يقلل من حساسية الأنسولين، مما يؤدي إلى توزيع الدهون في وسط البطن، ومن المعروف أن زيادة الدهون تعد من أحد عوامل الخطر للإصابة بالسكري. [3]

نظام تشخيص آلي لمرض السكري باستخدام خوارزميات التصنيف، صالح.

وتشمل أعراض الاضطراب النفسي:

1. الشعور بالإحباط والشعور بالذنب والحزن المستمر
2. انعدام الرغبة للقيام بالأنشطة والهوايات المحببة
3. قلة الطاقة والشعور بالتعب
4. صعوبة التركيز والتذكر
5. الأرق والنهوض مبكراً في الصباح أو النوم لساعات طويلة
6. سرعة التهيج والتوتر

6.2 مقدمات السكري:

يُعرف ما قبل السكري بالحالة التي تكون فيها معدلات السكر في الدم أعلى من السوية وأقل من المعدل التشخيصي للسكري.

تتميز مرحلة ما قبل السكري بأن تركيز الجلوكوز عند الصيام (mg/dl 124-110) وتسمى بضعف تحمل الجلوكوز، أو عندما يكون تركيز الجلوكوز في بلازما الدم في فحص تحمل الجلوكوز عن طريق الفم لمدة ساعتين ضمن المجال (mg/dl 198-140)، وهذا عامل خطر رئيسي للسكري من النوع الثاني، ولكن ذلك لا يقود إلى حدوث السكري، ولكن وجد أن 33% إلى 66% من الأشخاص المصابين بمقدمات السكري يستمر ارتفاع الجلوكوز لديهم ويتطور وضعهم إلى السكري من النوع الثاني خلال أعوام. [8]

ليس هناك سبب معين لحالة ما قبل السكري (كما في النوع الثاني للسكري)، ولكن هناك مجموعة من العوامل المصاحبة لخطر الإصابة بذلك وهي:

• التاريخ العائلي للسكري.

نظام تشخيص آلي لمرض السكري باستخدام خوارزميات التصنيف، صالح.

- زيادة الوزن والسمنة.
- ارتفاع ضغط الدم.
- ارتفاع مستوى الكوليسترول والدهون الثلاثية في الدم.
- الإصابة بالسكري الحملي بالنسبة للمرأة.
- عدم ممارسة الرياضة وقلة الحركة [3]

أعراض هذه المرحلة:

عادةً لا توجد أعراض لحالة ما قبل السكري، وتوصي الجمعية الأمريكية للسكري بعمل تحليل لمستوى السكر في الدم لأي شخص عمره 45 عاماً أو أكثر، وخاصة الذين يعانون من السمنة أو لديهم تاريخ عائلي للسكري. [3]

7.2 عوامل الخطورة للسكري من النوع الثاني:

على الرغم من أن الأسباب الدقيقة للسكري من النوع الثاني لا تزال غير واضحة، إلا أن هناك العديد من عوامل الخطر المهمة التي تزيد من احتمالات الإصابة بالسكري من النوع الثاني، وهي العوامل نفسها التي تؤدي إلى زيادة احتمال الإصابة بمقدمات السكري أيضاً، وتتضمن أبرزها:

1.7.2 سوابق الإصابة بالسكري الحملي أو إنجاب طفل أو عدة أطفال بحجم كبير يفوق 4 كغ:

تشير الدراسات إلى أن حوالي ثلث إلى نصف النساء المصابات بالسكري الحملي قد يصبين بالنوع الثاني من السكري في غضون عشر سنوات. قد يحدث ارتفاع السكر لدى المرأة الحامل بسبب إعاقة هرمونات المشيمة، المرتبطة بنمو الجنين، قدرة جسم الأم الحامل على استخدام الأنسولين على الوجه الصحيح، مما قد يجهد جزر لانغرهانس التي تقوم بإفرازه أو إلى مقاومة الأنسولين، ويتم تشخيصه عادة بالفحص الروتيني للحامل أو بولادة طفل ذات وزن كبير. ويعبر فائض الجلوكوز من الأم إلى

الطفل عبر المشيمة، ونتيجة لذلك يبدأ بنكرياس الطفل بإفراز المزيد من هرمون الأنسولين للتخلص من زيادة الغلوكوز، وبما أن الجنين كان يعتمد في غذائه على دم الأم وفقد هذا المصدر بعد الولادة، فقد يؤدي ذلك لانخفاض الغلوكوز في دمه وهذا قد يعرض الطفل لخطر حدوث:

1. تشوه خلقي

2. زيادة كبيرة في النمو

3. فشل في نمو الرئة وما يسببه ذلك من متاعب التنفس للطفل عند الولادة

4. زيادة احتمال خطر الإصابة بالسكري في المستقبل وزيادة الوزن أثناء الطفولة. [7]

2.7.2 اختلاف السلالات والأجناس

إن خطر الإصابة بالنوع الثاني من السكري، وكذلك شدة مضاعفات السكري يختلف بين سلالات البشر المختلفة وقد يكون للوراثة والعوامل الاقتصادية والاجتماعية دورا مهما في الاختلاف بين سلالات البشر. وقد بينت دراسة أمريكية أن الرجال الأمريكيين من أصل إفريقي أكثر عرضة للإصابة بالنوع الثاني من السكري بما يعادل مرة ونصف، والنساء الأمريكيات من أصل إفريقي أكثر عرضة لخطر الإصابة بما يعادل مرتين للأمريكيين من أصل أبيض، وقد يعزى ذلك إلى زيادة معدلات ارتفاع ضغط الدم والتدخين، وكذلك قلة العناية الصحية لدى السود، وقد تلعب الوراثة دورا مهما في ذلك وعلى سبيل المثال قد بينت بعض الأبحاث أن الأمريكيين من أصل إفريقي لديهم خلل في إفراز الأنسولين ليس له علاقة بالتغذية أو العوامل الأخرى، وقد وجد أن معدل حدوث السكري من النوع الثاني لدى الأمريكيين الأصليين أكبر بما يعادل 19 مرة من معدل حدوثه لدى البيض، وكذلك فإن معدل الإصابة بالنوع الثاني لدى الأمريكيين من أصل إسباني أكبر من ذلك الموجود لدى البيض بما يعادل الضعف [7]

3.7.2 السمنة:

تعرف السمنة بأنها تراكم السعرات الحرارية الزائدة عن احتياج الإنسان، وتخزينها على هيئة دهون مترسبة في أماكن مختلفة تحت الجلد مثلا الأرداف والبطن والذراعين، مما يؤدي لزيادة الوزن عن الحد السوي. ويرجع السبب الرئيسي للسمنة وزيادة الوزن إلى الإفراط في تناول السكريات والدهون مع قلة الحركة، بينما تشارك الوراثة والاضطرابات الهرمونية بنسبة ضئيلة.

سجلت الاحصاءات الطبية وجود علاقة طردية بين السمنة ومرض السكري، وجد أن ثلث البدينين مصابون بالسكري، وأن 80% من مرضى السكري النوع الثاني هم بمن البدينين، ويرجع السبب لذلك إلى عدم استجابة الخلايا لمفعول الأنسولين بشكل فعال وبالتالي ارتفاع مستوى الغلوكوز في الدم، ويمكن فك هذا الارتباط وتراجع معدل الخطورة من خلال إنقاص الوزن إلى حدوده السوية. [4]

توزع الدهون في الجسم:

تشير التقديرات إلى أن (80 - 95%) من الزيادة الحالية في النوع الثاني من السكري هي بسبب السمنة وزيادة الدهون في منطقة البطن، حيث أن زيادة الدهون تلعب دورا مهما في مقاومة الأنسولين، ولكن طريقة توزع الدهون في الجسم أيضا مهمة، فزيادة الدهون حول البطن والجزء العلوي من الجسم كشكل التفاحة له علاقة بمقاومة الأنسولين وحدوث السكري، ومرض القلب وزيادة ضغط الدم والجلطة الدماغية وزيادة معدلات الكوليسترول غير الصحية، وتفسير ذلك هو أن الخلايا الدهنية الموجودة في وسط البطن تطلق الدهون في الدم بسرعة تفوق تلك الخلايا الموجودة في أجزاء الجسم المختلفة، وذلك لتزويد الجسم بالطاقة الضرورية لتمكين الإنسان من الهرب عند الخطر أو الخوف، ولمنع تجمع الدهون في وسط البطن فإنه ينبغي الحد من العوامل التي تؤدي إلى ذلك والتي تشمل:

1- المشروبات الكحولية

2- التدخين

3- الضغوطات اليومية

4- ازدياد الوزن

5- قلة الرياضة

6- أكل أقل من ثلاث وجبات في اليوم.

بينما الدهون التي تكون على شكل إجابة حول الأوراك قد لا يكون له علاقة بالمرض. [7]

طرق تشخيص السمنة:

يمكن تشخيص السمنة بطرق كثيرة مثل قياس محيط الخصر، أو قياس سمك الجلد، ولكن أشهر هذه القياسات هي قياس مؤشر كتلة الجسم مع العلم أن هذا المؤشر صالح للتطبيق على البالغين (أكثر من 18 عاماً)، وهناك جداول خاصة لتقييم الأطفال والمراهقين، كما أن هناك محاذير لاستعمالها لدى الحوامل والرياضيين الذين يمارسون رياضات كمال الأجسام وبناء العضلات. [4]

1. مؤشر كتلة الجسم:

يعرف مؤشر كتلة الجسم (نسبة وزن الجسم بالكيلوغرام إلى مربع الطول بالأمتار).

قيم مؤشر كتلة الجسم:

• الوزن السوي: $(18.5-24.9) kg/m^2$

• الوزن الزائد: $(25-29.9) kg/m^2$

• السمنة: أكبر من $(30) kg/m^2$.

ترتبط فئات المخاطر السريرية لمؤشر كتلة الجسم بزيادة تدريجية في خطر الإصابة بالسكري. حيث

أشارت أحد الدراسات أن الخطر النسبي للسكري بين النساء ذوات مؤشر كتلة الجسم (23-

kg/m^2 (24.9) هو أكبر ب 2.67 مقارنة بالنساء ذوات مؤشر كتلة الجسم أقل من kg/m^2 (23).

كما أن غالبية حالات مرض السكري تتطور لدى الأفراد ذوي الوزن الطبيعي والوزن الزائد، وليس بين

الأشخاص الذين يعانون من السمنة المفرطة فقط. [9]

2. محيط الخصر:

أشارت العديد من الدراسات إلى أن الخطر النسبي المرتبط بمحيط الخصر الأعلى كان أقوى قليلاً من ذلك المرتبط بارتفاع مؤشر كتلة الجسم. لذلك يعد محيط الخصر بديلاً صالحاً لمؤشر كتلة الجسم عند تقييم خطر الإصابة بالسكري من النوع الثاني في بيئة سريرية أو على مستوى السكان. ومع ذلك فإن الجمع بين مؤشر كتلة الجسم ومحيط الخصر يمكن أن يكون مفيداً، حيث أن كلا المقياسين يرتبطان بشكل مستقل بالمخاطر، ونظراً للارتباط القوي بين مؤشر كتلة الجسم ومحيط الخصر، فإن الأشخاص ذوي الوزن السوي الذين يعانون من السمنة في منطقة البطن أو البدينين الذين لا يعانون من السمنة في منطقة البطن، غير شائعين نسبياً. وعلى الرغم من ذلك، بالنسبة للأشخاص الذين يعانون من زيادة الوزن، فإن قياس محيط الخصر بالإضافة إلى مؤشر كتلة الجسم يسمح بمزيد من التقسيم الطبقي لمخاطر الإصابة بالسكري. اقترحت دراسة أن الأفراد الذين يعانون من زيادة الوزن والسمنة في منطقة البطن (محيط الخصر ≤ 102 سم بين الرجال، ≤ 88 سم بين النساء) لديهم خطر مماثل مقارنة بالأفراد الذين يعانون من السمنة المفرطة (مؤشر كتلة الجسم $\leq 30 kg/m^2$). [9]

4.7.2 أمراض القلب والأوعية الدموية:

تعد الإصابة بأمراض القلب والأوعية الدموية مؤشراً قوياً للإصابة بالسكري من النوع الثاني، حيث يصاب (25%-36%) من مرضى القلب والأوعية الدموية بالسكري، ويمكن تفسير ذلك من خلال حقيقة أن السكري يشترك في عوامل الخطر مع الأمراض القلبية الوعائية كمقاومة الأنسولين، السمنة،

التدخين، قلة النشاط البدني. [8]

5.7.2 الخمول البدني

يعرف الخمول البدني بغياب النشاط الفيزيائي المنتظم على الأقل ثلاث مرات أسبوعياً بمعدل ثلاثين دقيقة، حيث أشارت أحد الدراسات أن المشي السريع المنتظم (الذي يُعرف بأنه ≤ 2.5 ساعة أسبوعياً) يؤدي إلى انخفاض خطر الإصابة بالسكري من النوع الثاني بنسبة 30%، مقارنة بعدم المشي، بالإضافة إلى التمارين الرياضية الهوائية (مثل المشي السريع والركض والجري وركوب الدراجات والسباحة والتنس والتجديف)، و تدريب الأثقال لمدة ≤ 150 دقيقة أسبوعياً يؤدي إلى انخفاض خطر الإصابة بالسكري من النوع الثاني بنسبة 34%، حيث تزيد التمارين الهوائية من كثافة الجسيمات الكوندرية ونشاط الإنزيم المؤكسد، مما يحسن أكسدة الأحماض الدهنية، ويحفز امتصاص الجلوكوز في العضلات الهيكلية المستقلة عن الأنسولين، وكذلك المعتمدة على الأنسولين، كما يزيد التدريب كتلة العضلات الهيكلية وقدرة تحلل الجلوكوز، مما يعزز امتصاص الجلوكوز. [9]

6.7.2 النظام الغذائي:

ليس هناك طعام صحي مخصص لمرضى السكري، ولكن المهم هو الحصول على غذاء صحي متوازن ومنتظم، وذلك لتفادي الكثير من الأمراض كالسمنة وارتفاع ضغط الدم، وكثير من أمراض القلب. لكن يجب أن يتضمن الغذاء الصحي على ألياف كثيرة مصدرها الخضراوات والفواكه، وتأتي السكريات فيه عن طريق الفواكه والنشويات المركبة، وتقل فيه الدهون وخاصة المشبعة منها، والبروتينات، مع تناول كمية كافية من الماء، والإقلال من ملح الطعام إلى ما دون ملعقة صغيرة، وتقل إلى نصف ملعقة عند الإصابة بارتفاع ضغط الدم. [4]

يرتبط تناول العديد من المواد الغذائية الفردية أو المجموعات الغذائية بمخاطر الإصابة بمرض السكري، حيث أن تناول الحبوب الكاملة باستمرار يخفض خطر الإصابة بالسكري، على العكس من ذلك، ارتبط تناول كميات أكبر من الأرز الأبيض، بزيادة خطر الإصابة بمرض السكري من النوع الثاني خاصة بين السكان الآسيويين الذين يستهلكون كميات أكبر بشكل ملحوظ من الأرز الأبيض.

تشير الدراسات إلى أن تناول المشروبات المحلاة بالسكر يزيد من خطر الإصابة بمرض السكري من النوع الثاني عن طريق تغيير استقلاب الجلوكوز عند تناول كميات كبيرة من الكربوهيدرات سريعة الامتصاص. حيث تحتوي المشروبات المحلاة بالسكر، مثل المشروبات الغازية ومشروبات الفاكهة ومشروبات الطاقة، على مواد تحلية ذات سرعات حرارية مشتقة بشكل طبيعي، بما في ذلك السكروز.

شراب الذرة الفركتوز، أو عصير الفاكهة المركز. [9]

أشارت أحد الدراسات أن اتباع الشخص لنظام غذائي معين يقلل من خطر الإصابة بمرض السكري من النوع الثاني، كنمط النظام الغذائي للبحر الأبيض المتوسط، والذي يتميز باستهلاك مرتفع من الأطعمة النباتية المعالجة بالحد الأدنى، زيت الزيتون كمصدر رئيسي للدهون، الاستهلاك المعتدل لمنتجات الألبان والأسمك والدواجن، انخفاض استهلاك اللحوم الحمراء والاستهلاك المنخفض إلى المتوسط من النبيذ مع وجبات الطعام، أدى التقييد إلى انخفاض كبير في خطر الإصابة بالسكري مع مكملات زيت الزيتون البكر وانخفاض كبير في المخاطر مع مكملات الجوز المختلطة مقارنة بنظام غذائي منخفض الدهون. [9]

7.7.2 التدخين:

يرتبط التدخين بشكل إيجابي مع زيادة خطر الإصابة بمرض السكري، حيث دلت إحدى الدراسات أنه تزداد الإصابة بالسكري لدى 45% من المدخنين الحاليين مقارنة مع غير المدخنين. [9]

كما دلت بعض الدراسات أن التدخين يؤثر على طريقة عمل الانسولين في الجسم، ويزيد من صعوبة التحكم بالسكري وحصول مضاعفات، حيث يزيد من خطر اعتلال الكلية السكري، واعتلال الشبكية، والاعتلال العصبي، وأن المواد الكيميائية الموجودة في دخان السجائر تعمل على تضيق الأوعية الدموية مسببة زيادة لزوجة الصفائح الدموية مما يؤدي إلى الإصابة بالسكري وارتفاع معدل الكوليسترول، ويؤدي لزيادة سرطان البنكرياس لدى المدخنين أكثر من غيرهم، إذ يتسبب سرطان البنكرياس بخلل في نسبة السكر في الدم، وبالتالي الإصابة بالسكري. [6]

ولا تعتمد الإصابة بالسكري على نوع الدخان بل على كميته، حيث أشارت أحد الدراسات أن خطر الإصابة بالسكري أكبر بالنسبة للمدخنين الشريين (20سيجارة / يوميا) مقارنة بالمدخنين الخفيفين وأقل بالنسبة للمدخنين السابقين مقارنة مع المدخنين النشطين، بما يتوافق مع ظاهرة الاستجابة للجرعة. [11]

كما أظهرت مجموعة متنوعة من الدراسات الوبائية الأخرى وجود ارتباطات بين تدخين السجائر وتطور مرض السكري من النوع الثاني، في هذه الدراسات كان لدى الرجال الذين يدخنون 25 سيجارة أو أكثر يوميا نسبة خطر مرتفعة للإصابة بمرض السكري مقارنة بغير المدخنين. [12]

8.7.2 الوراثة أو السوابق العائلية خاصة من الدرجة الأولى عند الآباء والإخوة:

تلعب الوراثة دورا مهما في الإصابة بالسكري خاصة النوع الثاني، وتتوقف احتمالية الإصابة لأحد الأبناء بالسكري خاصة إذا كان أحد والديه أو كليهما مصابا بالمرض على عدة أمور منها:

- أحد الأبوين مصابا بالسكري أو كليهما
- نوع السكري لدى الأب والأم

- عمر الأب والأم المصابين عند إنجاب الطفل.

وبناء على ذلك فإن احتمالية إصابة الأبناء بالسكري تصل إلى 65% لدى إصابة الوالدين معا بالنوع

الثاني من مرض السكري [4].

بينت الدراسات أنه حوالي 25 إلى 30% من المرضى المصابين بالنوع الثاني من السكري لديهم أفراد

عائلة مصابين بالمرض، والأشخاص الذين لديهم مصابين بالسكري في العائلة يكونون أكثر عرضة

للإصابة بالسكري في سن مبكرة وتكون أعراضه شديدة. [5]

9.7.2 ارتفاع ضغط الدم:

يوصى بأن يكون ضغط الدم المستهدف أقل من (140/90) ملم زئبق للمرضى الذين يعانون من ارتفاع

ضغط الدم والسكري، وينبغي التحكم في ضغط الدم إلى أقل من (130/80) ملم زئبق لدى المرضى

الذين يعانون من مرض السكري الذين لديهم سمات سريرية عالية الخطورة. [13]

يعد ارتفاع ضغط الدم أحد عوامل الخطر القلبية الوعائية المهمة التي يجب السيطرة عليها وهو شائع

بين مرضى السكري. حيث تشير الدراسات إلى أن 26% من جميع مرضى ارتفاع ضغط الدم الذين

تزيد أعمارهم عن 20 عامًا مصابين بالسكري ويتلقون علاجًا لارتفاع ضغط الدم ولمرض السكري معًا،

إضافة لذلك فإن 58.6% من مرضى السكري البالغين يعانون من ارتفاع ضغط الدم، و 55.5% منهم

فقط يمكن التحكم في ارتفاع ضغط الدم لديهم نظرًا لأن ارتفاع ضغط الدم غير المنضبط يعد عامل

خطر مهم لفشل القلب (HF)، والرجفان الأذيني، أمراض الكلى المزمنة، وأمراض القلب الصمامية،

والخرف، وأمراض القلب التاجية (CHD)، والسكتة الدماغية، لذلك فمن الضروري السيطرة بشكل صحيح

على ضغط الدم. [13]

10.7.2 العمر:

هو عامل خطر مهم للإصابة بمرض السكري، حيث يتزايد مع التقدم في السن، وذروته تكون بين الثلاثين والأربعين عاماً، وتزداد خطورة الإصابة به وخاصة بعد سن الـ 45 عاماً نتيجة زيادة مقاومة الأنسولين وتناقص وظيفة البنكرياس. وبالإضافة إلى ذلك، يميل الناس إلى ممارسة الرياضة أقل، وفقدان كتلة العضلات وزيادة الوزن مع تقدم العمر، وكلها عوامل خطر إضافية لمرض السكري من النوع الثاني. [2-3-4-7]

11.7.2 تكرار الحمل:

السكري الحملي هو أحد الاضطرابات التي تتصف بضعف تحمل الغلوكوز لحظة تشخيصه خلال الحمل، يتطور هذا النوع من السكري عندما تقشل خلايا بيتا في سد زيادة طلب الانسولين نتيجة النقص الحاصل في حساسية الانسولين خلال فترة الحمل، يؤثر الحمل على المرأة المصابة بزيادة إفراز هرمون الانسولين في الدم، ونسبة السكر في الدم بعد تناول الطعام وامتصاصه، وترجع هذه التغيرات الى زيادة تحلل هرمون الانسولين في المشيمة والتي تفرز مواد شبيهة بهرمون النمو والذي يعمل ضد عمل الانسولين، تحفز هذه العوامل غدة البنكرياس على إفراز أكبر كمية من هرمون الانسولين وتسبب مضاعفات على المرأة الحامل منها التهاب المجاري البولية، تسمم الحمل، ارتفاع ضغط الدم، ولادة الجنين قبل موعده، زيادة نسبة التشوهات الخلقية، وموت الجنين داخل الرحم أو بعد الولادة بسبب انخفاض نسبة السكر في الدم، ويترك هذا الأمر للسيدة ورغبتها بالإنجاب مع وجوب اتخاذ الاحتياطات اللازمة عند تكرار الحمل. [6]

12.7.2 اضطرابات النوم

النوم هو فترة من الراحة الجسدية والعقلية. باعتباره حاجة إنسانية لا غنى عنها، فإن النوم أمر حيوي للصحة الجيدة ونوعية الحياة. ومع ذلك، فإن اضطرابات النوم، وكذلك اضطرابات المزاج تؤثر سلباً

على جودة النوم.

تعد نوعية النوم السيئة، واضطرابات النوم، وخاصة الأرق، من أكثر مشاكل النوم شيوعاً لدى الأشخاص المصابين بداء السكري من النوع الثاني. أشارت بعض الدراسات إلى أن النوم غير المنتظم ليلاً يؤدي إلى زيادة وظيفة هرمون الكورتيزول وهرمونات النمو، مما يؤثر أيضاً على الاستقلاب ويقلل من تحمل الجلوكوز وبالتالي التسبب في الإصابة بالسكري. [14]

بالإضافة إلى ذلك، ارتبطت مدة النوم القصيرة (من أربع إلى ست ساعات/ليلة) بزيادة خطر الإصابة بالسكري من النوع الثاني، في حين ارتبطت مدة النوم الطويلة (أكثر من تسع ساعات/ليلة) أيضاً المرتبطة بزيادة المخاطر، وينصح أن تكون مدة النوم المثالية بين سبع وتسع ساعات في الليلة، كما زاد خطر الإصابة بالسكري من النوع الثاني بين أولئك الذين يعانون من صعوبة في الحفاظ على النوم. بالإضافة إلى ذلك، قد يؤدي العمل بنظام الورديات، الذي يعطل التزامن الطبيعي لدورة الضوء والظلام والنوم، إلى عدم تطابق إيقاعات الساعة البيولوجية، مما يؤدي إلى سلسلة من التأثيرات المتعلقة بالغدد الصماء، لذلك من المهم لمرضى السكري الحصول على فترات نوم مناسبة للتحكم بشكل أفضل بالسكري. [9]

13.7.2 معدل ضربات القلب:

يتراوح معدل ضربات القلب أثناء الراحة للبالغين بين (60-100) ضربة في الدقيقة. السكري وأمراض القلب يسيران جنباً إلى جنب، حيث يتعرض البالغون المصابون بداء السكري لخطر متزايد للإصابة بأمراض القلب، كما يؤدي ارتفاع مستويات الجلوكوز في الدم وارتفاع مستويات الأنسولين المنتشرة في مجرى الدم إلى ارتفاع معدل ضربات القلب أثناء الراحة. بالإضافة إلى ذلك يؤثر السكري على نظام التوصيل الكهربائي للقلب مما قد يؤدي إلى عدم انتظام ضربات القلب. [26]

8.2 مضاعفات السكري:

يسهم السكري إذا ترك دون علاج في العديد من المشاكل الصحية الخطيرة التي تؤثر على كل من الأوعية الصغيرة والكبيرة، أي مضاعفات الأوعية الدموية الدقيقة والأوعية الدموية الكبيرة على الترتيب:

1.8.2 مضاعفات الأوعية الدموية الدقيقة للسكري:

تؤثر مخاطر الأوعية الدموية الدقيقة على الكلى، وتؤدي إلى الفشل الكلوي المستمر (اعتلال الكلية). بالإضافة إلى الأعصاب (الاعتلال العصبي)، مما يزيد من احتمالية الإصابة بتقرحات القدم السكرية وعمليات بتر الأطراف. علاوة على ذلك، يمكن أن تؤدي إلى إصابة العين (اعتلال الشبكية) وبالتالي تؤدي إلى العمى. [15]

1. الاعتلال الكلوي:

يعد السكري من أهم العوامل المسببة لعجز الكلى النهائي، حيث يوجد حوالي 20-40 % من مرضى السكري بنوعيه الأول والثاني قد ينتج لديهم اعتلال الكلية السكري مع مرور الوقت.

العلامة المبكرة لاعتلال الكلى السكري هو ظهور كمية بسيطة ولكن غير طبيعية من الزلال (البروتين في البول بما يسمى البيلة الألبومينية الصغيرة) (microalbuminuria). وتقدر كمية البروتين في البول في هذه الحالة بأكثر من 30 مغ في ال 24 ساعة)، وتعد علامة لزيادة فرصة الإصابة بمشاكل القلب والأوعية الدموية. وفي هذه الحالة إذا لم يتم التدخل العلاجي السريع فإن 80 % من مرضى السكري سيزيد لديهم معدل البروتين البولي إلى إن يصل المصاب إلى مرحلة اعتلال الكلية السكري

حيث تزيد كمية البروتين في البول إلى أكثر من 300 مغ في ال 24 ساعة بما يسمى المتلازمة

الكلائية، ومنها إلى الفشل الكلوي وترتفع الأملاح في الدم مثل حمض البول و الكرياتين ونسبة

البوتاسيوم ويرتفع ضغط الدم ويصعب تنظيمه وبذلك يصل المريض إلى المرحلة النهائية من الفل الكلوي، ويشعر المريض بالضعف والهزال ونقص الشهية وتورم في القدمين. [3]

1. الاعتلال العصبي:

هو مجموعة من الاختلالات العصبية تحدث بسبب السكري وتصيب الأعصاب الطرفية التي تتجه من المخ والنخاع الشوكي إلى أجزاء الجسم المختلفة وتختلف الأعراض تبعاً لنوع العصب ومكانه في الجسم وتكون الأعصاب إما حسية (تنقل الإحساس من الجسم إلى المخ) أو حركية (تنقل الأوامر من المخ إلى العضلات) أو لارادية (تغذي القلب والأحشاء). [3]

2. اعتلال الشبكية:

الشبكية هي الجزء المتحسس للضوء في العين وتقع بمحاذاة الجدار الخلفي للعين وهي تحتوي على خلايا عصبية ومخروطية الشكل تتحسس الضوء وترسل إشارة إلى الدماغ. تحتوي الشبكية على كثير من الأوعية الدموية التي تزودها بالدم والأكسجين والغذاء. وهناك نوعان من اعتلال الشبكية التي تصيب مريض السكري وهما:

• اعتلال الشبكية غير المتشعب

يحدث بسبب توسع الشعيرات الدموية في الشبكية وبسبب ذلك تفقد جداران الأوعية الدموية قدرتها على حفظ السوائل داخلها وينتج عن ذلك حدوث نزف بسيط وارتشاح وتورم في شبكية العين. وفي هذه المرحلة ممكن ألا يشعر المريض بأي أعراض لذا يجب الفحص الدوري سنوياً لقاع العين لاكتشاف هذه التغيرات لأنه من الممكن أن يتطور إلى النوع الثاني الذي يعد أكثر خطورة على البصر. [3]

• اعتلال الشبكية المتشعب

يتطور اعتلال الشبكية غير المتشعب فتقل كمية الأكسجين والمواد الغذائية التي تصل إلى الشبكية فتقوم بتعويض ذلك بتكوين أوعية دموية جديدة وهذه الأوعية لا تزود الشبكية بالدم بل تعترض الطريق

وهي ضعيفة وسريعة النزف مسببه ضعفا شديدا في الإبصار وعمتة السائل الزجاجي في الغرفة الخلفية للعين. وقد تنزف الأوعية الدموية وتسحب معها شبكية العين مؤدية إلى انفصال الشبكية وفي هذه الحالة يشعر المريض بأن هناك ستارة أمام عينيه. [3]

2.8.2 مضاعفات الأوعية الدموية الكبيرة للسكري:

1. إصابة شرايين القلب التاجية:

أمراض شرايين القلب التاجية هي السبب الرئيسي للوفاة في الدول الصناعية وكثير من دول العالم الأخرى، ولدى مرضى السكري تزداد نسبة الإصابة بأمراض شرايين القلب التاجية، يؤدي تأثر الشرايين التاجية للقلب بالتصلب وإلى نقص التروية لعضلة القلب، مما يصيب المريض بالذبحة الصدرية وحدوث جلطة بشرايين القلب مما يؤدي إلى احتشاء عضلة القلب، وضعف عضلة القلب هو النتيجة النهائية التي قد تحدث نتيجة احتشاء العضلة وهذا يجعل المريض يعاني من ضيق بالتنفس مع أقل مجهود وفي المراحل المتقدمة يكون ضيق التنفس حتى في حال عدم بذل الجهد، مع هبوط عضلة القلب وتزداد نسبة الوفاة المفاجئة نتيجة زيادة واضطراب في دقات القلب. [3]

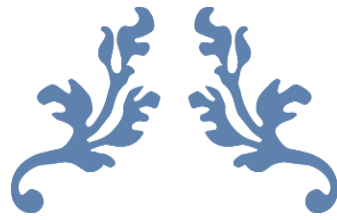
2. الجلطة الدماغية:

يؤدي تصلب شرايين الدماغ إلى حدوث جلطة دماغية، وهو أيضاً أحد مضاعفات مرض السكري نتيجة عوامل كثيرة أهمها زيادة نسبة الكوليسترول بالدم وارتفاع ضغط الدم، وتفاوت حدة الجلطة الدماغية من ضعف أو شلل مفاجئ يصيب الشق الأيمن أو الأيسر من الجسم ويستمر لدقائق أو أيام، أو قد تكون هناك إصابة شديدة تؤدي إلى حدوث غيبوبة لأيام ثم يحدث تحسن تتفاوت درجاته.

قد يستعيد كامل قدراته وقد يكون هناك شلل نصفي شديد أو بسيط حسب الحالة. [3]

9.2 الخاتمة:

تم في هذا الفصل تغطية الجانب النظري للبحث، حيث استعرض المعلومات المهمة عن السكري، وأنواعه، ومضاعفاته وأعراضه، إضافة إلى أهم عوامل الخطورة المحتملة للتنبؤ بالسكري من النوع الثاني.



الفصل الثالث

تعلم الآلة

لتشخيص السكري



1.3 مقدمة:

يقدم هذا الفصل بعض المعلومات المهمة عن التعريف بتعلم الآلة والأسس النظرية لخوارزميات التصنيف التي تم استخدامها في هذه الدراسة لبناء نماذج التصنيف، وكيفية إيجاد مقاييس تقييم الأداء لهذه النماذج وكفاءتها، بالإضافة إلى بعض الاختبارات الإحصائية وخوارزميات اختيار السمات التي تم استخدامها لاستخراج أهم السمات للتنبؤ بالسكري.

2.3 تعلم الآلة وأنواعه:

تعلم الآلة: هو أحد فروع الذكاء الاصطناعي الذي يركز على تطوير الأنظمة التي تمكن أجهزة الحاسوب من التعلم والتكيف وتحسين أدائها من خلال التجربة، ويعتمد على البيانات، حيث يتم استخراج المعلومات الهامة والتعلم منها مما يعطي الحاسوب قدرة على اتخاذ القرارات والتنبؤ بنتائج مستقبلية، ويوجد عدة تعريفات لتعلم الآلة، حيث يمكن تعريفه بأنه مجال الدراسة الذي يمنح الحواسيب القدرة على التعلم دون برمجتها صراحة. [16]

يتضمن تعلم الآلة تغذية النظام بالبيانات مثل (الجدول والصور والنصوص)، ثم تبدأ عملية التعلم عن طريق تحديد الأنماط والعلاقات بين البيانات باستخدام خوارزميات تعلم مختلفة، بعد ذلك يتم إنشاء نماذج مختلفة اعتماداً على الأنماط المكتسبة، ثم يتم استخدامها في التنبؤ والتصنيف باستخدام بيانات جديدة، وفي النهاية يمكن تطوير النموذج وتحسين أدائه من خلال تلقي المزيد من البيانات والتغذية الراجعة على تنبؤاته، وتصنف خوارزميات تعلم الآلة بشكل عام إلى ثلاثة أنواع وهي:

1. التعلم الموجه (Supervised Learning): وهو أحد الأنواع الأساسية لتعلم الآلة الذي تتوفر فيه

المدخلات والمخرجات في مجموعة البيانات المستخدمة في التدريب، ويمكن اعتبار هذه البيانات بمثابة معلم يوجه الخوارزمية، ويشرف على عملية التعلم من خلال المخرجات الصحيحة المعروفة مسبقاً، والتي يجب أن تسعى الخوارزمية الوصول إليها في مخرجاتها، وبالتالي فإن الخوارزمية تقوم

بإجراء التنبؤات بشكل متكرر على بيانات التدريب، وتصحيح الأوزان والمحددات بناء على الأخطاء الناتجة عن هذه التنبؤات، وتستمر عملية التعلم حتى تصل الخوارزمية إلى مستوى مقبول من الأداء، وهذا ما يسمح لها بالتعميم بشكل جيد على بيانات جديدة لم تستخدم في التدريب.

2. التعلم غير الموجه (Unsupervised Learning): في هذا النوع من التعلم لا تتوفر المخرجات الصحيحة ولا المعلومات المعروفة مسبقاً، بدلاً من ذلك تحاول الخوارزمية اكتشاف الأنماط والهيكليات الموجودة في البيانات من خلال تحليل المدخلات الموجودة فقط، وهو بعكس التعلم الموجه فلا يوجد معلم يوجه عملية التعلم، وإنما تقوم الخوارزمية بتحديد أوجه التشابه والاختلاف بين المدخلات، مما يسمح بجمع العناصر المتشابهة في مجموعات أو أصناف معينة تعتمد عليها الخوارزمية في اتخاذ القرارات بالنسبة للبيانات الجديدة. [16]

3. التعلم المعزز (Reinforcement Learning): يجمع هذا النوع من التعلم بين التعلم الموجه وغير الموجه، حيث تتلقى الخوارزمية مكافآت أو عقوبات بناء على أداؤها، فيتم مكافأتها عندما تحقق نتائج صحيحة، ومعاقبتها عندما تقوم بارتكاب الأخطاء، ولكن دون توجيهها بشكل مباشر حول كيفية تصحيح الأخطاء، ويتوجب على الخوارزمية استكشاف بيئتها وتجربة احتمالات مختلفة حتى تتوصل إلى كيفية الحصول على الإجابة الصحيحة، أي أن التعلم المعزز يعتمد على مفهوم

التجربة والخطأ، مما يسمح للخوارزمية بتطوير استراتيجيات فعالة مع مرور الوقت لتحقيق أهداف مرجوة. [16]

1.3.3 الحاجة إلى تعلم الآلة لتشخيص السكري:

على مدى السنوات القليلة الماضية، انتشرت خوارزميات التعلم الآلي للكشف عن السكري وذلك للعديد من الميزات:

1. الكشف المبكر عن السكري في مراحله الأولى، والتدبير العلاجي له يعد خطوة مساعدة وداعمة لمقدمي الرعاية الصحية من أجل التدخل المبكر والحصول على تشخيص أفضل وتدبير علاجي فعال. [17]

2. ضمان تحقيق نتائج صحية أفضل، وللمحد من عبء المضاعفات على المدى البعيد.

3. تساعد هذه التقنية على تطوير أداة دقيقة وفعالة تصل إلى طاولة الأطباء لمساعدتهم على اتخاذ قرارات أفضل بشأن حالة المرض.

4. استخدام خوارزمية التعلم الآلي لبناء إطار يوفر نتائج موثوقة مع تقليل الجهد البشري. [18]

5. علاوة على ذلك، من خلال التنبؤ الدقيق بالسكري والرعاية الصحية يمكن للمهنيين تنفيذ التدابير الوقائية وتقديم خطط رعاية شخصية للأفراد المعرضين للخطر. يمكن أن يشمل هذا تعديلات نمط الحياة، والتدخلات الغذائية، وأنظمة التمارين الرياضية، وإدارة الدواء ومراقبة مستويات السكر في الدم بشكل فعال. [17]

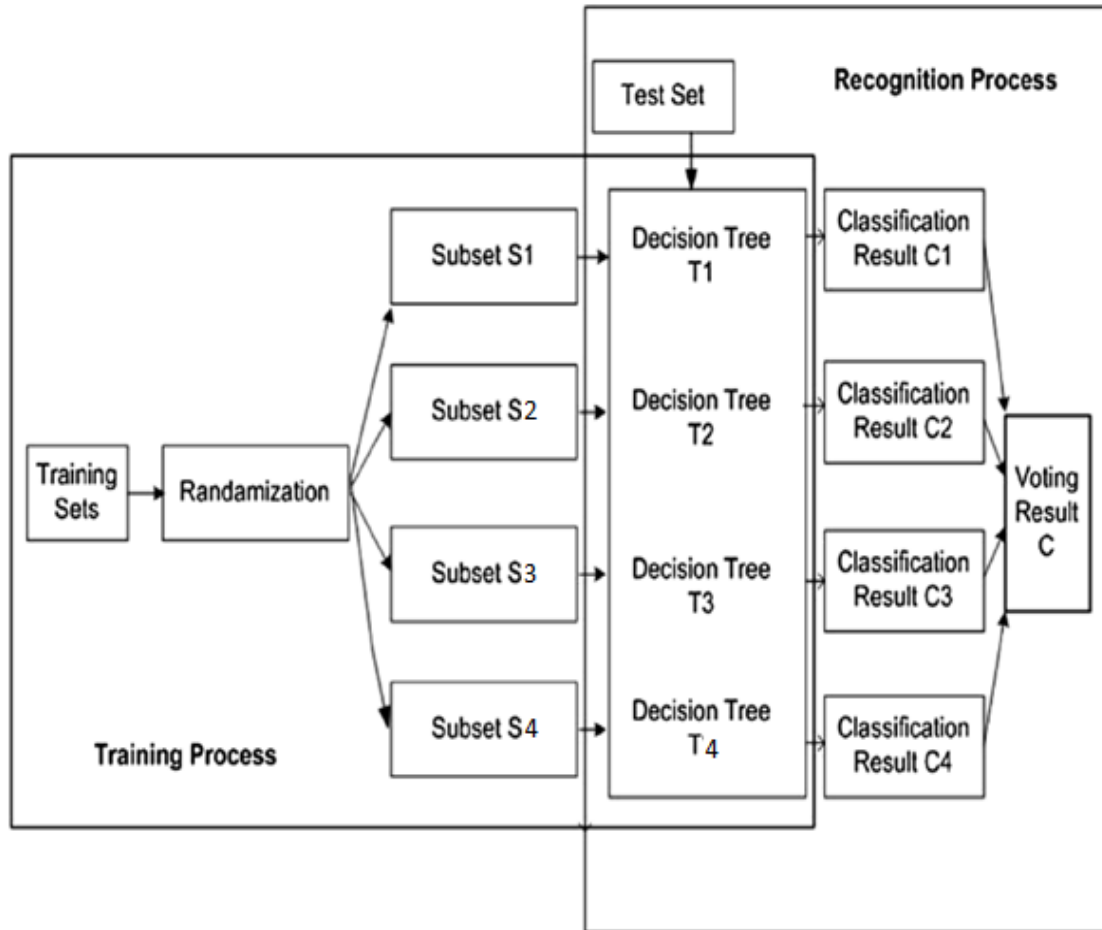
2.3.3 خوارزميات تعلم الآلة الموجه:

1. الغابة العشوائية (Random Forest):

أول اقتراح لخوارزمية (RF) كان من قبل Breiman في جامعة كاليفورنيا في عام 2001، وهي عبارة عن مجموعة مصنفات أساسية وهي أشجار قرار DT التي تكون مستقلة عن بعضها البعض بشكل كامل، حيث عند اختيار عينة جديدة يتم تحديد صنفها اعتماداً على تصويت كل من تلك المصنفات التي تعطي قرارها بشكل مستقل، فالصنف الذي يحصل على أكبر عدد من الأصوات يكون هو صنف تلك العينة، أي أن النتيجة النهائية لخوارزمية الغابة العشوائية يعتمد على التصويت (Voting).

إن تسمية Forest جاءت كون هذه الخوارزمية تتألف من عدد كبير من أشجار القرار، وبالنسبة إلى Random نتيجة للعملية العشوائية أثناء عملية البناء عند اختيار مجموعة بيانات التدريب الفرعية من مجموعة بيانات التدريب الكاملة، واختيار مجموعات السمات الفرعية وذلك لكل شجرة قرار مكونة في هذه الخوارزمية، وهذا ما يحقق الاستقلالية للمصنفات ويحسن من دقة التصنيف ويساعد على التعميم بشكل أفضل، ويعطي سرعة كبيرة في البناء ووقت أقل في التدريب أي تجنب فرط التدريب، وهذا ما يعطي هذه الخوارزمية شعبيتها وزيادة استخدامها يوماً بعد يوم. [19]

يوضح الشكل (2) آلية عمل خوارزمية الغابة العشوائية: [19]



الشكل (2) آلية عمل خوارزمية الغابة العشوائية [19]

توضح المعادلة (1) عمل خوارزمية الغابة العشوائية: [20]

$$H(x) = \arg \max_Y \sum_{i=1}^k I(h_i(x)) \quad \dots\dots\dots(1)$$

حيث:

Y: خرج النموذج، K: عدد أشجار القرار، h: نموذج شجرة القرار رقم i.

2. الانحدار اللوجستي (Logistic Regression)

هو أسلوب إحصائي يختص بدراسة العلاقة بين متغير تابع ومتغير أو عدة متغيرات مستقلة، ناتجاً عن ذلك معادلة رياضية حيث تمثل هذه العلاقة أفضل تمثيل، يعتبر نموذج الانحدار اللوجستي حالة خاصة من النموذج الخطي العام.

يستخدم الانحدار الخطي للتنبؤ بالقيم المستمرة، بينما يمكن استخدام الانحدار اللوجستي للتنبؤ بالقيم الفئوية والمستمرة على حد سواء، ولكنه يستخدم على نطاق واسع كخوارزمية تصنيف.

الاختلاف الرئيسي الذي يجعل كلاهما مختلفين عن بعضهما البعض هو عندما تكون المتغيرات التابعة ثنائية أو ثلاثية بمعنى آخر فئوية فعندئذ يتم استخدام الانحدار اللوجستي، أما عندما تكون المتغيرات التابعة مستمرة عندئذ يتم استخدام الانحدار الخطي.

ينقسم الانحدار اللوجستي إلى ثلاثة أنواع حسب عدد فئات المتغير التابع وهي الانحدار اللوجستي الثنائي حيث يكون للمتغير التابع فئتين، والانحدار اللوجستي متعدد الفئات حيث يكون للمتغير التابع أكثر من فئتين والانحدار اللوجستي الترتيبي حيث تكون فئات المتغير التابع ذات مقياس ترتيبي، تم التطرق في هذا البحث إلى دراسة الانحدار اللوجستي الثنائي فقط.

الانحدار اللوجستي الثنائي:

هو أحد خوارزميات التعلم الآلي الأكثر استخدامًا في مجال التعلم المُوجه، وهو نموذج يستخدم للتنبؤ باحتمالية وقوع حدث ما من خلال تحليل العلاقة بين مجموعة من المتغيرات المستقلة الغير خطية أيا كان نوعها والمتغير التابع الذي يكون ثنائي القيمة، ويعد من أكثر النماذج شيوعا والمستخدمه في المجال الطبي عندما يكون لمتغير المرض قيمتين (0 غير مصاب، 1 مصاب)، إذ تم تصميم نموذج الانحدار اللوجستي لوصف احتمال حيث يكون الناتج المتوقع قيمة احتمالية بين 0 و 1.

عادة يستخدم عتبة قرار افتراضية لمسائل التصنيف الثنائي والتي تكون قيمتها عادة 0.5، وبالتالي فإنه يمكن التنبؤ بانتماء عينة إلى فئة ما عن طريق مقارنة الاحتمال مع القيمة 0.5، حيث أنه إذا كانت قيمة الاحتمال أكبر من 0.5 فإن العينة تنتمي إلى الفئة الإيجابية ذات الترميز 1، وإذا كانت قيمة الاحتمال أصغر من 0.5 فإن العينة تنتمي إلى الفئة السلبية ذات الترميز 0، لذلك تُعتبر قيمة العتبة 0.5 نقطة انطلاق جيدة، إذ توازن بين حساسية النموذج وخصوصيته. [29]

والشكل العام لمعادلة نموذج الانحدار اللوجستي ثنائي الحدين تكتب بالصيغة الموضحة في المعادلة (2) :

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad \dots\dots\dots(2)$$

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \dots \dots + \beta_k x_k$$

حيث:

المعامل β_0 يمثل الحد الثابت في معادلة الانحدار اللوجستي، والمعاملات $\beta_1, \beta_2, \dots, \beta_k$ تسمى معاملات النموذج، و x_1, x_2, x_k هي المتغيرات المستقلة، و z هو المتغير التابع الثنائي.

4.3 تحويل البيانات إلى بيانات قياسية:

يتم تحويل البيانات إلى بيانات قياسية عندما يكون لميزات مجموعة بيانات الإدخال اختلافات كبيرة بين نطاقاتها، أو ببساطة عندما يتم قياسها بوحدات مختلفة، تسبب هذه الاختلافات في نطاقات الميزات الأولية مشكلة للعديد من نماذج التعلم الآلي. لذلك فهي تقنية مهمة يتم إجراؤها في الغالب كخطوة معالجة مسبقة قبل إدخال البيانات في العديد من نماذج التعلم الآلي، لتوحيد نطاق ميزات مجموعة بيانات الإدخال. [30]

ويمكن إجراؤها عن طريق طرح المتوسط من قيمة البيانات والقسمة على الانحراف المعياري، كما هو موضح في المعادلة (3-3). [30]

$$z = \frac{x_i - \mu}{\sigma} \quad \dots\dots\dots(3)$$

حيث Z: الدرجة المعيارية.

σ : الانحراف المعياري للبيانات

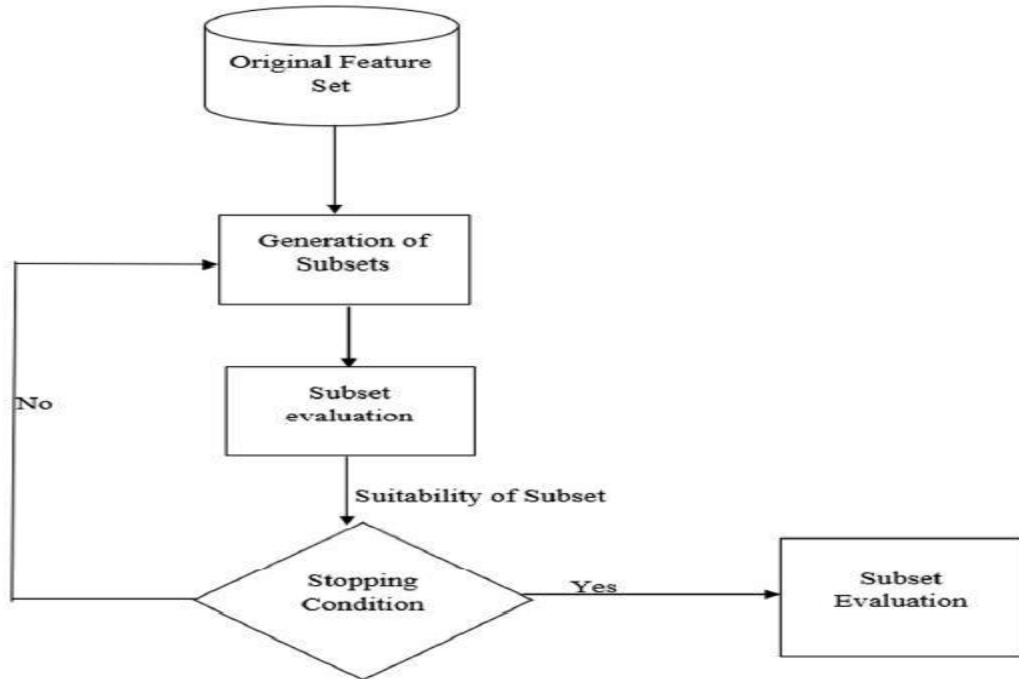
μ : المتوسط الحسابي للبيانات.

X: قيمة كل عينة.

بمجرد الانتهاء من التوحيد القياسي، سيكون لكل الميزات متوسط صفر وانحراف معياري قدره واحد، حيث يمكن للبيانات الموحدة أن تسهل مهام معالجة البيانات وتخزينها، بالإضافة إلى تحسين الدقة أثناء تحليل البيانات، كما يمكن أن يقلل التكاليف ويوفر الوقت. [30]

5.3 اختيار السمات (feature selection):

هي إحدى جوانب هندسة الميزات التي تهدف إلى تخفيض عدد الميزات التي لا تؤثر بشكل مباشر وكموني بمتغير التابع، أي إيجاد أفضل مجموعة فرعية من السمات ذات الصلة من مجموعة سمات البيانات الأصلية باستخدام معايير تقييم محددة، وذلك لتقليل أبعاد المشكلة التي يتم معالجتها، مما يسهل عملية التصنيف ويعطي أداء أفضل في عملية التعلم، ويوضح الشكل (3) آلية عمل اختيار السمات العامة.



الشكل (3) آلية اختيار الخصائص العامة. [21]

هناك عدة طرق لاختيار السمات الفرعية منها:

1. طريقة المرشح (Filter Approach): تتضمن مقياساً مستقلاً لتقييم المجموعات الفرعية من

السمات دون إشراك إحدى خوارزميات تعلم الآلة، وهي فعالة وسريعة حسابياً ولكنها يمكن أن تهمل

ميزات ليست مفيدة بحد ذاتها ويمكن أن تكون مفيدة جداً عند دمجها مع ميزات أخرى، أي أن هذه

الطريقة لا تأخذ بعين الاعتبار الارتباط البيئي لمتغيرات الدراسة.

2. طريقة التغليف (Wrapper Approach): يستخدم إحدى خوارزميات تعلم الآلة لتقييم

المجموعة الفرعية من السمات بعد تدريب الخوارزمية عدة مرات، وينتقى مجموعة السمات الفرعية

المناسبة لخوارزمية تعلم الآلة بالشكل الأفضل، لذلك يكون أدائه أفضل عادة. وتعد طريقة استبعاد

السمات العودية ((Recursive Feature Elimination (RFE)) تابعة لهذا النهج، حيث

تتطلب تحديد خوارزمية تعلم الآلة غالبا (الغابة العشوائية أو الانحدار اللوجستي) لتكون المقياس

لتقييم المجموعات الفرعية من السمات، وكذلك تحديد عدد السمات المطلوب الوصول إليها. [25]

3. الطريقة المضمنة (Embedded Approach): وهي دمج بين طريقتي المشرح والتغليف، حيث

يتم اختيار مجموعة السمات الفرعية المناسبة بشكل مباشر أثناء عملية التدريب. [25]

6.3 ضبط المعاملات الفائقة (Hyper Parameter Tuning):

يعد ضبط المعاملات الفائقة خطوة مهمة في عملية التعلم الآلي لأنه يتضمن إيجاد القيم المثلى للمعاملات

الفائقة للنموذج لتحسين أدائه على بيانات غير مرئية، وبالتالي تعزيز قدرته على التعميم. [24]

وعلى عكس المعاملات التي يحددها النموذج أثناء التدريب، يجب إسناد قيم للمعاملات الفائقة قبل بدء

التدريب، و يستغرق تحديدها يدويا الكثير من الوقت خاصة عند الحاجة إلى ضبط عدد كبير من

المعاملات، لذلك فإن استخدام الطرق الآلية توفر الكثير من الوقت والجهد، وتبحث في مجموعة قيم

أكبر للوصول إلى أفضل تركيبة للمعاملات الفائقة. [24]

1.6.3 طريقة البحث الشبكي مع التحقق المتقاطع (Grid search with cross validation):

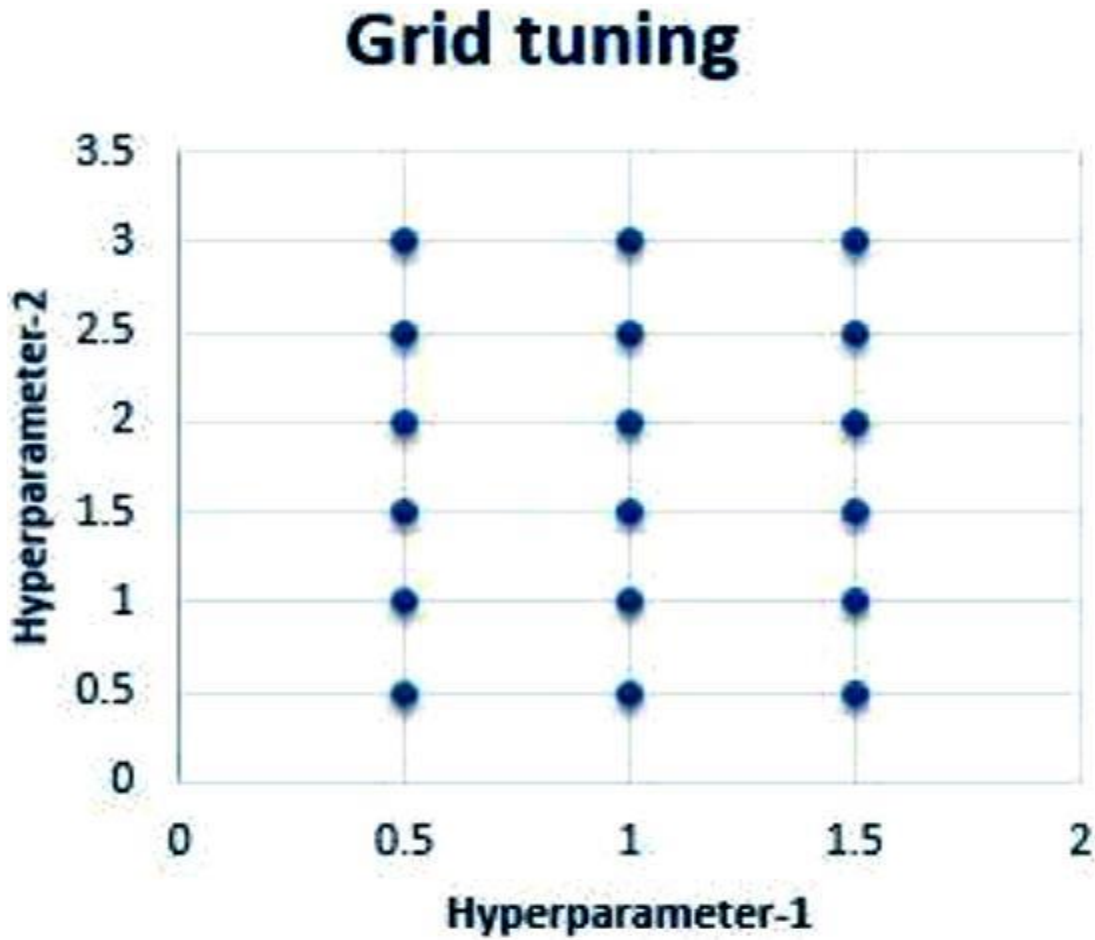
وهي خوارزمية استكشافية مستقلة شاملة تأخذ بالاعتبار كل مجموعة ممكنة من القيم المعطاة لاختيار

المعاملات الفائقة منها، حيث تختبر وتقيم أداء كل مجموعة فريدة من المعاملات في مساحة البحث

لتحديد المجموعة التي تعطي الأداء الأفضل، إضافة إلى أن هذه الطريقة تعد بسيطة ويمكن تنفيذها

بسهولة.

ويوضح الشكل (4) طريقة البحث الشبكي: [21]



الشكل (4) طريقة البحث الشبكي لضبط المعاملات الفائقة. [24]

2.6.3 التحقق المتبادل باستخدام k-fold cross validation:

هي طريقة تستخدم لتقييم نماذج التعلم الآلي على مجموعة بيانات محدودة، وتعد هذه الطريقة الأكثر

استخداماً لأنها تعتبر حلاً وسطاً بحيث لا تسمح بفرط التقدير من خلال اختيار مجالات قصيرة الطول،

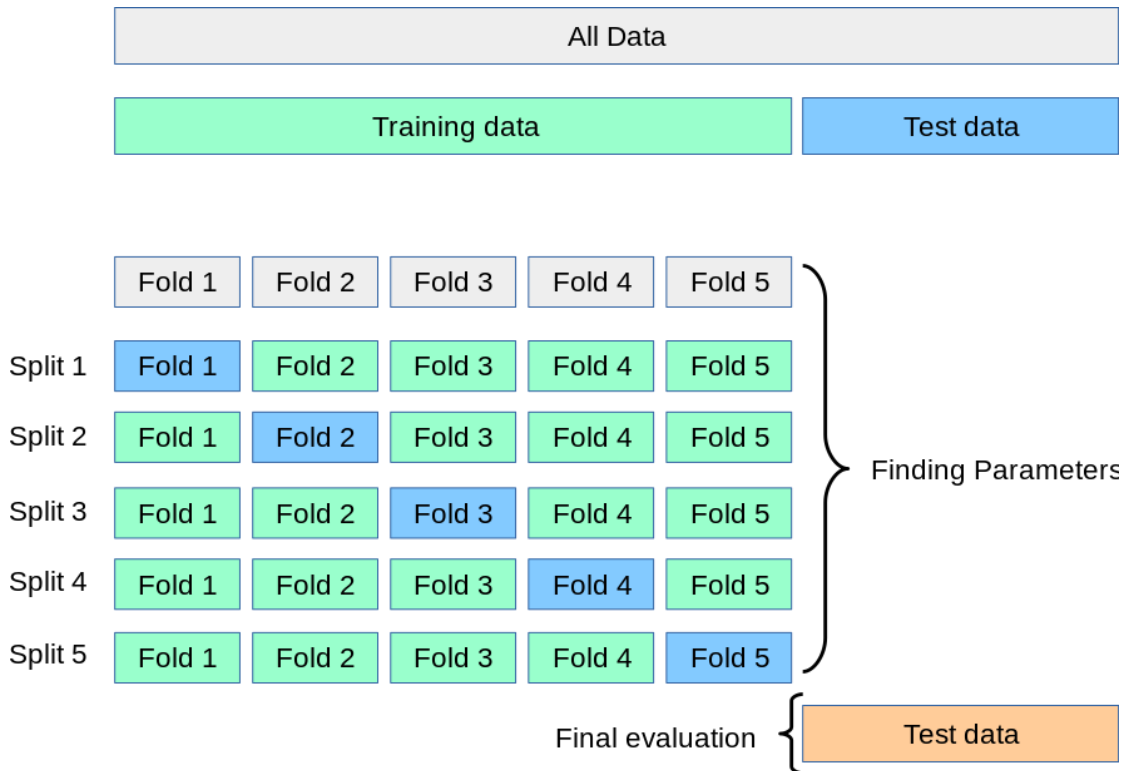
وكذلك لا تسمح بتجاوز التغيرات الصغيرة الواقعية من خلال اختيار مجالات طويلة، والذي يمكن أن

يؤدي إلى ضعف الأداء المعمم للنموذج على مجموعة الاختبار. [31]

الإجراء هو كما يوضح الشكل (5):

1. تخطط مجموعة البيانات عشوائياً، ثم يجري تقسيم مجموعة البيانات إلى مجموعات k (يشير k - fold إلى عدد الطيات التي تنقسم إليها مجموعة البيانات)، يمكن حساب قيمة k أو اختيارها، والتي عادة تكون قيمتها (5,10).

2. لكل مجموعة فريدة تعد مجموعة بيانات اختبار والمجموعات المتبقية كمجموعة بيانات تدريب، ثم يتم بناء النموذج على مجموعة بيانات التدريب وتقييمه على مجموعة الاختبار، ثم يجري تلخيص نتائج عملية التحقق المتبادل باستخدام k -fold بمتوسط نتائج النموذج جميعها وتعد متوسط النتيجة هي دقة النموذج.



الشكل (5) آلية التحقق المتبادل. [31]

7.3 التقنيات المستخدمة لمعالجة اختلال توازن البيانات

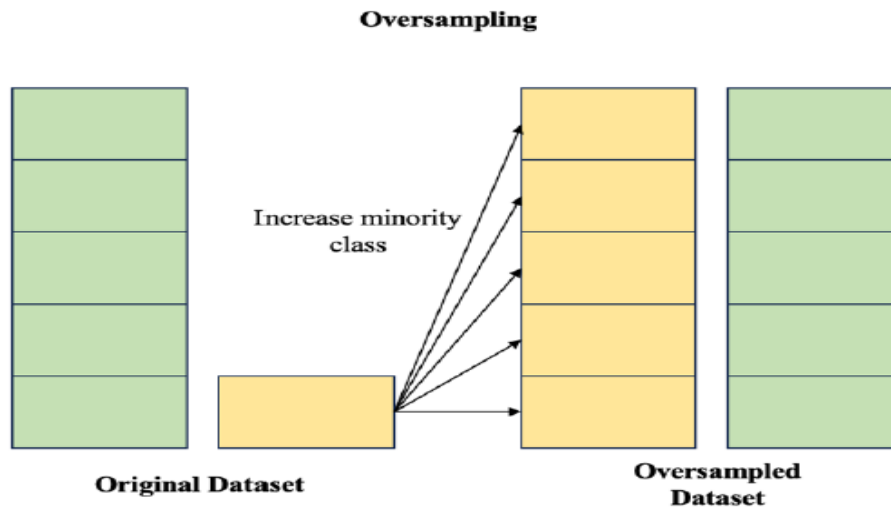
يحدث اختلال التوازن بين الفئات عندما تتفوق فئة واحدة في مشكلة التصنيف بشكل كبير على الفئة الأخرى، ويعد تحدياً شائعاً وحاسماً في مجال التعلم الآلي. حيث تُعد مجموعة البيانات غير المتوازنة مشكلة لأنها قد تؤدي إلى نماذج متحيزة وتوقعات غير دقيقة يمكن أن يؤدي إلى تحريف عملية تعلم النموذج لأنه قد يعطي الأولوية لفئة الأكثرية، مما يؤدي إلى ضعف الأداء في فئة الأقلية، مما يسبب حدوث أخطاء في التصنيف والتي من الممكن أن يكون لها تأثيرات سلبية كبيرة. [32]

ولمعالجة التحديات التي يفرضها اختلال توازن البيانات، تم تطوير تقنيات واستراتيجيات مختلفة، التي تهدف إلى إعادة توازن مجموعة البيانات وتعديل عملية تعلم النموذج، والحصول على أداء أفضل له، وأهم هذه التقنيات:

1.7.3 الزيادة العشوائية للعينات (Random Over Sampler)

في هذه الطريقة كما يوضح الشكل (6)، يتم تكرار الحالات العشوائية من فئة الأقلية حتى يتم تحقيق توزيع أكثر توازناً. وفي حين أن هذا يمكن أن يوازن توزيع الفئة، فقد يؤدي إلى الإفراط في التدريب.

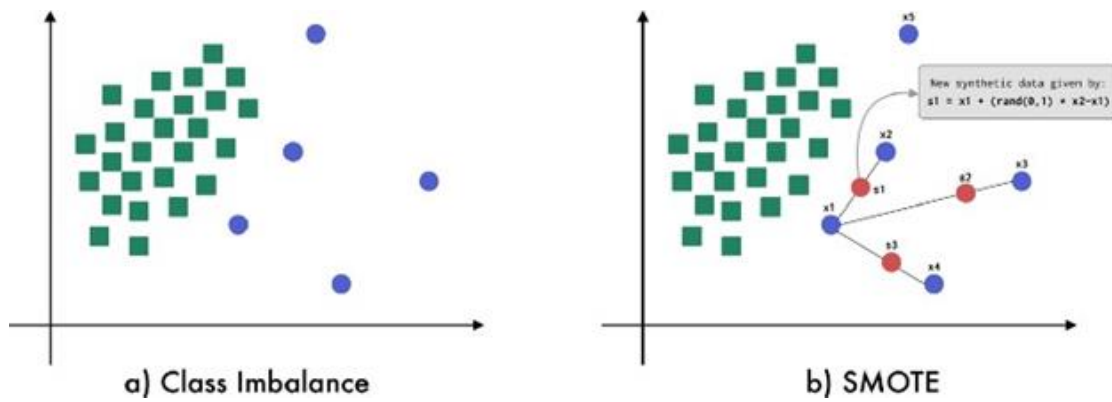
[32]



الشكل (6) آلية الزيادة العشوائية للعينات باستخدام تقنية ROS. [32]

2.7.3 الإفراط في أخذ العينات من الأقليات الاصطناعية (SMOTE)

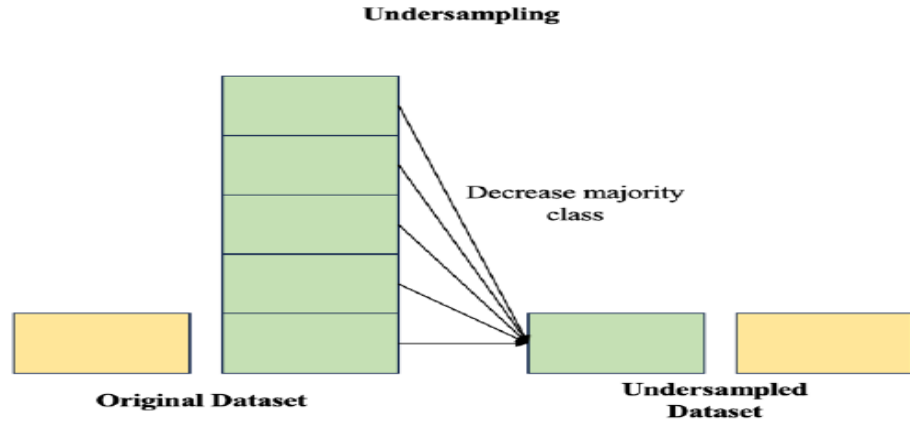
وهي طريقة رائدة في معالجة مجموعات البيانات غير المتوازنة، على عكس أساليب أخذ العينات الزائدة التقليدية، تعمل استراتيجية الزيادة هذه على تعزيز تنوع تمثيلات فئة الأقلية وتساعد في منع الإفراط في التدريب، وبالتالي تحسين أداء نماذج التعلم الآلي، وذلك من خلال توليد بيانات تركيبية (synthetic data) بين كل عينة من فئة الأقلية و k أقرب جيرانها (افتراضياً $k=5$))، حيث تُستخدم المسافة الإقليدية للعثور على أقرب الجيران، كما هو موضح في الشكل (7). [32]



الشكل (7) آلية الإفراط في أخذ العينات باستخدام تقنية SMOTE. [32]

3.7.3 الإنقاص العشوائي للعينات (Random Under Sampler(RUS))

في هذه الطريقة كما يوضح الشكل (8) يتم تحديد مجموعة فرعية عشوائية من الحالات من فئة الأغلبية والاحتفاظ بها، مما يقلل بشكل فعال من عدد حالات فئة الأغلبية، وبينما يمكن أن تساعد هذه التقنية في موازنة توزيع الفئة، فقد تؤدي إلى فقدان معلومات ذات قيمة محتملة من فئة الأغلبية ويواجه النموذج صعوبة في التعميم نظراً لمحدودية عدد الحالات في كل فئة. [32]



الشكل (8) آلية الإنقاص العشوائي للعينات باستخدام تقنية **RUS**. [32]

8.3 مقاييس الأداء :

تعد مقاييس الأداء مفيدة للغاية لتقييم ومقارنة نماذج تعلم الآلة المختلفة، واختبار قدرة المصنف وتحليل سلوك نفس النموذج عن طريق ضبط معاملاته الفائقة المختلفة، وتستند جميع المقاييس على مصفوفة الارتباك لأنها تحتوي على جميع المعلومات ذات الصلة حول أداء النموذج وقاعدة التصنيف، وأهم هذه المقاييس الدقة، الحساسية، الاحكام ودرجة $F1$ ، والمساحة تحت منحنى الخواص (Area Under Curve). [23]

1.8.3 مصفوفة الارتباك **Confusion Matrix**:

عبارة عن جدول متقاطع يسجل عدد مرات ظهور الحالات بين الأصناف الحقيقية بحسب مخرجات البيانات المعطاة للخوارزمية والأصناف التي قامت الخوارزمية بالتنبؤ بها، ويقع على القطر الرئيسي الحالات التي يتوافق فيها الصنف المتنبئ به مع الحقيقي. [23]

وبين الشكل (9) مصفوفة الارتباك لمشكلة ثنائية التصنيف:

		Predicted Condition	
		Positive	Negative
True Condition	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

الشكل (9) مصفوفة الارتباك ثنائية التصنيف. [23]

1. حيث True Positive (TP) هو عدد الإيجابيات الحقيقية الفعلية (أي الحالات الإيجابية الفعلية

التي تنبأت الخوارزمية بتصنيفها بشكل صحيح على أنها إيجابية، مما يعني أن التصنيف الذي قامت به الخوارزمية صحيح ومتطابق مع التصنيف الحقيقي).

2. True Negative (TN) هو عدد السلبيات الحقيقية الفعلية (أي الحالات السلبية الفعلية التي

تنبأت الخوارزمية بتصنيفها بشكل صحيح على أنها سلبية).

3. False Positive (FP) هو عدد الإيجابيات الخاطئة (تشير إلى الحالات السلبية الفعلية التي

تنبأت بها الخوارزمية بشكل خاطئ على أنها تنتمي إلى الصنف الإيجابي، بينما هي في الواقع

تنتمي إلى الصنف السلبي، مما يعني أن التصنيف الذي قامت به الخوارزمية غير متطابق مع الصنف الحقيقي).

4. False Negative (FN) هو عدد السلبيات الخاطئة (تعبّر عن عدد الحالات الإيجابية الفعلية

التي توقعها النموذج بشكل غير صحيح على أنها سلبية) [23]

2.8.3 الدقة Accuracy

هي مقياس أساسي في تقييم أداء نماذج التصنيف، وتمثل نسبة جميع التصنيفات الصحيحة الإيجابية والسلبية التي قام بها النموذج إلى إجمالي التصنيفات الإيجابية والسلبية الصحيحة والخاطئة التي قام بها. بمعنى آخر (هي نسبة الحالات السليمة والمريضة التي نجح النموذج بتصنيفها بشكل صحيح من جميع الحالات الحقيقية السليمة والمريضة)، تشير الدقة العالية إلى أن النموذج يحتوي على معدل منخفض من الإيجابيات الكاذبة.

ويتم حساب الدقة وفق المعادلة (4): [23]

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \dots\dots\dots(4)$$

3.8.3 الحساسية أو الاستدعاء Recall:

يقيس قدرة النموذج على تحديد العينات الإيجابية بشكل صحيح، وتمثل نسبة التنبؤات الإيجابية الصحيحة من إجمالي الحالات الإيجابية الحقيقية. تعتبر الحساسية مهمة عندما يكون فقدان الحالات الإيجابية الحرجة مشكلة، كما في تشخيص الأمراض.

ويتم حساب الحساسية وفق المعادلة (5): [23]

$$Recall = \frac{TP}{TP + FN} \quad \dots\dots\dots(5)$$

4.8.3 الإحكام Precision: نسبة التنبؤات الإيجابية الصحيحة من إجمالي التنبؤات الإيجابية

(الصحيحة والخاطئة). تعد الدقة النوعية مهمة عندما تكون الأخطاء الإيجابية الكاذبة مكلفة، كما في

اكتشاف الاحتيال وحسب وفق المعادلة (6): [23]

$$Precision = \frac{TP}{TP + FP} \quad \dots\dots\dots(6)$$

5.8.3 درجة F1:

هي المتوسط التوافقي للإحكام والحساسية، حيث يوازن بين المقياسين وهو مفيد عند التعامل مع بيانات

غير متوازنة. علاوة على ذلك يتم حساب درجة F1 كما هو موضح في المعادلة (3-7):[23]

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad \dots\dots\dots(7)$$

6.8.3 النوعية Specificity

يقيس قدرة النموذج على تحديد العينات السلبية بشكل صحيح، وتمثل نسبة التنبؤات السلبية الصحيحة

من إجمالي الحالات السلبية الحقيقية، ويعبر عنها وفق المعادلة (8):[23]

$$Specificity = \frac{TN}{TN + FP} \quad \dots\dots\dots(8)$$

7,8,3 منحنى ROC:

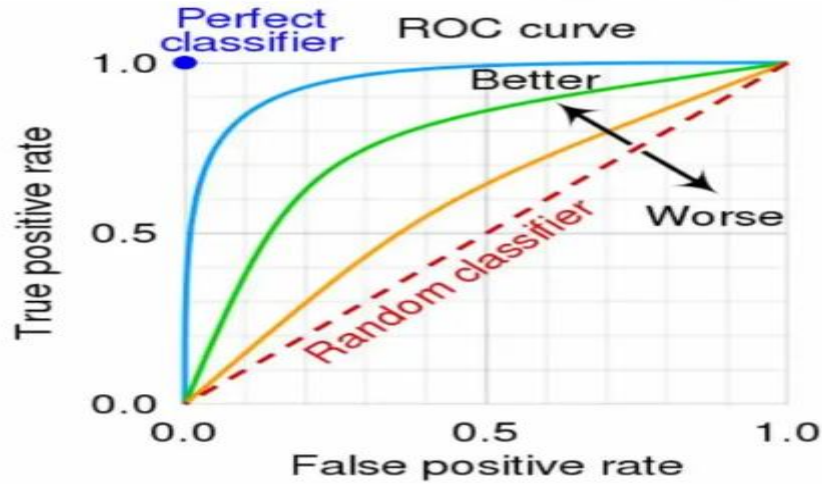
يحظى منحنى AUC_ROC المبين في الشكل (10) بشعبية كبيرة لأنه سهل الفهم نسبياً، فهو يوفر

رؤية شاملة لقدرة النموذج على التمييز بين الفئات، خاصةً في مهام التصنيف الثنائي، كما يُساعد في

فهم التوازنات بين الحساسية والنوعية عند عتبات قرار مختلفة.

فهو رسم بياني لمعدل القيم الإيجابية الصحيحة (TPR) التي تعرف بالحساسية مقابل معدل القيم الإيجابية الخاطئة (FPR). ويتميز بقدرته على تحديد قيمة P_value التي لها دلالة إحصائية، كما أنه قادر على تحديد نقطة القطع Cut-off التي تقابل أعلى قيمة للحساسية مع أعلى قيمة للPrecision، ويقاس هذا المنحنى (AUC التي تمثل المنطقة الواقعة تحت منحنى ROC) وهي مقياس لقدرة المصنف على التمييز بين الفئات الإيجابية والسلبية. كلما زادت المساحة تحت منحنى (AUC-ROC)، كان المصنف أفضل، المصنف الذي تكون فيها المساحة أعلى من 0.5 تشير إلى مقياس جيد لقابلية الفصل، بينما إذا كانت أقل من 0.5، فهي تشير إلى مقياس ضعيف في التمييز، المصنف المثالي تكون AUC تساوي 1.

إذا كان أداء النموذج جيدًا، فسيتم الحصول على مصنف جيد باختيار قيمة العتبة التي تجعل TPR قريبًا من 1 مع إبقاء FPR قريبًا من 0. [33]



الشكل (10) منحنى ROC. [33]

TPR هي نسبة الحالات الإيجابية التي تم تصنيفها بشكل صحيح. بينما FPR هي نسبة الحالات السلبية التي تم تصنيفها بشكل غير صحيح على أنها إيجابية والمبينتين في العلاقتين التاليتين (9).

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN} \quad \dots\dots\dots(9)$$

9.3 التحليل الإحصائي:

يُعد استعمال التحليلات الإحصائية أمراً بالغ الأهمية لفهم البيانات البحثية، فهي تساعد الباحثين في العثور على الاختلافات المهمة بين المجموعات، والاختبارات الإحصائية التي استعملت في هذه الدراسة:

1.9.3 اختبار t-test

اختبار T هو أداة إحصائية تستخدم لتحديد دلالة الفروق بين متوسطين لمجموعتين. يفحص ما إذا كان هذا الاختلاف يعتد به إحصائياً أم لا. يستعمل في حال ضيق عينة الدراسة أو عدم توفر معلومات عن التباين، ويعد هذا الاختبار أداة أساسية في البحوث العلمية، فهو يساعد في اتخاذ قرارات مبنية على نتائج إحصائية واضحة. [26]

اختبار T لعينتين مستقلتين Two Sample T-Test

يُحدد ما إذا كان هناك فرق يعتد به إحصائياً بين المتوسطات في مجموعتين غير مرتبطتين أي (إذا كان الفرق بين متوسطي العينتين المستقلتين يختلف اختلافاً معنوياً (حقيقياً) عن الفرق بين متوسطي المجتمع). [26]

شروط الاختبار:

1. ان يكون المتغير المستقل اسمي مرتب Ordinal.

2. يجب أن تكون العينتين مختارة بشكل عشوائي.

3. يجب أن تكون العينتين مستقلتين.

فرضيات الاختبار:

- الفرضية الصفرية: عدم وجود فرق بين متوسطي المجموعتين.
- الفرضية البديلة: وجود فرق بين متوسطي المجموعتين.
- يمكن صياغة قيمة T الإحصائية بالشكل التالي كما هو مبين في المعادلة (10):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}} \dots\dots\dots(10)$$

T: قيمة تي الإحصائية.

$\bar{x}_1, \bar{x}_2$: المتوسط الحسابي للمجموعتين الأولى والثانية.

$S(\bar{x}_1 - \bar{x}_2)$: الخطأ المعياري للفرق بين المتوسطات.

يتم حساب الخطأ المعياري للفرق بين متوسطات العينة بواسطة المعادلة (3-11): [26]

$$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \dots\dots\dots(11)$$

حيث:

S_1, S_2 : الانحراف المعياري للمجموعتين.

(n_1, n_2) : عدد العينات في المجموعتين.

2.9.3 اختبار مربع كاي chi-square test

يستعمل اختبار مربع كاي في تحليل البيانات النوعية، حيث تكون المتغيرات مصنفة ومقاسة بمقياس اسمي، وهو اختبار يستعمل للموازنة بين التوزيعات التكرارية للمتغيرات، ويصلح لمعالجة البيانات النوعية التي تكون على شكل تكرارات لمجموعات أو أصناف معينة، ويستعمل لتحديد وجود علاقة أو ارتباط بين متغيرين مصنفيين. [28]

ويجب أن يكون التوزيع الفعلي للتكرارات كما يلي:

1. أن تكون البيانات على شكل تكرارات وليست نسبا مئوية أو كسور.
2. ألا يقل مجموع التكرارات الفعلية عن 20 تكرارا ويفضل أن يزيد عددها عن 40 تكرارا.
3. ألا يقل مجموع التكرارات المتوقعة في أي فئة من فئات التصنيف عن خمسة تكرارات.

خطوات اختبار مربع كاي:

1. صياغة الفرضية الصفرية والفرضية البديلة، وتحديد مستوى الدلالة.
2. التحقق من افتراضات الاختبار.
3. إيجاد إحصائية الاختبار وذلك من خلال الخطوات التالية:
4. جمع البيانات وتنظيمها في جدول (وهو ترتيب يتم فيه تصنيف البيانات وفقاً لمتغيرين فئويين، وتظهر فئات أحد المتغيرين في الصفوف، وتظهر فئات المتغير الآخر في الأعمدة، ويجب أن يحتوي كل متغير على فئتين أو أكثر ويتم حساب حجم العينة، ومجموع الصفوف، ومجموع الأعمدة)، حيث تمثل هذه البيانات التكرارات الملاحظة.
5. حساب التكرارات المتوقعة من خلال ضرب احتمالية وجود تلك الخلية بالمجموع الإجمالي للبيانات.
6. حساب القيمة الاحصائية لاختبار مربع كاي من خلال حساب الفروق بين التكرارات المتوقعة والتكرارات الملاحظة، ومن ثم تربيعها وجمعها، ثم نقسمها على إجمالي التكرارات المتوقعة.

7. تحديد القيمة الجدولية لكاي مربع (وهي القيمة التي تتقاطع فيها درجة الحرية مع مستوى الدلالة)

من خلال حساب درجة الحرية من الصيغة التالي:

$$\text{درجة الحرية} = (\text{عدد الصفوف} - 1) * (\text{عدد الأعمدة} - 1)$$

8. تفسير النتائج من خلال المقارنة بين قيمة كاي مربع الجدولية والقيمة المحسوبة:

- إذا كانت القيمة المحسوبة أصغر من القيمة الجدولية نقبل الفرضية الصفرية أي (لا يوجد علاقة أو ارتباط بين المتغيرين الفئويين).

- إذا كانت القيمة المحسوبة أكبر من القيمة الجدولية نرفض الفرضية الصفرية أي (يوجد علاقة أو

ارتباط بين المتغيرين الفئويين). [28]

فرضيات اختبار مربع كاي:

الفرضية الصفرية: لا يوجد علاقة بين أو ارتباط بين متغيرين فئويين.

الفرضية البديلة: يوجد علاقة أو ارتباط بين متغيرين فئويين.

يتم تمثيل صيغة مربع كاي بالمعادلة (12):

$$\chi^2 = \sum \frac{(O - E)^2}{E} \dots\dots\dots(12)$$

O : التكرارات التجريبية التي تعتمد على العينة المدروسة.

E: التكرارات النظرية التي تعتمد على تابع التوزيع.

χ^2 : قيمة مربع كاي الإحصائية. [28]

3.9.3 اختبار مان-ويتني (Mann-Whitney U test)

هو اختبار إحصائي غير وسيطي (لا بارامتري) يستعمل من أجل مقارنة وسيطين عندما تكون المتغيرات المدروسة لا تتوزع توزيعاً طبيعياً، وهو بديل فعال لاختبار t المستقل، ويمكن استخدامه بشكل مناسب مع المتغيرات المتصلة والمتقطعة ويستخدم بشكل فعال مع العينات الصغيرة، حيث يعد هذا الاختبار من أدوات التحليل الإحصائي القوية التي لا تتطلب فرضيات توزيع محددة، مما يجعله خياراً مثالياً عند التعامل مع بيانات غير متوزعة طبيعياً، أو في حال كانت البيانات غير متوازنة أو غير متماثلة. [27]

تُصاغ فرضيات اختبار مان-ويتني على النحو التالي:

الفرضية الصفرية: لا يوجد فرق بين المجموعتين ($H_1 = H_2$).

الفرضية البديلة: هذا فرق بين المجموعتين ($H_1 \neq H_2$). [27]

خطوات إجراء اختبار مان ويتني:

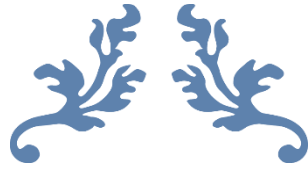
1. تجميع البيانات وتقسيمها إلى مجموعتين مستقلتين،
2. ترتيب جميع البيانات من الأصغر إلى الأكبر.
3. تحديد الرتب: اعط كل قيمة في البيانات رتبة بناءً على ترتيبها.
4. جمع رتب إحدى المجموعتين: اجمع رتب نقاط البيانات من إحدى المجموعتين.
5. حساب إحصائية U ، ثم استخدام الجداول الإحصائية لاختبار مان-ويتني U لإيجاد احتمالية ملاحظة قيمة U الأقل.
6. تفسير النتائج: عند إجراء اختبار مان ويتني، ستحصل على قيمة إحصائية تُعرف بـ " U " وقيمة p . قيمة " p " تساعد في تحديد ما إذا كانت الفروق بين المجموعتين معند به إحصائياً أم لا

نظام تشخيص آلي لمرض السكري باستخدام خوارزميات التصنيف، صالح.

إذا كانت قيمة p أقل من مستوى الأهمية (عادة 0.05)، ترفض الفرضية الصفرية، وتقبل الفرضية البديلة أي أن الفرق بين المجموعتين يعتد به إحصائياً. إذا كانت قيمة p أكبر من 0.05، فإنه لا يوجد دليل على وجود فرق دال بين المجموعتين. [27]

10.3 خاتمة:

جرى في هذا الفصل تغطية الجانب النظري للبحث، حيث استعرضت المعلومات المهمة عن تعلم الآلة والأسس النظرية للنماذج التي جرى استخدامها في هذه الدراسة، وكيفية إيجاد مقاييس تقييم أدائها، بالإضافة إلى الاختبارات الإحصائية وخوارزمية اختيار السمات التي استخدمت لتحديد عوامل الخطورة للتنبؤ بالسكري.



الفصل الرابع

أدبيات البحث



1.4 مقدمة:

جذب انتشار مرض السكري من النوع الثاني اهتمام الباحثين في جامعات دول العالم، والذين يعملون على تطوير منهجيات ذات أداء عالي للتشخيص والعلاج وتقديم خطط للرعاية الصحية، وذلك انطلاقاً من أهمية الكشف المبكر عن السكري الذي من الممكن أن يكون مفتاح العلاج الفعال، أو على الأقل رفع سوية حياة المرضى، ويستعرض هذا الفصل أهم الدراسات المرجعية.

2.4 الدراسات المرجعية:

1.2.4 استخدام طرق تعلم الآلة للتنبؤ عن مرض السكري باستخدام قاعدة البيانات PID عند عدم موازنة البيانات:

هناك العديد من الدراسات التي استخدمت قاعدة البيانات (PID) للكشف أو التنبؤ عن مرض السكري، حيث استخدمت هذه الدراسات خوارزميات تعلم الآلة المختلفة للكشف عن مرض السكري، وبشكل خاص خوارزمية الانحدار اللوجستي وخوارزمية الغابة العشوائية كما في الدراسات [29-34-35-36-37]، حيث تراوحت دقة الكشف باستخدام خوارزمية الانحدار اللوجستي في هذه الدراسات بين (73%-83.1%)، بينما تراوحت دقة الكشف عن مرض السكري باستخدام خوارزمية الغابة العشوائية بين (73.38%-85.53%).

وفيما يلي سنتحدث عن أهم هذه الدراسات:

في دراسة تهدف إلى التنبؤ بمرض السكري باستخدام قاعدة البيانات PID، قام Guptaa وزملاؤه عام 2023 على مقارنة تحليل أداء خوارزميات التصنيف المستخدمة في التنبؤ، وتعزيز أداء المصنفات الوظيفي من خلال اختيار طريقة مناسبة لمعالجة البيانات وضبط معلمات كل نموذج، حيث تم

استخدام أربع خوارزميات للتصنيف (KNN-DT-SVM-RF)

في مرحلة ما قبل المعالجة من منهجية البحث استخدم الباحثون ثلاث طرق لمعالجة البيانات الأولى هي ملء القيم المفقودة بالمتوسط الحسابي لكل قيم العامود، والطريقة الثانية إزالة كل الأسطر التي تحوي على القيم المفقودة، والطريقة الثالثة إزالة كل الأسطر التي تحوي على القيم المفقودة والقيم المتطرفة، ثم جرى تقسيم البيانات بنسبة 80% من البيانات كبيانات لتدريب النموذج، 20% من البيانات كبيانات لاختبار النموذج، ثم تم ضبط المعاملات الخاصة بكل خوارزمية تصنيف، ثم تدريبها واختبارها وتقييم أدائها، واختيار أفضلها. [34]

الجدول (1) نتائج الدقة ودرجة F1 لدراسة Gupta. [34]

طريقة المعالجة المسبقة	خوارزميات التصنيف	F1_score	Accuracy
حذف أسطر القيم المفقودة	RF	75.68%	88.61%
	DT	75.56%	86.1%
	SVM	75%	84.8%
	KNN	61.53%	81.01%
حذف أسطر القيم المفقودة والمتطرفة	RF	58%	74%
	DT	61.5%	71%
	SVM	62.5%	82.35%
	KNN	61%	73%
استبدال القيم المفقودة بقيمة المتوسط الحسابي	RF	73.39%	81.17%
	DT	68%	79%
	SVM	56%	77%
	KNN	67.5%	78%

عند المقارنة بين أداء نماذج التصنيف كما هو موضح في الجدول (1) باستخدام طرق المعالجة

المسبقة تم اختيار نموذج مجموعة البيانات عند حذف القيم المفقودة كأفضل نموذج تتبؤ مقارنة بطرق

نظام تشخيص آلي لمرض السكري باستخدام خوارزميات التصنيف، صالح.

المعالجة الأخرى، لأن جميع خوارزميات التصنيف أبدت أداءً جيداً.

حيث تبين أن خوارزمية الغابة العشوائية هي الأفضل من حيث الأداء مقارنة مع خوارزميات التصنيف الأخرى، وذلك في حالتي المعالجة المسبقة للبيانات (ملء القيم المفقودة بالمتوسط الحسابي وإزالة القيم المفقودة)، لكن كان أداء خوارزمية الغابة العشوائية الأعلى عند إزالة القيم المفقودة حيث حققت أعلى دقة مقارنة مع جميع خوارزميات التصنيف والتي تساوي 88.61%، بينما كان أداء خوارزمية آلة أشعة الدعم الأفضل عند إزالة القيم المفقودة والقيم المتطرفة، حيث حققت دقة 82.35% [34].

قامت هذه الدراسة بالمقارنة بين خوارزميات التصنيف بعد استخدام عدة طرق للمعالجة المسبقة، ولكن لم تقم باستخدام تقنيات لتوزيع البيانات لمنع انحياز النماذج.

وفي دراسة أخرى قام Al-Ahmed وزملاؤه في عام 2023 بإظهار أهمية ضبط معاملات خوارزميات تعلم الآلة لتحسين أداء نماذج التصنيف للكشف عن مرض السكري، من خلال استخدام ثلاث خوارزميات للتصنيف وهي (SVM، LR، RF).

تتكون منهجية البحث من عدة خطوات وهي:

1. تحميل قاعدة البيانات البيما
2. استكشاف البيانات عن طريق إنشاء تمثيلات رسومية للبيانات لتسهيل فهم وتفسير مجموعة بيانات بيما الهندية لمرض السكري (PIDD).
3. المعالجة المسبقة للبيانات: من خلال التحقق من القيم المفقودة، ومعالجة هذه القيم المفقودة عن طريق احتساب القيمة المتوسطة للعمود، ثم حذف القيم المتطرفة والمكررة، ثم توحيد نطاق البيانات.

4. ضبط المعاملات الفائقة لخوارزميات التصنيف باستخدام طريقة البحث الشبكي للحصول على أفضل نموذج.

5. تقييم أداء النماذج باستخدام معايير الأداء ومقارنة النتائج وتحديد النموذج الأفضل.

تشير النتائج كما هو موضح في الجدول (2) التي توصل إليها الباحثون إلى أن نموذج الانحدار اللوجستي هو الأفضل في الأداء، كما تحسن أداء هذه النماذج بعد ضبط المعاملات الفائقة، حيث كانت دقة خوارزمية الانحدار اللوجستي 79%، ودقة خوارزمية الغابة العشوائية 77% [29]

الجدول (2) نتائج الدقة ودرجة F1 لدراسة AL-Ahmed [29]

خوارزمية التصنيف	F1_score	Accuracy
RF	74%	77%
SVM	74%	78%
LR	74%	79%

في هذه الدراسة لم يتم استخدام تقنية لتوزيع البيانات، ينجم عنه أن النموذج يعاني من مشكلة التحيز لفئة الأكثرية، وقد يكون هناك حاجة لتغيير مجالات المعاملات الفائقة لتحسين أداء النماذج.

وفي دراسة أخرى استخدم Al_Zain وزملاؤه في عام 2024 أربع خوارزميات للتصنيف (RF،

GB، LR، SVM) للتعليق عن مرض السكري. [35]

في البداية تم تحميل قاعدة البيانات، ثم تم تقسيم البيانات بنسبة 70% من البيانات لتشكيل مجموعة بيانات التدريب، 30% من البيانات لتشكيل مجموعة بيانات الاختبار، ثم جرى بناء النماذج باستخدام

خوارزميات التصنيف الأربعة بدون معالجة مسبقة للبيانات وتقييم أدائها باستخدام معايير الأداء، وبعد ذلك، تم إجراء طرق المعالجة المسبقة (التوحيد القياسي للبيانات، وملء البيانات المفقودة) على مجموعة البيانات، ثم بناء النماذج باستخدام خوارزميات التصنيف نفسها وتقييم أدائها باستخدام معايير الأداء مرة أخرى، ثم مقارنة النتائج للنماذج قبل وبعد المعالجة المسبقة للبيانات.[35]

من خلال مقارنة قيم النتائج كما هو موضح في الجدول (3) التي تم الحصول عليها قبل وبعد المعالجة المسبقة للبيانات كان هناك تحسناً، حيث حقق مصنف الانحدار اللوجستي والغابات العشوائية أفضل النتائج بدقة (83.1%)

الجدول (3) نتائج الدقة ودرجة F1 في دراسة Al-Zain. [35]

خوارزميات التصنيف	استخدام المعالجة المسبقة للبيانات	F1_score	Accuracy
RF	لا	57.1%	76.61%
RF	نعم	72.9%	83.1%
LR	لا	58.1 %	74.1%
LR	نعم	72.3%	83.1%
SVM	لا	53.6%	72%
	نعم	58.8 %	77.2%
GB	لا	54.7%	69.4%
	نعم	61.5%	79.8%

في هذه الدراسة لم يتم موازنة البيانات، كما لم تصرح عن طريقة المعالجة المسبقة للبيانات بشكل واضح، ولم يتم أيضاً استخدام ضبط المعاملات الفائقة التي تعزز من أداء النماذج.

2.2.4 استخدام طرق تعلم الآلة عند توزيع البيانات للتنبؤ عن السكري باستخدام قاعدة البيانات

PID:

في دراسة Mushtaq وزملاؤه في 2022 جرى التنبؤ عن السكري باستخدام خوارزميات التعلم الآلي عند استخدام تقنيتي توزيع البيانات (TOMEK،SMOTE) .

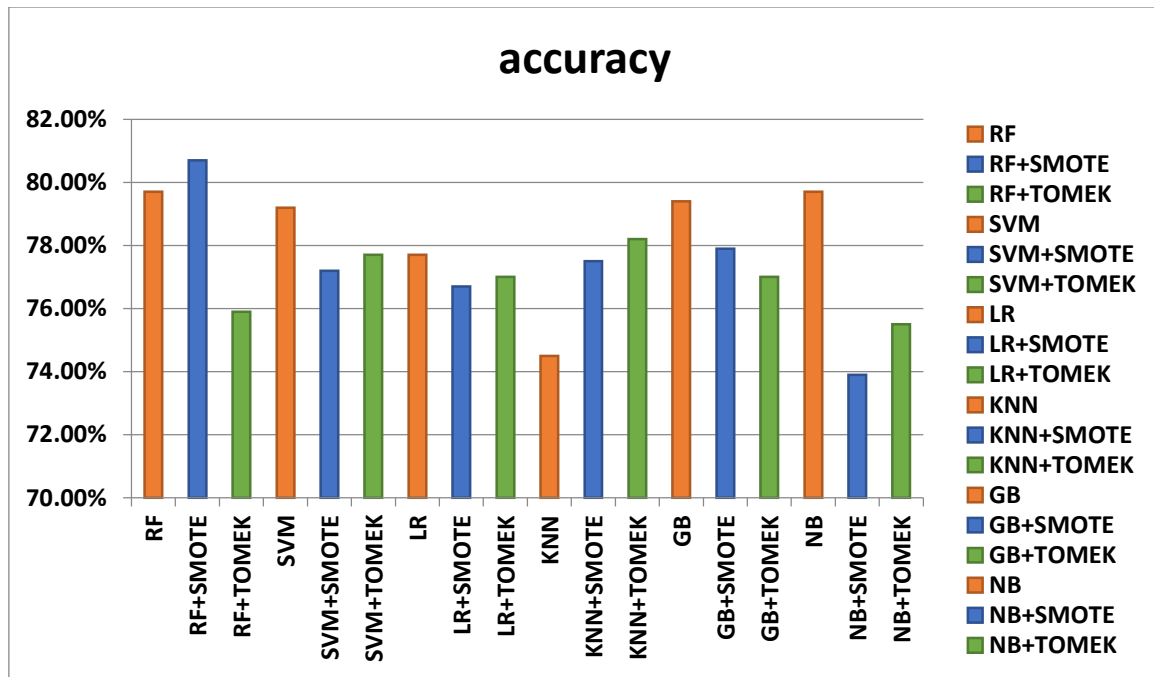
حيث تتضمن منهجية البحث المتبعة تحميل بيانات السكري PID واستكشاف البيانات، ثم الكشف عن القيم المتطرفة (باستخدام النطاق الربيعي IQR والمخطط الصندوقي Boxplot)، وحذفها، وتوزيع البيانات باستخدام تقنيتين أحدهما لزيادة عدد العينات الأقلية (SMOTE) وأخرى لخفض عدد العينات الأكثرية (Tomek)، كما تم التحقق من صحة النتائج باستخدام التحقق المتبادل K_fold cross validation ، واستخدم الباحثون ستة خوارزميات وهي (Naïve Bayes ،KNN ،RF،LR)، وبعد تدريب النماذج واختبارها تم تقييم أدائها واختيار النموذج الأفضل، يوضح الجدول (4) نتائج معايير الأداء لهذه الدراسة.

الجدول (4) نتائج الدقة والمساحة تحت منحنى الخواص في دراسةMustaq[36].

Accuracy	AUC	خوارزمية التصنيف
79.7%	0.73	RF
80.7%	0.80	RF+SMOTE
75.9%	0.69	RF+TOMEK
79.2%	0.71	SVM
77.2%	0.70	SVM+SMOTE
77. 7%	0.70	SVM+TOMEK
77. 7%	0.68	LR
76. 7%	0.76	LR+SMOTE
77 %	0.69	LR+TOMEK
74.5%	0.66	KNN
77.5%	0.78	KNN+SMOTE
78.2%	0.73	KNN+TOMEK
79.4%	0.71	GB
77.9%	0.78	GB+SMOTE

77 %	0.72	GB+TOMEK
79.7%	0.73	NB
73.9%	0.74	NB+SMOTE
75.5%	0.72	NB+TOMEK

عند المقارنة بين نتائج هذه الدراسة قبل موازنة البيانات وبعد استخدام تقنيتي التوزين (SMOTE)،
(TOMEK) كما هو مبين في الشكل (11) تبين أن:



الشكل (11) قيمة الدقة قبل وبعد موازنة البيانات في دراسة [36]. Mustaq

1. أفضل أداء لخوارزمية الغابة العشوائية قبل موازنة البيانات حيث كانت الدقة تساوي 79.7%، وأسوأ أداء لخوارزمية الجار الأقرب حيث كانت الدقة تساوي 74.5%.
2. نموذج الغابة العشوائية، والجار الأقرب قد تحسن أدائهما بعد زيادة عدد البيانات باستخدام تقنية (SMOTE)، لكن كان أداء نموذج الغابة العشوائية أفضل، حيث حقق دقة 80.7%، بينما تراجع أداء بقية الخوارزميات، والأداء الأسوأ كان لخوارزمية NB
3. تراجع أداء جميع النماذج في حال استخدام تقنية إنقاص البيانات (TOMEK)، باستثناء خوارزمية

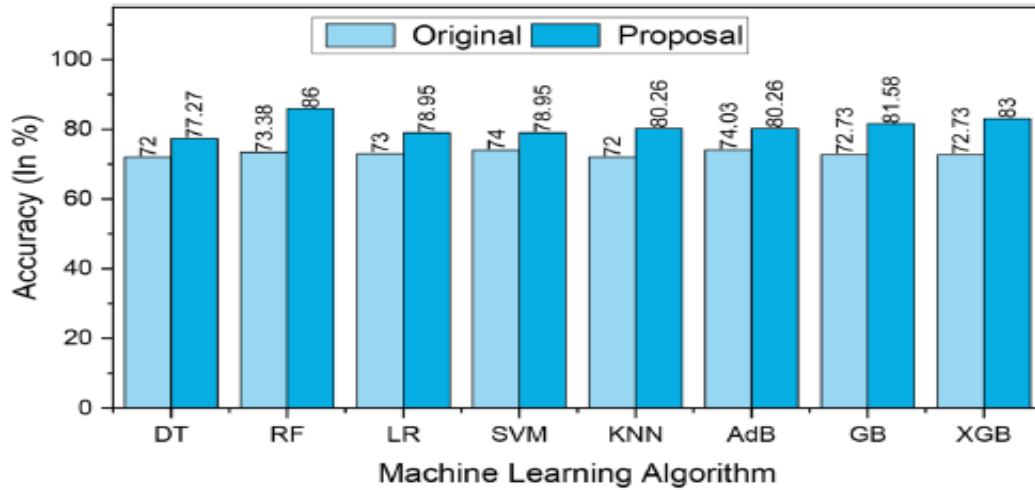
الجار الأقرب التي تحسن أدائها حيث حققت أفضل دقة مقارنة مع بقية خوارزميات التصنيف والتي تساوي 78.2%. [36]

تعد نتائج هذه الدراسة مشجعة لاستخدام تقنيات توزيع البيانات (زيادة البيانات باستخدام SMOTE)، لكن الباحثين لم يصرحوا بالمنهجية المتبعة لمعالجة القيم المفقودة التي قد تؤثر سلباً على أداء نماذج التصنيف.

في دراسة Talukder وزملاؤه في عام 2024 تم التنبؤ عن السكري باستخدام خوارزميات التعلم الآلي عند استخدام تقنية توزيع العينات (ROS).

حيث تتضمن منهجية البحث المتبعة تحميل بيانات مرض السكري PIMA واستكشاف البيانات، ثم الكشف عن القيم المتطرفة (باستخدام النطاق الربيعي IQR والمخطط الصندوقي Boxplot) وحذفها، واستبدال القيم المفقودة بالمتوسط الحسابي، ثم جرى استخدام التوحيد القياسي للبيانات وتوزيع البيانات باستخدام تقنية الزيادة العشوائية للبيانات (ROS)، كما جرى التحقق من صحة النتائج باستخدام التحقق المتبادل K_fold cross validation، واستخدم الباحثون ستة خوارزميات وهي (KNN، RF، LR، Gradient Boost، DT، SVM)، وبعد ضبط معاملات النماذج تم تدريبها واختبارها، ثم تقييم أدائها واختيار النموذج الأفضل.

حيث وجد وفقاً لنتائج هذه الدراسة الموضحة في الشكل (12) أن جميع نماذج التصنيف قد تحسن أدائها بعد توزيع البيانات باستخدام تقنية الزيادة العشوائية للبيانات، وكان نموذج الغابة العشوائية كما هو مبين في الجدول (5) هو الأفضل، حيث حقق دقة 85.53%، بينما كانت خوارزمية شجرة القرار هي الأقل أداءاً مقارنة ببقية خوارزميات التصنيف بعد موازنة البيانات وذلك بدقة تساوي 77.27%. [37]



الشكل (12) مقارنة بين دقة النماذج قبل وبعد التوزيع باستخدام ROS. [37]

الجدول (5) نتائج الدقة ودرجة F1 لنماذج التصنيف في دراسة Talukder. [37]

Accuracy	F1-Score	خوارزمية التصنيف
73.38%	70.89%	RF
86%	85.99%	RF+ROS
72%	71.96%	DT
77.27%	75.72%	DT+ROS
73%	73%	LR
78.95%	75.06%	LR+ROS
74%	73.96%	SVM
78.95%	75.06%	SVM+ROS
72%	71.99%	KNN
80.26%	75.68%	KNN+ROS
74.03%	71.48%	Ada Boost
80.26%	77.87%	Ada Boost+ROS
72.73%	70.54%	GB
81.58%	79.89%	GB+ROS
72.73%	70.75%	XGBoost
83%	83%	XGBoost+ROS

تعد نتائج هذه الدراسة مشجعة أيضا لاستخدام تقنية توزيع البيانات (ROS).

3.2.4 استخدام طرق تعلم الآلة للتنبؤ عن مرض السكري باستخدام خوارزميات تقليل السمات

اعتمادا على قواعد بيانات مختلفة

عمل Viswanatha وزملاؤه في عام 2023 على تصميم نموذج يتنبأ فيما إذا كان الشخص

سيصاب بالسكري، بناءً على بعض التدابير التشخيصية الواردة في مجموعة البيانات واستكشاف تقنيات

مختلفة لتحسين الأداء والدقة، حيث تم جمع البيانات من قاعدة البيانات بيما وقاعدة فاندر بيلت التي

تتكون من 162 رجلا و228 امرأة، منهم 60 شخصا مصابا بالسكري، 330 شخص غير مصاب

بالسكري، وتحتوي القاعدة على 13 سمة، إحداها (الكوليسترول الكلي، تركيز سكر الدم عند الصيام،

الكوليسترول الجيد HDL، العمر، الطول، الوزن، مؤشر كتلة الجسم، ضغط الدم الانقباضي، ضغط

الدم الانبساطي، الجنس، محيط الخصر، محيط الورك، نسبة محيط الخصر إلى الورك). [38]

منهجية البحث:

جرت إزالة القيم المفقودة، ثم التوحيد القياسي للبيانات، ولتدريب النموذج واختباره، قسمت البيانات بنسبة

70:30 لمجموعتي التدريب والاختبار، تم استخدام خوارزمية اختيار السمات أحادي المتغير

univariate feature selection مع اختبار chi-square وطبق على قاعدة فاندربيلت لاختيار

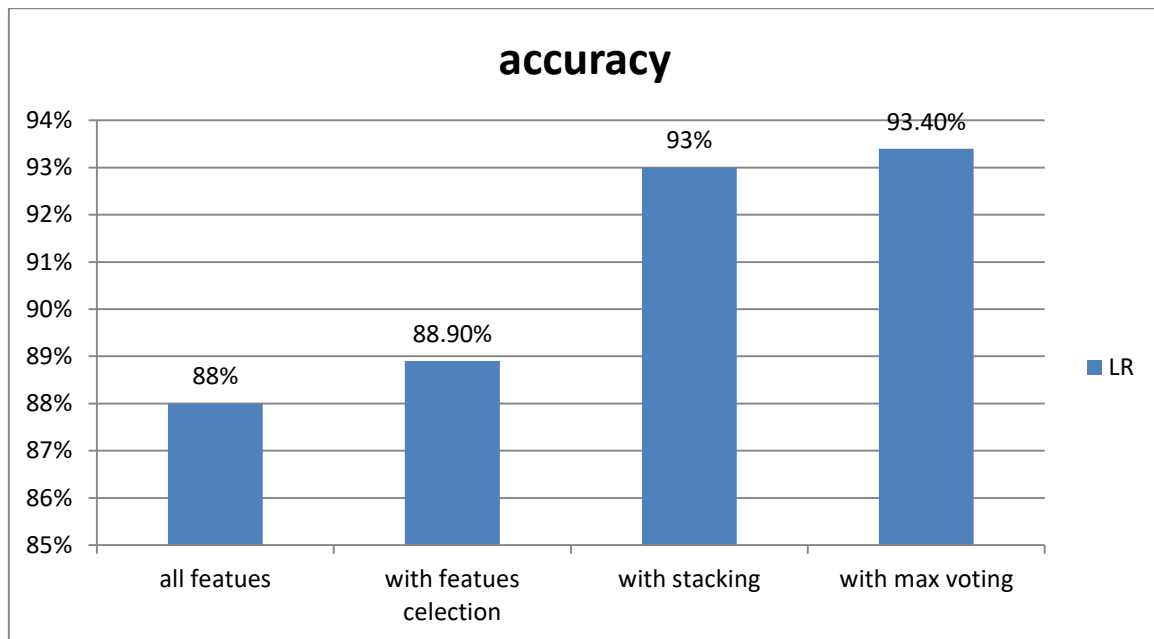
أفضل السمات، حيث تُقارن كل سمة بالمتغير المستهدف لتحديد أي علاقات معتمد بها إحصائياً بينهما،

وتم استخلاص ثماني سمات من 13 سمة، ثم طبقت خوارزمية الانحدار اللوجستي LR للتنبؤ بالسكري،

ولتحسين أداء النموذج جرى استخدام تقنيتي التجميع: الحد الأقصى للتصويت (max voting)، وتقنية

التراص (stacking)، وطبق النموذج على القاعدتين السابقتين، وكانت النتائج عند استخدام قاعدة

البيانات فاندربيلت كما في الشكل (13): [38]



الشكل (13) نتائج دراسة [38].Viswanatha

كانت دقة النموذج عند عدم استخدام خوارزمية اختيار السمات الأحادية وعدد المتغيرات 13 تساوي

88%، عند استخدام خوارزمية اختيار السمات تم تقليل عدد السمات إلى ثماني سمات حيث تحسن

أداء الانحدار اللوجستي وقد كانت الدقة تساوي 88.8 %، وعند استخدام الانحدار اللوجستي مع

استخدام خوارزمية اختيار السمات وتقنية التكديس كانت أفضل قيمة للنموذج وتساوي 93.4% [38].

في هذه الدراسة لم يتم تحديد الميزات المختارة بعد تطبيق خوارزمية اختيار السمات، كما لم يتم

معالجة القيم المتطرفة والتي قد تؤثر على أداء النموذج.

أما بالنسبة لدراسة Kakoly وزملاؤه في عام 2023 أجرى التنبؤ بعوامل خطر الإصابة بمرض

السكري من خلال تطبيق خوارزميات التعلم الآلي، ثم استخدمت عوامل الخطورة المتنبأ بها في التنبؤ

بالسكري. [39]

جمعت هذه الدراسة بيانات مع استبيان مسحي من 738 مشاركاً من مختلف المناطق الحضرية والريفية في بنغلاديش، مع تحديد الفئة العمرية بين (19-78) عاماً، يوجد منهم (256) شخصاً مصاباً بالسكري متوسط أعمارهم 50 عاماً و 483 شخصاً غير مصاباً بالسكري متوسط أعمارهم 42 عاماً.

كانت أسئلة الاستطلاع تغطي العوامل الديموغرافية والسريية وغير السريية، حيث جرى اختيار 18 سمة في البداية. [39]

تفاصيل البيانات:

- 1- عوامل ديموغرافية: وتشمل (العمر، الجنس، مستوى التعليم، الحالة المادية)
- 2- عوامل تتعلق بجودة الحياة والعادات الغذائية: وتشمل (النشاط الفيزيائي، ضغوط العمل، تناول الخضار والفواكه واللحوم والأسماك والمشروبات الغازية)
- 3- عوامل تتعلق ببيانات سريرية: وتشمل (ضغط الدم الانقباضي، ضغط الدم الانبساطي، مؤشر كتلة الجسم، مستوى السكر العشوائي، عدد مرات الحمل، التاريخ العائلي للإصابة بالسكري)

منهجية البحث:

تم إجراء منهجية البحث في أربع خطوات:

- 1- المعالجة المسبقة للبيانات: تم حذف جميع القيم المفقودة والمتطرفة ثم استُخدم MinMaxScaler لتوحيد قيم البيانات، ثم جرى تقسيم مجموعة البيانات بنسبة 80:20 حيث تم استخدام 80% من البيانات لمجموعة التدريب وتم استخدام بيانات 20% المتبقية لمجموعة الاختبار.
- 2- اختيار السمات: تم استخراج السمات باستخدام تقنية تحليل المكونات الرئيسية (PCA) واستخراج السمات باستخدام تقنية نسبة الكسب (IG). وذلك لتحقيق دقة أعلى في التنبؤ.

3- تدريب نموذج التصنيف مع تحسين المعاملات الفائقة: تم استخدام خمس خوارزميات تصنيف

(LR، KNN، RF، DT، SVM)، والتي تم تدريبها والتحقق من صحتها بعد ضبط معاملاتها

الفائقة مع التحقق المتقاطع باستخدام البحث الشبكي.

4- تقييم الأداء بعد اختبار نموذج التعلم الآلي من خلال معايير الأداء المعروفة. [39]

وفقاً لنتائج هذه الدراسة الموضحة في الجدول (6) تبين أنه:

عند استخدام جميع السمات والتي عددها 18 سمة في التنبؤ عن السكري باستخدام خوارزميات

التصنيف، تراوحت دقة النماذج بين (70.2%-81.4%)، حيث كان أفضل أداء يعود لخوارزمية

الانحدار اللوجستي حيث كانت الدقة 81.40%، بينما خوارزمية أشعة الدعم كانت الأسوأ أداء بين

خوارزميات التصنيف حيث بلغت الدقة 70.2%.

الجدول (6) نتائج الدقة ودرجة F1 لنماذج التصنيف في دراسة Kakoly. [39]

خوارزميات التصنيف		كل السمات		تقنية PCA		تقنية IG مع أفضل 10 سمات		تقنية IG مع أفضل 5 سمات	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
LR	81.4%	70%	70.9%	82%	67.9%	80.8%	70.6%	82.2%	70.6%
RF	81%	69.8%	65.8%	78.3%	68.5%	79.9%	68.7%	80.2%	68.7%
KNN	73.3%	54.9%	55%	73.2%	63.9%	77.8%	63.3%	77.4%	63.3%
DT	78%	68.4%	67.7%	78%	66.9%	78%	65.8%	77.5%	65.8%
SVM	70.2%	62.2%	59.3%	71%	58.5%	69.6%	63.4%	71.7%	63.4%

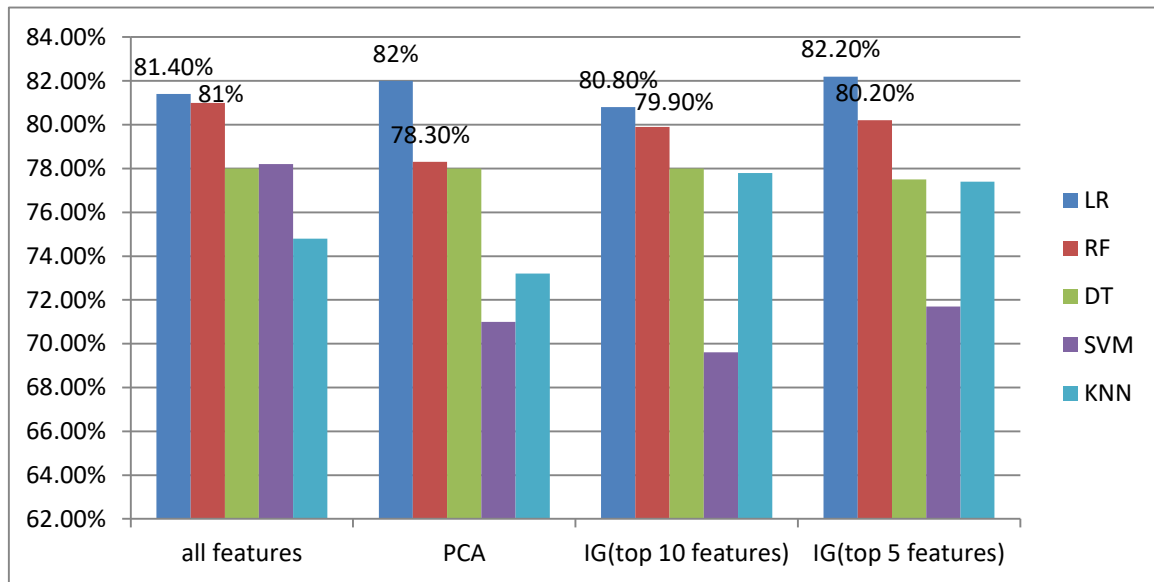
عند استخدام تقنية تحليل المكونات الرئيسية لاختيار السمات، تم اختيار 15 سمة، حيث تحسن أداء

نموذج الانحدار اللوجستي وكانت الدقة تساوي 82% وتراجع أداء بقية النماذج، وكان الأداء الأسوأ

لخوارزمية أشعة الدعم، حيث كانت الدقة تساوي 71%.

عند استخدام تقنية اكتساب المعلومات لاختيار السمات، تم اختيار أفضل 10 سمات (تركيز الجلوكوز، تناول اللحوم والفواكه والسمك، العمر، مؤشر كتلة الجسم، التاريخ العائلي للإصابة بالسكري، النشاط الفيزيائي، المنطقة، مستوى التعليم)، حيث تراجع أداء جميع المصنفات.

عند استخدام تقنية اكتساب المعلومات لاختيار أفضل 5 سمات، كان أداء نموذج الانحدار اللوجستي والغابة العشوائية كما هو موضح في الشكل (14) قد تحسن مقارنة مع أدائهما عند اختيار السمات باستخدام تقنية PCA، وكان الانحدار اللوجستي هو الأفضل مقارنة مع بقية الخوارزميات وكانت الدقة تساوي 82.2% [39].



الشكل (14) مقارنة بين نتائج نماذج دراسة Kakoly [39].

على الرغم من استخدام خوارزميات اختيار السمات اكتساب المعلومات وتقنية تحليل المكونات الرئيسية لتقليل عدد السمات إلا أن أداء النماذج لم يكن جيداً، قد يعود السبب إلى أن هاتين التقنيتين غير متناسبتين مع مجموعة السمات المختارة من قبل الباحث.

أما بالنسبة لدراسة Maniruzzaman وزملاؤه في عام 2020 فقد قاموا بتحديد أهم عوامل الإصابة بمرض السكري باستخدام القيم الاحتمالية ونسبة الأرجحية القائمة على نموذج LR [40].

كما طوروا نظام قائم على تقنيات تعلم الآلة لتصنيف دقيق لخطر الإصابة بالسكري باستخدام أربعة خوارزميات تصنيف (NB، DT، AB،RF)، ثم تم اختيار أفضل نموذج.

استُخدمت في هذه الدراسة مجموعة بيانات السكري للفترة 2009-2012، والمستمدة من المسح الوطني NHANES. حيث كان عدد السمات الكلي 14 (العمر، العرق، الجنس، مستوى التعليم، الوضع العائلي، الطول، الوزن، مؤشر كتلة الجسم، الضغط الانقباضي، الضغط الانبساطي، الكوليسترول المباشر، الكوليسترول الكلي، النشاط الفيزيائي، نوع العمل)، حيث تم حذف جميع القيم المفقودة والمتطرفة من مجموعة البيانات، وجرى تحليل الفروق في المتغيرات بين مرضى السكري ومجموعة السمات باستخدام اختبار t المستقل للمتغيرات المستمرة، واختبار Chi-square للمتغيرات الفئوية، ثم جرى استخراج السمات باستخدام الانحدار اللوجستي Multiple logistic regression اعتماداً على القيم الاحتمالية ونسبة الأرجحية (OR) والتي كان عددها سبعة (العمر، والتعليم، ومؤشر كتلة الجسم، وضغط الدم الانقباضي، وضغط الدم الانبساطي، والكوليسترول المباشر، والكوليسترول الكلي)، ثم جرى تقسيم البيانات باستخدام التحقق المتقاطع، وبعد ذلك تم تدريب خوارزميات التصنيف، وتقييم أداء هذه النماذج، وكانت أهم النتائج كما هو مبين في الجدول (7):

تبين من التحليل الإحصائي أن جميع السمات المستمرة والفئوية مرتبطة ارتباطاً وثيقاً إحصائياً بالسكري، حيث كانت قيم P أقل من 0.05.

الجدول (7) نتائج الدقة ودرجة F1 في دراسة Maniruzaman. [40]

خوارزمية التصنيف	F1_score	Accuracy
RF	96.3%	94.15%
DT	94.68%	89.97%
NB	92.53%	86.61%
Ada Boost	95.99%	92.72%

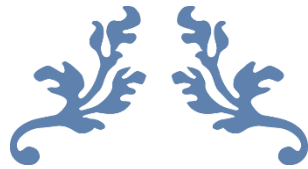
كانت أفضل خوارزمية تصنيف هي الغابة العشوائية وذلك بدقة 94.15%، درجة F1 يساوي 96.3%،

بينما خوارزمية NB كانت أقل أداء من بقية الخوارزميات، حيث كانت الدقة 86.61%. [40]

تعد نتائج هذه الدراسة مشجعة من أجل استخدام الاختبارات الإحصائية لاستخلاص أهم السمات للتنبؤ بالسكري.

3.4 خاتمة:

تعددت الدراسات التي تعمل على تسخير تقنيات تعلم الآلة لاستخراج عوامل الخطورة المحتملة لاستخدامها في التنبؤ والكشف المبكر عن السكري، لما في ذلك خدمة لجميع المرضى الذين تظهر عليهم الأعراض السريرية بعد فترة طويلة من الإصابة، حيث أن التشخيص المبكر قد يكون مفتاحاً لعلاج هؤلاء المرضى، والوقاية من مرض السكري.



الفصل الخامس
منهجية البحث وإجراءاته



1.5 مقدمة :

سيقدم هذا الفصل الجانب العملي من البحث، حيث سيتضمن منهجية البحث وخطوات العمل المتبعة لإتمام هذا البحث من تأمين قواعد البيانات، وتهيئتها ومعالجتها وتقسيمها، وبناء النماذج، وضبط معاملاتها الفائقة، وتدريبها واختبارها، وتقييم أدائها للوصول إلى النموذج الأفضل دقة وكفاءة.

2.5 المنهجية:



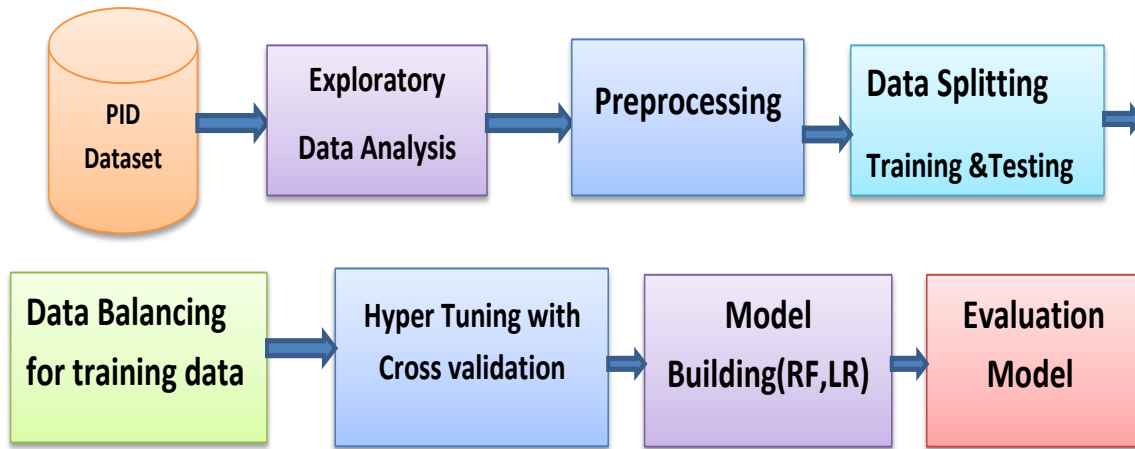
الشكل (15) المخطط العام لبناء نماذج التصنيف.

يوضح الشكل (15) المخطط العام لبناء نماذج التصنيف للتنبؤ بمرض السكري، في المرحلة الأولى من البحث سيتم بناء نموذج تصنيف للتنبؤ بمرض السكري باستخدام قاعدة البيانات الهندية PID، وذلك من خلال جمع البيانات ثم تحليلها ومعالجتها، وضبط معاملات نماذج التصنيف ثم تريها واختبارها وتقييم أدائها، ثم الحصول على النموذج الأفضل، وفي المرحلة الثانية سيتم بناء نموذج التنبؤ بمرض السكري باستخدام قاعدة بيانات محلية يتم بناءها ثم يتم تدريب النموذج واختباره بعد ضبط معاملاته الفائقة، ثم تقييم أداء النموذج.

1.2.5 منهجية البحث لبناء نموذج التنبؤ بمرض السكري باستخدام قاعدة البيانات PID

نفذ نموذج التنبؤ عن مرض السكري بلغة بايثون وطبق على قاعدة البيانات PID، حيث يبين الشكل (16) المخطط الصندوقي لمنهجية البحث المتبعة والتي تتألف من عدة خطوات:

تم في الخطوة الأولى جمع البيانات الخاصة بقاعدة البيانات الهندية (Pima Indian Diabetes) من مواقع الانترنت المتاحة، وفي الخطوة الثانية تم التحليل الاستكشافي للبيانات للكشف عن القيم المفقودة والمتطرفة، واستكشاف بعض القيم الاحصائية، والقيم المفقودة والمتطرفة.



الشكل (16) المخطط الصندوقي لبناء نموذج التنبؤ باستخدام قاعدة البيانات PID.

تعد الخطوة الثالثة من الخطوات الضرورية والأساسية للوصول إلى الهدف المطلوب، حيث تتضمن هذه الخطوة معالجة وتهيئة البيانات من خلال معالجة القيم المفقودة والمتطرفة، بالإضافة إلى التوحيد القياسي للبيانات وذلك من خلال تحويل نطاقات البيانات جميعها إلى نطاق واحد.

الخطوة الرابعة هي تقسيم مجموعة البيانات إلى مجموعتين من أجل تدريب النماذج واختبارها، حيث قسمت مجموعة البيانات إلى 80% من البيانات لتشكيل مجموعة بيانات التدريب، و20% من البيانات لتشكيل مجموعة بيانات الاختبار.

في الخطوة الخامسة تم استخدام تقنيات التوزين لمنع تحيز النموذج نحو الفئة الأغلبية، ولتحسين أداء النماذج.

أما بالنسبة للخطوة السادسة فقد كانت بناء نماذج تعلم الآلة، حيث تم بناء عدة نماذج، ومن ثم تم ضبط المعاملات الفائقة الخاصة بكل نموذج وذلك للوصول إلى أعلى كفاءة باستخدام طريقة البحث الشبكي والتحقق المتقاطع.

في الخطوة السابعة تم تدريب النماذج واختبارها باستخدام المعاملات التي نتجت من عملية الضبط. في الخطوة الثامنة تم تقييم أداء النماذج لتحديد النموذج الأفضل.

توصيف قاعدة البيانات (Pima Indian Diabetic Description)

مجموعة البيانات هذه مأخوذة من مستودع التعلم الآلي التابع لـ UCI، وهي في الأصل من المعهد الوطني للسكري وأمراض الجهاز الهضمي والكلية، الهدف من مجموعة البيانات هذه هو التنبؤ التشخيصي فيما إذا كان المريض يعاني من السكري أم لا، بناءً على قياسات تشخيصية معينة مدرجة في مجموعة البيانات، جميع المرضى هم من الإناث بعمر 21 عامًا على الأقل من سكان البيما

الخطوة الأولى: تحميل قاعدة البيانات البima وتوصيفها [10]

تتكون مجموعات البيانات من 8 متغيرات تنبؤية طبية مستقلة ومتغير مستهدف واحد هو النتيجة، وهي

كالتالي:

1. عدد مرات الحمل (number of pregnancy).
2. الغلوكوز (Glucose): تركيز الغلوكوز في البلازما بعد ساعتين في اختبار تحمل الغلوكوز عن طريق الفم
3. ضغط الدم الانبساطي (Diastolic Blood Pressure): يقاس بوحدة (mmHg)
4. سماكة الجلد (Skin Thickness): سماكة طيات جلد ثلاثية الرؤوس تقاس بوحدة (mm)، وهي توفر تقديرًا جيدًا للسمنة وتوزيع الدهون في الجسم.
5. الأنسولين (Insulin): أنسولين المحقون لمدة ساعتين (mu/ml)
6. مؤشر كتلة الجسم (BMI).
7. وظيفة نسب شجرة عائلة السكري (Diabetes Pedigree Function): سجل احتمالية الإصابة بالسكري بناءً على تاريخ العائلة.
8. العمر (Age).

النتيجة (Outcome): متغير الفئة (0 أو 1)، حيث يوجد 268 شخصاً مصاباً بالسكري ونرمز لهذا المتغير بالرمز 1، و 500 شخص غير مصاب بمرض السكري ويرمز له بالرمز 0.

الخطوة الثانية: تحليل البيانات الاستكشافية Exploratory Data Analysis

التحليل الاستكشافي للبيانات هي أول مرحلة في عملية تحليل البيانات، وهي مهمة لمعرفة شكل البيانات ووصفها وعرضها بشكل بياني واكتشاف العلاقات بين القيم. هذه المرحلة تساعد على فهم البيانات واستخلاص الأسئلة التي ستجيب عنها هذه البيانات، ويمكن القيام بذلك عن طريق إلقاء

نظرة شاملة على الأنماط، والقيم المتطرفة، والنتائج غير المتوقعة في البيانات الموجودة عبر التمثيل البصري والكمي.

1. استكشاف بعض القيم الإحصائية للبيانات:

تم إجراء التحليل الكمي الأساسي في الجدول الموضح في الجدول (8)، حيث جرى حساب بعض القيم الإحصائية للمتغيرات الخاصة بقاعدة البيانات الهندية وقد شملت (المتوسط، الوسيط، الانحراف المعياري، الربيع الأول والثالث)، كما تبين أن العديد من هذه المتغيرات أدنى قيمة لها تبلغ 0، وهو أمر غير بديهي من الناحية الواقعية. على سبيل المثال، من المستحيل أن يكون ضغط الدم عند الأشخاص صفرًا، لذلك من المحتمل أن تكون هذه القيم الصفرية في البيانات هي القيم المفقودة.

الجدول (8) استكشاف بعض القيم الإحصائية للبيانات.

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768	3.84	3.37	0.00	1.00	3.00	6.00	17.00
Glucose	768	120.89	31.97	0.00	99.00	117.00	140.25	199.00
BloodPressure	768	69.11	19.36	0.00	62.00	72.00	80.00	122.00
SkinThickness	768	20.54	15.95	0.00	0.00	23.00	32.00	99.00
Insulin	768	79.80	115.24	0.00	0.00	30.50	127.25	846.00
BMI	768	31.99	7.88	0.00	27.30	32.00	36.60	67.10
DiabetesPedigree Function	768	0.47	0.33	0.08	0.24	0.37	0.62	2.42
Age	768	33.24	11.76	21.00	24.00	29.00	41.00	81.00
Outcome	768	0.35	0.48	0.00	0.00	0.00	1.00	1.00

2. استكشاف القيم المفقودة:

تبين من خلال التحليل الاستكشافي باستخدام لغة البايثون لقاعدة البيانات البيما كما يظهر الشكل (5-3) أن هناك 6 سمات تحتوي على قيم صفرية لبعض العينات والتي هي (عدد مرات الحمل، مستوى الغلوكوز، ضغط الدم، سماكة الجلد، مستوى الأنسولين، ومؤشر كتلة الجسم، وعدد هذه القيم الصفرية

نظام تشخيص آلي لمرض السكري باستخدام خوارزميات التصنيف، صالح.

لهذه السمات الستة هي كالتالي (111,5,35,227,374,11)، حيث أغلب هذه السمات التي تحتوي على قيم صفرية قد تكون نتيجة لعملية جمع بيانات خاطئة خلال عملية جمع المعلومات، وهذه القيم يمكن أن تؤثر على أداء المصنف، لذلك يمكن اعتبار هذه القيم بأنها مفقودة، ويجب التعامل معها بشكل مناسب أثناء عملية بناء النموذج.

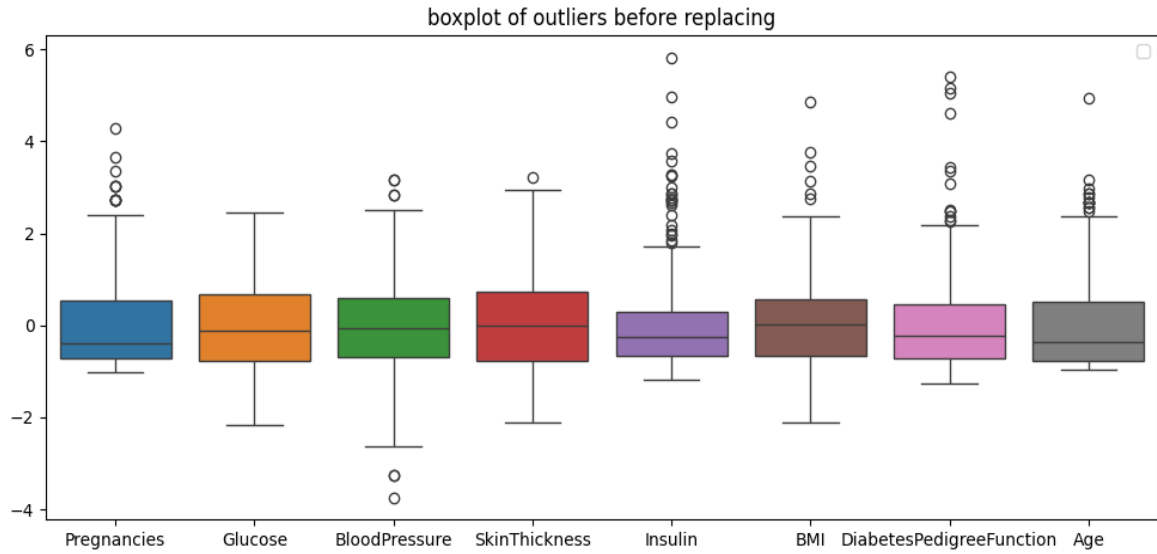
```
✓ [13] # 10 find howmany zero values are included in each variables:
0s print((diabetes_dataset == 0).sum())
```

Pregnancies	111
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11
DiabetesPedigreeFunction	0
Age	0
Outcome	500
dtype:	int64

الشكل (17) عدد القيم المفقودة في كل عمود

3. استكشاف القيم المتطرفة:

تبين عند استخدام المخططات الصندوقية كما يظهر الشكل (18) إلى وجود بعض القيم المتطرفة المحتملة في بعض السمات مثل عدد مرات الحمل، الغلوكوز، ضغط الدم، سماكة الجلد والأنسولين، حيث يمكن أن تؤثر هذه القيم المتطرفة على عملية النقييس، مما يؤدي إلى حدوث أخطاء، وتؤدي إلى الأداء السيء للنموذج وتقلل من دقته في التصنيف.



الشكل (18) المخطط الصندوقي لاستكشاف القيم المتطرفة لجميع السمات قبل معالجتها.

الخطوة الثالثة: المعالجة المسبقة للبيانات Data preprocessing

الهدف من المعالجة المسبقة هو تنظيف البيانات وإعدادها لمزيد من التحليل، مثل هندسة الميزات وبناء النماذج، وهي أمر ضروري من أجل تحسين البيانات لإنشاء نماذج موثوقة للتعليم الآلي، والتي تعمل في النهاية على تحسين الدقة، حيث يمكن تطبيق العديد من تقنيات المعالجة المسبقة الشائعة على PIDD وهي:

1. معالجة القيم المفقودة:

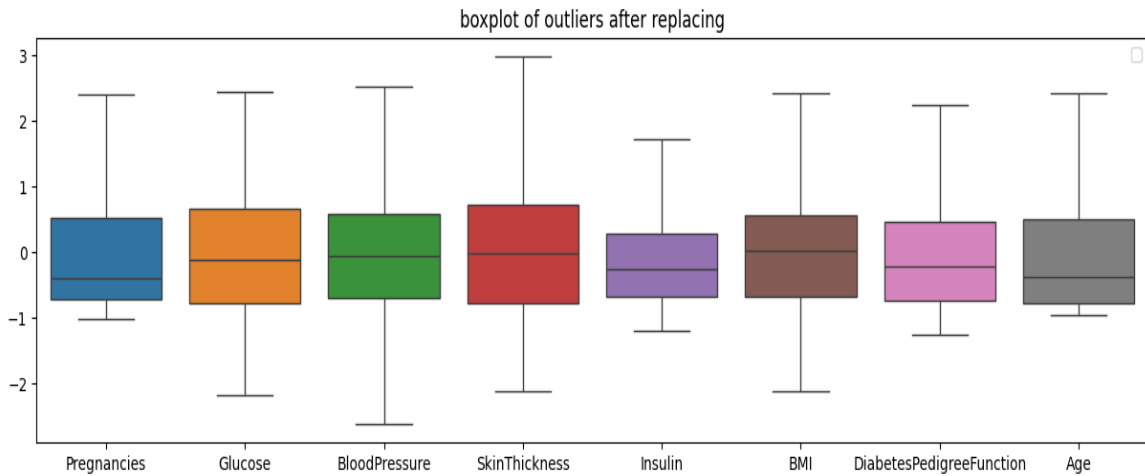
يمكن معالجة هذه القيم المفقودة بطرق عدة، إما باستبدالها بقيم المتوسط أو الوسيط أو حذفها، لكن ما تم القيام به هو من حذف كافة الأسطر التي تحوي على هذه القيم المفقودة PID، مما أدى إلى تناقص عدد أسطر قاعدة البيانات وأصبح عددها 394 سطرًا (130 منها تعود لأشخاص مصابين بالسكري، و264 تعود لأشخاص سليمين) و9 أعمدة.

2. تحويل البيانات إلى بيانات قياسية standardization:

بالإضافة إلى خطوات المعالجة المسبقة للبيانات التي تم ذكرها سابقاً، استخدمت طريقة تحويل البيانات إلى بيانات قياسية على مجموعة بيانات بيما الهندية بحيث تم تحويل البيانات إلى نطاق واحد، مما يمكن أن يساعد في تثبيت التباين وتقليل تأثير القيم المتطرفة.

3. معالجة القيم المتطرفة:

في حالة إنشاء نموذج تعلم الآلة يمكن أن تؤدي القيم المتطرفة إلى ضعف أداء النموذج و لمنع حدوث ذلك، بالإمكان محاولة إزالة القيم المتطرفة وتدريب النموذج باستخدام بيانات أقل في حال كان هناك الكثير من البيانات وعدد قليل من القيم المتطرفة، لكن إذا كانت القيم المتطرفة عددها كبير وكمية البيانات قليل فيمكن استبدال القيم المتطرفة بالوسيط مثلاً، أو إسناد لها قيمة كالحده الأقصى أو الأدنى وهذا ما تم تطبيقه على القيم المتطرفة الموجودة في سمات قاعدة البيانات PID، لأننا لا نريد فقدان نقاط بيانات أكثر، وهذا ما أظهره المخطط الصندوقي لجميع السمات بعد معالجة القيم المتطرفة في الشكل (19)، حيث تم التخلص من القيم المتطرفة من هذه السمات بعد إسناد قيم الحد الأقصى والحد الأدنى إلى القيم المتطرفة.



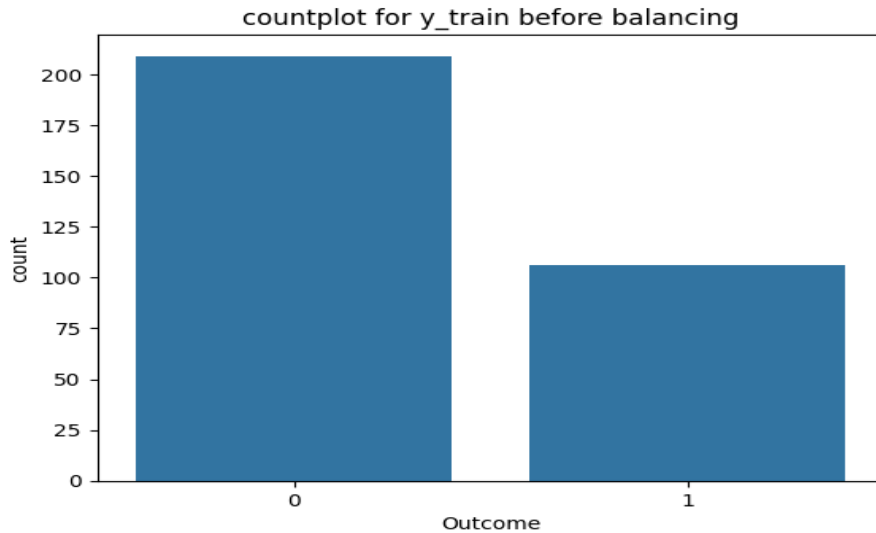
الشكل (19) المخطط الصندوقي لاستكشاف القيم المتطرفة لجميع السمات بعد استبدال القيم المتطرفة.

الخطوة الرابعة: تقسيم البيانات

بعد المعالجة المسبقة وتصفية البيانات أصبح عدد العينات 394 عينة منها 264 عينة تعود لأشخاص سليمين و 130 عينة تعود لأشخاص مرضى، ثم أصبحت مجموعة البيانات جاهزة للتدريب والاختبار حيث قسمت مجموعة البيانات إلى 80% منها لتشكيل مجموعة بيانات التدريب، و 20% منها لتشكيل مجموعة بيانات الاختبار، حيث تتضمن مجموعة بيانات التدريب 315 عينة (209 منها تعود لأشخاص سليمين و 106 تعود لأشخاص مرضى) وعدد السمات 8، بينما مجموعة الاختبار تتضمن 79 عينة (55 منها تعود لأشخاص سليمين و 24 تعود لأشخاص مرضى).

الخطوة السادسة: التحقق من توازن البيانات وتوزين بيانات التدريب

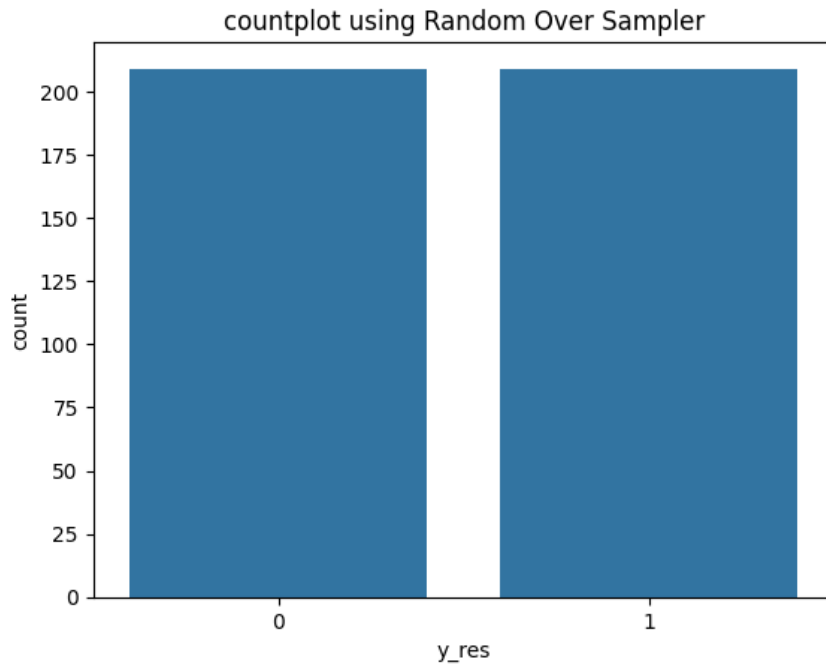
بعد تقسيم البيانات إلى مجموعتي التدريب والاختبار، تبين أن مجموعة بيانات التدريب كما يظهر الشكل (20) غير متوازنة، فقد بلغ عدد الأشخاص غير المصابين بالسكري 209 والذين يتفوق عددهم على عدد الأشخاص المصابين بالسكري الذين بلغ عددهم 130 مريض، وهذا يعني أن مصنف البيانات متحيز، لذلك فإن النماذج قد يكون لديها أداء تنبؤي ضعيف، خاصة بالنسبة لفئة الأقلية،



الشكل (20) عدد بيانات التدريب السليمة والمريضة قبل موازنتها.

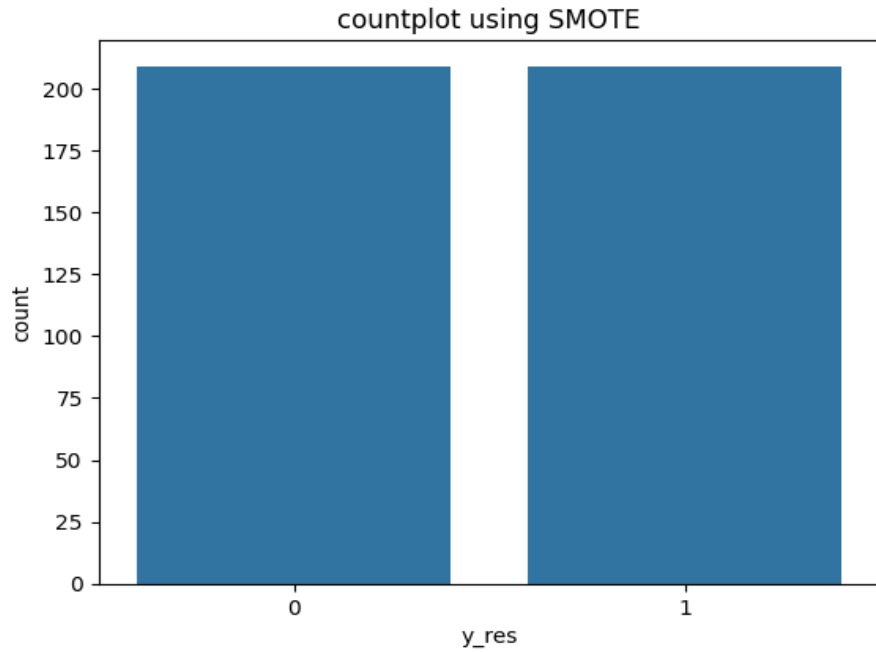
وبما أن البيانات غير متوازنة وعينة البيانات ليست ضخمة، فقد تم استخدام عدة تقنيات لموازنة بيانات التدريب للتعامل مع هذه المشكلة، والتي أعطت نتائج مشجعة في بعض الدراسات المرجعية السابقة [36-37] وهذه التقنيات هي:

1. **الزيادة العشوائية للعينات (ROS):** تم أخذ العينات من مجموعة التدريب (حيث تم زيادة البيانات العشوائية المريضة لتصبح مساوية عدد العينات السليمة وتساوي (209) كما هو مبين في الشكل (21).



الشكل (21) موازنة بيانات التدريب باستخدام تقنية ROS.

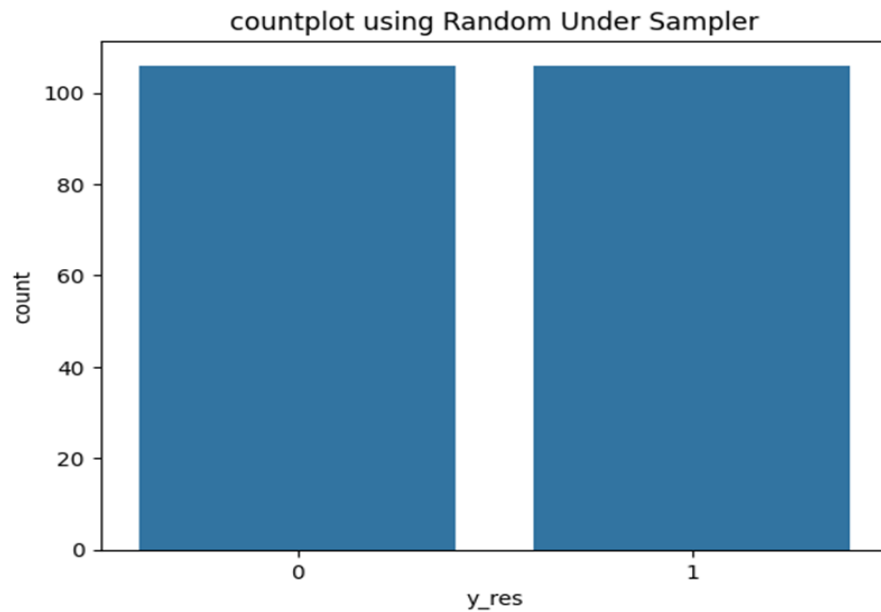
2. **تقنية الافراط في أخذ العينات الأقلية الاصطناعية (SMOTE)** أصبح عدد العينات لكل من الأشخاص السليمين والأشخاص المرضى متوازن ويساوي (209) كما هو مبين في الشكل (22).



الشكل (22) موازنة بيانات التدريب باستخدام تقنية SMOTE

3. تقنية الإنقاص العشوائي للعينات (RUS) تم إنقاص عدد العينات السليمة لتصبح مساوية لعدد

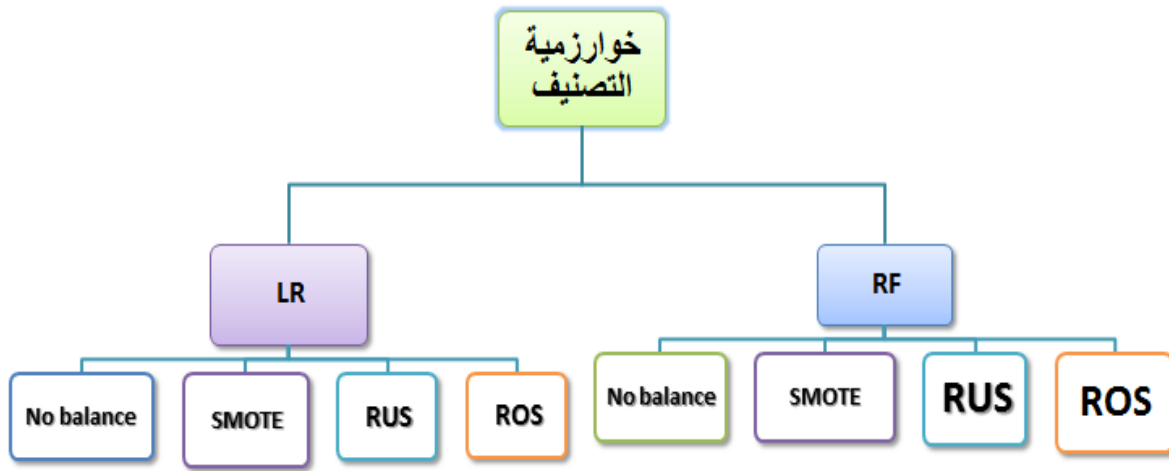
العينات المريضة (106) كما هو مبين في الشكل (23).



الشكل (23) موازنة بيانات التدريب باستخدام تقنية RUS

الخطوة السابعة: تحديد خوارزميات التصنيف وضبط المعاملات الفائقة

عند بناء نماذج التصنيف تم استخدام تقنيات تعلم الآلة وهما خوارزميتي الانحدار اللوجستي والغابة العشوائية لبناء ثمانية نماذج تصنيف قادرة على التنبؤ بالسكري كما يظهر في الشكل (5-9)، وذلك عند استخدام تقنيات التوزين السابقة، ومن ثم تم ضبط المعاملات الفائقة لهذه النماذج باستخدام طريقة البحث الشبكي، حيث يوضح الشكل (20) نماذج التصنيف الثمانية التي تم بناءها عند عدم توزيع البيانات وعند استخدام تقنيات التوزين السابقة:



الشكل (24) بناء نماذج التصنيف باستخدام خوارزميتي (LR, RF)

وسنوضح فيما يلي المعاملات الفائقة التي تم ضبطها للنماذج بهدف الوصول إلى النموذج الأفضل والأكثر كفاءة.

1. المعاملات الفائقة لخوارزمية الانحدار اللوجستي التي تم ضبطها:

يوضح الجدول (9) المعاملات الفائقة والقيم التي أسندت إليها للاختيار منها باستخدام البحث الشبكي والتحقق المتقاطع من أجل الحصول على أفضل معاملات لخوارزمية الانحدار اللوجستي، وبالتالي الحصول على النموذج الأمثل.

الجدول (9) المعاملات الفائقة لخوارزمية الانحدار اللوجستي.

Model	Hyper parameters	Range
LR	C	Range between [0.01,10]
	Max_iter	1000
	Solver	["Lbfgs", "saga", "sag", "newton_cg"]
	Class Weight	["None", "Balanced"]
	Penalty	[L2, elastic net]

حيث جرى ضبط المعلمة C التي تساعد على تحقيق التوازن بين البساطة والدقة لتحقيق الأداء الأمثل للنموذج وتشير القيمة الأصغر لها إلى تنظيم أقوى (أي أن النموذج يصبح مبسطاً للغاية غير قادر على ملائمة البيانات وبالتالي سوء التصنيف)، وعندما تكون قيمة C مرتفعة جداً فإنها تسمح للنموذج بملاءمة بيانات التدريب بشكل أوثق لكنه يفشل في التعميم، ويصبح النموذج أكثر تعقيداً، حيث تم اختيار قيمها في دراستنا ضمن المجال [0.01,10].

كما تشير (Penalty) إلى التنظيم وتهدف إلى تقليل خطأ تعميم النموذج، حيث تمنع هذه التقنية تعلم نماذج أكثر تعقيداً من أجل منع فرط التعلم (Overfitting).

بينما (solver) فهي تمثل الحلول أي الخوارزميات التي تستخدم لتحسين أداء النموذج من خلال إيجاد أفضل المعلمات الملائمة التي تقلل من الخطأ، وقد تم اختيار عدة أنواع منها موضحة في الجدول السابق. كما تقوم المعلمة (class weight) بضبط وزن كل فئة في الانحدار اللوجستي.

أما بالنسبة إلى المعلمة (max_iter) فهي تحدد الحد الأقصى لعدد التكرارات التي يُسمح للحل بتنفيذها للوصول إلى حل، والتي تم إسناد لها القيمة 1000

2. المعاملات الفائقة لخوارزمية الغابة العشوائية التي تم ضبطها:

يوضح الجدول (10) المعاملات الفائقة والقيم التي أسندت إليها للاختيار منها باستخدام البحث

الشبكي والتحقق المتقاطع من أجل الحصول على أفضل معاملات لنموذج الغابة العشوائية.

الجدول (10) المعاملات الفائقة لخوارزمية الغابة العشوائية.

Model	Hyper parameters	Range
RF	Criterion	"Gini"، "entropy"
	Max depth	Range between(1,15)
	N_estimators	Range between(20-200)
	Max Features	"Sqrt"، "Log2"

حيث تم ضبط المعيار (criterion) وهو الأسلوب الذي يحدد أفضل الانقسامات عند بناء كل شجرة في النموذج واختيار أفضل السمات للاستخدام في كل انقسام، وذلك من خلال تقييم جودة الانقسامات مما يساهم في تحسين دقة وأداء النموذج، وتم اختيار المعيار من ضمن المعيارين (Entropy، Gini). بالإضافة إلى ضبط عدد الأشجار (n_estimators) التي تحدد عدد أشجار القرار التي سوف يجري بناء نموذج الغابة العشوائية عليها، حيث كلما زاد عدد الأشجار زاد استقرار النموذج، وقد تم في هذه الدراسة اختيار عدد أشجار القرار ضمن المجال (20-200).

كما تم ضبط العمق الأعظمي (max_depth): والتي تحدد الحد الأقصى للعمق الذي يمكن أن تصل إليه الأشجار في النموذج (أي الحد الأقصى لعدد المستويات أو التقسيمات)، وقد تم اختيار العمق الأعظمي من ضمن المجال (1-15).

وكذلك تم ضبط (max_features) وهي تحدد العدد الأعظمي للسمات المناسبة لبناء كل شجرة وعقدها (أي التحكم بالعشوائية عند بناء الأشجار واختيار السمات)، حيث تم اختيار عدد السمات الأعظمي (Log2، Sqrt).

الخطوة الثامنة: تدريب النماذج وتقييمها

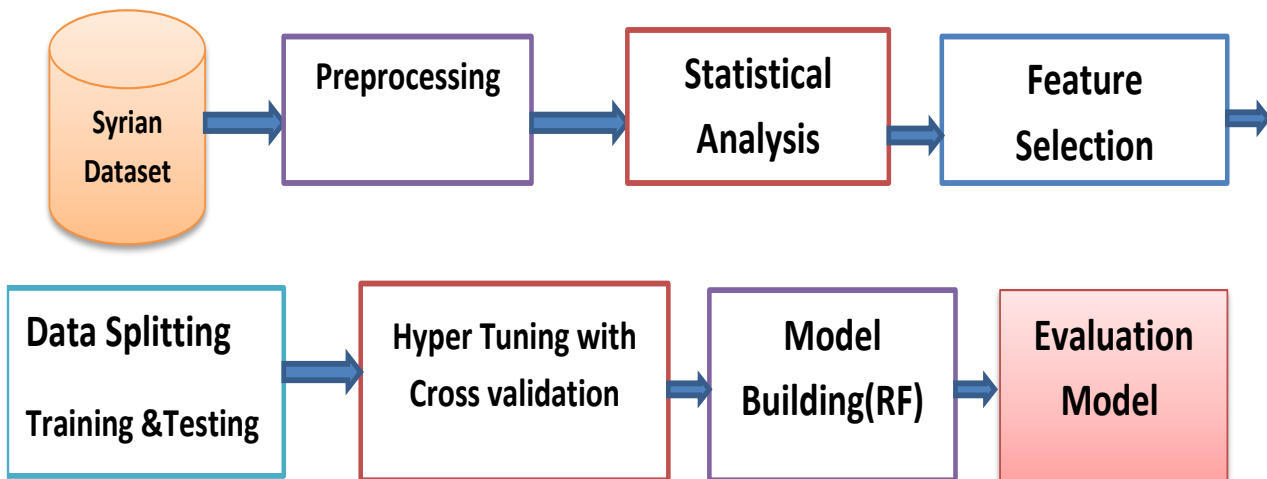
بعد الانتهاء من الخطوات السابقة تم بناء النماذج، ثم تم تدريب هذه النماذج واختبارها وذلك باستخدام

السمات الناتجة من المراحل السابقة والتي عددها (8)، وفقا لتقسيم البيانات المحدد (20/80) وذلك للتنبؤ بالمتغير المستهدف (مريض سكري أو سليم)، ثم تقييم كفاءة هذه النماذج باستخدام مقاييس تقييم الأداء (الدقة، الحساسية، الاحكام، درجة F1، المساحة تحت منحنى الخواص ROC) والمقارنة فيما بينها، وبناء على هذه المقاييس تم تحديد أفضل نموذج.

2.2.5 منهجية البحث لبناء نموذج التنبؤ بمرض السكري من النوع الثاني باستخدام قاعدة بيانات

محلية

تم العمل في هذا البحث وفق المخطط الصندوقي لمنهجية البحث الموضح في الشكل (25) لبناء نموذج التنبؤ بمرض السكري باستخدام قاعدة بيانات محلية، وتتألف هذه المنهجية من عدة خطوات: حيث كانت الخطوة الأولى تصميم استبيان يتعلق بالعوامل المؤثرة بمرض السكري، ومن ثم جمع البيانات من أشخاص سليمين ومرضى سكري من أجل بناء قاعدة البيانات التي ستبنى عليها باقي الخطوات، حيث تم إنشاء قاعدة بيانات خاصة بالجمهورية العربية السورية من أجل العمل عليها.



الشكل (25) المخطط الصندوقي لبناء نموذج التنبؤ بمرض السكري باستخدام قاعدة بيانات محلية.

الخطوة الثانية هي من الخطوات الأساسية والضرورية للوصول إلى الهدف المطلوب، وتتضمن معالجة وتهيئة البيانات من خلال معالجة القيم المفقودة والتوحيد القياسي للبيانات ومعالجة القيم المتطرفة.

الخطوة الثالثة: جرى في هذه الخطوة إجراء تحليل وصفي لتقييم توزيع المتغيرات الكمية، كما تم استخدام عدة اختبارات إحصائية لتحديد العلاقة بين المتغيرات الفئوية والمتغير المستهدف.

الخطوة الرابعة: تم استخدام خوارزميات اختيار السمات لاختيار السمات الأكثر أهمية واستخدامها لتدريب النموذج واختباره، وذلك من أجل الحصول على أفضل نموذج، مع أفضل مجموعة سمات تم الوصول إليها.

الخطوة الخامسة هي تقسيم مجموعة البيانات إلى مجموعتين من أجل تدريب النماذج واختبارها، حيث قسمت مجموعة البيانات إلى 80% من البيانات لتشكيل مجموعة بيانات التدريب، و20% من البيانات لتشكيل مجموعة بيانات الاختبار.

أما بالنسبة للخطوة السادسة فقد كانت بناء نماذج تعلم الآلة باستخدام خوارزميات التصنيف، ومن ثم تم ضبط المعاملات الفائقة الخاصة بكل نموذج وذلك للوصول إلى أعلى كفاءة باستخدام طريقة البحث الشبكي والتحقق المتقاطع.

في الخطوة السابعة تم تدريب النماذج واختبارها باستخدام المعاملات التي نتجت من عملية الضبط.

الخطوة الثامنة تم تقييم أداء النماذج لتحديد النموذج الأفضل.

الخطوة الأولى: بناء قاعدة البيانات

مراحل بناء قاعدة البيانات:

يوجد عدة مراحل لإنشاء قاعدة البيانات الخاصة بالجمهورية العربية السورية وهي كالتالي:

نظام تشخيص آلي لمرض السكري باستخدام خوارزميات التصنيف، صالح.

1. تصميم استبيان مسحي يتضمن أهم العوامل الأكثر خطورة المسببة للسكري من النوع الثاني.
2. تحديد عينة الدراسة، والمراكز المراد التوجه إليها لجمع البيانات.
3. جمع البيانات من الأشخاص السليمين والمرضى.
4. تعريف المتغيرات، وتخزين البيانات التي جرى جمعها ضمن ملف إكسل، ثم ترميزها.

1. تصميم الاستبيان المسحي:

جرى إعداد استبيان مُحدّد بدقة لإجراء الدراسة البحثية، حيث يتضمن هذا الاستبيان أسئلة عددها 42 سؤالاً يغطي المعلومات الشخصية وعوامل الخطر المرتبطة بالسكري من النوع الثاني، ومضاعفاته، وأعراضه.

أقسام الاستبيان ومحتواه:

يوضح الجدول (11) أن الاستبيان يتألف من ثلاثة أقسام:

1. القسم الأول: يتعلق بعوامل الخطورة المحتملة للإصابة بمرض السكري الموزعة ضمن ست عوامل كما هي موضحة في الجدول (11).
2. القسم الثاني: يتألف من قسمين يتعلقان بأعراض السكري من النوع الثاني ومضاعفاته.
3. القسم الثالث: موجه فقط للنساء المتزوجات ويتضمن العوامل المتعلقة بالحمل والولادة والتي تشير إلى احتمال الإصابة بالسكري (عدد مرات الحمل، ولادة طفل يفوق وزنه أربعة كيلو غرام ، إصابة المرأة بسكري الحمل).

نظام تشخيص آلي لمرض السكري باستخدام خوارزميات التصنيف، صالح.

الجدول (11) أقسام الاستبيان المصمم لبناء قاعدة البيانات المحلية.

القسم الأول: عوامل الخطورة المحتملة للإصابة بالسكري:	اسم المتغير:
1. العوامل الديموغرافية:	العمر، الجنس، مستوى التعليم
2. عوامل تتعلق بالبيانات السريرية:	الطول، الوزن، مؤشر كتلة الجسم، محيط الخصر، ضغط الدم الانقباضي، ضغط الدم الانبساطي، معدل ضربات القلب، مستوى السكر عند الصيام
3. عوامل جودة الحياة	التدخين، عدد ساعات النوم، النشاط الفيزيائي
4. عوامل تتعلق بالعادات الغذائية:	تناول الخضراوات، تناول الفاكهة، تناول الأطعمة الدسمة، تناول الكربوهيدرات، تناول المشروبات المحلاة بالسكر، اتباع نظام غذائي صحي، نوع الزيت المستخدم في الطهي ، طريقة الطهي
5. عامل الحالة الطبية:	أمراض القلب والأوعية الدموية أمراض ارتفاع ضغط الدم تناول الأدوية بشكل منتظم
6. العامل الوراثي	التاريخ العائلي للإصابة بمرض السكري (قربة من الدرجة الأولى أو الثانية أو الثالثة)
7. العامل النفسي والاجتماعي	القلق، التوتر، الغضب، الخوف، الأرق
القسم الثاني أعراض مرض السكري ومضاعفاته:	

الشعور المستمر بالجوع، الشعور المستمر بالعطش، تكرار التبول، فقدان مفاجئ للوزن، تعب عام	أعراض مرض السكري:
ارتفاع نسبة الكوليسترول، مشاكل في الكلى، مشاكل في العين أو الرؤية، الغثيان والقيء والإمساك، أمراض الغدة الدرقية، تنميل/ألم/وخز في اليدين/القدمين، ضيق التنفس	مضاعفات السكري:
عدد مرات الحمل، ولادة طفل يفوق وزنه أربعة كيلو غرام، الإصابة بسكري الحمل	<u>القسم الثالث خاص بالنساء المتزوجات:</u>

2. تحديد عينة الدراسة والمراكز التي تم التوجه إليها لجمع البيانات:

تم إجراء دراسة مقطعية صمم من خلالها استبيان للتنبؤ بسمات التنبؤ بمرض السكري من النوع الثاني، وكان هذا الاستبيان موجه لأشخاص سليمين ومصابين بمرض السكري من النوع الثاني وأعمارهم تتراوح بين ال (19-88) عاماً من النساء والرجال في عدة مراكز صحية ومشافي تابعة لمدينة دمشق وريفها وذلك بعد الحصول على موافقات أخلاقية من الجهات المسؤولة في الجمهورية العربية السورية.

المراكز الصحية والمشافي التي جرى من خلالها جمع البيانات:

- مركز السكري التخصصي في منطقة الزاهرة.
- المركز الصحي في مدينة التل.
- مستشفى التل الوطني.

3. جمع البيانات:

وُزِعَ الاستبيان على المشاركين المُختارين، حيث أبدى المشاركون في هذا الاستبيان موافقتهم على مشاركتهم في البحث، وكان مجموعهم 159 شخصاً من كلا الجنسين تتراوح أعمارهم بين (19-88)

عاما، وكان منهم 63 شخصا سليما لا يعاني من مرض السكري، و97 شخصا يعاني من السكري من النوع الثاني.

تعريف المتغيرات وترميزها:

يبين الملحق (1) متغيرات الدخل الخاصة بقاعدة البيانات المحلية والتي كان عددها 63، وتنقسم إلى ثلاثة أنواع:

1. المتغيرات الكمية (Numerical or Continuous): وقد بلغ عددها 9، والتي تمثل (العمر، الطول، الوزن، الضغط الانقباضي والانبساطي، محيط الخصر، تركيز سكر الدم عند الصيام، مؤشر كتلة الجسم، معدل نبضات القلب)، وقد تم الحصول عليها باستخدام أدوات القياس (ميزان الكتروني لقياس الوزن، متر لقياس الطول، جهاز الضغط لقياس الضغط الانقباضي والانبساطي ومعدل ضربات القلب، جهاز قياس سكر الدم لقياس تركيز الجلوكوز عند الصيام).
2. المتغيرات الاسمية الترتيبية (Ordinal): بلغ عددها 2 والتي هي التاريخ العائلي للإصابة بمرض السكري، مقدار الشعور بالتوتر والضغطات.
3. المتغيرات الاسمية غير الترتيبية (Nominal): بلغ عددها 52 متغير.

الخطوة الثانية: المعالجة المسبقة للبيانات

تعد مرحلة المعالجة الأولية للبيانات وتهيئتها (Data Preprocessing) مرحلة مهمة لتحسين جودة البيانات وتجنب الأخطاء، كما تساعد في تحسين عملية التحليل وتساعد في تحسين أداء النماذج، فهي مرحلة أساسية للوصول إلى نتائج دقيقة.

بعد حذف الأعمدة الإدارية وأعمدة خط البداية تم حذف كافة الأسطر التي تحوي على القيم المفقودة، وذلك لتبسيط قاعدة البيانات، والتقليل من الضجيج وفي نهاية هذه الخطوة كان عدد البيانات في عامود

المتغير المستهدف (158)، ثم تم توحيد نطاق جميع المتغيرات باستخدام Standardization

نظام تشخيص آلي لمرض السكري باستخدام خوارزميات التصنيف، صالح.

ثم تمت معالجة القيم المتطرفة لبناء نموذج أكثر دقة وأفضل أداء من خلال استبدالها بتحديد الحدود الدنيا والعليا باستخدام تقنية الربيعيات (الربيع الأول والثالث والفرق بينهما)، وتوضح المعادلتين (1-5)،

(2-5) العلاقتين المستخدمتين لتحديد الحدود الدنيا والعليا

$$Lower_{bound} = Q1 - 1.5 \times IQR \quad \dots\dots\dots(1-5)$$

$$Upper_{bound} = Q3 + 1.5 \times IQR \quad \dots\dots\dots(2-5)$$

Q1: الربيع الأول.

Q3: الربيع الثالث.

IQR: المدى بين الربيع الأول والثالث.

الخطوة الثالثة: التحليل الإحصائي (Statistical Analysis)

جرى تطبيق أربعة اختبارات إحصائية وصفية لتحليل بيانات عينة الدراسة لاستخلاص أهم السمات التي تنتبأ بالسكري من النوع الثاني، حيث تعتمد هذه الاختبارات على نوع البيانات وعلى طبيعتها توزيعها، وهذه الاختبارات هي:

1. اختبار شابيرو Shapiro_test: طبق على المتغيرات الرقمية للتأكد من توزيع البيانات الرقمية هل

هو طبيعي أم لا، حيث وجد أن جميع السمات الرقمية تتوزع بشكل غير طبيعي، باستثناء الطول فقد كان توزيعها طبيعي.

2. اختبار مان ويتني (Mann Whitney test) : طبق على المتغيرات الرقمية ذات التوزيع الغير طبيعي.

نظام تشخيص آلي لمرض السكري باستخدام خوارزميات التصنيف، صالح.

3. اختبار T_test: طبق على البيانات الرقمية ذات التوزيع الطبيعي، والذي يعود إلى المتغير طول الشخص.

4. اختبار كاي مربع: طبق على متغيرات الدخل الاسمية النوعية.

بعد تطبيق هذه الاختبارات كان عدد المتغيرات التي لها دلالة إحصائية للإصابة بمرض السكري 35 متغير، بينما بقية المتغيرات ليس لها دلالة إحصائية ذات معنى.

الخطوة الرابعة: تقسيم البيانات (Data Splitting)

تقسم مجموعة البيانات إلى مجموعتين من أجل تدريب النماذج واختبارها، حيث قسمت إلى 80% من البيانات لتشكيل مجموعة بيانات التدريب، و20% منها لتشكيل مجموعة بيانات الاختبار. حيث تضمنت مجموعة بيانات التدريب 126 عينة وعدد متغيرات الدخل 35 متغير، بينما مجموعة الاختبار تضمنت 32 عينة وعدد متغيرات الدخل 35 متغير.

الخطوة الخامسة: بناء النموذج وضبط معاملاته الفائقة

تم بناء نموذج باستخدام تقنيات تعلم الآلة وهي خوارزمية الغابة العشوائية وسنوضح فيما يلي المعاملات الفائقة التي تم ضبطها للنموذج بهدف الوصول إلى النموذج الأفضل والأكثر كفاءة.

المعاملات الفائقة لخوارزمية الغابة العشوائية

يوضح الجدول (12) المعاملات الفائقة والقيم التي أسندت إليها للاختيار منها باستخدام البحث الشبكي والتحقق المتقاطع من أجل الحصول على أفضل معاملات لخوارزمية الغابة العشوائية، وبالتالي الحصول على النموذج الأمثل.

نظام تشخيص آلي لمرض السكري باستخدام خوارزميات التصنيف، صالح.

الجدول (12) المعاملات الفائقة لخوارزمية الغابة العشوائية.

Model	Hyper parameters	Range
RF	Criterion	"Gini"، "entropy"
	Max depth	Range between[1،15]
	N_estimators	Range between[20،200]
	Max Features	"Sqrt"، "Log2"

حيث جرى ضبط المعيار (criterion) الذي يحدد أفضل الانقسامات عند بناء كل شجرة في النموذج، وجرى اختيار المعيار من ضمن النوعين (Entropy، Gini).

بالإضافة إلى ضبط عدد الأشجار (n_estimators): وهي تحدد عدد أشجار القرار التي سوف يتم بناء نموذج الغابة العشوائية عليها، وقد جرى في هذه الدراسة اختيار عدد أشجار القرار من ضمن المجال (20-200)، كما جرى ضبط العمق الأعظمي (max_depth) من ضمن المجال (1-15).

وكذلك تم ضبط (max_features) التي تحدد العدد الأعظمي للسمات المناسبة لبناء كل شجرة وعقدها (أي التحكم بالعشوائية عند بناء الأشجار واختيار السمات)، حيث تم اختيار عدد السمات الأعظمي (Log2، Sqrt).

الخطوة السادسة: اختيار السمات (Feature Selection)

جرت إضافة هذه الخطوة لتقليل عدد السمات التي تم الوصول إليها في المراحل السابقة، وذلك للوصول إلى المجموعة الفرعية للسمات التي تحقق التوازن بين عدد السمات الأقل ودقة النماذج في الوقت نفسه، فالغرض الأساسي من تقليل السمات هو تقليل الأبعاد وزيادة دقة النماذج بأن وحد.

استبعاد السمات العودية (RFE)

يوضح الجدول (13) المحددات الفائقة والقيم التي أسندت إليها من أجل ضبط خوارزمية RFE

للوصول إلى أفضل كفاءة في اختيار مجموعة السمات

الجدول (13) المعاملات الفائقة لخوارزمية RFE .

Feature selection	Hyper parameters	Value
RFE	Model	RF
	n_feature_to select	10,8,6

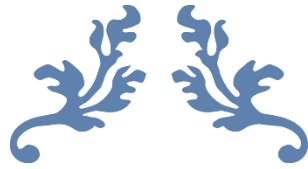
في هذه الطريقة يجب تحديد أحد نماذج تعلم الآلة لإعطاء أوزان للسماوات ليتم الاختيار فيما بينها، وعادة ما يكون الغابة العشوائية أو الانحدار اللوجستي، وبما أن نموذج الغابة العشوائية أدائه أعلى في هذا البحث فقد تم اختياره مع طريقة تقليل السماوات العودية، وتم اختيار (n_feature_to select) المسؤولة عن تحديد عدد السماوات المراد اختيارها وكانت تساوي (6,8,10) لتحديد أفضل نموذج.

الخطوة السابعة: تدريب النماذج وتقييمها:

بعد الانتهاء من الخطوات السابقة تم تدريب هذا النموذج واختباره وذلك باستخدام السماوات الناتجة من المراحل السابقة والتي عددها (6,8,10)، وفقاً لتقسيم البيانات المحدد (20/80)، وذلك للتنبؤ بالمتغير المستهدف (مريض سكري أو سليم)، ثم تقييم كفاءة هذه النماذج باستخدام مقاييس تقييم الأداء (الدقة، الحساسية، الاحكام، درجة F1، المساحة تحت منحنى الخواص ROC)، وبناء على هذه المقاييس تم تحديد أفضل نموذج.

3.5 خاتمة:

تناول هذا الفصل القسم العملي من البحث بكل مراحله وخطواته بدءاً من تصميم الاستبيان المسحي، وبناء قاعدة البيانات، ومن ثم تهيئة ومعالجة البيانات وتحليلها باستخدام الأساليب الإحصائية الوصفية لاستخلاص عوامل التنبؤ بالسكري من النوع الثاني، ثم تقسيم مجموعة البيانات لتدريب واختبار النموذج بعد ضبط محدداته الفائقة وانتهاء بتقييم أداء هذه النماذج للوصول إلى أفضل نموذج منها، كما جرى توضيح خطوة تقليل السماوات للوصول إلى أفضل نموذج مع أقل عدد من السماوات.



الفصل السادس

النتائج والمناقشة



1.6 مقدمة:

نتيجة انتشار السكري بكافة أنواعه في الآونة الأخيرة أصبح محط اهتمام الباحثين في جامعات العالم، والذين يعملون على تطوير منهجيات ذات أداء عالي للتشخيص والعلاج والوقاية، وذلك انطلاقاً من أهمية الكشف المبكر عن السكري الذي من الممكن أن يكون مفتاح العلاج الفعال والحد من مضاعفاته الخطيرة، أو على الأقل التخطيط المستقبلي لرفع مستوى الرعاية الصحية لجميع المرضى، سيقدم هذا الفصل نتائج البحث ومناقشتها.

2.6 نتائج نماذج التصنيف للتنبؤ بالسكري باستخدام قاعدة البيانات PID:

1.2.6 نتائج نماذج خوارزمية الانحدار اللوجستي:

تم بناء أربعة نماذج باستخدام خوارزمية الانحدار اللوجستي للتنبؤ بالسكري باستخدام قاعدة البيانات PID، وهذه النماذج هي كالتالي:

- 1- (LR) نموذج الانحدار اللوجستي عند عدم موازنة البيانات.
- 2- (LR+ROS) نموذج الانحدار اللوجستي في حال استخدام تقنية الزيادة العشوائية للبيانات ROS.
- 3- (LR+RUS) نموذج الانحدار اللوجستي في حال استخدام تقنية الإنقاص العشوائي للبيانات RUS
- 4- (LR+SMOTE) نموذج الانحدار اللوجستي في حال استخدام تقنية توليد بيانات تركيبية مصطنعة.

1. المعاملات الفائقة لنموذج الانحدار اللوجستي:

يوضح الجدول (14) القيم الأفضل للمعاملات الفائقة لنموذج الانحدار اللوجستي في حال عدم موازنة البيانات، وفي حال استخدام تقنيات التوزين (ROS، RUS، SMOTE)، والتي تعطي أفضل

نظام تشخيص آلي لمرض السكري باستخدام خوارزميات التصنيف، صالح.

كفاءة وأداء لهذا النموذج مقارنة مع باقي نطاق القيم المسندة للمعاملات الفائقة التي اختير منها القيم الأفضل باستخدام البحث الشبكي والتحقق المتقاطع.

الجدول (14) المعاملات الفائقة لنماذج الانحدار اللوجستي قبل وبعد موازنة البيانات.

المعاملات الفائقة	النموذج
C:0.05,class_weight:none,max_iter:1000,penalty:l2,solver:newton-cg	LR
C:0.06,class_weight:none,max_iter:1000,penalty:l2,solver:lbfgs	LR+ROS
C:0.02,class_weight:none,max_iter:1000,penalty:l2,solver:lbfgs	LR+SMOTE
C:0.05,class_weight:none,max_iter:1000,penalty:l2,solver:lbfgs	LR+RUS

حيث تم اختيار التنظيم L2 مع جميع النماذج ليتناسب مع الحل الذي تم اختياره وهو Newton_cg لنموذج الانحدار اللوجستي في حال عدم موازنة البيانات، بينما تم اختيار الحل lbfgs مع نماذج الانحدار اللوجستي في حالة موازنة البيانات باستخدام تقنيات التوزين السابقة. بينما قيمة (Max_iter) الحد الأقصى من التكرارات للوصول إلى حل فقد تم اختيارها لجميع النماذج 1000.

أما بالنسبة لقيمة C لهذه النماذج تراوحت بين (0.06-0.02) ويمكن اعتبارها مقبولة كسمادية لضبط الأخطاء.

وهذه كانت أفضل قيم للمعاملات الفائقة المختارة التي استخدمت في بناء هذا النموذج للوصول إلى النموذج ذي الكفاءة والأداء الأفضل مقارنة مع النماذج الأخرى المبنية باستخدام نطاق القيم

نظام تشخيص آلي لمرض السكري باستخدام خوارزميات التصنيف، صالح.

المختارة بواسطة البحث الشبكي والتحقق المتقاطع التي ستفضي إلى النموذج الأعلى دقة وكفاءة.

2. تقييم أداء نماذج خوارزمية الانحدار اللوجيستي:

جرى تقييم أداء النماذج النهائية للانحدار اللوجيستي من خلال معايير الأداء التالية (مصفوفة الارتباك الدقة، درجة F1، المساحة تحت منحنى الخواص AUC) المبينة في الجدول (15)، وذلك بهدف الوصول إلى النموذج الأعلى أداء وكفاءة.

الجدول (15) نتائج معايير الأداء لنماذج التنبؤ باستخدام LR.

Recall	Specificity	Confusion matrix	AUC	F1_score	Accuracy	Model
81%	81%	$\begin{bmatrix} 52 & 3 \\ 8 & 16 \end{bmatrix}$	0.81	82%	86%	LR
81%	81%	$\begin{bmatrix} 48 & 7 \\ 6 & 18 \end{bmatrix}$	0.81	81%	84%	LR+ROS
81%	81%	$\begin{bmatrix} 48 & 7 \\ 6 & 18 \end{bmatrix}$	0.81	81%	84%	LR+SMOTE
80%	80.5%	$\begin{bmatrix} 45 & 10 \\ 5 & 19 \end{bmatrix}$	0.80	79%	81%	LR+RUS

يبين الجدول (15) أن دقة نماذج الانحدار اللوجيستي قد تراوحت بين (81%-86%)، حيث حققت أعلى قيمة لها في حال عدم موازنة البيانات وكانت تساوي 86%، أما بالنسبة للقيمة الأقل منها فقد كانت في حال موازنة البيانات باستخدام تقنيتي (SMOTE،ROS)، بينما أقل قيمة للدقة كانت في حال موازنة البيانات باستخدام تقنية (RUS) والتي تساوي 81%.

كما تبين أن درجة F1 لنماذج الانحدار اللوجيستي قد تراوحت بين (79%-82%)، حيث حققت أعلى قيمة لها في حال عدم موازنة البيانات وكانت تساوي 82%، أما بالنسبة للقيمة الأقل منها فقد كانت في حال موازنة البيانات باستخدام تقنيتي (ROS,SMOTE)، بينما أقل قيمة للدقة كانت في حال موازنة البيانات باستخدام تقنية (RUS) والتي تساوي 81%.

أما بالنسبة لقيمة المساحة تحت منحنى الخواص لنماذج الانحدار اللوجيستي قد حققت أعلى قيمة لها في حال عدم موازنة البيانات وموازنة البيانات باستخدام تقنيتي (SMOTE، ROS) حيث بلغت 0.81، وكانت أفضل عتبة تصنيف لنموذج الانحدار اللوجيستي في حال عدم موازنة البيانات تساوي 0.42، وكانت أفضل عتبة تصنيف لنموذج الانحدار اللوجيستي في حال موازنة البيانات باستخدام تقنية ROS تساوي 0.578، أما عند موازنة البيانات باستخدام تقنية SMOTE فقد كانت أفضل عتبة تصنيف لنموذج الانحدار اللوجيستي تساوي 0.582.

بينما أقل قيمة للمساحة كانت في حال موازنة البيانات باستخدام تقنية (RUS) والتي تساوي 0.80، حيث كانت أفضل عتبة تصنيف لنموذج الانحدار اللوجيستي قيمتها تساوي 0.62.

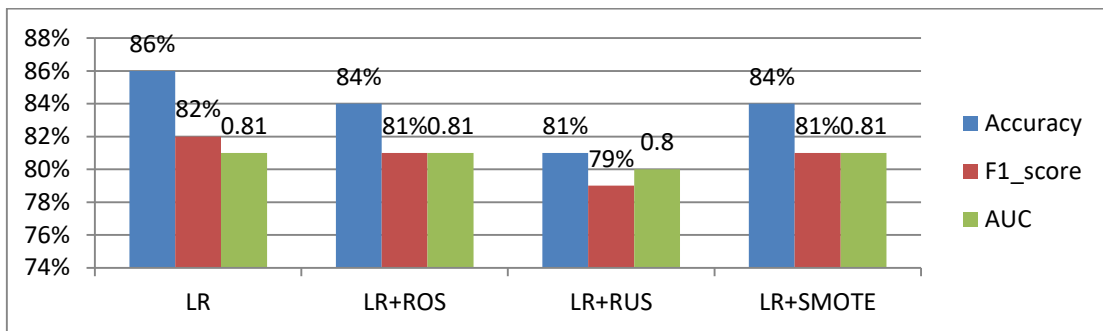
أما بالنسبة لمصفوفة الارتباك استطاع نموذج الانحدار اللوجيستي في حال عدم موازنة البيانات التنبؤ بشكل صحيح عن الحالات السليمة والمريضة، فمن عينات الاختبار لمرضى السكري نجح بالتنبؤ عن 16 حالة مريضة بشكل صحيح وأخطأ بمقدار 8، كما تنبأ عن 52 حالة سليمة بشكل صحيح وأخطأ بمقدار 3 حالات، بينما استطاع كل من نموذجي الانحدار اللوجيستي في حال موازنة البيانات باستخدام تقنيتي (SMOTE، ROS) التنبؤ بشكل صحيح عن الحالات المريضة بمقدار أعلى من النموذج الأول، حيث نجحاً بالتنبؤ بمقدار (18) حالة مريضة وأخطأ هذين النموذجين بمقدار 6 حالات، وفي حال موازنة البيانات باستخدام تقنية RUS استطاع نموذج الانحدار اللوجيستي في هذه الحالة التنبؤ عن مرضى السكري بشكل صحيح بمقدار أعلى من بقية النماذج، حيث تنبأ النموذج عن 19 حالة مريضة بشكل صحيح وأخطأ بمقدار 5 حالات .

3. مقارنة بين نماذج الانحدار اللوجيستي:

عند المقارنة بين أداء نموذج الانحدار اللوجيستي في حال عدم موازنة البيانات وفي حال موازنة البيانات باستخدام تقنيات التوزين كما هو موضح في الشكل (26):

1. أبدى النموذج أداء أفضل في حال عدم التوزين مقارنة مع أداء النموذج في حال استخدام تقنيات

التوزين الثلاثة السابقة



الشكل (26) نتائج نماذج LR للتنبؤ بمرض السكري قبل وبعد استخدام تقنيات التوزين.

2. نموذج الانحدار اللوجيستي أبدى نفس الأداء عند استخدام تقنيتي توزيع البيانات (ROS)

و (SMOTE)، حيث كان لهما نفس قيم معايير الأداء، فقد كانت قيمة الدقة لهما 84%، ودرجة

F1 بلغت 81%.

3. أبدى النموذج في حال توزيع البيانات باستخدام تقنية RUS أداء أقل مقارنة مع أدائه في حالتي

(SMOTE, ROS).

4. أفضل نموذجين تم الحصول عليهما لخوارزمية الانحدار اللوجيستي: نموذج الانحدار اللوجيستي في

حال عدم توزيع البيانات، نموذج الانحدار اللوجيستي في حال توزيع البيانات باستخدام

تقنيتي (ROS, SMOTE).

5. تقنيتي توزيع البيانات ROS، SMOTE أفضل مقارنة مع تقنية التوزين RUS.

2.2.6 نتائج نماذج خوارزمية الغابة العشوائية

تم بناء أربعة نماذج باستخدام خوارزمية الغابة العشوائية لتشخيص مرض السكري باستخدام قاعدة

البيانات PID، وهذه النماذج هي كالتالي:

1. (RF) نموذج الغابة العشوائية عند عدم موازنة البيانات.
2. (RF+ROS) نموذج الغابة العشوائية في حال استخدام تقنية الزيادة العشوائية للبيانات.
3. (RF+RUS) نموذج الغابة العشوائية في حال استخدام تقنية الإنقاص العشوائي للبيانات.
4. (RF+SMOTE) نموذج الغابة العشوائية في حال استخدام تقنية توليد بيانات تركيبية مصطنعة.

1. المعاملات الفائقة لنموذج الغابة العشوائية:

يوضح الجدول (16) القيم الأفضل للمعاملات الفائقة لنموذج الغابة العشوائية في حال عدم موازنة البيانات، وفي حال استخدام تقنيات التوزيع (SMOTE،RUS،ROS)، والتي تعطي أفضل كفاءة وأداء لهذا النموذج مقارنة مع باقي نطاق القيم المسندة للمعاملات الفائقة التي اختير منها القيم الأفضل باستخدام البحث الشبكي والتحقق المتقاطع.

الجدول (16) المعاملات الفائقة لنماذج الغابة العشوائية قبل وبعد موازنة البيانات.

المعاملات الفائقة	النموذج
'Criterion':gini,'max_depth':13,'max_features':sqrt('n' : 40	RF
'Criterion':gini,'max_depth':11,'max_features':log2, 'n' : 20	RF+ROS
'Criterion':gini,'max_depth':13,'max_features':sqrt, 'n' :140	RF+SMOTE
'Criterion' :entropy,'max_depth':4,'max_features':log2, 'n' :20	RF+RUS

حيث تم اختيار معيار criterion ليكون من نوع Gini لنموذج RF الغابة العشوائية في حال عدم موازنة البيانات، وفي حالتي توزيع البيانات باستخدام تقنيتي التوزيع (SMOTE،ROS)، بينما في حال توزيع البيانات باستخدام تقنية RUS تم اختيار المعيار من نوع Entropy.

حيث تم اختيار معيار criterion ليكون من نوع Gini لنموذج الغابة العشوائية في حال عدم موازنة البيانات، وفي حالتي توزيع البيانات باستخدام تقنيتي التوزيع (SMOTE، ROS)، بينما في حال توزيع البيانات باستخدام تقنية RUS تم اختيار المعيار من نوع Entropy.

وبالنسبة إلى قيمة max_depth فتم اختيار أعلى قيمة 13 مع نموذجي الغابة العشوائية عند عدم موازنة البيانات وعند التوزيع باستخدام تقنية SMOTE، حيث كانت هذه القيمة أقل بقليل في حال موازنة البيانات باستخدام تقنية ROS وتساوي 11، بينما كانت أقل قيمة لها عند توزيع البيانات باستخدام تقنية RUS وتساوي 4.

في حين تم اختيار Sqrt لتكون قيمة max_features لنموذج الغابة العشوائية في حال عدم موازنة البيانات، وعند توزيع البيانات باستخدام تقنية التوزيع (SMOTE)، وجرى انتقاء Log2 لتكون قيمة max_features لنموذج الغابة العشوائية في حال موازنة البيانات باستخدام تقنيتي التوزيع (ROS، RUS).

كما كانت أكبر قيمة لعدد الأشجار المختارة هي 140 وذلك لنموذج الغابة العشوائية عند موازنة البيانات باستخدام تقنية SMOTE، والقيمة الأقل بقليل تساوي 40 عند عدم موازنة البيانات، بينما أقل قيمة لها كانت عند توزيع البيانات باستخدام (ROS، RUS) وكانت تساوي 20.

وهذه كانت أفضل قيم للمعاملات الفائقة المنتقاة التي استخدمت في بناء هذه النماذج للوصول إلى RF النهائي ذي الكفاءة والأداء الأفضل مقارنة مع نماذج RF الأخرى المبنية باستخدام نطاق القيم المسندة للمعاملات الفائقة المنتقاة بواسطة البحث الشبكي والتحقق المتقاطع التي ستقضي إلى النموذج الأعلى دقة وكفاءة.

2. تقييم أداء نماذج الغابة العشوائية:

جرى تقييم أداء النماذج النهائية للغابة العشوائية من خلال معايير الأداء التالية: (مصفوفة الارتباك، الدقة، درجة F1، المساحة تحت منحنى الخواص AUC) المبينة في الجدول (4-6)، وذلك بهدف الوصول إلى النموذج الأعلى أداء وكفاءة.

الجدول (17) نتائج معايير الأداء لنماذج التنبؤ بمرض السكري باستخدام الغابة العشوائية.

Recall	specificity	Confusion matrix	Roc	F1 score	Accuracy	Model
82%	81%	$\begin{bmatrix} 53 & 2 \\ 8 & 16 \end{bmatrix}$	0.81	84%	87%	RF
82%	81%	$\begin{bmatrix} 49 & 6 \\ 6 & 18 \end{bmatrix}$	0.82	82%	85%	RF+ROS
78%	78%	$\begin{bmatrix} 49 & 6 \\ 8 & 16 \end{bmatrix}$	0.77	79%	82%	RF+SMOTE
84%	84%	$\begin{bmatrix} 44 & 11 \\ 3 & 21 \end{bmatrix}$	0.83	81%	82%	RF+RUS

يبين الجدول (17) أن دقة نماذج الغابة العشوائية قد تراوحت بين (82%-87%)، حيث حققت أعلى قيمة لها في حال عدم موازنة البيانات وكانت تساوي 87%، أما بالنسبة للقيمة الأقل منها فقد كانت في حال موازنة البيانات باستخدام تقنية (ROS)، والأقل بقليل كانت في حال موازنة البيانات باستخدام ROS وتساوي 85%، بينما أقل قيمة للدقة كانت في حال موازنة البيانات باستخدام تقنيتي (RUS, SMOTE) والتي تساوي 82%.

كما تبين أن درجة F1 لنماذج الانحدار اللوجستي قد تراوحت بين (79%-84%)، حيث حققت أعلى قيمة لها في حال عدم موازنة البيانات وكانت تساوي 84%، أما بالنسبة للقيمة الأقل منها فقد كانت في حال موازنة البيانات باستخدام تقنيتي (ROS) وتساوي 82%، بينما أقل قيمة للدقة كانت في حال موازنة البيانات باستخدام تقنية (SMOTE) والتي تساوي 79%.

أما بالنسبة لقيمة المساحة تحت منحنى الخواص لنماذج الغابة العشوائية فقد حققت أعلى قيمة لها في حال موازنة البيانات باستخدام تقنية (RUS) والتي تساوي 0.83.

بينما عند موازنة البيانات باستخدام تقنيتي (ROS) فقد بلغت المساحة تحت منحنى الخواص 0.82، والأقل بقليل كانت عند عدم موازنة البيانات وتساوي 0.81، بينما أقل قيمة للمساحة كانت في حال موازنة البيانات باستخدام تقنية (SMOTE) والتي تساوي 0.77.

أما بالنسبة لمصفوفة الارتباك استطاع نموذج الغابة العشوائية في حال عدم موازنة البيانات التنبؤ بشكل صحيح عن الحالات السليمة والمريضة، فمن عينات الاختبار لمرضى السكري نجح بالتنبؤ عن 16 حالة مريضة بشكل صحيح وأخطأ بمقدار 8، كما تنبأ عن 53 حالة سليمة بشكل صحيح وأخطأ بمقدار حالتين، بينما استطاع نموذج الغابة العشوائية في حال موازنة البيانات باستخدام تقنية (ROS) التنبؤ بشكل صحيح عن الحالات المريضة بمقدار أعلى من النموذج الأول، حيث نجح بالتنبؤ بمقدار (18) حالة مريضة وأخطأ بمقدار 6 حالات، وفي حال موازنة البيانات باستخدام تقنية RUS استطاع نموذج الغابة العشوائية في هذه الحالة التنبؤ عن مرضى السكري بشكل صحيح بمقدار أعلى من بقية النماذج، حيث تنبأ النموذج عن 21 حالة مريضة بشكل صحيح وأخطأ بمقدار 3 حالات

3. مقارنة بين نماذج الغابة العشوائية:

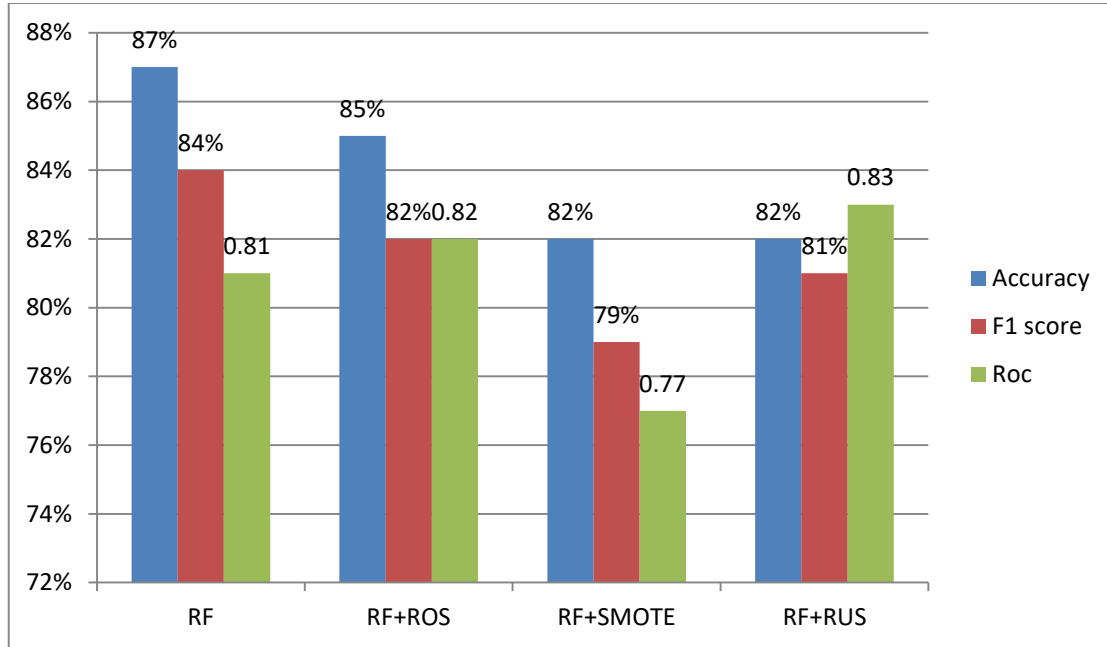
عند لمقارنة بين معايير الأداء لخوارزمية الغابة العشوائية قبل موازنة البيانات وبعد استخدام تقنيات التوزين (RUS، ROS، SMOTE) تبين كما هو موضح بالشكل (27) مايلي:

1. أبدى النموذج أداء أفضل في حال عدم التوزين مقارنة مع أدائه في حال استخدام تقنيات التوزين

الثلاثة السابقة، حيث بلغت قيمة الدقة 87%، بينما كانت قيمة درجة F1 تساوي 84%.

2. أبدى النموذج في حال توزيع البيانات باستخدام تقنية ROS أداء أفضل مقارنة مع أدائه في حالتي

(SMOTE،RUS)، حيث بلغت قيمة الدقة 85%، بينما كانت قيمة درجة F1 تساوي 82%.



الشكل (27) نتائج نماذج RF للتنبؤ بمرض السكري قبل وبعد استخدام تقنيات التوزيع.

3. أبدى النموذج في حال توزيع البيانات باستخدام تقنية SMOTE أداء أقل مقارنة مع أدائه في

حالتي (ROS،RUS).

4. أفضل نموذجين تم الحصول عليهما لخوارزمية الغابة العشوائية في حال عدم توزيع البيانات، في

حال توزيع البيانات باستخدام تقنيتي (ROS).

3.2.6 مقارنة بين خوارزميتي التصنيف:

عند المقارنة بين خوارزميتي التصنيف (الغابة العشوائية والانحدار اللوجستي) كما هو موضح في

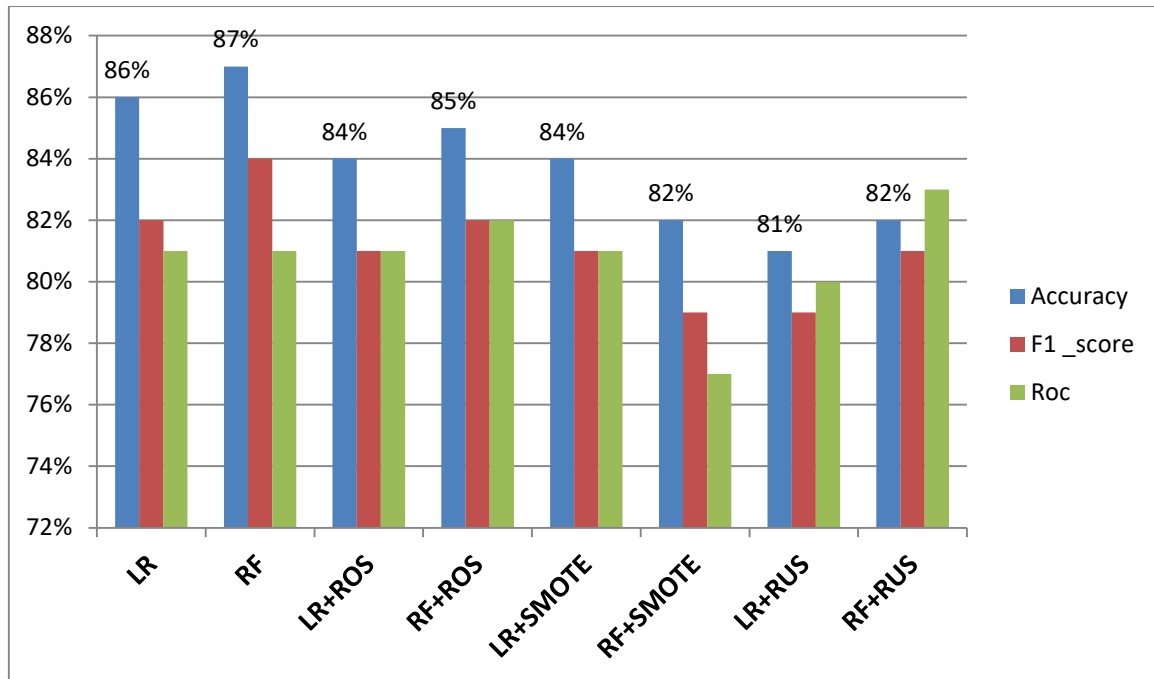
الشكل (28) نجد أن:

1. خوارزمية الغابة العشوائية هي الأفضل مقارنة مع خوارزمية الانحدار اللوجستي في حالة عدم

موازنة البيانات.

2. خوارزمية الغابة العشوائية أفضل في حال استخدام تقنيتي التوزين (ROS, RUS) مقارنة مع

خوارزمية الانحدار اللوجستي.



الشكل (28) مقارنة بين نماذج التنبؤ بمرض السكري باستخدام قاعدة البيانات PID.

3. أسوأ أداء للخوارزميتين:

عند استخدام تقنية التوزين (RUS) مع الانحدار اللوجستي.

عند استخدام تقنية (SMOTE) مع خوارزمية الغابة العشوائية.

4. أفضل تقنية توزين هي ROS.

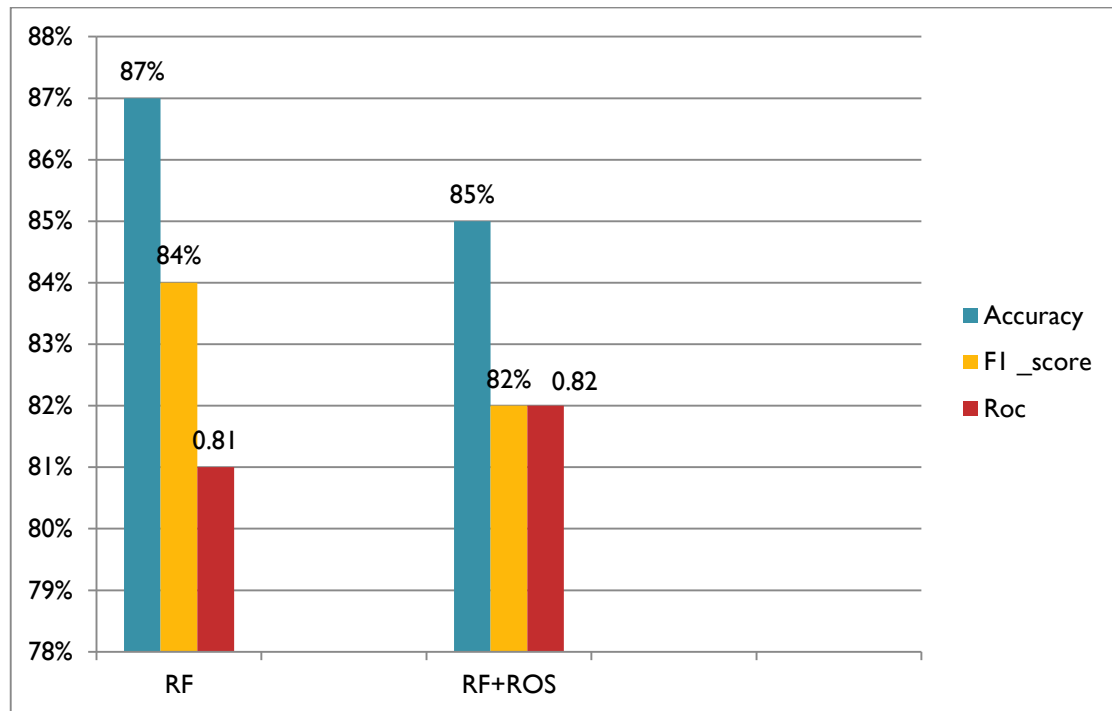
4.2.6 تحديد النموذج الأفضل:

أفضل نموذج كما هو مبين في الشكل (29)، هو خوارزمية الغابة العشوائية في حال عدم موازنة

البيانات وذلك بدقة 87%، ودرجة F1 تساوي 84%، والمساحة تحت منحنى الخواص 0.81، عند

توزين البيانات كانت الغابة العشوائية هي الأفضل باستخدام تقنية ROS، وذلك بدقة 85%،

ودرجة F1 82%، والمساحة تحت منحنى الخواص 0.82.



الشكل (29) تحديد النموذج الأفضل للتنبؤ بمرض السكري باستخدام قاعدة البيانات PID

5.2.6 مقارنة مع الدراسات المرجعية:

عند مقارنة نتائج هذه الدراسة مع نتائج الدراسات المرجعية فقد تبين من الجدولين (18) و (19) أن نتائج خوارزمية الانحدار اللوجستي والغابة العشوائية لدراستنا أفضل من نتائج الدراسات المرجعية قد يعود ذلك لعدة أسباب:

1. طريقة المعالجة المسبقة للبيانات (إزالة القيم المفقودة واستبدال القيم المتطرفة) لدراستنا تزيد من أداء

النموذج مقارنة مع جميع الدراسات المرجعية.

2. استخدام طريقة البحث الشبكي لضبط المعاملات الفائقة يحسن من أداء النموذج وهذا ما نراه في

(Al_Ahmed,2023) (Guptaa,2023)، حيث كانت معايير الأداء أفضل عند استخدام طريقة

البحث الشبكي لضبط المعاملات.

الجدول (18) مقارنة نتائج نماذج الانحدار اللوجستي للبحث مع نتائج الدراسات المرجعية.

References	Data set	Classification algorithms	Grid search	K_fold	Accuracy	F1_score	Roc
Al_Ahmed et al.2023	PID	LR	Yes	Yes	79%	74%	___
Al_Zain et al.2024	PID	LR	No	No	83.1%	72.3%	___
Mushtaq et al.2022	PID	LR	No	Yes	77.7%	___	0.683
	PID	LR+SMOTE	No	Yes	76.7%	___	0.765
	PID	LR+Tomek	No	Yes	77%	___	0.697
Talukder et al.2024	PID	LR	No	Yes	73%	73%	__
	PID	LR+ROS	No	Yes	78.95%	75.06%	___
Our study	PID	LR	Yes	Yes	86%	82%	0.81
Our study	PID	LR +ROS	Yes	Yes	84%	81%	0.81
Our study	PID	LR+SMOTE	Yes	Yes	84%	81%	0.81
Our study	PID	LR+RUS	Yes	Yes	81%	79%	0.80

الجدول (19) مقارنة نتائج نماذج الغابة العشوائية للبحث مع نتائج الدراسات المرجعية.

References	Data set	Classification algorithms	Grid search	K_fold	Accuracy	F1_score	Roc
Guptaa et al. 2023	PID	RF	Yes	Yes	88.61%	75.68%	___
Al_Ahmed et al.2023	PID	RF	Yes	Yes	77%	74%	___
Al_Zain et al.2024	PID	RF	No	No	83.1%	72.9%	___
Mushtaq et al.2022	PID	RF	No	Yes	79.7%	___	0.735
	PID	RF+SMOTE	No	Yes	80.7%	___	0.806
	PID	RF+Tomek	No	Yes	75.9%	___	0.698
Talukder et al.2024	PID	RF	No	Yes	73.38%	70.89%	__
	PID	RF +ROS	No	Yes	85.53%	83.06%	___
Our study	PID	RF	Yes	Yes	87%	84%	0.81
Our study	PID	RF +ROS	Yes	Yes	85%	82%	0.82
Our study	PID	RF+SMOTE	Yes	Yes	82%	79%	0.77
Our study	PID	RF+RUS	Yes	Yes	82%	81%	0.83

3. أداء نموذج الانحدار اللوجستي والغابة العشوائية في حال توزيع البيانات في دراستنا كانت أفضل

مقارنة مع نتائج الدراسات المرجعية أيضا [37-38] لأن أغلب الدراسات السابقة لم تستخدم طريقة

البحث الشبكي لضبط المعاملات الفائقة التي أدت إلى زيادة أداء نموذج دراستنا.

3.6 نتائج نماذج التصنيف للتنبؤ بمرض السكري باستخدام قاعدة البيانات المحلية:

1.3.6 نتائج التحليل الإحصائي:

1. نتائج اختبار شابيرو ويلك:

وفقا لنتائج تحليل اختبار شابيرو ويلك الذي طبق على المتغيرات الرقمية المبينة في الجدول (20) لتحديد طبيعة توزيع البيانات الرقمية، وجد أن جميع السمات الرقمية (العمر، الوزن، مؤشر كتلة الجسم، محيط الخصر، معدل ضربات القلب، مستوى سكر الصيام، ضغط الدم الانقباضي، ضغط الدم الانبساطي) موزعة بشكل غير طبيعي، بينما السمة الوحيدة الرقمية ذات التوزيع الطبيعي هي الطول.

الجدول (20) نتائج اختبار شابيرو ويلك

Shapiro-statistic	Shapiro-p_value	Factors
0.95	0.00	Age
0.97	0.0017	Weight
0.98	0.0103	BMI
0.97	0.001	Waist
0.84	0.000	Fasting Glucose
0.96	0.0002	DBP
0.98	0.007	SBP
0.97	0.0012	RHR

2. نتائج اختبار Mann Whitney:

عند تطبيق اختبار مان ويتني على البيانات الرقمية غير الموزعة بشكل طبيعي لوحظت فروق معند بها إحصائياً ($P < 0.05$) بين مجموعتي النتائج لمعظم السمات الرقمية الموضحة في الجدول (21)، مما يشير إلى قوة تنبؤية محتملة للإصابة بمرض السكري من النوع الثاني، وتحديدًا لوحظت قيم P منخفضة جدا للسمات (العمر، الوزن، مؤشر كتلة الجسم، محيط الخصر، تركيز الجلوكوز عند الصيام، ضغط الدم الانقباضي والانبساطي).

3. نتائج اختبار T-test:

عند تطبيق اختبار T-test على البيانات الرقمية الموزعة بشكل طبيعي وهي سمة الطول، لوحظ فروق معتد بها إحصائياً ($P < 0.05$) بين مجموعتي النتائج الموضحة في الجدول، مما يشير إلى وجود قوة تنبؤية محتملة بين الطول والاصابة بالسكري، وهذا ما يوضحه الجدول (21).

الجدول (21) نتائج التحليل الاحصائي باستخدام اختبار (T_test , Mannwithney)

significance	p_value	Factors
yes	0.00	Age
yes	0.00	Weight
yes	0.0012	High
yes	0.00483	BMI
yes	0.00	waist
yes	0.00	Fasting Glucose
yes	0.00	DBP
yes	0.00	SBP
yes	0.0012	RHR

4. نتائج اختبار Chi-square:

طبق اختبار كاي مربع لدراسة العلاقة بين السمات الفئوية والاصابة بمرض السكري من النوع الثاني ، وجد أن بعض السمات التي لها فروق يعتد بها إحصائياً عند قيم منخفضة جداً ($P < 0.05$) الموضحة في الملحق العلمي هي: (الجنس، مستوى التعليم، التدخين، عدد ساعات النوم، تناول الخضار، تناول الفواكه، تناول الأرز، تناول المشروبات المحلاة بالسكر، استهلاك زيت الزيتون، الأطعمة المقلية والمشوية، ارتفاع ضغط الدم، ارتفاع الكوليسترول، الشعور بالعطش والجوع، هبوط مفاجئ في الوزن، عدم وضوح الرؤية، تكرار التبول، وجود مشاكل بالقلب، تناول الأدوية بانتظام، اتباع نظام غذائي صحي ومتوازن، التاريخ العائلي للإصابة بمرض السكري، تميل الأطراف، ضيق التنفس، عدد مرات الحمل)،

بينما السمات الأخرى ليس لها فروق معتد بها إحصائياً عند قيم كبيرة جداً ($P > 0.05$) الموضحة في الجدول، مما تشير على عدم وجود قوة تنبؤية محتملة للإصابة بالسكري، وهذه السمات هي: (تناول البطاطا والمعجنات والخبز الأبيض، استهلاك زيت الذرة أو دوار الشمس، استهلاك الدهون الحيوانية أو النباتية، استهلاك اللحوم أو الدجاج أو السمك، استهلاك مشروبات الطاقة أو الغازية أو الحليب، تناول الأطعمة المطبوخة أو النيئة، الشعور بالغضب أو الخوف، وجود أمراض في الغدة الدرقية، وجود مشاكل بالكلية، الشعور بالغثيان أو الإمساك، الإصابة بالسكري الحلمي، ولادة طفل يزن أربع كيلو غرامات، المشي المنتظم لمدة نصف ساعة يومياً أو أكثر، عدد السجائر اليومية)، كما يوضح الملحق تكرارات كل متغير فنوي لكلا مجموعتي الخرج.

2.3.6 تحديد سمات التنبؤ بمرض السكري من النوع الثاني:

اعتماداً على نتائج الاختبارات الإحصائية الأربعة (اختبار شابيرو ويلك، اختبار مان ويتني، اختبار Chi-square، T_test) فقد تم استخلاص السمات الأكثر أهمية المتعلقة بمرض السكري، والتي عددها 35 سمة وهي موضحة في الجدول (22):

الجدول (22) تحديد سمات التنبؤ بالسكري من النوع الثاني.

Column name	P_Values	Test name	significances
Age	0.000	U_test	yes
Weight	0.000	U_test	yes
High	0.0012	T_test	yes
BMI	0.0048	U_test	yes
Waist circumferences	0.000	U_test	yes
Fasting Glucose	0.000	U_test	yes
DBP	0.000	U_test	yes
SBP	0.000	U_test	yes
RHR	0.0119	U_test	yes
Gender	0.0033	Chi test	Yes
Education Level	0.00272	Chi test	Yes
Smoking	0.01229	Chi test	Yes
Cigarettes	0.1399	Chi test	No
Sleep Hours	0.000	Chi test	Yes
Physical activity	0.350	Chi test	No
number of times	0.3465	Chi test	No
Eating vegetable	0.000	Chi test	Yes
Eating Fruit	0.0378	Chi test	Yes
Fish	1	Chi test	No
Chicken	0.2167	Chi test	No
Fat meat	0.1066	Chi test	No
Low fat meat	0.741	Chi test	No
Soft drinks	0.395	Chi test	No
Sugar drink	0.01953	Chi test	Yes
whole milk	0.386	Chi test	No
skim milk	0.683	Chi test	No
energy drink	1	Chi test	No
potato	0.084	Chi test	No
Rice	0.00190	Chi test	Yes
Noodles	0.267	Chi test	No
Pasta	0.549	Chi test	No
White bread	0.7355	Chi test	No
olive oil	0.00064	Chi test	Yes
Corn oil	0.608	Chi test	No
Sun oil	1	Chi test	No
Animal fat	0.157	Chi test	No
Plant fat	0.879	Chi test	No
Fried food	0.0009	Chi test	Yes
Cooked food	1	Chi test	No
Grilled food	0.0044	Chi test	Yes
Boiled food	0.246	Chi test	No

Un Cooked food	1	Chi test	No
healthy diet	0.00070	Chi test	Yes
Heart decease	0.0000	Chi test	Yes
High Blood Pressure	0.0000	Chi test	Yes
Regularly medications	00.000	Chi test	Yes
Family History	0.000	Chi test	Yes
Angry feelings	0.738	Chi test	No
Stress	0.725	Chi test	No
Urination frequency	0.000	Chi test	Yes
weight loss	0.00351	Chi test	Yes
permanent fatigue	0.00284	Chi test	Yes
Hungry	0.000	Chi test	Yes
Thirsty	0.000	Chi test	Yes
High cholesterol	0.000	Chi test	Yes
kidney problem	0.127	Chi test	No
Thyroid disease	0.283	Chi test	No
Shortness of breath	0.00151	Chi test	Yes
Eye problem	0.000	Chi test	Yes
Numbness	0.000	Chi test	Yes
Nausea a and constipation	0.467	Chi test	No
Birth of baby more than 4 kilogram	0.183	Chi test	No
Number of pregnancy	0.0307	Chi test	Yes
Gestational diabetes	0.271	Chi test	No

يبين الجدول (22) أن جميع السمات الرقمية والتي عددها 9 لها دلالة إحصائية للإصابة بمرض السكري،

إضافة إلى عدد كبير من المتغيرات الفئوية والتي عددها 26 (منها الجنس، مستوى التعليم، التدخين،

عدد ساعات النوم، تناول الخضار، تناول الفواكه، تناول الأرز، تناول المشروبات المحلاة بالسك،

استهلاك زيت الزيتون، الأطعمة المقلية والمشوية، ارتفاع ضغط الدم، ارتفاع الكوليسترول، الشعور

بالعطش والجوع، هبوط مفاجئ بالوزن، عدم وضوح الرؤية، تكرار التبول، وجود مشاكل بالقلب، تناول

الأدوية بانتظام، اتباع نظام غذائي صحي ومتوازن، التاريخ العائلي للإصابة بالسكري، تتميل الأطراف،

ضيق التنفس، عدد مرات الحمل)

3.3.6 تقييم أداء النماذج:

بعد الانتهاء من استخلاص السمات الأكثر أهمية المتعلقة بالسكري من النوع الثاني، فقد تم استخدام هذه السمات لبناء نماذج تصنيف قادرة على التنبؤ بالسكري باستخدام خوارزمية الغابة العشوائية، وذلك بعد تطبيق طريقة تقليل السمات العودية RFE للحصول على أفضل (6،8،10) سمات يتم اختيارها لبناء نماذج التصنيف وتدريبها، ومن ثم استخدام مقاييس الأداء للمقارنة فيما بينها واختيار أفضل نموذج مع أقل عدد من السمات، وهذه النماذج هي:

1. نموذج الغابة العشوائية مع أفضل عشر سمات:

يوضح الجدول (23) السمات العشرة الأفضل التي تم تحديدها باستخدام خوارزمية تقليل السمات العودية RFE، وهذه السمات هي: (العمر، محيط الخصر، مستوى الجلوكوز عند الصيام، الوزن، ارتفاع الكوليسترول، معدل نبضات القلب، ارتفاع ضغط الدم، ضغط الدم الانبساطي، عدد ساعات النوم، وجود مشاكل في العين

الجدول (23) تحديد أفضل عشر سمات للتنبؤ بالسكري باستخدام خوارزمية RFE

Top RFE 10 Features	أفضل عشر سمات
Age	العمر
Waist	محيط الخصر
Fasting Glucose	مستوى الجلوكوز عند الصيام
Weight	الوزن
High cholesterol	ارتفاع الكوليسترول
Eye Problem	وجود مشاكل في العين
DBP	ضغط الدم الانبساطي
High Blood Pressure	ارتفاع ضغط الدم
RHR	معدل نبضات القلب
Sleep Hours	عدد ساعات النوم في الليل

1. المعاملات الفائقة للنموذج:

يوضح الجدول (24) القيم الأفضل للمعاملات الفائقة لنموذج الغابة العشوائية عند استخدام طريقة تخفيض السمات العودية والتي تم من خلالها انتقاء أفضل عشر سمات الموضحة في الجدول (23)، والتي تعطي أفضل كفاءة وأداء لهذا النموذج مقارنة مع باقي نطاق القيم المسندة للمعاملات الفائقة التي اختير منها القيم الأفضل باستخدام البحث الشبكي والتحقق المتقاطع.

الجدول (24) نتائج أداء نموذج RF مع أفضل 10 سمات للتنبؤ بالسكري.

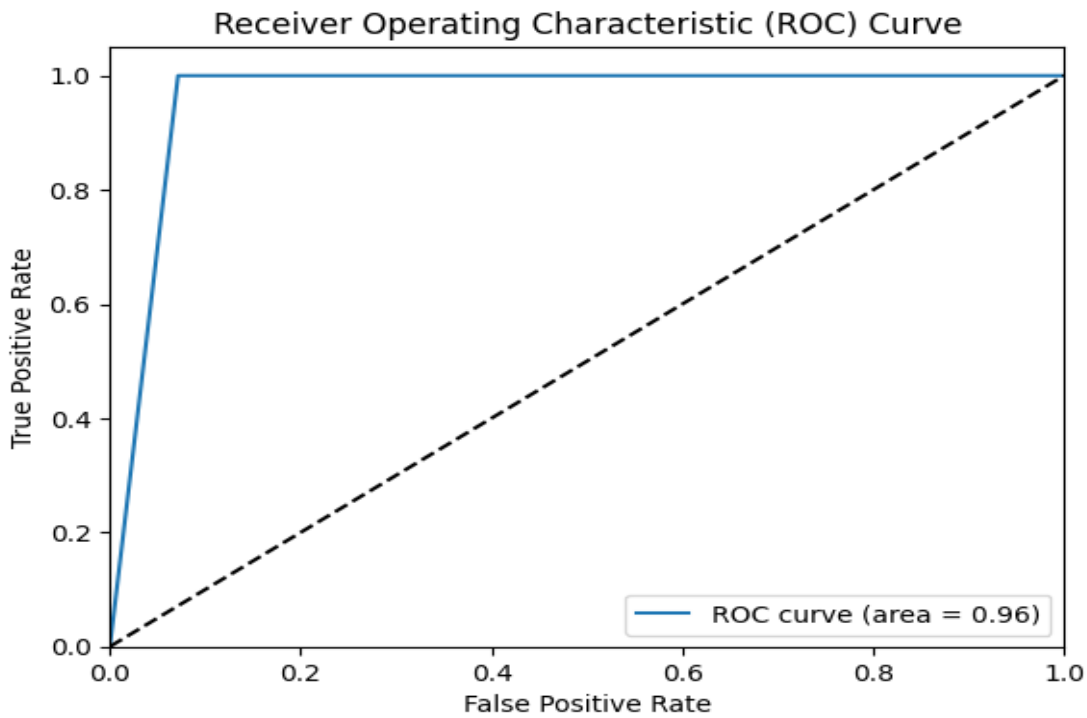
Confusion matrix	Roc	F1_score	Accuracy	Hyper parameters
$\begin{bmatrix} 13 & 1 \\ 0 & 18 \end{bmatrix}$	0.96	97%	97%	Criterion: gini, max_depth:3 Max_features :Sqrt,n:80

حيث تم اختيار معيار criterion ليكون من نوع Gini لنموذج RF، وبالنسبة إلى قيمة max_depth فتم اختيار قيمة 3 له، في حين تم اختيار Sqrt لتكون قيمة max_features، كما كانت قيمة عدد الأشجار المختارة هي 80.

2. معايير الأداء للنموذج:

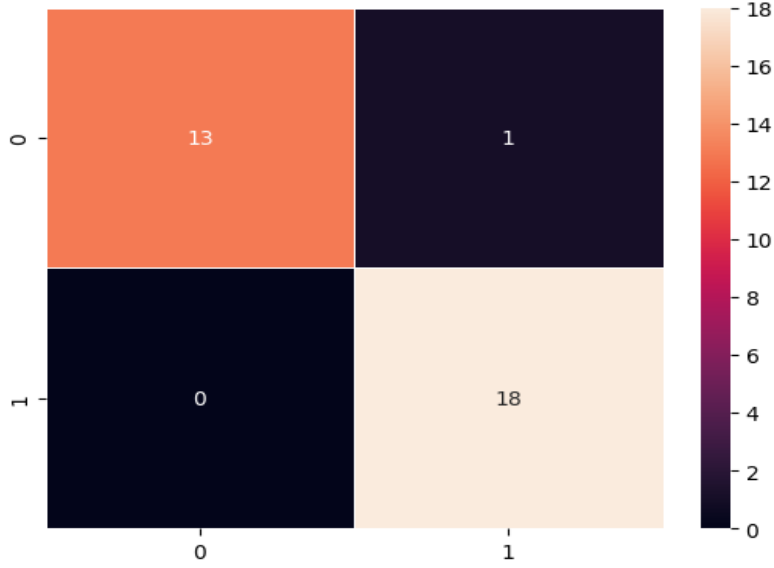
في النهاية لابد من تقييم أداء وكفاءة النموذج كاملاً حيث جرى حساب الدقة، والمساحة تحت منحنى الخواص ودرجة F1 التي تعد قيمة شاملة لتحديد كفاءة النموذج في كل صنف كونها تدمج بين الحساسية والاحكام للنموذج ككل، ويوضح الجدول (24) قيم هذه المقاييس، حيث تم حساب قيم هذه المقاييس كنسب مئوية فقد كانت درجة F1 تساوي 97%، وقيمة الدقة تساوي 97%.

أما بالنسبة للمساحة تحت منحنى الخواص التي تعتبر مقياساً لقدرة النموذج على التمييز بين الحالات السليمة والمريضة فقد بلغت 0.96، كما هو موضح في الشكل (30) حيث تعد هذه القيمة مثالية أي أن النموذج قادر على التمييز بين الحالات السليمة والمريضة بشكل كبير وذلك عند أفضل عتبة تصنيف قيمتها تساوي 0.659، وذلك عند أفضل قيمة للحساسية والتي تساوي 94% وأقل قيمة للحالات السلبية FPR والتي تساوي 0.



الشكل (30) منحنى الخواص ROC لنموذج RF للتنبؤ بالسكري باستخدام 10 سمات.

كما تبين من مصفوفة الارتباك الموضحة في الشكل (31) أن النموذج نجح في تصنيف 31 حالة من جميع عينات الاختبار، وأخطأ النموذج في التصنيف بين الحالات السليمة والمريضة بالسكري بمقدار حالة واحدة، فمن عينات الاختبار لمرضى السكري صنف 18 حالة صحيحة، ولم يخطئ في تصنيف الحالات المريضة بالسكري، ومن عينات الاختبار السليمة صنف 13 حالة صحيحة، وأخطأ بمقدار 1 حيث صنفها على أنها مريضة لكنها في الحقيقة حالة سليمة.



الشكل (31) مصفوفة الارتباك لنموذج RF مع 10 سمات للتنبؤ.

2- نموذج الغابة العشوائية مع أفضل ثمانية سمات:

يوضح الجدول (25) السمات الثمانية الأفضل التي تم تحديدها باستخدام خوارزمية تقليل السمات

العودية RFE، وهذه السمات هي: (العمر، محيط الخصر، مستوى الغلوكوز الصيامي، الوزن، ارتفاع

الكوليسترول، ارتفاع ضغط الدم، ضغط الدم الانبساطي، وجود مشاكل في العين.

الجدول (25) تحديد أفضل ثمانية سمات للتنبؤ بالسكري باستخدام خوارزمية RFE

Top RFE 8 Features	أفضل ثمانية سمات
Age	العمر
Waist	محيط الخصر
Fasting Glucose	مستوى الغلوكوز الصيامي
Weight	الوزن
High cholesterol	ارتفاع الكوليسترول
Eye Problem	وجود مشاكل في العين
DBP	ضغط الدم الانبساطي
High Blood Pressure	ارتفاع ضغط الدم

1. المعاملات الفائقة للنموذج:

يوضح الجدول (26) القيم الأفضل للمعاملات الفائقة لنموذج الغابة العشوائية عند استخدام طريقة تقليل السمات العودية والتي جرى من خلالها اختيار أفضل ثمانية سمات الموضحة في الجدول (25)، والتي تعطي أفضل كفاءة وأداء لهذا النموذج مقارنة مع باقي نطاق القيم المسندة للمعاملات الفائقة التي اختير منها القيم الأفضل باستخدام البحث الشبكي والتحقق المتقاطع.

الجدول (26) نتائج أداء نموذج RF مع أفضل 8 سمات للتنبؤ بالسكري.

Confusion matrix	Roc	F1_score	Accuracy	Hyper parameters
$\begin{bmatrix} 12 & 2 \\ 0 & 18 \end{bmatrix}$	0.93	94%	94%	Criterion:gini ,max_depth:5, n:40 Max_features :Sqrt

حيث تم اختيار معيار criterion ليكون من نوع Gini لنموذج RF، وبالنسبة لقيمة max_depth فتم اختيار قيمة 5 له، في حين تم اختيار Sqrt لتكون قيمة max_features، كما كانت قيمة عدد الأشجار المختارة هي 40.

وهذه كانت أفضل قيم للمعاملات الفائقة المختارة التي استخدمت في بناء هذا النموذج للوصول إلى RF النهائي ذي الكفاءة والأداء الأفضل مقارنة مع نماذج RF الأخرى المبنية باستخدام نطاق القيم المسندة للمعاملات الفائقة.

2. معايير الأداء للنموذج:

في النهاية لابد من تقييم أداء وكفاءة النموذج كاملاً، حيث تم حساب الدقة، والمساحة تحت منحنى الخواص ودرجة F1 التي تعد قيمة شاملة لتحديد كفاءة النموذج في كل صنف كونها تدمج بين الحساسية

والاحكام للنموذج ككل، ويوضح الجدول (26) قيم هذه المقاييس، حيث تم حساب قيم هذه المقاييس

كنسب مئوية فقد كانت درجة F1 تساوي 94%، وقيمة الدقة تساوي 94%

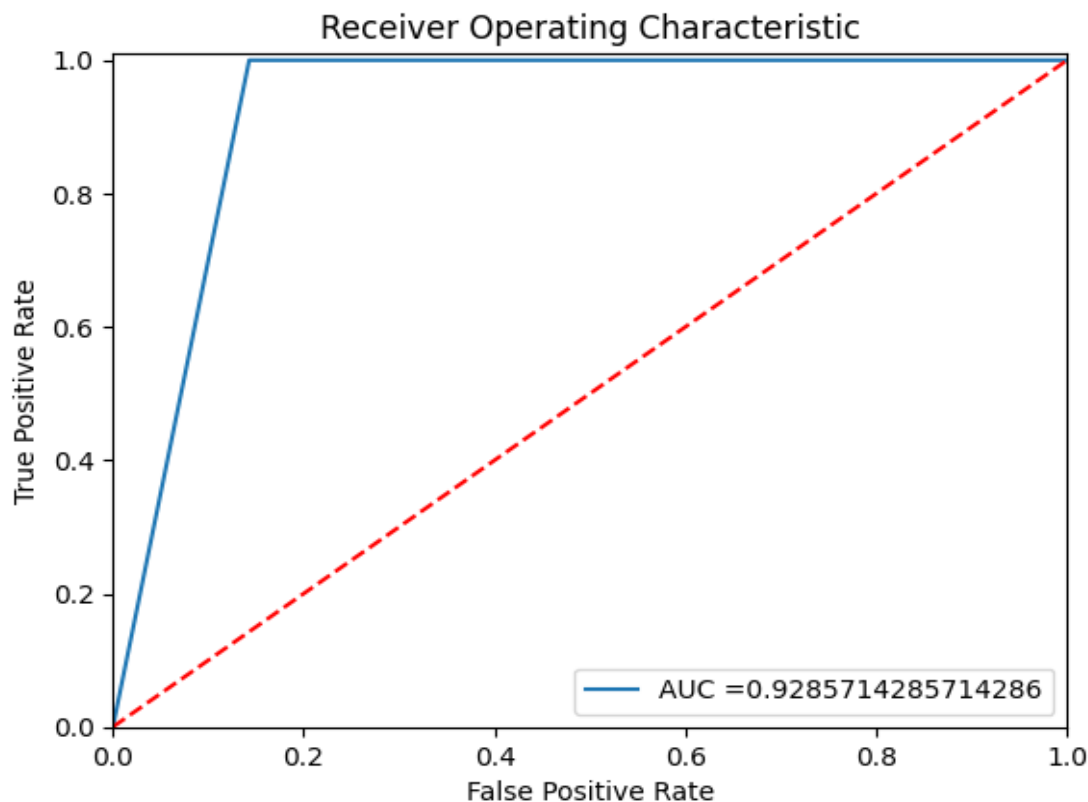
أما بالنسبة للمساحة تحت منحنى الخواص التي تعد مقياساً لقدرة النموذج على التمييز بين الحالات

السليمة والمريضة فقد بلغت 0.93، كما هو موضح في الشكل (32) حيث تعتبر هذه القيمة مثالية أي

أن النموذج قادر على التمييز بين الحالات السليمة والمريضة بشكل كبير عند أفضل عتبة تصنيف

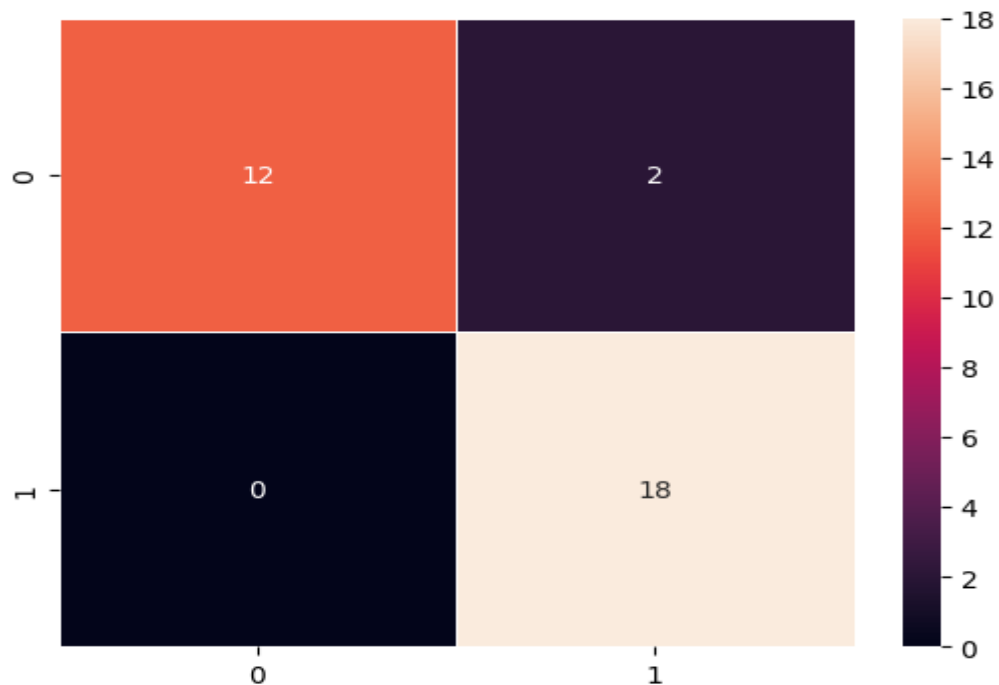
قيمتها تساوي 0.577، وذلك عند أفضل قيمة للحساسية والتي تساوي 94% وأقل قيمة للحالات

السلبية FPR والتي تساوي 0.



الشكل (32) منحنى الخواص ROC لنموذج RF للتنبؤ بالسكري باستخدام 8 سمات.

وتبين من مصفوفة الارتباك الموضحة في الشكل (33) أن النموذج نجح في تصنيف 30 حالة من جميع عينات الاختبار، وأخطأ التصنيف بين الحالات السليمة والمريضة بالسكري بمقدار 2، فمن عينات الاختبار لمرضى السكري صنف 18 حالة مريضة بشكل صحيح، ولم يخطئ في تصنيف الحالات المريضة بالسكري، ومن عينات الاختبار السليمة صنف 12 حالة سليمة بشكل صحيح، وأخطأ بمقدار 2 حيث صنفها على أنها مريضة لكنها في الحقيقة حالات سليمة.



الشكل (33) مصفوفة الارتباك لنموذج RF مع 8 سمات للتنبؤ

3- نموذج الغابة العشوائية مع أفضل ست سمات:

يوضح الجدول (27) السمات الستة الأفضل التي تم تحديدها باستخدام خوارزمية تقليل السمات العودية RFE، وهذه السمات هي: (العمر، محيط الخصر، مستوى الغلوكوز عند الصيام، الوزن، ارتفاع الكوليسترول، ارتفاع ضغط الدم، ضغط الدم الانبساطي، وجود مشاكل في العين)

الجدول (27) تحديد أفضل ستة سمات للتنبؤ بمرض السكري باستخدام خوارزمية RFE

Top RFE 6 Features	أفضل ستة سمات
Age	العمر
Waist	محيط الخصر
Fasting Glucose	مستوى الجلوكوز عند الصيام
Weight	الوزن
High cholesterol	ارتفاع الكوليسترول
Eye Problem	وجود مشاكل في العين

1. المعاملات الفائقة للنموذج:

يوضح الجدول (28) القيم الأفضل للمعاملات الفائقة لنموذج الغابة العشوائية عند استخدام طريقة تقليل السمات العودية والتي تم من خلالها اختيار أفضل ستة سمات الموضحة في الجدول (27)، والتي تعطي أفضل كفاءة وأداء لهذا النموذج مقارنة مع باقي نطاق القيم المسندة للمعاملات الفائقة التي اختير منها القيم الأفضل باستخدام البحث الشبكي والتحقق المتقاطع.

الجدول (28) نتائج أداء نموذج RF مع أفضل 6 سمات للتنبؤ بمرض السكري.

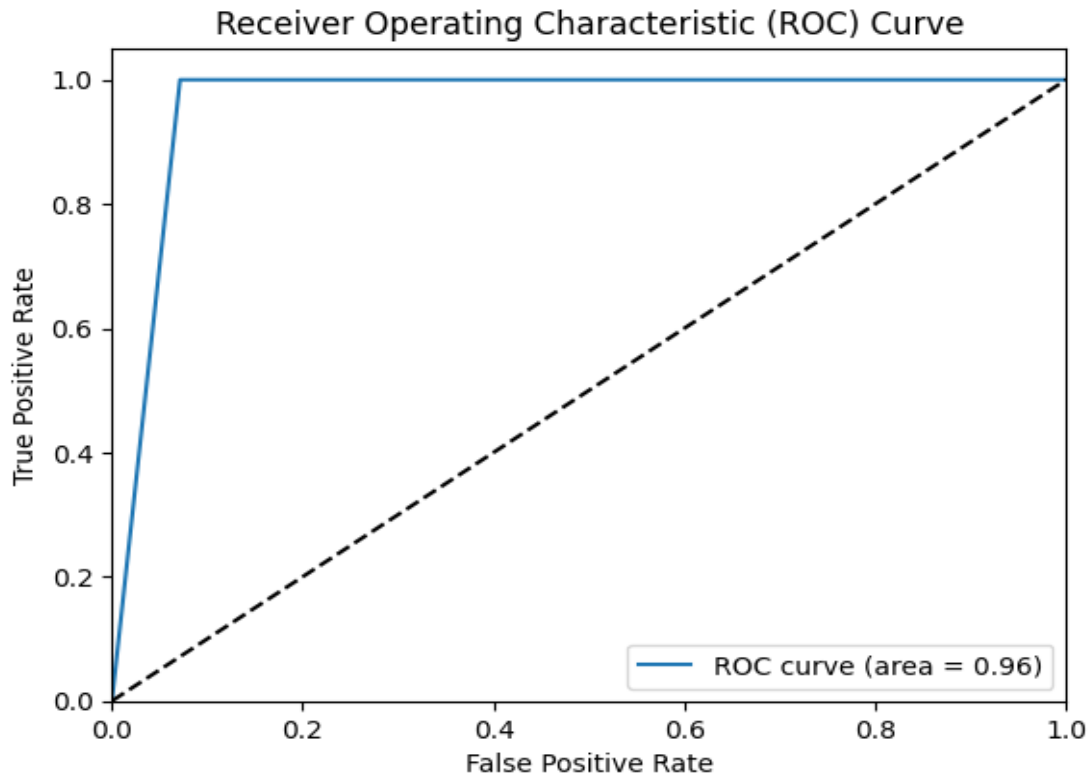
Confusion matrix	Roc	F1_score	Accuracy	Hyper parameters
$\begin{bmatrix} 13 & 1 \\ 0 & 18 \end{bmatrix}$	0.96	97%	97%	Criterion:gini,max_depth:4, n:40 Max_features :Sqrt

حيث تم اختيار معيار criterion ليكون من نوع Gini لنموذج RF، وبالنسبة إلى قيمة max_depth فتم اختيار قيمة 4 له، في حين تم اختيار Sqrt لتكون قيمة max_features، كما كانت قيمة عدد الأشجار المختارة هي 40.

وهذه كانت أفضل قيم للمعاملات الفائقة المختارة التي استخدمت في بناء هذا النموذج للوصول إلى RF النهائي ذي الكفاءة والأداء الأفضل مقارنة مع نماذج RF الأخرى المبنية باستخدام نطاق القيم المسندة للمعاملات الفائقة.

1. معايير الأداء للنموذج:

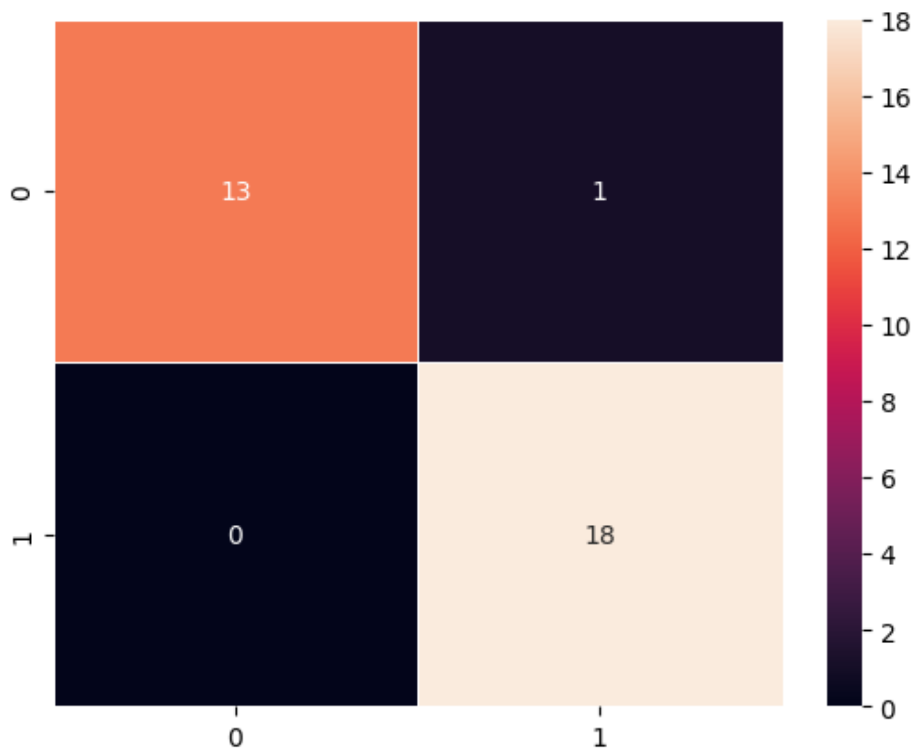
في النهاية لابد من تقييم أداء وكفاءة النموذج كاملاً حيث جرى حساب الدقة، والمساحة تحت منحنى الخواص ودرجة F1 التي تعتبر قيم شاملة لتحديد كفاءة النموذج في كل صنف كونها تدمج بين الحساسية والاحكام للنموذج ككل، ويوضح الجدول (34) قيم هذه المقاييس، حيث جرى حساب قيم هذه المقاييس كنسب مئوية فقد كانت درجة F1 تساوي 97%، وقيمة الدقة تساوي 97%.



الشكل (34) منحنى الخواص ROC لنموذج RF للتنبؤ بالسكري باستخدام 6 سمات.

أما بالنسبة للمساحة تحت منحنى الخواص التي تعتبر مقياساً لقدرة النموذج على التمييز بين الحالات السليمة والمريضة فقد بلغت 0.96، كما هو موضح في الشكل (34) حيث تعتبر هذه القيمة مثالية أي أن النموذج قادر على التمييز بين الحالات السليمة والمريضة بشكل كبير عند أفضل عتبة تصنيف قيمتها تساوي 0.653، وذلك عند أفضل قيمة للحساسية والتي تساوي 94% وأقل قيمة للحالات السلبية FPR والتي تساوي 0.

كما تبين من مصفوفة الارتباك الموضحة في الشكل (35) أن النموذج نجح في تصنيف 31 حالة من جميع عينات الاختبار، وأخطأ النموذج في التصنيف بين الحالات السليمة والمريضة بالسكري بمقدار حالة واحدة، فمن عينات الاختبار لمرضى السكري صنف 18 حالة صحيحة، ولم يخطأ في تصنيف الحالات المريضة بالسكري، ومن عينات الاختبار السليمة صنف 13 حالة صحيحة، وأخطأ بمقدار 1 حيث صنفها على أنها مريضة لكنها في الحقيقة حالات سليمة.



الشكل (35) مصفوفة الارتباك لنموذج RF مع 6 سمات للتنبؤ

4.3.6 مقارنة بين أداء نماذج الغابة العشوائية مع استخدام طريقة تقليل السمات العودية:

يظهر الجدول ملخص عن مقاييس الأداء التي جرى إيجادها للنموذج ككل مع سمات RFE لجميع النماذج التي تم استخدامها في هذه الدراسة بهدف المقارنة فيما بينها من أجل تحديد أفضل نموذج والذي يبدي أفضل أداء وكفاءة.

الجدول (29) مقارنة بين نماذج RF مع تقليل سمات التنبؤ بطريقة RFE

Confusion matrix	Roc	F1_score	Accuracy	Features	Hyper parameters
$\begin{bmatrix} 12 & 2 \\ 0 & 18 \end{bmatrix}$	0.93	94%	94%	Age, waist ,DBP ,Fasting Glucose High BP, High cholesterol ,Eye problem, weight	Criterion: gini, max_depth:5 Max_features:Sqrt,n:40
$\begin{bmatrix} 13 & 1 \\ 0 & 18 \end{bmatrix}$	0.96	97%	97%	Age ,waist , Fasting Glucose, High cholesterol ,Eye problem weight	Criterion: gini, max_depth:4 Max_features:Sqrt,n:40
$\begin{bmatrix} 13 & 1 \\ 0 & 18 \end{bmatrix}$	0.96	97%	97%	Age ,waist ,RHR ,Fasting Glucose ,High BP, High cholesterol ,Eye problem , Sleep Hours ,DBP, weight	Criterion: gini, max_depth:3 Max_features :Sqrt,n:80

يوضح الجدول (29) أن أداء نموذج الغابة العشوائية متقارب عند استخدام طريقة تقليل السمات العودية وذلك في حالتي تقليل السمات إلى ست سمات أو عشر سمات، حيث كانت قيمة الدقة ودرجة F1 لكل النموذجين تساوي 97%، وكذلك المساحة تحت منحنى الخواص لها نفس القيمة وتساوي 0.96، كما أن النموذجين قادرين على تصنيف الحالات المريضة بشكل صحيح دون حدوث أية أخطاء، لكنهما أخطأ في تصنيف الحالات السليمة بمقدار 1.

كما أبدى النموذج أداء أقل عند تقليل السمات إلى ثمانية سمات قد يعود السبب إلى المعلومات الفائقة التي تم ضبطها باستخدام طريقة البحث الشبكي والتحقق المتقاطع وخصوصا العمق الأعظمي، حيث

كانت الدقة ودرجة F1 تساوي 94%، أما بالنسبة للمساحة تحت منحنى الخواص فقد بلغت 0.93، كما أن النموذج قادر على تصنيف الحالات المريضة بدون حدوث أية أخطاء، لكنه أخطأ بالكشف عن الحالات السليمة بمقدار 2.

تبين أيضاً أن المعيار criterion الذي تم اختياره لهذه النماذج جميعها من نوع Gini، وكذلك الأمر بالنسبة لعدد السمات الأعظمي فقد كانت من نوع Sqrt، أما بالنسبة للعمق الأعظمي لشجرة القرار فقد تراوح بين القيم ثلاثة وخمسة، وعدد الأشجار كان قليل.

3.3.6 مقارنة مع الدراسات المرجعية:

عند مقارنة نتائج دراستنا بنتائج الدراسات المرجعية وجد أن خوارزمية اختيار السمات العودية RFE مع نموذج الغابة العشوائية أعطت نتائج أفضل مقارنة مع خوارزميات اختيار السمات المستخدمة مع خوارزمية الغابة العشوائية في الدراسات المرجعية.

نتائج التحليل الإحصائي لدراستنا متوافقة مع نتائج التحليل الإحصائي للدراسات المرجعية (وجود فروق ذات دلالة إحصائية لجميع السمات الرقمية، والسمات الفئوية التي تم تحديدها سابقاً [41]، لكنها غير متوافقة مع بعض السمات مثل (تناول الفواكه، النشاط الفيزيائي، تناول اللحوم والأسماك) كما في الدراسة [40].

بعض سمات التنبؤ التي تم تحديدها باستخدام خوارزمية تقليل السمات RFE المستخدمة في دراستنا متوافقة مع الدراسات المرجعية (العمر، ضغط الدم الانبساطي، مستوى الجلوكوز الصيامي في الدم) [41 - 40]، وبقيّة السمات مختلفة عما تم استخدامه في هذه الدراسة للكشف عن السكري، ويوضح

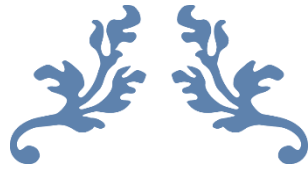
الجدول (30) مقارنة نتائج البحث مع نتائج الدراسات العالمية المماثلة.

الجدول (30) مقارنة نتائج البحث مع نتائج الدراسات المرجعية.

Referances	Dataset	Risk Factors	Feature selection	Accuracy	F1
Jahanur et al .2022	NHANCES	Age , Education, marital status ,SBP, Smoking, BMI	RF + multiple logistic regression	94.8%	94.1%
		Age , Race, marital status ,SBP, DBP , Smoking ,BMI , Direct cholesterol, Physical activity		95.9%	95.3%
Maniruzza man et al.2020	NHANCES	Age , SBP, DBP , BMI ,Direct cholesterol, total cholesterol	RF + multiple logistic regression	94.15%	96.3%
Viswanatha et al.2023	Vanderbilt	Not defined	LR+ univariate Feature selection	89%	80%
Kakoly et al.2023	Bangladesh	Age ,Gender, Education, marital status ,SBP, DBP , Glucose, BMI, eating fruit, eating meat, soft drinks ,physical activity, Number of pregnancy, Family history	RF+PCA	78.3%	65.8%
		Age, Area, Education , Glucose, BMI, Eating meat, eating fruit, eating fish, physical activity, Family history	RF+IG	79.9%	68.5%
		Glucose, Eating meat, eating fruit, Age, Area	RF+IG	80.2%	68.7 %
Our study	Syrian dataset	waist, RHR ,Fasting ,Age Glucose, High BP, High cholesterol, Eye problem, Sleep Hours, weight	RF+RFE	97%	97%
Our study	Syrian dataset	Age, waist, Fasting Glucose, High cholesterol, Eye problem, weight	RF+RFE	97%	97%

4.6 خاتمة:

تعددت الدراسات التي تعمل على تسخير تقنيات تعلم الآلة للوصول إلى التشخيص المبكر للسكري من النوع الثاني، لما في ذلك خدمة لهؤلاء المرضى الذين تظهر الأعراض السريرية عليهم بعد فترة طويلة من الإصابة، ومعظم الدراسات تؤكد على أن التشخيص المبكر قد يكون مفتاح العلاج الذي يقي من مضاعفات السكري.



الفصل السابع

الخاتمة:



1.7 مقدمة:

يقدم هذا الفصل ختام هذه الدراسة ويستعرض الاستنتاجات كملخص للنتائج والمناقشة وكذلك مميزات هذا البحث، وي طرح بعض التوصيات والمقترحات للعمل بها كخطوات مستقبلية.

2.7 الاستنتاجات:

من الاستنتاجات التي تم الوصول إليها:

1. خوارزمية الغابة العشوائية هي الأفضل مقارنة مع خوارزمية الانحدار اللوجستي في حالة عدم توازن البيانات عند استخدام قاعدة البيانات PID.

2. أفضل تقنية توزيع هي ROS.

3. أسوأ أداء للخوارزميتين:

- عند استخدام تقنية التوزيع (RUS) مع الانحدار اللوجستي.

- عند استخدام تقنية (SMOTE) مع خوارزمية الغابة العشوائية

4. أفضل مجموعة سمات للتنبؤ بالسكري والتي تتسجم مع عوامل الاختطار المؤثرة على استقلاب

العضوية للغلوكوز (العمر، ضغط الدم الانبساطي، مستوى الغلوكوز عند الصيام، محيط الخصر،

الوزن، عدم وضوح الرؤية، ارتفاع الكوليسترول، ارتفاع الضغط، عدد ساعات النوم، معدل نبضات

القلب).

5. خوارزمية اختيار السمات العودية RFE مع نموذج الغابة العشوائية تعد هي الأفضل مقارنة مع

خوارزميات اختيار السمات المستخدمة مع خوارزمية الغابة العشوائية في الدراسات المرجعية.

3.7 مميزات البحث:

1. المنهجية المقترحة للبحث.

نظام تشخيص آلي لمرض السكري باستخدام خوارزميات التصنيف، صالح.

2. ضبط المعاملات الفائقة للنماذج باستخدام طريقة البحث الشبكي والتحقق المتقاطع.
3. طريقة معالجة القيم المتطرفة (استبدال القيم المتطرفة)، طريقة تقليل عدد سمات التنبؤ بالسكري (خوارزمية السمات العودية RFE).
4. تفوق نموذج خوارزمية الغابة العشوائية والانحدار اللوجستي في أدائهما عند التنبؤ بالسكري باستخدام قاعدة البيانات PID مقارنة مع الدراسات المرجعية في حالة عدم موازنة البيانات، حيث كانت الدقة (86-87%) لكل منهما.
5. تفوق نموذج خوارزمية الغابة العشوائية والانحدار اللوجستي في أدائهما عند التنبؤ بالسكري باستخدام قاعدة البيانات PID مقارنة مع الدراسات المرجعية في حالة موازنة البيانات باستخدام تقنية ROS، حيث كانت الدقة (84-85%) لكل منهما.
6. تصميم قاعدة بيانات سورية خاصة بمرض السكري من النوع الثاني.
7. تفوق نموذج خوارزمية الغابة العشوائية في أدائه عند التنبؤ بمرض السكري من النوع الثاني باستخدام قاعدة البيانات المحلية مقارنة مع الدراسات المرجعية، حيث كانت الدقة (97%) مع أفضل ست وعشر سمات.

4.7 التوصيات والمقترحات:

من التوصيات والمقترحات لهذا البحث:

1. نشر التوعية بشكل أكبر بالسكري من أجل التقليل من مخاطر وتأخير الإصابة به.
2. تجربة تقنيات تعلم آلة أخرى (كتقنيات التعلم العميق)، استخدام أساليب التجميع للحصول على نتائج أكثر دقة وموثوقية.
3. إضافة المزيد من السمات السريرية التي تزيد من دقة خوارزميات التصنيف كسماكة طيات الجلد

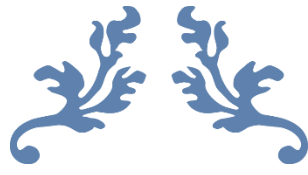
نظام تشخيص آلي لمرض السكري باستخدام خوارزميات التصنيف، صالح.

4. التوجه إلى عدد أكبر من المراكز الصحية الخاصة بمرض السكري لزيادة عدد العينات وتصحيح التجربة أكثر عمومية وعشوائية.

5. دعم مادي مقابل الاختبارات السريرية المكلفة والتي تعد أحد المشاكل التي جرت مواجهتها خلال سير عملية البحث.

5.7 خاتمة:

عمل هذا البحث على بناء نموذج للتنبؤ والكشف المبكر لمرض السكري من النوع الثاني اعتماداً على قاعدتي البيانات PID، وقاعدة البيانات المحلية، حيث أن استخدام تقنيات تعلم الآلة والتحليل الإحصائي لتحديد عوامل التنبؤ بمرض السكري، والتنبؤ والكشف المبكر عن هذا المرض حققت نتائج واعدة لذلك يجب العمل على تطويره في المستقبل.



المراجع:



المراجع العربية:

- 1- عادل، بورغدة ورضا، شريدي.(2021). "أثر النشاط البدني الرياضي المكيف في استقرار نسبة السكر في الدم ووظائف الجسم لدى مرضي السكري النوع الثاني". رسالة ماجستير، جامعة محمد بوضياف، معهد علوم وتقنيات النشاط البدني والرياضي، الجزائر
- 2- حداد، فاطمة الزهرة.(2023). "دراسة إحصائية لعوامل خطر الإصابة بداء السكري من النوع الثاني" مجلة روافد، المجلد(7)، العدد(2)
- 3- ميمش، زياد أحمد وآخرون.(2011). "المرجع الوطني السعودي لتتقيف مرضى داء السكري"، الإصدار(1) ، وزارة الصحة، السعودية
- 4- لطفي، منير.(2018). "رحلتي مع مرض السكري"، (الإصدار الطبعة 1)، المنصورة، مصر: دار اليقين
- 5- بيلوس، رودي. (2013). مرض السكري (الإصدار الطبعة 1). (هنادي مزبودي، المترجمون) الرياض، السعودية: دار المؤلف.
- 6- الزيايدي، هناء كريم بردان، وآخرون.(2018). "العوامل المؤثرة في الإصابة بداء السكري وتحليلها احصائياً في محافظة النجف". مجلة البحوث الجغرافية، العدد(30)
- 7- سعد الحميد، محمد. (2008). "داء السكري أسبابه ومضاعفاته وعلاجه". الرياض، السعودية.

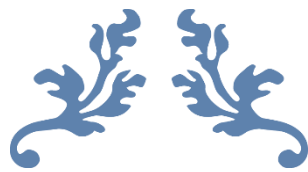
المراجع الأجنبية:

- 8- Evan, M., Morgan, A.R., et al,(2021). Risk Prediction of the Diabetes Missing Million Identifying Individuals at High Risk of Diabetes and Related Complications. *Diabetes Ther*
- 9- Ley, S.H. , Meigs, J. B., et al, (2015). RISK FACTORS FOR TYPE 2 DIABETES. *DIABETES IN AMERICA 3rd Edition*
- 10- Dua, D., Graff, C., (2017). UCI machine learning repository. <http://archive.ics.uci.edu/ml>
- 11- Willi, C., Bodenmann, P., et al, (2007). Active Smoking and the Risk of Type 2 Diabetes. *The Journal of the American Medical Association* .298(22).2654-2664

- 12- Maddatu, J., Molina, C. E., et al, (2017). Smoking and the Risk of Type 2 Diabetes. *Translational Research*.
- 13- Kim, H.J., Kim, K.I., (2022). Blood Pressure Target in Type 2 Diabetes Mellitus. *Diabetes Metab* . 46:667-674
- 14- Kia, NS, Gharib, E., et al, (2023). Factors affecting sleep quality in patients with type 2 diabetes: a cross-sectional study in Iran. *Middle East Current Psychiatry* . 30:40
- 15- Mezil, S.A., Abed, B.A., (2021). Complication of Diabetes Mellitus. *Annals of R.S.C.B. ISSN*. Vol. 25, Issue 3, Pages. 1546 – 1556
- 16- Murphy, K. P., et al., (2012). Machine learning: a probabilistic perspective. *MIT press*.
- 17- Ganie, J.M., Mallik, S., et al, (2023). An ensemble learning approach for diabetes
- 18-** prediction using boosting techniques. *Frontiers in Genetics*.
- 19-** Chako, A., et al., (2021). diabetes prediction using machine learning classification techniques. *Computer Science and engineering karunya institute of technology and science*.
- 20- Parmar, A., Katariya, R., & Patel, V., (2019). A review on random forest: An ensemble classifier. *In International conference on intelligent data communication technologies and internet of things (ICICI)* .
- 21- Sahour, H., Gholami, V., Torkaman, J., Vazifedan, M., & Saeedi, S., (2021). Random forest and extreme gradient boosting algorithms for streamflow modeling using vessel features and tree-rings. *Environmental Earth Sciences*. 80. 1-14.
- 22- Ogunsanya, M., Isichei, J., & Desai, S., (2023). Grid search hyperparameter tuning in additive manufacturing processes. *Manufacturing Letters*. 35. 1031-1042.
- 23- Akinola, O., Ezugwu, A., Agushaka, J., Zitar, R., & Abualigah, L., (2022). Multiclass feature selection with metaheuristic optimization algorithms: a review. *Neural Computing and Applications*. 34(22). 19751-19790.
- 24- Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv*.
- 25- Hossain, R., & Timmer, D., (2021). Machine learning model optimization with hyper parameter tuning approach. *Global Journal of Computer Science and Technology: D Neural & Artificial Intelligence*. 21(2). 31-39.
- 26- Kumar, V., & Minz, S., (2014). Feature selection: A literature review. *Smart Computing Review*. 4(3). 211-229.

- 27- Moore, D. S., Notz, W. I, & Flinger, M. A., (2013). *The basic practice of statistics (6th ed.)*. New York, NY: W. H. Freeman and Company.
- 28- Susdarwono, E., (2021). Testing through the Mann-Whitney Method and the Wilcoxon Signed Level on Learning Cube and Perfect Cube Roots. *Journal of Education and Learning Mathematics Research (JELMaR)*, 2(1), 26-36.
- 29- Aggarwal, K., (2021). Comparison of Feature Selection Techniques for Improved Diabetes Prediction using Random Forest. *International Journal of Mechanical Engineering*. Vol. 6 No. 3.
- 30- Al-Ahmed, I.I., Al-Haj, Y.A., et al., (2023). Optimal Using of Machine Learning Algorithms Hyper parameters for Diabetes Prediction. *Al-Andalus university for science and technology*
- 31- Padmavathi, M.S., Sumathi, C.S., (2019). New Method of Data Preparation for Classifying Diabetes Dataset. *Indian Journal of Science and Technology*. Vol 12
- 32- Noori, N.A., Yassin, A.A., (2021). A Comparative Analysis for Diabetic Prediction Based on Machine Learning Techniques, *Journal of Basrah Researches ((Sciences))*. Vol. 1. (47).
- 33- Ruslan, R.A., Arbaiy, N., (2024). Handling Imbalanced Datasets in Machine Learning: Challenges, Approaches, and Best Practices. *JASTEC*, Vol. 2, p. 20-27
- 34- Gupta, S., Verma, H.K., & Bhardwaj, D., (2021). Classification of diabetes using naive bayes and support vector machine as technique, *Lecture Notes on Multidisciplinary Industrial Engineering*. Springer. Singapore. pp. 365–376
- 35- Gupta, S.C., et al, (2023). Predictive modeling and analytics for diabetes using Hyperparameter tuned Machine Learning Techniques. *Procedia Computer Science* 218 1257–1269
- 36- Alzain, M.A., Alzain, H.M., (2024). Using Machine Learning Techniques to Predict Type 2 Diabetes. *Scientific Journal of Faculty of Education*. Misurata University-Libya.
- 37- Mushtaq, Z., Ali, S., et al, (2022). Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques". *Hindawi Mobile Information Systems*. Article ID6521532, <https://doi.org/10.1155/2022/6521532>
- 38- Talukder, Md, Al., Islam, Md. M., et al, (2024). Toward reliable diabetes prediction: Innovations in data engineering and machine learning applications. *DIGITAL HEALTH*. Volume 10.

- 39- Viswanatha, V., Ramachandra, A., et al. (2023). Diabetes Prediction Using Machine Learning Approach. Vol. 10 .IS 8. <http://stradresearch.org>
- 40- Kakoly, I.j., Hoque, Md.,Hasan, Naj.,(2023). Data-Driven Diabetes Risk Factor Prediction Using Machine Learning Algorithms with Feature Selection Technique. Sustainability15, 4930. <https://doi.org/10.3390/su15064930>
- 41- Maniruzzaman, Md., Ahammed, B., et al., (2020).classification and prediction of diabetes desease using machine learning paradigm. *Health Information and System*. <https://doi.org/10.1007/s13755-019-0095-z>



الملحق
العلمي:



الملحق (1) جدول متغيرات الدخل الخاصة بقاعدة البيانات المحلية وترميزها.

اسم المتغير:	ترميز المتغير:	نوع المتغير:
1-العمر	رقمي	رقمي
2- الجنس: اسمي (ذكر: 0, أنثى: 1)	[0] ذكر [1] أنثى	اسمي غير ترتيبى
3- مستوى التعليم:	[1] ابتدائي [2] إعدادي [3] ثانوي [4] جامعي	اسمي غير ترتيبى
4- الطول	رقمي	رقمي
5- الوزن	رقمي	رقمي
6- محيط الخصر	رقمي	رقمي
7- ضغط الدم الانقباضي	رقمي	رقمي
8- ضغط الدم الانبساطي	رقمي	رقمي
9- تركيز سكر الدم في حالة الصيام	رقمي	رقمي
10- معدل ضربات القلب عند الراحة	رقمي	رقمي
11- حالة التدخين:	[0] غير مدخن [1] مدخن [2] مدخن سابق	اسمي غير ترتيبى
12- عدد السجائر في اليوم :	[0] ولا واحدة [1] من 1 إلى ثلاثة [2] من ثلاثة إلى عشرة [3] حتى عشرين سيجارة [4] بين عشرين حتى الأربعين [5] أكثر من أربعين سيجارة	اسمي غير ترتيبى
13- عدد ساعات النوم في الليل:	[0] بين 4 إلى ست ساعات [1] بين 8 إلى تسع ساعات [2] أكثر من تسع ساعات	اسمي غير ترتيبى
14- عدد مرات ممارسة الرياضة نصف ساعة أو أكثر في الأسبوع:	[0] ولا مرة [1] مرة واحدة [2] من مرتين إلى خمس مرات	اسمي غير ترتيبى
15- تناول الخضار أكثر من ثلاث حصص في اليوم	نعم [1] لا [0]	اسمي غير ترتيبى
16- تناول الفواكه أكثر من ثلاث حصص في اليوم	نعم [1] لا [0]	اسمي غير ترتيبى
17- تناول اللحوم الحمراء الدسمة	نعم [1] لا [0]	اسمي غير ترتيبى
18- تناول اللحوم الحمراء قليلة الدسم	نعم [1] لا [0]	اسمي غير ترتيبى
19- تناول الدجاج	نعم [1] لا [0]	اسمي غير ترتيبى
20- تناول السمك	نعم [1] لا [0]	اسمي غير ترتيبى
21- تناول المشروبات الغازية	نعم [1] لا [0]	اسمي غير ترتيبى
22- تناول المشروبات المحلاة بالسكر	نعم [1] لا [0]	اسمي غير ترتيبى
23- تناول الحليب كامل الدسم	نعم [1] لا [0]	اسمي غير ترتيبى

اسمي غير ترتيبى	نعم [1] لا [0]	24- تناول الحليب خالي الدسم
اسمي غير ترتيبى	نعم [1] لا [0]	25- تناول المشروبات الغازية
اسمي غير ترتيبى	نعم [1] لا [0]	26- تناول البطاطا
اسمي غير ترتيبى	نعم [1] لا [0]	27- تناول الأرز
اسمي غير ترتيبى	نعم [1] لا [0]	28- تناول المعكرونة
اسمي غير ترتيبى	نعم [1] لا [0]	29- تناول المعجنات
اسمي غير ترتيبى	نعم [1] لا [0]	30- تناول الخبز الأبيض
اسمي غير ترتيبى	نعم [1] لا [0]	31- استخدام زيت الزيتون
اسمي غير ترتيبى	نعم [1] لا [0]	32- استخدام زيت الذرة
اسمي غير ترتيبى	نعم [1] لا [0]	33- استخدام زيت دوار الشمس
اسمي غير ترتيبى	نعم [1] لا [0]	34- استخدام السمن النباتي
اسمي غير ترتيبى	نعم [1] لا [0]	35- استخدام السمن الحيواني
اسمي غير ترتيبى	نعم [1] لا [0]	36- طهي الطعام بشكل مقلي
اسمي غير ترتيبى	نعم [1] لا [0]	37- طهي الطعام بشكل مطبوخ
اسمي غير ترتيبى	نعم [1] لا [0]	38- طهي الطعام بشكل مشوي
اسمي غير ترتيبى	نعم [1] لا [0]	39- طهي الطعام بشكل مسلوق
اسمي غير ترتيبى	نعم [1] لا [0]	40- تناول الطعام بشكل نئى
اسمي غير ترتيبى	نعم [1] لا [0]	41- اتباع نظام غذائي صحي
اسمي غير ترتيبى	نعم [1] لا [0]	42- أمراض القلب والأوعية الدموية
اسمي غير ترتيبى	نعم [1] لا [0]	43- ارتفاع ضغط الدم
اسمي غير ترتيبى	نعم [1] لا [0]	44- أمراض الكلى
اسمي غير ترتيبى	نعم [1] لا [0]	45- أمراض الغدة الدرقية
اسمي غير ترتيبى	نعم [1] لا [0]	46- ارتفاع نسبة الكوليسترول
اسمي غير ترتيبى	نعم [1] لا [0]	47- ضيق التنفس
اسمي غير ترتيبى	نعم [1] لا [0]	48- تدمير القدمين أو اليدين
اسمي غير ترتيبى	نعم [1] لا [0]	49- الشعور بالغثيان أو القيء أو الإمساك
اسمي غير ترتيبى	نعم [1] لا [0]	50- تكرار التبول
اسمي غير ترتيبى	نعم [1] لا [0]	51- هبوط مفاجئ بالوزن
اسمي غير ترتيبى	نعم [1] لا [0]	52- التعب الدائم
اسمي غير ترتيبى	نعم [1] لا [0]	53- عدم وضوح الرؤية
اسمي غير ترتيبى	نعم [1] لا [0]	54- الشعور بالعطش المستمر
اسمي غير ترتيبى	نعم [1] لا [0]	55- الشعور بالجوع المستمر
اسمي ترتيبى	[0] لا يوجد [1] قرابة من الدرجة الثالثة [2] قرابة من الدرجة الثانية [3] قرابة من الدرجة الأولى	56- العامل الوراثي
اسمي غير ترتيبى	نعم [1] لا [0]	57- الشعور بالغضب أو الخوف أو التوتر
اسمي ترتيبى	[0] ليس كثيرا [1] أحيانا [2] غالبا [3] دائما	58- مقدار التوتر
اسمي غير ترتيبى	نعم [1] لا [0]	59- ولادة طفل بوزن أعلى من 4 كيلو غرام
اسمي غير ترتيبى	[0] ولا مرة [1] مرة واحدة [2] اثنان أو أكثر	60- عدد مرات الحمل
اسمي غير ترتيبى	نعم [1] لا [0]	61- الإصابة بسكري الحمل

اسمي غير ترتيبى	نعم [1] لا [0]	62- الحمل حاليا
اسمي غير ترتيبى	نعم [1] لا [0]	63- تناول الأدوية بانتظام

الملحق (2) نتائج التحليل الاحصائي باستخدام Chi-Squared test

Features	Outcome							Chi square	p_values
		0		1		Total			
	coding	Count	Column N %	Count	Column N %	Count	Column N %		
Gender	0	12	19.4%	42	43.3%	54	34.0%	8.630734	0.0033
	1	50	80.6%	55	56.7%	105	66.0%		
	Total	62	100.0%	97	100.0%	159	100.0%		
Education Level	1	19	31.1%	47	51.6%	66	43.4%	14.136	0.0027
	2	11	18.0%	26	28.6%	37	24.3%		
	3	13	21.3%	6	6.6%	19	12.5%		
	4	18	29.5%	12	13.2%	30	19.7%		
	Total	61	100.0%	91	100.0%	152	100.0%		
Smoking	0	55	88.7%	70	73.7%	125	79.6%	8.7974	0.0122
	1	7	11.3%	13	13.7%	20	12.7%		
	2	0	0.0%	12	12.6%	12	7.6%		
	Total	62	100.0%	95	100.0%	157	100.0%		
cigarettes	0	55	88.7%	71	82.6%	126	85.1%		0.14
	1	0	0.0%	4	4.7%	4	2.7%		
	2	0	0.0%	2	2.3%	2	1.4%		
	3	5	8.1%	7	8.1%	12	8.1%		
	4	2	3.2%	2	2.3%	4	2.7%		
	Total	62	100.0%	86	100.0%	148	100.0%		
Sleep Hours	0	30	48.4%	77	79.4%	107	67.3%	26.0671	0.00
	1	31	50.0%	13	13.4%	44	27.7%		
	2	1	1.6%	7	7.2%	8	5.0%		
	Total	62	100.0%	97	100.0%	159	100.0%		
physical activity	0	7	11.3%	19	19.6%	26	16.4%	2.0977	0.35
	1	15	24.2%	24	24.7%	39	24.5%		
	2	40	64.5%	54	55.7%	94	59.1%		
	Total	62	100.0%	97	100.0%	159	100.0%		
number of times	0	22	35.5%	43	44.3%	65	40.9%	0.8859	0.35
	2	40	64.5%	54	55.7%	94	59.1%		
	Total	62	100.0%	97	100.0%	159	100.0%		

Eating Vegetables	0	39	62.9%	22	22.7%	61	38.4%	24.206	0.00
	1	23	37.1%	75	77.3%	98	61.6%		
	Total	62	100.0%	97	100.0%	159	100.0%		
Eating Fruit	0	50	80.6%	62	63.9%	112	70.4%	4.3113	0.037
	1	12	19.4%	35	36.1%	47	29.6%		
	Total	62	100.0%	97	100.0%	159	100.0%		
fish	0	62	100.0%	96	99.0%	158	99.4%	0	1
	1	0	0.0%	1	1.0%	1	0.6%		
	Total	62	100.0%	97	100.0%	159	100.0%		
chicken	0	10	16.1%	25	25.8%	35	22.0%	1.5259	0.216
	1	52	83.9%	72	74.2%	124	78.0%		
	Total	62	100.0%	97	100.0%	159	100.0%		
little fat meat	0	55	88.7%	83	85.6%	138	86.8%	0.1093	0.74
	1	7	11.3%	14	14.4%	21	13.2%		
	Total	62	100.0%	97	100.0%	159	100.0%		
fat meat	0	54	87.1%	73	75.3%	127	79.9%	2.602	0.106
	1	8	12.9%	24	24.7%	32	20.1%		
	Total	62	100.0%	97	100.0%	159	100.0%		
soft drinks	0	55	88.7%	91	93.8%	146	91.8%	0.72	0.395
	1	7	11.3%	6	6.2%	13	8.2%		
	Total	62	100.0%	97	100.0%	159	100.0%		
sugar drinks	0	47	75.8%	88	90.7%	135	84.9%	5.45336	0.019
	1	15	24.2%	9	9.3%	24	15.1%		
	Total	62	100.0%	97	100.0%	159	100.0%		
whole milk	0	44	71.0%	76	78.4%	120	75.5%	0.75	0.386
	1	18	29.0%	21	21.6%	39	24.5%		
	Total	62	100.0%	97	100.0%	159	100.0%		
chicken	0	62	100.0%	95	97.9%	157	98.7%	0.1667	0.683
	1	0	0.0%	2	2.1%	2	1.3%		
	Total	62	100.0%	97	100.0%	159	100.0%		
energy drink	0	61	98.4%	95	97.9%	156	98.1%	0.00	1
	1	1	1.6%	2	2.1%	3	1.9%		
	Total	62	100.0%	97	100.0%	159	100.0%		
potato	0	20	32.3%	46	47.4%	66	41.5%	2.98	0.084
	1	42	67.7%	51	52.6%	93	58.5%		
	Total	62	100.0%	97	100.0%	159	100.0%		
rice	0	14	22.6%	47	48.5%	61	38.4%	9.6415	0.0019
	1	48	77.4%	50	51.5%	98	61.6%		
	Total	62	100.0%	97	100.0%	159	100.0%		
noodles	0	34	54.8%	63	64.9%	97	61.0%	1.227	0.267
	1	28	45.2%	34	35.1%	62	39.0%		
	Total	62	100.0%	97	100.0%	159	100.0%		
	0	50	80.6%	83	85.6%	133	83.6%	0.358	0.549

pasta	1	12	19.4%	14	14.4%	26	16.4%		
	Total	62	100.0%	97	100.0%	159	100.0%		
white bread	0	4	6.5%	9	9.3%	13	8.2%	0.114	0.735
	1	58	93.5%	88	90.7%	146	91.8%		
	Total	62	100.0%	97	100.0%	159	100.0%		
olive oil	0	46	74.2%	44	45.4%	90	56.6%	11.654	0.00006
	1	16	25.8%	53	54.6%	69	43.4%		
	Total	62	100.0%	97	100.0%	159	100.0%		
corn oil	0	59	95.2%	95	97.9%	154	96.9%	0.262	0.608
	1	3	4.8%	2	2.1%	5	3.1%		
	Total	62	100.0%	97	100.0%	159	100.0%		
sun oil	0	4	6.5%	7	7.2%	11	6.9%	0.00	1
	1	58	93.5%	90	92.8%	148	93.1%		
	Total	62	100.0%	97	100.0%	159	100.0%		
animal fat	0	61	98.4%	89	91.8%	150	94.3%	1.999	0.157
	1	1	1.6%	8	8.2%	9	5.7%		
	Total	62	100.0%	97	100.0%	159	100.0%		
plant fat	0	24	38.7%	40	41.2%	64	40.3%	0.022	0.879
	1	38	0.6129032	57	59%	95	60%		
	Total	62	1	97	100%	159	100%		
Fried	0	49	0.7903226	50	52%	99	62%	11.01997	0.00
	1	13	0.2096774	47	48%	60	38%		
	Total	62	1	97	100%	159	100%		
cooked	1	62	1	97	100%	159	100%	0.00	1
	Total	62	1	97	100%	159	100%		
grilled	0	62	1	83	86%	145	91%	8.0973	0.0044
	1	0	0	14	14%	14	9%		
	Total	62	1	97	100%	159	100%		
boiled	0	55	0.8870968	78	80%	133	84%	1.345	0.246
	1	7	0.1129032	19	20%	26	16%		
	Total	62	1	97	100%	159	100%		
uncooked	0	62	1	97	100%	159	100%	0	1
	Total	62	1	97	100%	159	100%		
healthy diet	0	58	0.9354839	67	70%	125	79%	11.469	0.0007
	1	4	0.0645161	29	30%	33	21%		
	Total	62	1	96	100%	158	100%		
Heart decease	0	58	0.9354839	56	58%	114	72%	22.179	0.00
	1	4	0.0645161	41	42%	45	28%		
	Total	62	1	97	100%	159	100%		
High Blood Pressure	0	50	0.8064516	22	23%	72	45%	48.977	0.00
	1	12	0.1935484	75	77%	87	55%		
	Total	62	1	97	100%	159	100%		

Regularly medications	0	32	0.516129	1	1%	33	21%	55.8	0.00
	1	30	0.483871	96	99%	126	79%		
	Total	62	1	97	100%	159	100%		
Family History	0	19	0.3064516	20	21%	39	25%	30.81	0.00
	1	17	0.2741935	2	2%	19	12%		
	2	4	0.0645161	5	5%	9	6%		
	3	22	0.3548387	68	72%	90	57%		
	Total	62	1	95	100%	157	100%		
angry feelings	0	15	0.2419355	20	21%	35	22%	0.111	0.738
	1	47	0.7580645	77	79%	124	78%		
	Total	62	1	97	100%	159	100%		
Stress	0	13	0.2096774	21	22%	34	21%	1.313	0.725
	1	24	0.3870968	43	44%	67	42%		
	2	10	0.1612903	10	10%	20	13%		
	3	15	0.2419355	23	24%	38	24%		
	Total	62	1	97	100%	159	100%		
Urination frequency	0	52	0.8387097	34	35%	86	54%	34.362	0.00
	1	10	0.1612903	63	65%	73	46%		
	Total	62	1	97	100%	159	100%		
weight loss	0	59	0.9516129	74	76%	133	84%	8.517	0.0035
	1	3	0.0483871	23	24%	26	16%		
	Total	62	1	97	100%	159	100%		
permanent fatigue	0	48	0.7741935	51	53%	99	62%	8.905	0.0025
	1	14	0.2258065	46	47%	60	38%		
	Total	62	1	97	100%	159	100%		
hungry	0	54	0.8709677	49	51%	103	65%	20.61	0.00
	1	8	0.1290323	48	49%	56	35%		
	Total	62	1	97	100%	159	100%		
thirsty	0	53	0.8548387	41	42%	94	59%	27.467	0.00
	1	9	0.1451613	56	58%	65	41%		
	Total	62	1	97	100%	159	100%		
high cholesterol	0	59	0.9516129	31	32%	90	57%	58.962	0.00
	1	3	0.0483871	66	68%	69	43%		
	Total	62	1	97	100%	159	100%		
kidney problem	0	60	0.9677419	86	89%	146	92%	2.324	0.127
	1	2	0.0322581	11	11%	13	8%		
	Total	62	1	97	100%	159	100%		
eye problems	0	46	0.7419355	24	25%	70	44%	35.554	0.00
	1	16	0.2580645	73	75%	89	56%		
	Total	62	1	97	100%	159	100%		

thyroid disease	0	56	0.9032258	93	96%	149	94%	1.149	0.283
	1	6	0.0967742	4	4%	10	6%		
	Total	62	1	97	100%	159	100%		
shortness of breath	0	56	0.9032258	65	67%	121	76%	10.0567	0.0015
	1	6	0.0967742	32	33%	38	24%		
	Total	62	1	97	100%	159	100%		
numbness	0	39	0.6290323	29	30%	68	43%	15.513	0.00
	1	23	0.3709677	68	70%	91	57%		
	Total	62	1	97	100%	159	100%		
nausea and constipation	0	59	0.9516129	88	91%	147	92%	0.526	0.467
	1	3	0.0483871	9	9%	12	8%		
	Total	62	1	97	100%	159	100%		
birth of baby more than 4 kilogram	0	54	0.8709677	75	77%	129	81%	1.766	0.183
	1	8	0.1290323	22	23%	30	19%		
	Total	62	1	97	100%	159	100%		
numb of preg	0	22	0.3548387	48	49%	70	44%	6.9646	0.031
	1	3	0.0483871	0	0%	3	2%		
	2	37	0.5967742	49	51%	86	54%		
	Total	62	1	97	100%	159	100%		
gestational diabetes	0	62	1	93	96%	155	97%	1.21	0.271
	1	0	0	4	4%	4	3%		
	Total	62	1	97	100%	159	100%		

Abstract

Diabetes mellitus is a chronic disease that elevates blood glucose levels, and if uncontrolled, can damage various body systems, particularly the heart, blood vessels, and nerves. Early diagnosis is challenging due to overlapping symptoms with other diseases.

This research focuses on building eight predictive models for early detection of type 2 diabetes, using machine learning algorithms: Random Forest and Logistic Regression.

Pima Indian Diabetes dataset (PID) was utilized and preprocessed to address missing and outlier values, normalize features, and balance class distributions using: random over sampler, random under sampler, and SMOTE techniques. Hyper parameters were tuned using grid search and cross-validation, and models were trained, tested, and evaluated to determine the optimal one. Additionally, a local dataset was built by the result of cross sectional study. Feature selection was conducted using statistical tests (Man-Whitney, T-test, Chi-Square) and Recursive Feature Elimination (RFE).

The model was retrained on this dataset and evaluated, Results showed that the Random Forest model outperformed Logistic Regression on the PID dataset (accuracy: 87%, F1-score: 84%, AUC: 0.81). On the local dataset, Random Forest achieved even better performance (accuracy: 97%, F1-score: 97%, AUC: 0.96).

Keywords: Type 2 Diabetes, Early Detection, Machine Learning, Hyper parameter Tuning, Logistic Regression, Random Forest

Syrian Arab Republic

Damascus University

Faculty of Mechanical & Electrical Engineering

Biomedical Engineering Department



An Automated Diagnosis System for Diabetes using Classification Algorithms

A dissertation submitted in partial fulfillment of the
requirements for the degree of Master in Biomedical Engineering

By

Eng. Mona Saleh

Supervisor

Prof. Eng. Rasha Massoud

2024-2025

