دراسة مقارنة بين النماذج الإحصائية التقليدية ونموذج الشبكات العصبية الاصطناعية في التصنيف والتنبؤ بالإصابة بمرض السكري

السيد محمد السيد محمد

إشراف

أ.د/ ممدوح عبد العليم سعد موافي
الأستاذ المساعد بقسم الإحصاء والرياضة والتأمين
كلية التجارة - جامعة عين شمس

أ.د/ أحمد الألفي سعيد محمد
الأستاذ المتفرغ بقسم الإحصاء التطبيقي والتأمين
كلية التجارة - جامعة قناة السويس

المستخلص:

يُعد مرض السكري أحد أكثر الأمراض المزمنة انتشارًا، ويُشكل تحديًا صحيًا واقتصاديًا كبيرًا نظرًا لما يترتب عليه من مضاعفات خطيرة قد تؤدي إلى الوفاة، وهو ما يُبرز الحاجة إلى أدوات تحليلية دقيقة لفهم العوامل المؤثرة في الإصابة وتعزيز فرص الوقاية والتدخل المبكر. وفي هذا السياق، هدفت الدراسة إلى مقارنة كفاءة ثلاثة نماذج إحصائية في تصنيف والتنبؤ بمرض السكري، وهي: الانحدار اللوجستي، التحليل التمييزي، والشبكات العصبية الاصطناعية، بالتطبيق على عينة مكوّنة من ٣٨٥ حالة. وقد تم تقييم أداء هذه النماذج باستخدام مجموعة متنوعة من مؤشرات الأداء، بالإضافة إلي تطبيق التحقق المتقاطع لضمان ثبات الأداء وموثوقية النتائج. أظهرت النتائج تفوقًا ملحوظًا لنموذج الشبكات العصبية الاصطناعية بدقة كلية بلغت ٩٧,٢%، مقارنة بـ ٩٥,٨% و٩٥,٣% لكل من الانحدار اللوجستي والتحليل التمييزي على التوالي. كما اتفقت النماذج الثلاثة في تحديد أبرز المتغيرات المؤثرة، وفي مقدمتها: HbA1c test، مؤشر كتلة الجسم، العمر، والتاريخ العائلي للمرض. وتوصي الدراسة باعتماد نموذج الشبكات العصبية الاصطناعية كأداة تنبؤية فعالة في البيئات السريرية، إلى جانب استخدام نموذج الانحدار اللوجستي كأداة مساعدة في التشخيص، خاصة في الحالات التي تتطلب تفسيرًا واضحًا للنتائج.

الكلمات المفتاحية: مرض السكري، التحليل التمييزي، الانحدار اللوجستي، الشبكات العصبية الاصطناعية، التصنيف، التنبؤ.

Abstract:

Diabetes mellitus stands among the most prevalent chronic diseases, posing a significant health and economic challenge due to its severe complications, which may lead to death. This highlights the need for precise analytical tools to understand the factors influencing the onset of the disease and to enhance opportunities for prevention and early intervention. In this context, the study aimed to compare the performance of three statistical models in classifying and predicting the incidence of diabetes: logistic regression, discriminant analysis, and artificial neural networks, applied to a sample of 385 cases. The models' performance was evaluated using a variety of performance indicators, in addition to applying cross-validation to ensure performance stability and result reliability. The results revealed a marked superiority of the artificial neural network model, which achieved an overall accuracy of 97.2%, compared to 95.8% and 95.3% for logistic regression and discriminant analysis, respectively. All three models concurred in identifying the most critical predictors, notably the HbA1c test, body mass index, age, and family history of diabetes. The study recommends adopting the artificial neural network model as an effective predictive tool in clinical settings, alongside utilizing the logistic regression

model as a supportive diagnostic tool, particularly in cases requiring clear interpretability of results.

Keywords: Diabetes Mellitus, Discriminant Analysis, Logistic Regression, Artificial Neural Networks, Classification, Prediction.

القسم الأول: الإطار العام البحث

١/١: المقدمة

تُواجه النظم الصحية حول العالم اليوم تحديًا متزايدًا في السيطرة على الأمراض المزمنة التي تتسم بتأثيراتها الممتدة ومضاعفاتها الخطيرة على صحة الإنسان وجودة الحياة. ويُعد مرض السكري في مقدمة هذه التحديات، ليس فقط بسبب انتشاره الواسع، ولكن أيضًا لما يسببه من مضاعفات صحية جسيمة تؤثر على القلب والأوعية الدموية والكلى والأعصاب والعينين، الأمر الذي ينعكس سلبًا على معدلات الوفاة والإعاقة. ووفقًا للاتحاد الدولي للسكري، بلغ عدد المصابين بهذا المرض أكثر من ٥٨٩ مليون شخص حول العالم في عام ٢٠٢٤، ومن المتوقع أن يصل هذا الرقم إلى ٨٥٣ مليون بحلول عام ٢٠٥٠، مما يجعله أحد أكثر الأمراض المزمنة شيوعًا وخطورة على الصعيد العالمي.

تكمن خطورة مرض السكري في طبيعته الصامتة خلال المراحل الأولى من الإصابة، حيث قد يظل المريض لفترة طويلة دون ظهور أعراض واضحة، وهو ما يؤدي إلى التأخر في التشخيص ويقلل من فرص التدخل المبكر للحد من المضاعفات. ومن هنا، أصبح الكشف المبكر والتنبؤ بالإصابة بالسكري أولوية قصوى في منظومة الرعاية الصحية الحديثة، إذ يُسهم في تحسين فرص الوقاية وتقليل الأعباء الصحية والاقتصادية الناتجة عن مضاعفات المرض. في ضوء ذلك، برزت الحاجة إلى الاعتماد على الأساليب الإحصائية والتقنيات التنبؤية القادرة على تحليل المؤشرات الصحية والديموغرافية للأفراد والتنبؤ بإمكانية إصابتهم بمرض السكري بدقة وفعالية.

وقد شهد هذا المجال تنوعًا في النماذج والأدوات المستخدمة، التي تتفاوت في فلسفتها وآليات عملها ومدى قدرتها على تحقيق نتائج دقيقة وموثوقة.

من بين هذه النماذج، يأتي **الانحدار اللوجستي (LR)** كأحد أبرز النماذج الإحصائية المستخدمة في التنبؤ بالمتغيرات التابعة التصنيفية، حيث يعتمد على الدوال الاحتمالية الشرطية ويتميز بمرونته في التعامل مع البيانات دون الحاجة إلى فرضيات صارمة حول توزيع المتغيرات المستقلة. كذلك يُعد **التحليل التمييزي (DA)** من الأساليب التقليدية واسعة الاستخدام في تصنيف المشاهدات إلى مجموعات، ويعتمد على حساب المسافات بين المشاهدات ومراكز الفئات المختلفة، مع ضرورة تحقق مجموعة من الافتراضات الإحصائية مثل التوزيع الطبيعي وتجانس التباين المشترك. ومع التقدم المتسارع في تقنيات الذكاء الاصطناعي وتعلم الآلة، ظهرت **الشبكات العصبية الاصطناعية (ANN)** كأحد النماذج الحديثة الفعّالة في مجالات التصنيف والتنبؤ، لما تتميز به من قدرة عالية على اكتشاف الأنماط غير الخطية والتفاعلات المعقدة بين المتغيرات، دون التقيد بالافتراضات التقليدية المتعلقة بطبيعة البيانات.

وانطلاقًا من ذلك، تهدف هذه الدراسة إلى إجراء مقارنة تحليلية بين هذه النماذج الثلاثة لتحديد النموذج الأكثر كفاءة وموثوقية في تصنيف الأفراد إلى مصابين وغير مصابين بمرض السكري، مع التعرف على العوامل الأكثر تأثيرًا في احتمالية الإصابة. كما تسعى المقارنة إلى تقديم رؤية علمية تدعم جهود الكشف المبكر عن المرض، وتوفر أدوات مساعدة لاتخاذ القرار الطبي تستند إلى أسس علمية دقيقة وموضوعية.

٢/١: مشكلة البحث

يُعد مرض السكري من أكثر الأمراض المزمنة انتشارًا حول العالم، ويُشكل تحديًا صحيًا واقتصاديًا كبيرًا نظرًا لما يترتب عليه من مضاعفات خطيرة قد تصل إلى الوفاة، إضافة إلى العبء المتزايد على الأفراد والمجتمعات وأنظمة الرعاية الصحية. وفي ظل هذه التحديات، تبرز أهمية النماذج الإحصائية كأدوات فعّالة لفهم

العوامل المؤثرة في الإصابة، والمساهمة في تصنيف المرضى والتنبؤ بإصابتهم بدقة، مما يعزز فرص الوقاية والتدخل المبكر. ورغم تنوع النماذج المستخدمة في هذا المجال، مثل الانحدار اللوجستي، والتحليل التمييزي، والشبكات العصبية الاصطناعية، إلا أن هناك تباينًا ملحوظًا في أدائها من حيث الدقة، والكفاءة، ومدى ملاءمتها لأنواع البيانات المختلفة. ويعود ذلك إلى الاختلاف في المنهجيات والافتراضات التي يستند إليها كل نموذج، وتفاوت قدرتها على التعامل مع البيانات المعقدة أو غير الخطية.

وبناءً على ذلك، تتبلور مشكلة الدراسة في الحاجة إلى مقارنة منهجية دقيقة بين هذه النماذج الثلاثة باستخدام مجموعة متنوعة من مقاييس الأداء وطرق التحقق، لتقييم كفاءة كل منها في تصنيف حالات الإصابة والتنبؤ بحدوث المرض، اعتمادًا على قاعدة بيانات تضم مجموعة متنوعة من المتغيرات الصحية والسلوكية والديموغرافية، بهدف تحديد النموذج الأكثر كفاءة وموثوقية في دعم جهود التشخيص المبكر، والمساعدة في الحد من الآثار السلبية للمرض.

وفي ضوء ما سبق عرضه تتبلور مشكلة البحث في الإجابة على التساؤل الرئيسي التالي: **أي من النماذج الثلاثة (التحليل التمييزي، الانحدار اللوجستي، والشبكات العصبية الاصطناعية) يُعد الأكثر فعالية ودقة في تصنيف المرضى والتنبؤ باحتمالية إصابتهم بمرض السكري؟**

وينبثق من التساؤل الرئيسي التساؤلات الفرعية التالية:

١. ما مدى دقة كل من نموذجي التحليل التمييزي والانحدار اللوجستي في تصنيف حالات الإصابة بمرض السكري؟

٢. ما مدى كفاءة نموذج الشبكات العصبية الاصطناعية في التنبؤ بالإصابة مقارنةً بالنماذج الإحصائية التقليدية؟

٣. ما هي أهم العوامل الصحية والديموغرافية والسلوكية المؤثرة في احتمالية الإصابة بمرض السكري وفقًا للنماذج الثلاثة؟

٤. كيف يمكن مقارنة أداء النماذج الثلاثة بهدف بناء نموذج شامل يدعم اتخاذ القرار الطبي ويُحسن من جهود التشخيص المبكر؟

١/٣: أهداف البحث

١- تقييم مدي فعالية ودقة النماذج الإحصائية المختلفة (التحليل التمييزي، الانحدار اللوجستي، والشبكات العصبية) في تصنيف حالات الإصابة بمرض السكري بناءً على مجموعة من العوامل الصحية، الديموغرافية، والسلوكية.
٢- تحديد العوامل الأكثر تأثيراً على احتمالية الإصابة بمرض السكري والتأكد من معنويتها باستخدام النماذج الإحصائية المختلفة.
٣- مقارنة أداء النماذج الإحصائية المستخدمة في الدراسة لتحديد النموذج الأكثر كفاءة ودقة في التنبؤ بالإصابة بمرض السكري، وتوضيح الفروقات في الأداء بين النماذج التقليدية (التحليل التمييزي، الانحدار اللوجستي) ونموذج الشبكات العصبية الاصطناعية باستخدام مجموعة متنوعة من طرق التقييم.
٤- تقديم توصيات لتحسين استراتيجيات التشخيص المبكر لمرض السكري من خلال توظيف النموذج الأمثل الذي يتم تحديده من الدراسة، مما يساهم في تعزيز الجهود الوقائية والعلاجية.

١/٤: أهمية البحث

- **الأهمية العلمية للبحث:**

تسعى هذه الدراسة إلى إثراء المعرفة العلمية في مجال الإحصاء التطبيقي وعلوم البيانات الطبية من خلال إجراء مقارنة منهجية بين ثلاثة من أهم النماذج الإحصائية وأساليب الذكاء الاصطناعي المستخدمة في التصنيف والتنبؤ، وهي: الانحدار اللوجستي، والتحليل التمييزي، والشبكات العصبية الاصطناعية. تساهم الدراسة في توضيح الفروق الجوهرية بين هذه النماذج من حيث الدقة والكفاءة والافتراضات الأساسية، مما يوفر مرجعًا علميًا يمكن أن يستند إليه الباحثون في اختيار النموذج الأنسب لمهام التصنيف الطبية.

- **الأهمية العملية للبحث:**

تنبع أهمية البحث من خطورة مرض السكري وما يترتب عليه من مضاعفات بالغة، مثل النوبات القلبية، والسكتات الدماغية، بالإضافة إلى اعتلال شبكية العين، وتلف العديد من الأعضاء الحيوية. وتسهم هذه المضاعفات في ارتفاع معدلات الوفيات وزيادة الأعباء على القطاع الصحي. ومن هنا تبرز أهمية التنبؤ بمرض السكري وتشخيصه مبكرًا باستخدام النماذج الإحصائية لتحديد العوامل الرئيسية المسببة له، مما يتيح سرعة التدخل للحد من تطور المرض أو الوقاية منه.

١/٥: فروض البحث

١- النماذج المستخدمة في الدراسة (الانحدار اللوجستي الثنائي، التحليل التمييزي، والشبكات العصبية الاصطناعية) ملائمة لتوفيق بيانات مرضي السكري.
٢- هناك علاقة ذات دلالة إحصائية بين المتغيرات الصحية، الديموغرافية، والسلوكية التي تم الاعتماد عليها في الدراسة، وبين احتمالية الإصابة بمرض السكري، ويتم اختبار هذه العلاقة باستخدام النماذج.
٣- توجد اختلافات معنوية في دقة وكفاءة النماذج الإحصائية المختلفة (التحليل التمييزي، الانحدار اللوجستي، والشبكات العصبية) في تصنيف الأفراد المصابين وغير المصابين بمرض السكري.
٤- يُظهر نموذج الشبكات العصبية الاصطناعية دقة أعلى في تصنيف حالات الإصابة بمرض السكري مقارنةً بالنماذج الإحصائية التقليدية (التحليل التمييزي والانحدار اللوجستي).

٦/١: مجتمع وعينة الدراسة

يتكون مجتمع الدراسة من جميع المرضى المترددين على المستشفى الجامعي بجامعة قناة السويس خلال الفترة من عام 2023 إلى عام 2025. ونظرًا لعدم توفر بيانات دقيقة حول الحجم الكلي للمجتمع، تم استخدام (Cochran's Formula) لتحديد الحد الأدنى لحجم العينة العشوائية البسيطة المطلوبة في حالة

المجتمعات غير المحدودة، وذلك لضمان التمثيل الإحصائي المناسب عند مستوى ثقة 95% وهامش خطأ 5%. وبناءً على ذلك، تم تحديد حجم العينة بمقدار385 حالة. وتضمنت العينة أفرادًا مصابين بمرض السكري، بالإضافة إلى آخرين غير مصابين، ولكن تظهر لديهم أعراض بدرجات متفاوتة.

٧/١: تنظيم البحث

القسم الأول: الإطار العام للبحث.

القسم الثاني: الإطار النظري للبحث.

القسم الثالث: الإطار التطبيقي للبحث .

القسم الرابع: نتائج وتوصيات البحث.

القسم الثاني: الإطار النظري للبحث

١/٢ نموذج التحليل التمييزي:

التحليل التمييزي هو أحد أساليب التحليل الإحصائي متعدد المتغيرات، حيث يُستخدم لتصنيف المفردات ضمن مجموعتين أو أكثر تم تحديدها مسبقًا، اعتمادًا على عدة متغيرات تنبؤية ملحوظة (Liberda et al., 2020)، يقوم هذا الاسلوب بتحليل المتغيرات الداخلة في النموذج بطريقة مترابطة، مع الأخذ في الاعتبار العلاقات المتداخلة بين هذه المتغيرات، كما يسعى إلى بناء نموذج إحصائي يوضح العلاقة المتبادلة بين المتغيرات المختلفة (Ahmed & Abd Elrazek, 2023).

وتكمن أهمية التحليل التمييزي بصفة أساسية في فاعليته في التمييز بين مفردات العينة، وذلك من خلال الاعتماد على تركيبة خطية للمتغيرات المستقلة تعرف بالدالة التمييزية، والتي تعمل على تعظيم الفروق والاختلافات بين متوسطات المجموعات التي نتجت عن التمييز، بحيث تزداد درجة التجانس بين المفردات داخل كل مجموعة وتقل درجة التجانس بين المجموعات وبعضها.(Johnson & Wichern, 2007; Tekić et al., 2021).

١/١/٢ أهداف نموذج التحليل التمييزي:

١- تصميم تركيبة خطية من المتغيرات المستقلة تتيح التمييز بشكل أفضل بين فئات المتغير التابع.

٢- التحقق فيما إذا كان هناك فروق ذات دلالة احصائية بين المجموعات فيما يتعلق بالمتغيرات المستقلة التوضيحية.

٣- تحديد المتغيرات التي تساهم بأكبر قدر من التمييز بين مجموعات المتغير التابع.

٤- تصنيف المشاهدات ضمن مجموعات المتغير التابع استنادًا إلى قيم المتغيرات المستقلة.

٥- تقييم دقة التصنيف كنسبة مئوية.

٦- يتمثل الهدف النهائي للتحليل التمييزي في استخدام قاعدة التصنيف لتحديد انتماء مشاهدة جديدة مجهولة إلى إحدى المجموعات المحددة مسبقًا، وذلك استنادًا إلى قيم خصائصها، و بأقل خطأ تصنيف ممكن (George et al., 2020; Kamel et al., 2022).

٢/١/٢ افتراضات نموذج التحليل التمييزي:

١- أن تكون المجتمعات محل الدراسة منفصلة وقابلة للتحديد.

٢- أن تحتوي البيانات المستخدمة في التحليل على عينة عشوائية من مفردات كل مجتمع من مجتمعات الدراسة، بحيث تكون هذه العينات ممثلة للمجتمعات محل التحليل.

٣- أن تكون المجتمعات الإحصائية محل الدراسة ذات توزيع طبيعي.

٤- عدم تساوي متوسطات المجموعات (فئات المتغير التابع).

٥- تساوي مصفوفة التباين والتغاير للمجموعتين.

٦- استقلالية المشاهدات، أي عدم وجود مشكلة الارتباط الخطي المتعدد (Multicollinearity).

٧- عدم وجود قيم متطرفة، حيث إن التحليل التمييزي حساس للغاية تجاه القيم الشاذة، والتي قد تؤثر على النتائج وتؤدي إلى انحراف توزيع البيانات عن التوزيع الطبيعي (بسيوني، ٢٠٢١).

٣/١/٢ اختبارات معنوية الدالة التمييزية الخطية:

بعد تقدير النموذج التمييزي، لا بد من اختبار معنويته لتحديد مدى قدرة الدالة التمييزية على التمييز والفصل بين المجموعات. بالإضافة إلى تحديد المتغيرات المعنوية التي ينبغي الإبقاء عليها في النموذج، مع استبعاد المتغيرات غير المعنوية، وذلك لضمان دقة النموذج عند استخدامه في تصنيف المفردات الجديدة والتنبؤ بانتمائها إلى إحدى المجموعات. ومن بين هذه الاختبارات:

١/٣/١/٢ اختبار F (F-test).

يُستخدم هذا الاختبار لتقييم قدرة الدالة التمييزية على التمييز، وذلك من خلال قياس الاختلافات بين المجموعات وداخلها باستخدام جدول تحليل التباين.

جدول (1) جدول تحليل التباين (ANOVA Table)

Source	SS	DF	MS	F
Between X's	SSB	k	MSB	$\frac{MSB}{MSE}$
Within X's	SSE	$n_1 + n_2 - k - 1$	MSE	
Total	SST	$n_1 + n_2 - 1$		

وتعتمد هذه العملية على اختبار الفرضية التالية:

H_0 : الدالة ليس لها القدرة على التمييز

H_1 : الدالة لها القدرة على التمييز

أولاً: إيجاد قيمة F المحسوبة:

يتم الحصول على قيمة F المحسوبة عن طريق قسمة متوسط مجموع المربعات بين المتغيرات على متوسط مجموع المربعات داخل المتغيرات، وذلك وفقًا للمعادلة التالية:

$$F = \frac{MSB}{MSE} = \left[\frac{n_1 n_2}{n_1 + n_2}\right]\left[\frac{n_1 + n_2 - K - 1}{(n_1 + n_2 - 2)K}\right](D^2) \quad \#(1)$$

ثانياً: إيجاد قيمة F الجدولية:

لحساب قيمة F الجدولية يتم استخدام الصيغة التالية: $F\ (k، n_1 + n_2 - k - 1، \alpha)$

ثالثا: مقارنة قيمة F المحسوبة مع قيمة F الجدولية:

إذا كانت القيمة المحسوبة أكبر من القيمة الجدولية، فهذا يعني رفض فرض العدم (H_0) وقبول الفرض البديل (H_1)، مما يشير إلى أن الدالة لها قدرة على التمييز. أما إذا كانت القيمة المحسوبة أقل من القيمة الجدولية، فهذا يدل علي أن الدالة ليس لها قدرة على التمييز وبالتالي يتم قبول فرض العدم (H_0) (الرواشدة، ٢٠٢٢).

٢/٣/١/٢ اختبار ويلكس لمدا (Wilk's lambda).

يُستخدم هذا الاختبار للتحقق من الفرض العَدَمي، الذي ينص على أن الدالة التمييزية لا تمتلك القدرة على التمييز، مقابل الفرض البديل، الذي يؤكد أن للدالة قدرة على التمييز.

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

يتم حساب إحصاء الاختبار (Λ) علي النحو التالي:

$$\Lambda = \prod_{i=1}^{k} \frac{1}{1+\lambda_i} \quad \#(2)$$

حيث إن λ_i تمثل القيم الذاتية (eigenvalues) لكل المتغيرات، بينما يشير k إلي عدد المتغيرات المستقلة. وتتراوح قيمة (Λ) بين الواحد الصحيح والصفر، حيث:

- كلما كانت قيمتها مساوية أو قريبة من الواحد الصحيح، دلّ ذلك على تساوي متوسطات المجموعات، مما يعني أن الدالة لا تمتلك القدرة على التمييز بين المجموعات.
- كلما كانت قيمتها مساوية أو قريبة من الصفر، دلّ ذلك على عدم تساوي متوسطات المجموعات، مما يعني أن الدالة تمتلك القدرة على التمييز بين المجموعات (Farouk & Bassiouni, 2024).

يُستخدم اختبار ويلكس لمدا أيضاً في اختيار المتغيرات التي سيتم إدخالها في الدالة التمييزية وكذلك تحديد المتغيرات التي سيتم استبعادها. ويتم ذلك من خلال اختيار

المتغيرات التي تمتلك أقل قيمة لإحصاء ويلكس لمدا Λ وأعلى قيمة لإحصائية F (Sarker & Chakraborty, 2024).

٤/١/٢ اختبارات جودة الملائمة للنموذج التمييزي:

تشير الملاءمة إلى مدى توافق النموذج الإحصائي مع بيانات عينة الدراسة، كما تُستخدم جودة الملاءمة لقياس مدى التقارب بين القيم المشاهدة والقيم المتوقعة للنموذج.

١/٤/١/٢ اختبار مربع كاي χ^2 Chi-Square Test:

يعد اختبار مربع كاي، من أهم إسهامات كارل بيرسون في النظرية الحديثة للإحصاء. وهو اختبارًا لا معلمي يستخدم لغرضين محددين: الأول لاختبار الفرضية بعدم وجود ارتباط بين مجموعتين أو أكثر (أي للتحقق من استقلالية المتغيرات)، والثاني لاختبار مدى توافق القيم المشاهدة مع القيم المتوقعة (أي لاختبار حسن المطابقة) (Salh et al., 2021). وتُحسب قيمته وفقًا للصيغة التالية:

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \#(3)$$

حيث: O_i: تمثل القيم المشاهدة. E_i: تمثل القيم المتوقعة. n: تمثل عدد المشاهدات.

٢/٤/١/٢ معامل الارتباط القانوني (Canonical Correlation).

يعادل الارتباط القانوني في التحليل التمييزي معامل إيتا (eta) في تحليل التباين. ويُستخدم كمؤشر لقياس جودة توفيق النموذج، فكلما ارتفعت قيمته، دلّ ذلك على جودة توفيق عالية، والعكس صحيح. ويتم حسابه من خلال أخذ الجذر التربيعي لنسبة مجموع مربعات التباينات بين المجموعات إلى مجموع مربعات التباينات الكلي (Hahs-Vaughn, 2024, p. 368; Verma, 2013, p. 394)، وذلك وفقاً للصيغة التالية:

$$\text{Canonical correlation} = \sqrt{\frac{SS_{Between\_groups}}{SS_{Total}}} \quad \#(4)$$

٣/٤/١/٢ معامل التحديد (R^2 Coefficient of determination).

هو مربع معامل الارتباط القانوني، ويُستخرج من جدول تحليل التباين الأحادي. يُستخدم هذا المقياس لاختبار قدرة النموذج التمييزي، أو لتحديد نسبة مساهمة العوامل المؤثرة التي يتضمنها النموذج في تفسير المتغير التابع، وذلك من خلال قيمة الجذر الكامن Eigenvalue (الدمرداش وبسيوني، ٢٠٢٢).

٥/١/٢ المعاملات التمييزية المعيارية (Standardized coefficients).

تُستخدم معاملات المعادلة التمييزية المعيارية لتحديد مدى أهمية المتغيرات، حيث تساهم المتغيرات ذات القيم المطلقة المرتفعة لمعاملاتها بدرجة كبيرة في تكوين المعادلة التمييزية. وتعكس إشارة المعامل التمييزي المعياري طبيعة مساهمة المتغير في التمييز، سواء كانت مساهمة موجبة أو سالبة (Hahs-Vaughn, 2024, p. 365)، ومن خلال المعادلة التمييزية المعيارية، يتم تحديد الحد الفاصل بين المعاملات التمييزية والمجموعات، والذي يمثل نقطة الوسط بين المجموعتين، حيث يُعبر هذا الحد عن الوسط الحسابي للمعاملات التمييزية المعيارية للمجموعات (أبو دومة، ٢٠١٩).

$$\hat{Y} = \hat{\alpha}_1 X_1 + \hat{\alpha}_2 X_2 + \hat{\alpha}_3 X_3 + \cdots + \hat{\alpha}_n X_n \quad \#(5)$$

حيث:

$\hat{Y}$: القيمة التمييزية المعيارية. X_n: المتغير التمييزي المعياري.

$\hat{\alpha}_n$: المعامل التمييزي المعياري. n: يمثل عدد المتغيرات التمييزية المعيارية.

٦/١/٢ المعاملات التمييزية غير المعيارية (Unstandardized coefficients).

يتم استخدام المعاملات التمييزية غير المعيارية في تكوين النموذج التمييزي بدلاً من المعاملات التمييزية المعيارية، وذلك لأن المتغيرات التمييزية للمجموعات تظهر

بالقيم الحقيقية والنسب، وليس بالقيم المعيارية. وتتمثل المعاملات التمييزية غير المعيارية بقيم (b_n) الظاهرة في المعادلة التالية (محمد وآخرون، ٢٠٢٠):

$$Y = f + b_1S_1 + b_2S_2 + \cdots + b_nS_n \#(6)$$

حيث:

Y: القيمة التمييزية غير المعيارية. f: الجزء الثابت.

b_n: المعاملات التمييزية غير المعيارية. S_n: المتغيرات التمييزية غير المعيارية.

٢/٢ نموذج الانحدار اللوجستي:

نموذج الانحدار اللوجستي هو أسلوب إحصائي لتحليل البيانات يهدف إلى وصف العلاقة بين المتغير التابع، الذي يحتوي على فئتين أو أكثر، وبين متغير واحد أو أكثر من المتغيرات المستقلة سواء كانت على مقياس مستمر أو فئوي. يمتلك أسلوب الانحدار اللوجستي تقنيات وإجراءات مشابهة لأسلوب الانحدار الخطي. ففي حين يعتمد الانحدار الخطي عادة على طريقة المربعات الصغرى العادية (OLS) لتقدير قيم المعلمات، يستخدم الانحدار اللوجستي طريقة تقدير الإمكان الأعظم (MLE). تهدف طريقة الإمكان الأعظم إلى تعظيم احتمال تصنيف المشاهدات المرصودة بشكل صحيح ضمن الفئات المناسبة. (Meloun & Militký, 2011, p.283; Wibowo & Wihayati, 2021).

١/٢/٢ افتراضات نموذج الانحدار اللوجستي:

على عكس التحليل التمييزي لا يتطلب الانحدار اللوجستي افتراضات محددة حول توزيع المتغيرات التنبؤية. حيث لا يشترط أن تكون المتغيرات التنبؤية موزعة توزيعًا طبيعيًا، أو أن تكون مرتبطة خطيًا، أو تجانس مصفوفتي التباين والتغاير داخل كل مجموعة. ومع ذلك أكدت دراسة Ajeel et al (2023) أن الاستخدام الفعّال للانحدار اللوجستي الثنائي يتطلب الالتزام بمجموعة من الافتراضات كما أشارت دراسة Beacom (2023) إلى أهمية مراعاة هذه الافتراضات لضمان دقة النتائج، ومن أبرزها:

١- أن يكون المتغير التابع ثنائي بطبيعته.

٢- عدم وجود مشكلة التعددية الخطية (Multicollinearity).

٣- عدم وجود قيم شاذة (متطرفة) في البيانات.
٤- وجود عدد كافٍ من المشاهدات لكل متغير.
٥- يُفترض أن تكون الفئات شاملة ومحددة بحيث ينتمي كل عنصر إلى فئة واحدة.
٦- وجود علاقة خطية بين المتغيرات المستقلة المستمرة ولوغاريتم المتغير التابع.

٢/٢/٢ بناء الشكل الرياضي لنموذج الانحدار اللوجستي الثنائي:

أولا: الانتقال من الانحدار الخطي إلى الانحدار اللوجستي.

الانحدار اللوجستي الثنائي هو نموذج إحصائي يُستخدم للتنبؤ بمخرجات ثنائية (binary outcome)، مثل النجاح/الفشل. وهو امتداد للانحدار الخطي يُستخدم عندما تكون النتيجة المراد التنبؤ بها غير متصلة، بل فئوية ثنائية. وبالتالي، يمكن أن تعمل مبادئ الانحدار الخطي كإطار مرجعي للانحدار اللوجستي. وعند التعامل مع مشاكل الانحدار، يمكن وصف الانحدار الخطي على النحو التالي:

$$\mathrm{E}(Y \mid x) = \beta_0 + \beta_1 x \#(7)$$

عندما يكون المتغير التابع y ثنائيًا، أي يأخذ احدي قيمتين مثل (مصاب أو غير مصاب)، يجب أن يكون المتوسط الشرطي (القيمة المتوقعة) $\mathrm{E}(Y \mid x)$ في النطاق [0 - 1] بمعني $(0 \le \mathrm{E}(Y \mid x) \le 1)$. وبالتالي فإن استخدام الانحدار الخطي يصبح غير مناسب. والسبب في ذلك هو أن التوقع الخطي $\mathrm{E}(Y \mid x)$ قد يأخذ قيمًا خارج النطاق [0 - 1]، مما لا يعكس بدقة احتمالية وقوع الحدث أو عدم وقوعه.

للتغلب على مشكلة التوقعات التي تقع خارج النطاق [صفر، واحد]، نستخدم الانحدار اللوجستي. حيث يتمثل الحل في تغيير العلاقة بين المتغير المستقل x والمتغير التابع y إلى علاقة غير خطية تسمح بالتنبؤات التي تقع دائمًا بين صفر و واحد. للتبسيط، نستخدم $\pi(x) = \mathrm{E}(Y \mid x)$ لتمثيل القيمة المتوقعة للمتغير التابع y بمعلومية المتغير المستقل x عندما يكون المتغير التابع ثنائياً. وتكون معادلة النموذج اللوجستي كالتالي:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \#(8)$$

$\pi(x)$: احتمال وقوع الحدث المستهدف بناءً علي قيمة المتغير المستقل x.

ثانياً: التحويل اللوغاريتمي للدالة اللوجستية (تحويلة اللوجيت) Logit Transformation:

للتغلب على قيود النماذج الخطية مع البيانات الثنائية، تم استخدام التحويل اللوغاريتمي للدالة اللوجستية (تحويلة اللوجيت). يساعد هذا التحويل في إنشاء علاقة خطية بين المتغيرات المستقلة ولوغاريتم نسبة الترجيح الخاصة بالمتغير التابع. ويتم توضيح ذلك باستخدام الصيغة التالية:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x \#(9)$$

يُظهر اللوجيت $g(x)$ علاقة خطية مع المتغيرات المستقلة، وهذا يعني أنه عند حدوث تغييرًا في المتغير المستقل x سيؤدي إلى تغيير في قيمة اللوغاريتم الطبيعي لنسبة الترجيح $g(x)$، مما يسمح بفهم تأثير المتغير المستقل على احتمالية وقوع الحدث (Lu & Wang, 2024, p.182).

٣/٢/٢ اختبارات معنوية نموذج الانحدار اللوجستي:

بعد تقدير نموذج الانحدار اللوجستي، من الضروري اختبار معنوية المتغيرات الداخلة في النموذج لتحديد المتغيرات التي يجب الاحتفاظ بها في النموذج وأيها غير معنوية ويجب إزالتها.

١/٣/٢/٢ اختبار نسبة الإمكان Likelihood ratio test:

توضح الملاءمة الكلية للنموذج مدى قوة العلاقة بين جميع المتغيرات المستقلة، مجتمعة، والمتغير التابع. ويمكن تقييمها من خلال مقارنة ملاءمة نموذجين، أحدهما يحتوي على المتغيرات المستقلة والآخر لا يحتوي عليها. يُقال إن النموذج اللوجستي الذي يحتوي على عدد K من المتغيرات المستقلة يوفر ملاءمة أفضل للبيانات إذا أظهر تحسنًا مقارنة بالنموذج الذي لا يحتوي على متغيرات مستقلة (النموذج الصفري).

$$\boldsymbol{D = -2\ln\left[\frac{(\text{ likelihood of the fitted model })}{(\text{likelihood of the saturated model })}\right]\#(10)}$$

الإحصاء D في المعادلة (10) تُسمى بالانحراف (deviance)، وفي الانحدار اللوجستي، تلعب هذه الإحصاء نفس الدور الذي يلعبه مجموع المربعات الخطأ (SSE) في الانحدار الخطي.

٢/٣/٢/٢ اختبار والد Wald Test:

يُعرف هذا الاختبار بإحصائية والد، وهي تُحدد كنسبة بين المُعامِل المقدَّر$(\hat{b})$ والخطأ المعياري له $(S.E_b)$. يُستخدم هذا الاختبار لتقييم الدلالة الإحصائية لمعاملات الانحدار اللوجستي، مما يساعد في تحديد ما إذا كانت هذه المعاملات ذات أهمية إحصائية أم لا. ويمكن التعبير عن إحصائية والد (w) بالصيغة التالية:

$$w = \left(\frac{\hat{b}}{S.E_b}\right)\#(11)$$

حيث تتبع الإحصائية w توزيع Z.

- إذا كانت قيمة (w) ذات دلالة إحصائية، فهذا يعني أن المتغير المستقل (x) له تأثير ملحوظ على التنبؤ بقيمة المتغير التابع.
- إذا كانت قيمة (w) ليست ذات دلالة إحصائية، فهذا يعني أن المتغير المستقل (x) لا يؤثر بشكل كبير على التنبؤ بالمتغير التابع، مما يجعله غير ضروري في النموذج (Al_Bairmani & Ismael, 2021)

٤/٢/٢ اختبارات جودة الملائمة للنموذج اللوجستي:

يعبر مقياس جودة التوفيق للنموذج عن مدى قرب القيم المشاهدة من خط التقدير، والملائمة تعني هل النموذج الإحصائي ملائم لبيانات عينة الدراسة أم لا، وتقيس جودة الملائمة درجة التقارب بين القيم المشاهدة والقيم المتوقعة في النموذج.

١/٤/٢/٢ اختبار Hosmer-Lemeshow لجودة المطابقة ($\hat{C}$):

يُستخدم اختبار Hosmer-Lemeshow لتقييم جودة ملاءمة نموذج الانحدار اللوجستي، حيث يساعد في تحديد مدى توافق النموذج مع البيانات. يقوم هذا الاختبار بتجميع حالات العينة بناءً على قيم الاحتمالات المقدرة، واقترح هوسمر وليمشو تطبيق واحدة من استراتيجيتين للتجميع ضمن هذا الاختبار وهما:

١- تجميع الحالات بناءً على النسب المئوية للاحتمالات المقدرة.

٢- تجميع الحالات بناءً على قيم ثابتة للاحتمالات المقدرة.

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)} \#(12)$$

حيث:

n'_k: العدد الكلي للحالات في المجموعة K. $o_k = \sum_{j=1}^{c_k} y_j$: عدد الاستجابات.

$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n'_k}$: متوسط الاحتمال المقدر.

تتبع إحصاء الاختبار $\hat{C}$ توزيع مربع كاي χ^2 بدرجات حرية (g - 2). فإذا كانت قيمة $\hat{C}$ مع درجات الحرية ومستوى الدلالة p-value أكبر من 0.05 فهذا يعني أن القيم المشاهدة تتساوى مع القيم المتوقعة، وهذا يدل على جودة التوفيق للنموذج (Hosmer & Lemeshow, 2000, pp.148-150).

٢/٤/٢/٢ معامل التحديد للنموذج اللوجستي (R^2Pseudo-):

يمكن معرفة القوة التفسيرية لنموذج الانحدار اللوجستي عن طريق حساب بعض المقاييس المشابهة لمعامل التحديد (R^2) في نماذج الانحدار الخطي، توضح هذه المقاييس نسبة التباين في المتغير التابع التي يمكن تفسيرها بواسطة المتغيرات المستقلة المدرجة في النموذج (Enad & Alrawi, 2022; Darnius & Siahaan, 2023).

١- معامل تحديد كوكس وسنيل ($\mathbf{Cox\ and\ Snell's\ R^2}$): وفقاً لهذا المقياس يتم مقارنة لوغاريتم احتمالية النموذج الجديد عند إضافة المتغيرات المستقلة (L_1) بلوغاريتم احتمالية النموذج الصفري (L_0). يتميز هذا المقياس بأن أقصى قيمة ممكنة له أقل من واحد، حتى بالنسبة للنموذج المثالي. ويحسب كالتالي:

$$R^2_{\text{cox \&snell}} = 1 - \left[\frac{L_0}{L_1}\right]^{\left(\frac{2}{n}\right)} \#(13)$$

٢- معامل تحديد ناغلكيرك ($\mathbf{Nagelkerke's\ R^2}$): هو نسخة معدلة من معامل التحديد الخاص بـ "كوكس وسنيل"، حيث يتم تعديل مقياس المؤشر ليغطي المدي الكامل من صفر إلى واحد، وذلك عن طريق قسمة معامل التحديد "كوكس وسنيل" على أقصى قيمة ممكنة له. ويحسب كالتالي:

$$R^2_{\text{Nagelkerke}} = \frac{R^2{}_{\text{cox \&snell}}}{1 - [L_0]^{\left(\frac{2}{n}\right)}} \#(14)$$

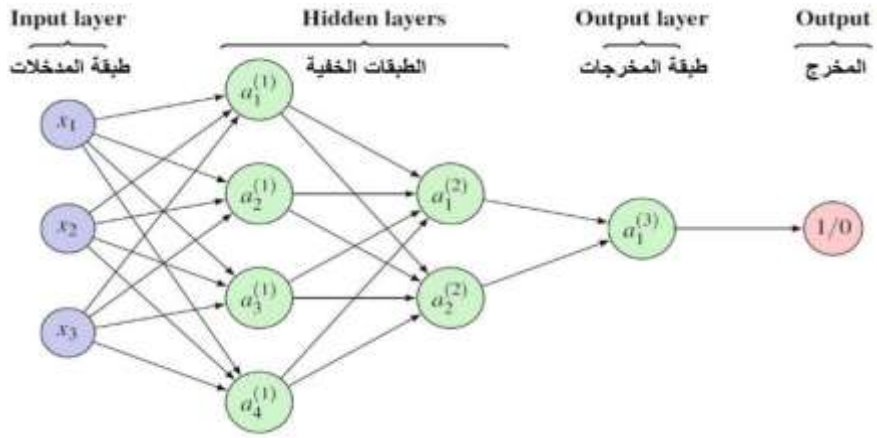
٣/٢ نموذج الشبكات العصبية الاصطناعية:

الشبكات العصبية الاصطناعية (ANNs) هي تقنية تصنيف قوية تتكون من مجموعة من الخلايا العصبية المتصلة في طبقات مختلفة. تتألف الشبكة العصبية من طبقة إدخال، وطبقة (أو أكثر) مخفية، وطبقة إخراج. يتم تمرير البيانات إلى العقد (أي الخلايا العصبية) من طبقة إلى أخرى، حيث تحتوي كل عقدة على دالة تنشيط تمثل تحويلًا رياضيًا للبيانات. يتم تعديل المدخلات من خلال الأوزان وإضافة التحيز، مما يحدد ما إذا كان سيتم تنشيط الخلية العصبية أم لا. تُستخدم هذه الخوارزمية في مجموعات البيانات المعقدة وغير المرتبطة (Alajramy & Jarrar, 2022, p.19).

الشبكة العصبية هي مجموعة مترابطة من عناصر معالجة بسيطة، تُعرف بالوحدات أو الخلايا العصبية، مرتبة في بنية تحاكي تنظيم الخلايا العصبية في أدمغة البشر. تهدف هذه الشبكة إلى معالجة المعلومات من خلال سلسلة من العمليات

الرياضية ذات الاستجابات الديناميكية، وذلك للتعرف على الأنماط في البيانات التي تكون معقدة للغاية بحيث لا يمكن للبشر تحديدها يدويًا (Tena et al., 2021).

١/٣/٢ مكونات الشبكة العصبية الاصطناعية:



شكل (1): معمارية الشبكة العصبية الاصطناعية

المصدر: (Nugues, 2024, p.198)

طبقة الإدخال (Input layer):

تُعد هذه الطبقة مسؤولة عن استقبال البيانات، والإشارات، والسمات، أو القياسات من البيئة الخارجية. عادةً ما يتم توحيد هذه المدخلات (العينات أو الأنماط) ضمن حدود القيم التي تنتجها دوال التنشيط. ويؤدي هذا التوحيد إلى تحسين الدقة العددية للعمليات الحسابية التي تُجريها الشبكة.

الطبقات الخفية أو الوسيطة (Hidden or intermediate layers):

تتكون هذه الطبقات من الخلايا العصبية المسؤولة عن استخراج الأنماط المرتبطة بالعملية أو النظام قيد التحليل. وتقوم هذه الطبقات بتنفيذ معظم عمليات المعالجة الداخلية في الشبكة؛ وقد تحتوي الشبكة العصبية على طبقة خفية واحدة أو أكثر.

طبقة الإخراج (Output layer):

هي الطبقة النهائية للشبكة العصبية والتي تتكون أيضًا من الخلايا العصبية، وبالتالي فهي مسؤولة عن إنتاج وعرض المخرجات النهائية للشبكة، والتي تنتج عن عمليات المعالجة التي أجرتها الخلايا العصبية في الطبقات السابقة (da Silva et al., 2017, p.21).

الوصلات البينية (Connections):

هي عبارة عن وصلات اتصال بين الطبقات المختلفة للشبكة العصبية، تقوم بربط الطبقات مع بعضها البعض أو الوحدات داخل كل طبقة عبر الأوزان التي تكون مرفقة أو مصاحبة لكل وصلة بينية، وتكمن وظيفة هذه الوصلات في نقل الإشارات بين وحدات المعالجة أو الطبقات (المهدي وآخرون، ٢٠٢٣).

وحدات المعالجة (Processing Elements):

وحدات المعالجة أو العصبونات هي الوحدات التي تقوم بعملية معالجة المعلومات في الشبكة العصبية، وتتصل هذه الوحدات بطرق مختلفة بواسطة الوصلات البينية. **وتتألف وحدة المعالجة أو العصبون من المكونات الأساسية التالية:**

معاملات الأوزان (weighting coefficients):

يُعتبر الوزن هو العنصر الرئيسي في الشبكات العصبية الاصطناعية، إذ يمثل الروابط التي يتم من خلالها نقل البيانات من طبقة إلى أخرى. ويعبر الوزن عن القوة النسبية أو الأهمية النسبية لكل مدخل إلى عنصر المعالجة (المهدي وآخرون، ٢٠٢٣).

دالة الجمع (Summation Function):

تقوم هذه الدالة بحساب متوسط الأوزان لكل المدخلات الواردة إلي وحدة المعالجة، ويتم ذلك عن طريق ضرب كل قيمة مدخلة (p_j) في الوزن المرتبط بها (w_{ij})، ليتم في النهاية الحصول علي مجموع المدخلات المرجحة (الموزونة)، ويمكن التعبير عن ذلك رياضياً بالمعادلة التالية (Mallik et al., 2024, p.74):

$$n = b + \sum_{j=1}^{R} w_{ij} p_j \ \#(15)$$

حيث: b يمثل التحيز (bias)، وهو يُعد وزن إضافي مرتبط بالخلية العصبية.

دالة التحويل (Transfer Function):

تعرف أيضاً بدالة التنشيط (Activation Function)، حيث تقوم بإجراء المعادلات الرياضية على القيم الخارجة من دالة الجمع، وتعديل الأوزان النسبية باستمرار طوال فترة تدريب الشبكة. ويوجد العديد من دوال التنشيط التي تم تقديمها من قبل الباحثين، والتي تختلف وفقاً لنوع المخرجات المطلوبة وأهداف الشبكة المراد تحقيقها (Negnevitsky, 2011, pp.169-170)، **ومن أبرز هذه الدوال ما يلي**:

- **دالة الخطوة (Step function)**
- **دالة الإشارة (sign function)**
- **الدالة اللوجستية (Logistic or Sigmoid Function)**
- **الدالة الخطية (Linear Function)**

دالة المخرجات (output function)

بعد أن تقوم دالة الجمع بجمع المدخلات المرجحة بالأوزان، ثم تقوم دالة التحويل (التنشيط) بمعالجة ناتج الجمع وتحويله إلى قيمة محصورة ضمن مدى معين. بعد ذلك، تُنتج دالة المخرجات المخرج النهائي للشبكة العصبية الاصطناعية بناءً على هذا التحويل (بسيوني، ٢٠٢٢).

٢/٣/٢ التركيب المعماري للشبكة العصبية الاصطناعية (ANN Architecture):

تتكون الشبكة العصبية من مجموعة من العصبونات المرتبطة داخلياً فيما بينها، ويتحدد نوع الشبكة بناء على نوعية الارتباط بين العصبونات المكونة لها، بالإضافة الى طبيعة هذه العصبونات. ويشير التركيب المعماري للشبكة العصبية إلي الطريقة التي ترتبط بها وحدات المعالجة مع بعضها البعض داخل كل طبقة أو بين الطبقات المختلفة في الشبكة. ويمكن

تصنيف التركيب المعماري للشبكة العصبية وفقاً لعدد الطبقات التي تتكون منها الشبكة، وطبيعة انتشار البيانات عبر طبقاتها (الإمام وآخرون، ٢٠٢٥).

١/٢/٣/٢ تصنيف الشبكات العصبية الاصطناعية وفقاً لعدد الطبقات:

الشبكات وحيدة الطبقة (Single-Layer Networks):

تُعد هذه الشبكات من أبسط أنواع الشبكات العصبية، حيث تتكوّن من طبقة واحدة من عناصر المعالجة. في هذا النوع، ترتبط مدخلات الشبكة مباشرة بالمخرجات، حيث تُجرى جميع العمليات الحسابية داخل طبقة المخرجات.

الشبكات متعددة الطبقات (Multilayer Networks):

يتركب هذا النوع من الشبكات من عدة طبقات من عناصر المعالجة، تربط بينها الوصلات البينية المعروفة بالأوزان. حيث تتكوّن الشبكة العصبية متعددة الطبقات من طبقة المدخلات، والتي لا تُحتسب ضمن عدد الطبقات الفعلية، وطبقة المخرجات، وبين الطبقتين السابقتين توجد الطبقة المخفية. ويمكن أن تحتوي الشبكة على أكثر من طبقة مخفية، وذلك وفقًا لنوع التطبيق الذي تُستخدم فيه.

٢/٢/٣/٢ تصنيف الشبكات العصبية وفقاً لطبيعة انتشار البيانات عبر طبقاتها:

الشبكات العصبية ذات التغذية الأمامية (Feed-Forward Neural Network):

الشبكة العصبية أمامية التغذية (FFNN) هي نوع من الشبكات العصبية الاصطناعية التي تتدفق فيها المعلومات في اتجاه واحد فقط: من طبقة الإدخال إلى طبقة الإخراج، مرورًا بالطبقات الخفية. تُعد هذه الشبكة واحدة من أبسط وأكثر أنواع الشبكات العصبية شيوعًا في مختلف مهام تعلم الآلة. في هذا النوع من الشبكات، لا توجد أي آلية ارتداد أو تغذية راجعة، مما يعني أن الروابط بين الخلايا العصبية لا تشكل أي حلقات دورية (Setiawan et al., 2024).

الشبكات العصبية ذات التغذية الخلفية (Feedback Neural Network):

يأخذ هذا النوع من الشبكات العصبية شكل الشبكة متعددة الطبقات، لكنه يتميز بوجود حلقة تغذية خلفية واحدة على الأقل. في هذه الشبكات، اما أن تعود مخرجات أحد العصبونات لتُستخدم كمدخلات لنفس العصبون، فيما يُعرف بالتغذية الخلفية الذاتية، أو أن تكون مدخلات لعصبون آخر داخل الشبكة (الإمام وآخرون، ٢٠٢٥).

٤/٣/٢ استخدام الشبكات العصبية في التصنيف:

أصبح بالإمكان استخدام الشبكات العصبية الاصطناعية (ANN) في مهام التصنيف، وذلك من خلال اختيار دالة تنشيط مخصصة للتصنيف. وهناك مجموعة متنوعة من دوال التنشيط والتي تختلف باختلاف النتائج المطلوبة والأهداف المرجوة من الشبكة العصبية. في هذا السياق، سيتم استخدام دالة الخطوة لأنها تتناسب مع عملية التصنيف، حيث تقدم مخرجات ثنائية القيم: صفر وواحد، على النحو التالي:

$$f(x) = \begin{cases} 0 \text{ if } x < 0 \\ 1 \text{ if } x \geq 0 \end{cases} \#(16)$$

وتستخدم دالة الخطوة في وحدات طبقة المخرجات، بينما تستخدم الدالة اللوجستية (Sigmoid) في الطبقة الخفية، وتأخذ الشكل الآتي:

$$f(x) = \frac{1}{1 + e^{-s}} \#(17)$$

حيث إن (S) هو المجموع الموزون للمدخلات مضافًا إليه حد التحيز (bias)الذي يرمز له بالرمز (θ)، ويعطي المجموع الموزون بالصيغة التالية:

$$S = \sum_{i=1}^{n} x_i w_j + \theta \ \#(18)$$

وبالتالي تكون مخرجات الشبكة العصبية إما صفر أو واحد صحيح بناءً على المدخلات. فإذا كانت النتيجة صفر، تُصنَّف المشاهدة ضمن المجموعة الأولى، أما إذا كانت النتيجة واحد، فتُصنَّف المشاهدة ضمن المجموعة الثانية (بسيوني، ٢٠٢٢).

٤/٢ معايير تقييم الأداء والمفاضلة بين النماذج.

تكمن أهمية هذه المعايير في قدرتها على إبراز جوانب القوة والضعف في كل نموذج، مما يُتيح للباحث إجراء مقارنة علمية دقيقة تُسهم في تحديد النموذج الأكثر كفاءة وملاءمة. ومن أبرز هذه المعايير ما يلي:

١/٤/٢ مصفوفة الارتباك Confusion Matrix:

تُعد مصفوفة الارتباك أداة أساسية للتحقق من أداء نموذج التصنيف. وهي تمثيل جدولي للمخرجات المتوقعة للنموذج مقارنة بالمخرجات الحقيقية لعينة البيانات محل الاختبار. وهي تأخذ شكل مصفوفة من الدرجة (n x n) حيث يمثل n عدد فئات المخرجات المستهدفة (Okwori et al., 2024).

جدول (2): مصفوفة الارتباك Confusion Matrix

مصفوفة الارتباك		الفئة الحقيقية	
		إيجابي	سلبي
الفئة المتوقعة	إيجابي	إيجابي حقيقي (TP)	إيجابي خاطئ (FP)
	سلبي	سلبي خاطئ (FN)	سلبي حقيقي (TN)

المصدر: (Selim et al., 2024).

ويمكن استخلاص مجموعة متنوعة من مؤشرات الأداء اعتماداً علي مصفوفة الارتباك كالتالي:

- **الدقة الكلية (Accuracy):** يقيس هذا المؤشر نسبة التنبؤات الصحيحة إلى إجمالي العينات التي تم تقييمها.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \#(19)$$

- **الحساسية (Sensitivity):** تقوم بقياس النسبة المئوية للحالات الإيجابية التي تم التعرف عليها بشكل صحيح بالنسبة إلى إجمالي الحالات الإيجابية الفعلية في مجموعة البيانات.

$$\text{Sensitivity} = \text{Recall} = \frac{TP}{TP + FN} \#(20)$$

- **النوعية (Specificity):** تقيس النسبة المئوية للحالات السلبية التي تم تحديدها بشكل صحيح بالنسبة إلى إجمالي الحالات السلبية الفعلية في مجموعة البيانات.

$$\text{Specificity} = \frac{TN}{TN + FP} \#(21)$$

- **الدقة التنبؤية (Precision):** يمثل نسبة الإيجابيات الحقيقية (TP) إلى مجموع الإيجابيات الحقيقية (TP) والإيجابيات الخاطئة (FP). وتعكس هذه النسبة مدي قدرة النموذج على التعرف على الحالات الإيجابية بشكل دقيق.

$$\text{Precision} = \frac{TP}{TP + FP} \#(22)$$

- **معدل الخطأ (Error rate):** يمثل النسبة المئوية لعدد التصنيفات أو التنبؤات الخاطئة التي قام بها النموذج إلي إجمالي عدد التنبؤات.

$$\text{Error rate} = \frac{FN + FP}{TP + TN + FP + FN} \#(23)$$

- **معدل الإيجابيات الخاطئة (FPR):** هو النسبة بين عدد الحالات السلبية التي تم تصنيفها بشكل خاطئ على أنها إيجابية إلى إجمالي عدد الحالات السلبية الفعلية.

$$\text{False Positive Rate } = \frac{FP}{FP + TN} \#(24)$$

- **F1 Score:** هي مقياس يمثل المتوسط التوافقي بين الدقة التنبؤية (Precision) والحساسية (Recall) للنموذج. ويتم حساب هذا المقياس كالتالي:

$$\text{F1 Score } = 2 \times \frac{\text{Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}} \#(25)$$

٢/٤/٢ منحنى خصائص تشغيل المستقبِل ROC Curve:

تُعتبر المساحة تحت منحنى روك مقياسًا ذا أهمية خاصة لتقييم ومقارنة أداء نماذج التصنيف، وتبلغ القيمة القصوى لهذه المساحة واحد صحيح، حيث يتراوح كلا المحورين، الإيجابيات الكاذبة والإيجابيات الحقيقية، بين [0 - 1]. إذا كانت المساحة قريبة من 0.5،

فهذا يشير إلى أن أداء النموذج لا يتفوق بشكل كبير على التصنيف العشوائي للمشاهدات، وعلي العكس كلما كانت قيمة المساحة تحت المنحني أعلى وتقترب من الواحد الصحيح، كلما كان أداء النموذج في التمييز أفضل (Acito, 2023, p.139).

٣/٤/٢ الجذر التربيعي لمتوسط مربعات الأخطاء (RMSE):

يُستخدم مؤشر الجذر التربيعي لمتوسط مربعات الأخطاء (RMSE) لتحديد الفروق بين القيم الفعلية والقيم المتوقعة، مما يساعد في تقييم جودة النموذج التنبؤي بناءً على مدى اقتراب القيم المتوقعة من القيم الفعلية. لذلك، يُعتبر RMSE مقياسًا فعّالًا للمقارنة بين أداء نماذج مختلفة. حيث عند مقارنة نموذجين لنفس المهمة، يكون النموذج الذي يمتلك قيمة RMSE أقل هو الأكثر دقة.

٤/٤/٢ التحقق المتقاطع باستخدام K-fold (K-Fold Cross Validation):

تعتمد هذه الطريقة على تقسيم مجموعة البيانات إلى عدد k من المجموعات الفرعية المتساوية. ومن خلال عملية تكرارية، يتم استخدام كل مجموعة فرعية كمجموعة اختبار بينما تُستخدم المجموعات المتبقية لتدريب النموذج. مع تطبيق هذه المنهجية، يتم ضمان مشاركة كل مشاهدة في كل من مجموعة الاختبار ومجموعة التدريب، مما يقلل من خطر التحيز ويوفر تقييماً أكثر دقة لأداء النموذج وقدرته على التعميم عند التعامل مع بيانات جديدة أو غير معروفة مسبقاً. وقد أصبحت هذه التقنية أداة أساسية لاختيار النماذج وتقييم الأداء في مختلف المجالات العلمية، مما يضمن اتخاذ قرارات مستنيرة مدعومة بأدلة تجريبية (Vergara-Lozano et al., 2023).

القسم الثالث: الاطار التطبيقي للبحث

١/٣ الهدف:

يتناول هذا الجانب من البحث التحليل الإحصائي للدراسة التي تهدف إلى المقارنة بين ثلاثة نماذج إحصائية في التصنيف والتنبؤ بمرض السكري. ترتكز هذه الدراسة على استخدام التحليل التمييزي، والانحدار اللوجستي الثنائي، والشبكات العصبية الاصطناعية لتحديد النموذج الأكثر فعالية ودقة في تصنيف حالات السكري، استناداً

إلى مجموعة من المتغيرات الصحية، الديموغرافية. تم استخدام برنامج SPSS كأداة رئيسية للتحليل الإحصائي لتنفيذ النماذج الثلاثة، كما تم الاستعانة بلغة البرمجة R لتطبيق طريقة التحقق المتقاطع (Cross Validation) لتقييم أداء هذه النماذج وقدرتها على التعميم.

٢/٣ متغيرات الدراسة:

اعتمدت الدراسة على ستة عشر متغيرًا مستقلًا، تضم تسعة متغيرات تصنيفية، وسبعة متغيرات كمية.

شملت المتغيرات التصنيفية: الجنس (Gender)، الحالة الوظيفية (Employment status)، التدخين (Smoking)، التاريخ العائلي للإصابة بمرض السكري (Family history)، مستوى النشاط البدني (Physical activity)، النظام الغذائي (Diet)، كثرة التبول (Polyuria)، الشعور المتكرر بالعطش (Polydipsia)، وضبابية الرؤية (Visual blurring). **أما المتغيرات الكمية فهي:** العمر (Age)، الضغط الانقباضي (Systolic BP)، الضغط الانبساطي (Diastolic BP)، مؤشر كتلة الجسم (BMI)، الكوليسترول الكلي (Total Cholesterol)، الدهون الثلاثية (Triglycerides)، ومستوى الهيموغلوبين السكري (HbA1c Test).

٣/٣ الفحص الأولي للبيانات:

١/٣/٣ فحص القيم المفقودة والقيم المتطرفة:

أظهرت نتائج فحص البيانات خلو جميع متغيرات الدراسة من القيم المفقودة، حيث بلغت نسبة الفقد 0.00%، ما يعكس دقة وجودة عملية جمع البيانات وإدخالها. كما تم التحقق من وجود القيم المتطرفة باستخدام طريقة نطاق الربيعات (IQR)، حيث تم تحديد القيم المتطرفة كتلك التي تقع خارج النطاق (Q1 - 1.5IQR, Q3 + 1.5 IQR). وأظهرت النتائج عدم وجود أي قيم متطرفة في المتغيرات الكمية أو النوعية.

٢/٣/٣ فحص التوزيع الطبيعي:

وفقاً للمعايير المعتمدة في الدراسات السابقة، تعتبر البيانات موزعة طبيعياً إذا تراوحت قيم الالتواء (Skewness) بين ±٢ وقيم التفرطح (Kurtosis) بين ±٧ (Kim, 2013; Sovey et al., 2022).وبناءً علي نتائج التحليل الإحصائي جدول (3)، يمكن الاستنتاج بثقة أن جميع المتغيرات الكمية في الدراسة تتبع التوزيع الطبيعي بشكل مقبول. هذا الاستنتاج يعزز صحة استخدام الأساليب الإحصائية المعلمية في التحليلات اللاحقة، مما يدعم موثوقية النتائج المتوقعة من الدراسة.

جدول (3): اختبار التوزيع الطبيعي.

Kurtosis		Skewness		N	Variable
Std. Error	Statistic	Std. Error	Statistic	Statistic	
0.248	-1.137	0.124	0.169	385	Age
0.248	-0.868	0.124	0.173	385	Systolic_BP
0.248	-0.741	0.124	-0.315	385	Diastolic_BP
0.248	-0.295	0.124	0.301	385	BMI
0.248	-1.218	0.124	0.118	385	Total_Cholesterol
0.248	-1.119	0.124	0.108	385	Triglycerides
0.248	-1.376	0.124	0.091	385	HbA1c_Test

المصدر: مخرجات برنامج SPSS

٣/٣/٣ تقييم مشكلة الازدواج الخطي:

عند تحليل نتائج الدراسة الحالية جدول (4)، نجد أن جميع المتغيرات المستقلة أظهرت قيماً مقبولة لمعاملي VIF و(Tolerance). وبالتالي يمكن الاستنتاج، أن البيانات المستخدمة في هذه الدراسة لا تعاني من مشكلة الازدواج الخطي، حيث لم تتجاوز أي من قيم VIF الحد الحرج (٥ أو١٠)، كما أن جميع قيم معامل (Tolerance) كانت مرتفعة عن (٠.١) بشكل مطمئن. هذا يعزز صحة استخدام هذه المتغيرات في النماذج الإحصائية المقترحة للدراسة.

جدول (4): اختبار الازدواج الخطي.

Collinearity Statistics		Variable
VIF	Tolerance	
2.057	0.486	Gender
1.814	0.551	Age
1.412	0.708	Employment_Status
1.762	0.568	Smoking
1.407	0.711	Family_history
1.427	0.701	Physical_activity
1.086	0.921	Diet
1.381	0.724	Polyuria
1.087	0.920	Polydipsia
1.146	0.873	Visual_blurring
1.045	0.957	Systolic_BP
1.038	0.963	Diastolic_BP
1.147	0.872	BMI
1.222	0.818	Total_Cholesterol
1.203	0.831	Triglycerides
2.059	0.486	HbA1c_Test

المصدر: مخرجات برنامج SPSS

٤/٣ تطبيق نموذج التحليل التمييزي:

في هذه الدراسة، تم توظيف التحليل التمييزي لتصنيف الأفراد إلى مصابين وغير مصابين بمرض السكري، اعتمادًا على مجموعة من المتغيرات الديموغرافية والصحية، مع ترميز الحالة المستهدفة (0 = غير مصاب، 1 = مصاب). وقد تم بناء النموذج باستخدام الأسلوب التدريجي (Stepwise Method).

١/٤/٣ الاختبارات الأولية للتحليل التمييزي:

يعرض جدول (5) نتائج الاختبارات الإحصائية للتحليل التمييزي. بداية، نلاحظ أن قيمة اختبار Box's M بلغت ٥٨.٩٢٥ بمستوى دلالة ٠.١٠١، وهي قيمة غير دالة إحصائياً عند مستوى ٠.٠٥، مما يشير إلى تحقق افتراض تجانس مصفوفات التباين-التغاير بين المجموعتين، وهو افتراض أساسي من افتراضات التحليل التمييزي.

جدول (5): اختبار التجانس.

Test Results		
Box's M		58.925
F	Approx.	1.277
	df1	45
	df2	468086.177
	Sig.	0.101

كما يظهر جدول (6) أن قيمة ويلكس لامدا (Wilks' Lambda) بلغت ٠.٢١٦ بمستوى دلالة ٠.٠٠٠، وهي قيمة دالة إحصائياً، مما يدل على وجود قدرة تمييزية عالية للنموذج في التفريق بين مجموعتي الدراسة.

جدول (6): اختبار القدرة التمييزية.

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	0.216	579.622	9	0.000

يوضح جدول (7) أن قيمة الجذر الكامن (Eigenvalue) للدالة التمييزية، بلغت ٣.٦٢٤، وهي قيمة مرتفعة تعكس قوة الدالة التمييزية. وبلغت نسبة التباين المفسر ١٠٠%، وهذا متوقع نظراً لوجود مجموعتين فقط ، مما يعني أن دالة تمييزية واحدة كافية لتفسير كامل التباين بين المجموعتين. كما بلغ معامل الارتباط القانوني ٠.٨٨٥، وهي قيمة مرتفعة جداً تشير إلى علاقة قوية بين الدالة التمييزية والانتماء للمجموعات. مربع هذه القيمة (٠.٧٨٣) يماثل معامل التحديد في تحليل الانحدار، ويشير إلى أن الدالة التمييزية تفسر حوالي ٧٨.٣% من التباين في المتغير التابع (الإصابة بالسكري).

جدول (7): الجذر الكامن والارتباط القانوني.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	3.624[a]	100.0	100.0	0.885
a. First 1 canonical discriminant functions were used in the analysis.				

٢/٤/٣ تقدير النموذج باستخدام التحليل التمييزي:

يعرض جدول (8) المتغيرات التي تم إدخالها في نموذج التحليل التمييزي باستخدام الطريقة التدريجية (Stepwise Method)، حيث يتم إدخال المتغيرات خطوة بخطوة وفقاً لقدرتها التمييزية. يمكن ملاحظة أن عملية إدخال المتغيرات تمت في تسع خطوات، بدأت بالمتغير الأكثر قدرة على التمييز بين المجموعتين وانتهت بالإبقاء على تسع متغيرات في النموذج النهائي من أصل المتغيرات الستة عشر التي تم دراستها. وهذه المتغيرات هي: HbA1c Test، التاريخ العائلي للمرض، العمر، كثرة التبول، مؤشر كتلة الجسم (BMI)، النظام الغذائي، التدخين، الدهون الثلاثية، كثرة العطش

جدول (8): اختبار F للعوامل المؤثرة في التحليل التمييزي.

Step	Entered	Wilks' Lambda							
		Statistic	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	HbA1c Test	0.307	1	1	383.000	863.817	1	383.000	0.000
2	Family history	0.279	2	1	383.000	493.903	2	382.000	0.000
3	Age	0.262	3	1	383.000	357.050	3	381.000	0.000
4	Polyuria	0.249	4	1	383.000	286.872	4	380.000	0.000
5	BMI	0.239	5	1	383.000	240.939	5	379.000	0.000
6	Diet	0.232	6	1	383.000	208.097	6	378.000	0.000
7	Smoking	0.227	7	1	383.000	183.687	7	377.000	0.000
8	Triglycerides	0.221	8	1	383.000	165.507	8	376.000	0.000
9	Polydipsia	0.216	9	1	383.000	151.020	9	375.000	0.000

باستخدام المعاملات التمييزية غير المعيارية يمكن كتابة معادلة التحليل التمييزي كالتالي:

LDA = -10.041 + 0.021 (Age) + 0.369 (Smoking) + 0.719 (Family_history) - 0.452 (Diet) + 0.539 (Polyuria) + 0.353 (Polydipsia) + 0.043 (BMI) + 0.003 (Triglycerides) + 1.001 (HbA1c_Test)

٣/٤/٣ جودة التصنيف للتحليل التمييزي:

يعرض جدول (9) مصفوفة الارباك (Confusion Matrix) التي توضح عدد الحالات المصنفة بشكل صحيح وخاطئ في كل مجموعة. يمكن ملاحظة أن النموذج نجح في تصنيف ١٧٣ حالة من أصل ١٨١ حالة غير مصابة بالسكري بشكل صحيح (بنسبة ٩٥.٦%)، في حين تم تصنيف ٨ حالات فقط بشكل خاطئ على أنها مصابة بالسكري. وبالمثل، نجح النموذج في تصنيف ١٩٤ حالة من أصل ٢٠٤ حالات مصابة بالسكري بشكل صحيح (بنسبة ٩٥.١%)، بينما تم تصنيف ١٠ حالات فقط بشكل خاطئ على أنها غير مصابة. بناءً على ذلك، بلغت نسبة التصنيف الصحيح الإجمالية ٩٥.٣%، وهي نسبة مرتفعة جداً تدل على كفاءة النموذج في التمييز بين المصابين وغير المصابين بالسكري.

جدول (9): تصنيف نموذج التحليل التمييزي.

Classification Results[a]					
Diabetes			Predicted Group Membership		Total
			غير مصاب بمرض السكري	مصاب بمرض السكري	
Original	Count	غير مصاب بمرض السكري	173	8	181
		مصاب بمرض السكري	10	194	204
	%	غير مصاب بمرض السكري	95.6	4.4	100.0
		مصاب بمرض السكري	4.9	95.1	100.0
Accuracy	Sensitivity	Specificity	Precision	Error_Rate	FPR
0.9532	0.9510	0.9558	0.9604	0.0468	0.0442
Area under curve = .991; Kappa = 0.906; F1_Score= 0.956					
a. 95.3% of original grouped cases correctly classified.					

٥/٣ تطبيق نموذج الانحدار اللوجستي الثنائي:

في هذه الدراسة، استُخدم الانحدار اللوجستي للتنبؤ بالإصابة بمرض السكري اعتمادًا على مجموعة من المتغيرات الديموغرافية والصحية، مع ترميز الحالة المستهدفة (0 = غير مصاب، 1 = مصاب). وقد تم بناء النموذج باستخدام أسلوب

الحذف التدريجي الخلفي (Backward Stepwise: LR)، بدءًا بإدراج ١٦ متغيرًا، ثم استُبعدت المتغيرات غير المعنوية إحصائيًا تدريجيًا، لضمان احتفاظ النموذج النهائي بالعوامل الأكثر تأثيرًا في احتمالية الإصابة.

١/٥/٣ تقدير النموذج باستخدام الانحدار اللوجستي:

يوضح جدول (10) نتائج اختبار Hosmer and Lemeshow تدعم جودة مطابقة النموذج، حيث بلغت قيمة كاي تربيع في الخطوة الأولى ٤.٨٤٨ بدرجات حرية ٨ ومستوى دلالة ٠.٧٧٤، وفي الخطوة الثامنة ٢.٠٦٠ بدرجات حرية ٨ ومستوى دلالة ٠.٩٧٩. هذه القيم غير الدالة إحصائياً تشير إلى عدم وجود فروق معنوية بين القيم المشاهدة والمتوقعة، وبالتالي جودة مطابقة النموذج للبيانات. ارتفعت قيمة Log likelihood 2- من ٨١.٥٧٩ في الخطوة الأولى إلى ٨٧.٥٢٨ في الخطوة الثامنة، وهذا الارتفاع الطفيف متوقع مع انخفاض عدد المتغيرات المستقلة. ومع ذلك، حافظت قيم معاملات التحديد على مستويات مرتفعة، حيث بلغت قيمة Cox & Snell R^2في الخطوة النهائية ٠.٦٨٥، وقيمة Nagelkerke R^2 ٠.٩١٥، مما يشير إلى أن النموذج النهائي يفسر نسبة كبيرة من التباين ٩١.٥% في المتغير التابع (وفقاً لمعامل Nagelkerke).

جدول (10): اختبارات جودة ومطابقة النموذج.

Hosmer and Lemeshow Test				Model Summary		
Step	Chi-square	df	Sig.	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	4.848	8	0.774	81.579[a]	0.690	0.921
8	2.060	8	0.979	87.528[b]	0.685	0.915

يعرض جدول (11) نتائج نموذج الانحدار اللوجستي النهائي في الخطوة الثامنة (Step 8) بعد تطبيق طريقة الحذف التدريجي الخلفي. يتضح من الجدول أن التحليل التدريجي قد استبعد سبع متغيرات من النموذج المبدئي (الجنس، العمر، الحالة الوظيفية، عدم وضوح

الرؤية، ضغط الدم الانقباضي، ضغط الدم الانبساطي، والدهون الثلاثية) لعدم معنويتها الإحصائية. أما المتغيرات التسعة المتبقية في النموذج النهائي هي (التدخين، التاريخ العائلي، النشاط البدني، النظام الغذائي، كثرة التبول، كثرة شرب الماء، مؤشر كتلة الجسم، الكوليسترول الكلي، واختبار HbA1c) جميعها ذات دلالة إحصائية عند مستوى معنوية ٠.٠٥ حسب قيم اختبار Wald وقيم الدلالة الإحصائية المقابلة (.Sig).

جدول (11): معاملات النموذج اللوجستي بطريقة Stepwise ونسبة الاحتمالات والحدود العليا والدنيا لفترات الثقة.

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 8[a]	Smoking	2.075	0.685	9.167	1	0.002	7.962	2.079	30.497
	Family history	1.917	0.595	10.375	1	0.001	6.800	2.118	21.830
	Physical activity	-1.903	0.830	5.256	1	0.022	0.149	0.029	0.759
	Diet	-1.918	0.639	9.023	1	0.003	0.147	0.042	0.513
	Polyuria	1.226	0.596	4.223	1	0.040	3.406	1.058	10.961
	Polydipsia	1.180	0.594	3.948	1	0.047	3.254	1.016	10.423
	BMI	0.189	0.067	8.013	1	0.005	1.208	1.060	1.378
	Total Cholesterol	0.016	0.007	5.252	1	0.022	1.017	1.002	1.031
	HbA1c_Test	2.682	0.378	50.272	1	0.000	14.613	6.963	30.670
	Constant	-27.834	4.463	38.901	1	0.000	0.000		

بناءً على قيم المعاملات في جدول (3-11)، يمكن كتابة معادلة الانحدار اللوجستي المقدرة كالتالي:

$$g(x) = -27.834 + 2.075\,(\text{Smoking}) + 1.917\,(\text{Family history}) - 1.903\,(\text{Physical activity}) - 1.918\,(\text{Diet}) + 1.226\,(\text{Polyuria}) + 1.180\,(\text{Polydipsia}) + 0.189\,(\text{BMI}) + 0.016\,(\text{Total Cholesterol}) + 2.682\,(\text{HbA1cTest}).$$

٢/٥/٣ جودة توفيق نموذج الانحدار اللوجستي:

جدول (12): يعرض جدول التصنيف في الخطوة الثامنة (Step 8) أداء النموذج النهائي للانحدار اللوجستي بعد استبعاد المتغيرات غير المعنوية. وتوضح مصفوفة الارباك (Confusion Matrix) توزيع الحالات المصنفة بشكل صحيح والخاطئ. من إجمالي ١٨١ شخصاً غير مصاب بمرض السكري، تم تصنيف ١٧٤ منهم بشكل صحيح، في حين تم تصنيف ٧ أشخاص خطأً على أنهم مصابون بالسكري. ومن إجمالي ٢٠٤ شخصاً مصاباً بمرض السكري، تم تصنيف ١٩٥ منهم بشكل صحيح، بينما تم تصنيف ٩ أشخاص خطأً على أنهم غير مصابين. بلغت الدقة الكلية للنموذج ٩٥.٨%، مما يعكس كفاءته العالية في التصنيف، في حين بلغت الحساسية ٠.٩٥٦، مما يعني أن النموذج قادر على تحديد ٩٥.٦% من الحالات الفعلية المصابة بالسكري.

جدول (12): تصنيف نموذج الانحدار اللوجستي (الدقة ومعدل الخطأ والحساسية والخصوصية).

Observed			Predicted		
			Diabetes		Percentage Correct
			غير مصاب بمرض السكري	مصاب بمرض السكري	
Step 8	Diabetes	غير مصاب بمرض السكري	174	7	96.1%
		مصاب بمرض السكري	9	195	95.5%
	Overall Percentage				95.8%
Accuracy	Sensitivity	Specificity	Precision	Error_Rate	FPR
0.958	0.956	0.961	0.965	0.042	0.039
Area under curve = .992; Kappa = 0.9166; F1_Score= 0.961					

٦/٣ تطبيق نموذج الشبكات العصبية الاصطناعية:

تعتبر الشبكات العصبية الاصطناعية من أهم تقنيات التعلم الآلي وأكثرها تطوراً في مجال التصنيف والتنبؤ. تستمد هذه الشبكات فكرتها الأساسية من محاكاة آلية عمل الخلايا العصبية في الدماغ البشري، حيث تتكون من طبقات متعددة من العقد (Nodes)المترابطة التي تعالج المعلومات بشكل متوازٍ. وقد تم تطبيقها في دراستنا الحالية لتحليل وتصنيف حالات مرض السكري. تتميز هذه التقنية بقدرتها على

التعامل مع العلاقات المعقدة وغير الخطية بين المتغيرات، مما يجعلها مناسبة بشكل خاص للتطبيقات الطبية.

١/٦/٣ تحليل البيانات باستخدام الشبكات العصبية الاصطناعية:

بالنظر إلى جدول (13) Case Processing Summary، نجد أن الشبكة العصبية تم تدريبها على ٢٧٧ حالة (٧١.٩% من إجمالي البيانات) بينما خُصصت ١٠٨ حالات (٢٨.١%) للاختبار. هذا التقسيم يتبع أفضل الممارسات في مجال التعلم الآلي، حيث يمثل نسبة ٧٠:٣٠ تقريباً لضمان تدريب كافٍ للنموذج مع الاحتفاظ بمجموعة اختبار مناسبة للتقييم. استبعاد صفر حالات يشير إلى أن البيانات كاملة وخالية من القيم المفقودة، وهو أمر مهم لفعالية الشبكة العصبية.

جدول (13): ملخص تدريب واختبار بيانات الشبكة العصبية.

		N	Percent
Sample	Training	277	71.9%
	Testing	108	28.1%
Valid		385	100.0%
Excluded		0	
Total		385	

يوضح جدول (14) Network Information التفاصيل الفنية للشبكة، حيث يؤكد استخدام طريقة Standardized لتنسيق وتهيئة المتغيرات المدخلة، وهو إجراء ضروري للشبكات العصبية لضمان تقارب فعال أثناء التدريب. تستخدم الشبكة دالة التنشيط Hyperbolic Tangent في الطبقات المخفية كما هو موضح في الجدول. تعتبر هذه الدالة فعالة لأنها تسمح بتدفق التدرجات بشكل أفضل مقارنة بدالة Sigmoid التقليدية، مما يساعد في تجنب مشكلة تلاشي التدرج (Vanishing Gradient Problem)خاصة في الشبكات العميقة (Glorot & Bengio, 2010).

في طبقة المخرجات، تستخدم الشبكة دالة Softmax لإنتاج احتمالات التصنيف للحالتين: مصاب بالسكري (Diabetes=1) وغير مصاب (Diabetes=0). وجود مخرجتين يمثل التصنيف الثنائي حيث يستخدم الناتج كقيمة احتمالية لانتماء الحالة إلى كل فئة. اختيار Softmax مناسب للغاية لهذا النوع من التصنيف لأنه يضمن أن

مجموع الاحتمالات يساوي واحد. كما يعد استخدام Cross-entropy كدالة خطأ متوافق مع دالة التنشيط Softmax ويعد أحد أفضل الممارسات في تدريب الشبكات العصبية للتصنيف.

تشير البنية العامة للشبكة، مع وجود طبقتين مخفيتين بعدد متناقص من العصبونات (١٤ و ٧)، إلى استخدام نهج "التدرج الهرمي" في معمارية الشبكة، حيث تقوم الطبقات الأولى بالتعرف على الخصائص المعقدة بينما تقوم الطبقات اللاحقة بتجريد هذه الخصائص في تمثيلات أبسط. هذا التصميم يساعد في تقليل التعقيد الحسابي مع الحفاظ على القدرة على تعلم الأنماط المعقدة.

جدول (14): معلومات ومتغيرات الشبكة العصبية المستخدمة في التحليل.

Network Information			
Input Layer	Covariates	1	Gender
		2	Age
		3	Employment_Status
		4	Smoking
		5	Family_history
		6	Physical_activity
		7	Diet
		8	Polyuria
		9	Polydipsia
		10	Visual_blurring
		11	Systolic_BP
		12	Diastolic_BP
		13	BMI
		14	Total_Cholesterol
		15	Triglycerides
		16	HbA1c_Test
	Number of Units[a]		16
	Rescaling Method for Covariates		Standardized
Hidden Layer(s)	Number of Hidden Layers		2
	Number of Units in Hidden Layer 1[a]		14
	Number of Units in Hidden Layer 2[a]		7
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	Diabetes
	Number of Units		2
	Activation Function		Softmax
	Error Function		Cross-entropy
a. Excluding the bias unit			

٢/٦/٣ جودة التصنيف باستخدام الشبكات العصبية الاصطناعية:

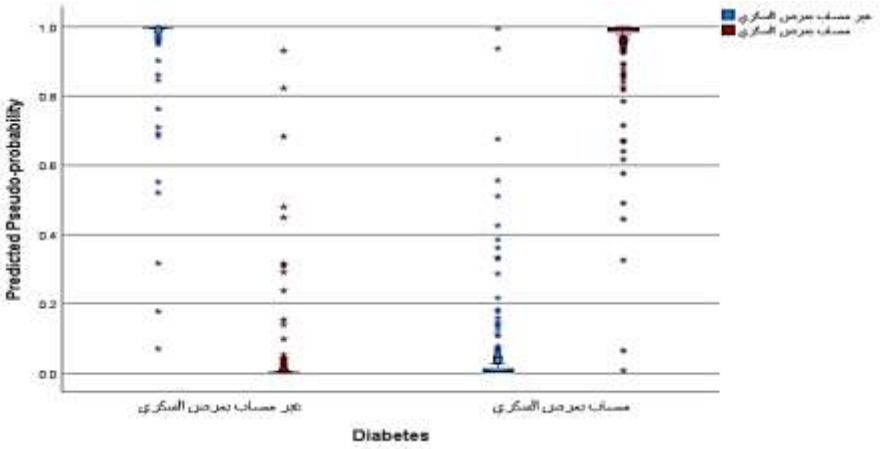
جدول (15): يعكس جدول التصنيف الأداء المتميز لنموذج الشبكة العصبية في تشخيص مرض السكري، محققًا دقة كلية بلغت ٩٨.٢% في بيانات التدريب و٩٧.٢% في بيانات الاختبار، مع فارق طفيف (١%) يدل على أن النموذج قد تعلم التعميم بشكل فعال دون أن يقع في مشكلة فرط التعلم. في بيانات التدريب، نجحت الشبكة في تصنيف ١٣٢ حالة من أصل ١٣٤ حالة غير مصابة (٩٨.٥%)، مع خطأين فقط. وفي حالات المصابين، صنفت بشكل صحيح ١٤٠ حالة من أصل ١٤٣ حالة (٩٧.٩%)، مما يُظهر توازنًا عاليًا في الأداء بين الفئتين. وقد حافظ على هذا التوازن في مجموعة الاختبار، إذ بلغت دقة التصنيف ٩٧.٩% لغير المصابين و٩٦.٧% للمصابين. تؤكد هذه النتائج قوة النموذج وموثوقيته كأداة مساعدة في التشخيص الطبي.

جدول (15): التصنيف حسب نموذج الشبكات العصبية.

Classification	Observed		Predicted		
			غير مصاب بمرض السكري	مصاب بمرض السكري	Percent Correct
	Training	غير مصاب بمرض السكري	132	2	98.5%
		مصاب بمرض السكري	3	140	97.9%
		Overall Percent	48.7%	51.3%	98.2%
	Testing	غير مصاب بمرض السكري	46	1	97.9%
		مصاب بمرض السكري	2	59	96.7%
		Overall Percent	44.4%	55.6%	97.2%
Accuracy	Sensitivity	Specificity	Precision	Error_Rate	FPR
0.972	0.967	0.979	0.983	0.028	0.021
Area under curve = .998; Kappa = 0.944; F1_Score= 0.975					

يعرض شكل (2) المخطط الصندوقي للاحتمالات المتوقعة (Predicted Pseudo-probability) الذي يُظهر توزيع الاحتمالات التي يخصصها النموذج لكل فئة. يُلاحظ وجود فصل واضح بين الفئتين، حيث تتجمع الحالات غير المصابة حول

احتمالات قريبة من 0، بينما تتجمع الحالات المصابة حول احتمالات قريبة من 1. مما يدل على كفاءة النموذج في التمييز. ويُلاحظ في المخطط الصندوقي تداخل طفيف جداً بين الفئتين، والذي يمثل الحالات الغامضة التي قد تكون في المنطقة الحدية بين التصنيفين. هذا التداخل المحدود يتوافق مع معدل الخطأ المنخفض الذي حققه النموذج (٢.٨%).



شكل (2): تقدير احتمالات التنبؤ للشبكات العصبية.

جدول (16): يُعد تحليل أهمية المتغيرات أداة أساسية لفهم كيفية عمل نماذج الشبكات العصبية، حيث يُسهم في تحديد العوامل الأكثر تأثيرًا في عملية التصنيف. ويُظهر الجدول أن اختبار HbA1c جاء في المرتبة الأولى من حيث الأهمية بنسبة ١٠٠%، مما يدل علي إدراك النموذج للقيمة التشخيصية العالية لهذا المؤشر. يليه مؤشر كتلة الجسم (٤٨%) والعمر (٣٧.٩%)، مما يؤكد دور السمنة والتقدم في العمر كعوامل خطورة رئيسية. كما برز كل من الكوليسترول الكلي (٣٠.٦%) والدهون الثلاثية (٢٦.٧%) كمؤشرات أيضية مهمة، بما يتوافق مع الأدبيات الطبية. في المقابل، أظهرت المتغيرات مثل التاريخ العائلي، النظام الغذائي، كثرة التبول، التدخين تأثيرًا متوسطًا، بينما كان تأثير الجنس والحالة الوظيفية أقل نسبيًا، ما يشير إلى دورها الثانوي في نموذج التنبؤ.

جدول (16): أهمية المتغيرات المستقلة في تأثيرها علي المتغير التابع.

Variable	Importance	Normalized Importance
Gender	0.031	12.9%
Age	0.090	37.9%
Employment_Status	0.021	8.7%
Smoking	0.043	18.1%
Family_history	0.055	23.3%
Physical_activity	0.039	16.3%
Diet	0.051	21.2%
Polyuria	0.047	19.8%
Polydipsia	0.041	17.2%
Visual_blurring	0.023	9.8%
Systolic_BP	0.034	14.4%
Diastolic_BP	0.036	15.1%
BMI	0.114	48.0%
Total_Cholesterol	0.073	30.6%
Triglycerides	0.064	26.7%
HbA1c_Test	0.238	100.0%

٧/٣ المقارنة بين النماذج محل الدراسة:

يستعرض هذا الجزء من البحث مقارنة شاملة بين هذه النماذج استناداً إلى مجموعة متنوعة من مقاييس التقييم، بهدف تحديد النموذج الأكثر دقة وفعالية في التنبؤ بالإصابة بمرض السكري.

١/٧/٣ المقارنة في حالة النموذج الأصلي:

تُظهر نتائج جدول (17) تفوقاً واضحاً لنموذج الشبكات العصبية الاصطناعية على النموذجين الآخرين في معظم مؤشرات الأداء. إذ حقق أعلى دقة كلية بلغت ٩٧.٢%، مقابل ٩٥.٨% للانحدار اللوجستي و٩٥.٣% للتحليل التمييزي. كما سجّل أعلى قيم للحساسية (٩٦.٧%)، والخصوصية (٩٧.٩%)، والدقة التنبؤية (٩٨.٣%)، مما يعكس فعاليته في الكشف عن الحالات الإيجابية والسلبية وتقليل الفحوصات غير الضرورية. وبلغ معدل الخطأ لديه ٢.٨%، وجذر متوسط مربع الخطأ (RMSE) ٠.١٣٣، وكِلاهما الأدنى بين النماذج. كذلك أظهر توازناً قويًا بين الحساسية والدقة من خلال F1 Score بقيمة ٠.٩٧٥، وأعلى معامل كابا (٠.٩٤٤)، مما يدل على قوة الاتفاق مع التصنيف الفعلي بعد استبعاد العشوائية. كما حقق أعلى قيمة للمساحة تحت منحنى ROC (AUC) بلغت ٠.٩٩٨، مما يشير إلى قدرة تمييزية عالية.

جدول (17) مؤشرات تقييم النماذج في حالة النموذج الأصلي				
Model	**Criteria**	**Discriminant Analysis**	**Logistic Regression**	**Neural Networks**
Original Model	**Accuracy**	**0.953**	**0.958**	**0.972**
	Sensitivity	**0.951**	**0.956**	**0.967**
	Specificity	**0.956**	**0.961**	**0.979**
	Precision	**0.96**	**0.965**	**0.983**
	Error_Rate	**0.047**	**0.042**	**0.028**
	FPR	**0.044**	**0.039**	**0.021**
	F1_Score	**0.956**	**0.961**	**0.975**
	Kappa	**0.906**	**0.917**	**0.944**
	AUC	**0.991**	**0.992**	**0.998**
	RMSE	**0.19**	**0.181**	**0.133**

٢/٧/٣ المقارنة بين النماذج في حالة التحقق المتقاطع (5-fold Cross Validation)

تشير نتائج التحقق المتقاطع (Cross Validation) في جدول (18) إلى استقرار أداء النماذج الثلاثة، مع فروق طفيفة مقارنة بالقيم الأصلية. ورغم هذا الاستقرار، واصل نموذج الشبكات العصبية تحقيق أفضل أداء، مسجلًا أعلى دقة كلية (٩٧.٤%)، ودقة تنبؤية بلغت ٩٧.٥%، وأقل معدل خطأ (٢.٦%). كما حافظ على أعلى قيمة لمؤشر F1 Score (٠.٩٧٥)، وأدنى RMSE (٠.١٦١)، وأعلى معامل كابا (٠.٩٤٨). إضافةً إلى ذلك، حقق أعلى قيمة للمساحة تحت المنحني AUC (٠.٩٩٧)، مما يؤكد قدرته على التمييز ويعزز الثقة في قدرته على التعميم عند التعامل مع بيانات جديدة لم يسبق له التدرب عليها.

جدول (18) مؤشرات تقييم النماذج في حالة التحقق المتقاطع				
Model	**Criteria**	**Discriminant Analysis**	**Logistic Regression**	**Neural Networks**
Cross validation	**Accuracy**	**0.951**	**0.958**	**0.974**
	Sensitivity	**0.946**	**0.961**	**0.975**
	Specificity	**0.956**	**0.956**	**0.972**
	Precision	**0.96**	**0.951**	**0.975**

	Error_Rate	0.049	0.042	0.026
	FPR	0.044	0.044	0.028
	F1_Score	0.953	0.956	0.975
	Kappa	0.901	0.917	0.948
	AUC	0.987	0.992	0.997
	RMSE	0.198	0.181	0.161

القسم الرابع: نتائج وتوصيات البحث

١/٤ نتائج البحث:

١- أظهرت نتائج الدراسة تفوقًا ملحوظًا لنموذج الشبكات العصبية الاصطناعية مقارنةً بالانحدار اللوجستي والتحليل التمييزي في معظم مؤشرات الأداء. إذ حقق أعلى دقة كلية (٩٧.٢%) مقابل ٩٥.٨% و٩٥.٣% على التوالي، كما سجل أعلى القيم في الحساسية (٩٦.٧%)، الخصوصية (٩٧.٩%)، الدقة التنبؤية (٩٨.٣%)، ومقياس F1 (٩٧.٥%)، إلى جانب أقل معدل خطأ (٢.٨%) وأدنى قيمة RMSE (١٣.٣%).

٢- رغم تفوق نموذج الشبكات العصبية، أظهرت النماذج الثلاثة أداءً قويًا في تصنيف حالات السكري، إذ حققت جميعها معدلات مرتفعة للدقة، الحساسية، والخصوصية، إلى جانب انخفاض في مؤشرات الخطأ. كما سجلت قيمًا مرتفعة جدًا للمساحة تحت منحنى ROC (AUC)، مما يعكس قدرة تمييزية ممتازة في التفريق بين الحالات المصابة وغير المصابة.

٣- كشفت الدراسة عن اتفاق جوهري بين النماذج الثلاثة في تحديد المتغيرات الأكثر أهمية في التنبؤ بالإصابة بمرض السكري، مع وجود بعض الاختلافات في الترتيب والأوزان. فقد ظهر اختبار HbA1c كأهم متغير في النماذج الثلاثة، مما يؤكد أهميته القصوى في تشخيص مرض السكري. كما برزت أهمية مؤشر كتلة الجسم والعمر والتاريخ العائلي للمرض كعوامل رئيسية مؤثرة في الإصابة بالمرض.

٤- أظهرت نتائج التحقق المتقاطع استقراراً ملحوظاً في أداء النماذج الثلاثة، مع وجود اختلافات طفيفة بين قيم النموذج الأصلي والتحقق المتقاطع. يعزز هذا الاستقرار الثقة في قدرة النماذج على التعميم على بيانات جديدة لم يسبق للنموذج التدرب عليها، وهو أمر بالغ الأهمية في التطبيقات السريرية.

٥- تتميز نماذج الانحدار اللوجستي والتحليل التمييزي بقابلية التفسير، حيث يمكن فهم تأثير كل متغير مستقل على المتغير التابع من خلال معاملات النموذج. في المقابل، تعاني الشبكات العصبية من صعوبة التفسير نظراً لتعقيد بنيتها الداخلية، رغم أن تحليل أهمية المتغيرات يوفر بعض التفسير لدور كل متغير.
٦- تشير النتائج أيضاً إلى أن العوامل الصحية والسلوكية والديموغرافية تلعب دوراً مهماً في التنبؤ بالإصابة بمرض السكري. فبالإضافة إلى المؤشرات المخبرية كاختبار HbA1c، تؤثر عوامل مثل مؤشر كتلة الجسم (السمنة) والعمر والتاريخ العائلي والتدخين والنظام الغذائي والنشاط البدني بشكل كبير على احتمالية الإصابة بالمرض.
٧- تؤكد النتائج العالية التي حققتها النماذج الثلاثة على إمكانية استخدام هذه النماذج كأدوات مساعدة في التشخيص المبكر لمرض السكري والتنبؤ باحتمالية الإصابة به، مما قد يسهم في تحسين استراتيجيات الوقاية والتدخل المبكر.

٢/٤ توصيات البحث:

١- نوصي باعتماد نموذج الشبكات العصبية الاصطناعية كأداة رئيسية في التنبؤ بحالات الإصابة بمرض السكري وتصنيفها، نظرًا لتفوقه في معظم مؤشرات الأداء. ويمكن توظيفه ضمن برامج الفحص المبكر لاستهداف الأفراد الأكثر عرضة للإصابة، بما يسهم في تعزيز جهود الوقاية والتدخل المبكر.
٢- نقترح تطوير تطبيق برمجي أو منصة إلكترونية تعتمد على نموذج الشبكات العصبية المستخدم، تتيح لمقدمي الرعاية الصحية إدخال بيانات المرضى والحصول على تنبؤات فورية لاحتمالية الإصابة بمرض السكري. يمكن أن تشمل هذه المنصة واجهة سهلة الاستخدام وإرشادات حول كيفية تفسير النتائج واتخاذ الإجراءات المناسبة.
٣- نوصي باستخدام نموذج الانحدار اللوجستي كأداة مساعدة في التشخيص، خاصة في الحالات التي تتطلب تفسيراً واضحاً للنتائج.

٤- نقترح تطوير نظام تصويت مجمع (Ensemble Voting System) يجمع بين النماذج الثلاثة (الشبكات العصبية، الانحدار اللوجستي، والتحليل التمييزي) للوصول إلى تنبؤ أكثر دقة وموثوقية.

٥- نقترح تطوير برامج توعية صحية تستهدف تعديل العوامل القابلة للتغيير التي ظهرت كمؤثرات مهمة في الإصابة بمرض السكري، مثل مؤشر كتلة الجسم (الوزن) والتدخين والنظام الغذائي والنشاط البدني. يمكن أن تستند هذه البرامج إلى الأوزان النسبية لهذه العوامل كما ظهرت في تحليل أهمية المتغيرات.

٦- نوصي بتوسيع نطاق الدراسة ليشمل عينات أكبر وأكثر تنوعاً من مختلف المناطق الجغرافية والفئات العمرية والمجموعات السكانية، للتحقق من عمومية النتائج وقابليتها للتطبيق في سياقات مختلفة. يمكن أن يساعد ذلك في تطوير نماذج أكثر دقة وقدرة على التعميم.

٧- نوصي بإجراء دراسات مقارنة إضافية تشمل تقنيات تعلم آلي أخرى مثل أشجار القرار والغابات العشوائية وآلات المتجهات الداعمة (SVM) وخوارزميات التعزيز التدريجي (Gradient Boosting) للتعرف على المنهجية الأكثر فعالية للتنبؤ بمرض السكري.

قائمة المراجع:

أولًا: المراجع العربية.

١. أبو دومة، حيدر جميل الله محمد. (٢٠١٩). استخدام أسلوبي تحليل الانحدار اللوجستي و التحليل التمييزي للعوامل المؤثرة على الإصابة بأمراض القلب. (رسالة دكتوراه غير منشورة). كلية الدراسات العليا، جامعة السودان للعلوم والتكنولوجيا.

٢. الإمام، سوزان الإمام محمد.، أبو ريا، محمد محمود نصر.، محمد، محمد إبراهيم. (٢٠٢٥). نموذج مقترح لتحسين منهجية بوكس - جينكنز اعتماداً على أسلوب الشبكات العصبية (دراسة تطبيقية). المجلة العلمية للدراسات والبحوث المالية والتجارية، ٦(١)، ٣١١-٣٣٨.

٣. بسيوني، عبد الرحيم عوض عبد الخالق. (٢٠٢١). استخدام التحليل التمييزي في التصنيف والتنبؤ (دراسة تطبيقية). التجارة والتمويل، ٤١(٣)، ٢٩٨ - ٣٢٥.

٤. بسيوني، عبد الرحيم عوض عبد الخالق. (٢٠٢٢). تحليل زمن البقاء باستخدام أسلوب الشبكات العصبية الاصطناعية ونموذجي انحدار كوكس والانحدار اللوجستي (دراسة تطبيقية). مجلة التجارة والتمويل، ٤٢(٤)، ٨٥٦-٨٩٦.
٥. الدمرداش، هاني محمد علي.، وبسيوني، عبد الرحيم عوض عبد الخالق. (٢٠٢٢). التحليل التمييزي في مُقابل تحليل التباين المتعدد: دراسة تطبيقية للتعرف على العوامل المحددة للركود التضخمي في مصر في الفترة ١٩٨٠- ٢٠٢١. مجلة البحوث المالية والتجارية، ٢٣(٤)، ٣٧٤-٤٢٢.
٦. الرواشدة، سكينة محمود سليمان. (٢٠٢٢). استخدام التحليل التمييزي وتحليل الانحدار اللوجستي للكشف عن العوامل التي تساهم في تصنيف الطلبة الموهوبين والمتفوقين: دراسة مقارنة (رسالة دكتوراه غير منشورة). جامعة اليرموك، إربد.
٧. محمد، ولاء محمد العربي.، مصطفى، مصطفى جلال.، وموافى، ممدوح عبد العليم سعد. (٢٠٢٠). دمج تحليل التمايز وتقنيات اختزال البيانات لتحسين دقة التصنيف: دراسة عملية على مرضى القلب. المجلة العلمية للاقتصاد والتجارة، (٤)، ٢٨١ - ٣٠٨.
٨. المهدي، عادل محمد أحمد.، صقر، عمر محمد عثمان.، والشافعي، أحمد صلاح. (٢٠٢٣). محددات معدل الصرف الأجنبي الحقيقي الفعال في الاقتصاد المصري باستخدام أسلوب الشبكات العصبية. المجلة العلمية للبحوث والدراسات التجارية، ٣٧(١)، ٩٧٩ - ١٠١٧.

ثانياً: المراجع الأجنبية.

1. Acito, F. (2023). Logistic Regression. In: Predictive analytics with KNIME. Analytics for citizen data scientists. Switzerland: Springer Nature, pp. 125-167.
2. Ahmed, K. R. A., & Abd Elrazek, A. H. (2023). Discriminant analysis of social capital in one of the graduate's villages. Journal of the Advances in Agricultural Researches, 28(2), 296-341.
3. Ajeel, S. M., Haji, J. A., & Jahwar, B. H. (2023). Using Multinomial Logistic Regression to Identify Factors Affecting Platelet. Journal of Duhok University, 26(2), 47-56.
4. Alajramy, L., & Jarrar, R. (2022). Using Artificial Neural Networks to Identify COVID-19 Misinformation. In: Multidisciplinary International Symposium on Disinformation in Open Online Media (pp. 16-26). Cham: Springer International Publishing.

5. Al-Bairmani, Z. A. A., & Ismael, A. A. (2021, March). Using Logistic regression model to study the most important factors which affects diabetes for the elderly in the City of Hilla/2019. In Journal of Physics: Conference Series (Vol. 1818, No. 1, p. 012016). IOP Publishing.
6. Beacom, E. (2023). Considerations for running and interpreting a binary logistic regression analysis–a research note. DBS Business Review, 5.
7. da Silva, I. N., Spatti, D. H., Flauzino, R. A., Liboni, L. H. B., & dos Reis Alves, S. F. (2017). Artificial neural network architectures and training processes. In: Artificial neural networks: A practical course (pp. 21–28). Springer International Publishing.
8. Darnius, O., & Siahaan, D. (2023). Modelling of risk factors that influence malaria infection using Binary Logistic Regression. In Journal of Physics: Conference Series (Vol. 2421, No. 1, p. 012001). IOP Publishing.
9. Enad, F. H., & Alrawi, Z. N. M. (2022). The use of logistic regression method in data classification with practical application of Covid-19 patients in Nasiriya General Hospital. University of Thi-Qar Journal, 17(2), 35-55.
10. Farouk, M., & Bassiouni, A. R. A. (2024). Classification of lung cancer data using support vector machine and discriminant analysis. Journal of Financial and Commercial Research, 25(2), 52–75.
11. George, D. S., Enegesele, D., Biu, O. E., & Dagogo, J. (2020). Discriminant Analysis of Unemployment and Literacy Rate by State in Nigeria. International Journal of Transformation in Applied Mathematics & Statistics. 3(1). 44-60.
12. Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh & M. Titterington (Eds.), Proceedings of the 13th International Conference on

Artificial Intelligence and Statistics (AISTATS 2010) (Vol. 9, pp. 249–256). JMLR: Workshop and Conference Proceedings.

13. Hahs-Vaughn, D. L. (2024). Discriminant analysis. In Applied multivariate statistical concepts (2nd ed., pp. 359-430). Routledge.
14. Hosmer, D.W. and Lemeshow, S. (2000). Assessing the Fit of the Model. In Applied Logistic Regression (2nd ed.). John Wiley & Sons, Inc., New York, pp 143–202.
15. Johnson, R. A., & Wichern, D. W. (2007). Discrimination and classification. In Applied multivariate statistical analysis (6th ed., pp. 575-670). Pearson.
16. Kamel, M. M., Salem, H. A., & Abdelgawad, W. (2022). An Application of Linear Programming Discriminated Analysis for Classification. The Scientific Journal of Commerce and Finance, 42(2), 89-105.
17. Kim, H. Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. Restorative dentistry & endodontics, 38(1), 52.
18. Liberda, E. N., Zuk, A. M., Martin, I. D., & Tsuji, L. J. (2020). Fisher's linear discriminant function analysis and its potential utility as a tool for the assessment of health-and-wellness programs in indigenous communities. International Journal of Environmental Research and Public Health, 17(21), 7894.
19. Lu, W., & Wang, Y. (2024). Logistic regression. In Textbook of Medical Statistics: For Medical Students (pp. 181-189). Singapore: Springer Nature Singapore.
20. Mallik, B. B., Ghosh, S., & Panem, C. (2024). Analyzing Artificial Neural Network Design Through Mathematical Principles. In: Proceedings of the Fifth International Conference on Emerging Trends in

Mathematical Sciences & Computing (IEMSC-24) (pp. 70–87). Cham: Springer Nature Switzerland.

21. Meloun, M., & Militký, J. (2011). Statistical analysis of multivariate data. In: Statistical Data Analysis. Woodhead Publishing India Pvt Ltd, pp 151-403.
22. Negnevitsky, M. (2011). Artificial neural networks. In Artificial intelligence: A guide to intelligent systems (3rd ed., pp. 165–216). Pearson Education Limited.
23. Okwori, O. A., Agana, M. A., Ofem, O. A., & Ofem, O. I. (2024). Prediction of Patient's Stroke Vulnerability Status Using Logistic Regression Machine Learning Model. Asian Basic and Applied Research Journal, 6(1), 70-82.
24. Salh, S. M., Abdalla, H. T., & Omer, Z. M. (2021). Using Multinomial Logistic Regression model to study factors that affect chest pain. Tikrit Journal of Administrative and Economic Sciences, 17(53, 2), 534–555.
25. Sarker, B., & Chakraborty, S. (2024). Discriminant analysis-based parametric study of an electrical discharge machining process. Scientia Iranica, 31(3), 186-205.
26. Setiawan, H., Firnanda, A., & Khair, U. (2024). Enhancing the Accuracy of Diabetes Prediction Using Feedforward Neural Networks: Strategies for Improved Recall and Generalization. Brilliance: Research of Artificial Intelligence, 4(1), 201-207.
27. Sovey, S., Osman, K., & Mohd-Matore, M. E. (2022). Exploratory and confirmatory factor analysis for disposition levels of computational thinking instrument among secondary school students. European Journal of Educational Research, 11(2), 639-652.
28. Tekić, D., Mutavdžić, B., Milić, D., Novković, N., Zekić, V., & Novaković, T. (2021). Credit risk assessment of agricultural enterprises

in the Republic of Serbia: Logistic regression vs discriminant analysis. Economics of Agriculture, 68(4), 881-894.

29. Tena, F., Garnica, O., Lanchares, J., & Hidalgo, J. I. (2021). Ensemble models of cutting-edge deep neural networks for blood glucose prediction in patients with diabetes. Sensors, 21(21), 7090.
30. Vergara-Lozano, V., Lagos-Ortiz, K., Chavez-Urbina, J., & Rochina García, C. (2023). Logistic Regression Model to Predict the Risk of Contagion of COVID-19 in Patients with Associated Morbidity Using Supervised Machine Learning. In International Conference on Technologies and Innovation (pp. 14-26). Cham: Springer Nature Switzerland.
31. Verma, J. P. (2013). Application of discriminant analysis: For developing a classification model. In Data analysis in management with SPSS software (pp. 389-412). Springer India.
32. Wibowo, F. W & Wihayati. (2021). Prediction modelling of COVID-19 outbreak in Indonesia using a logistic regression model. In Journal of Physics: Conference Series (Vol. 1803, No. 1, p. 012015). IOP Publishing.