



دراسة مقارنة بين الشبكات العصبية ونموذج آلة المتجه بالتطبيق على مرضى السكر بفلسطين

د. أشرف إسماعيل محمود أبو سمرة

باحث إحصائي، غزة- فلسطين

Ashrafsamra80@hotmail.com

المستخلص

يُعد مرض السكر مشكلة صحية مزمنة وعالمية. حيث تهدف هذه الدراسة لاختيار أفضل نموذج إحصائي للعوامل المؤثرة على مرض السكر في قطاع غزة (فلسطين)، من خلال المقارنة بين الشبكات العصبية، ونموذج آلة المتجه على بيانات حقيقية للمراجعين على عيادات الصحة. وفي هذه الدراسة أجريت مقارنة بين النماذج الإحصائية باستخدام أربعة أساليب مختلفة (Cross-validation with half of the observations, Leave-One-Out Cross-validation, Bootstrapping, ROC curve) للوصول إلى أفضل نموذج للبيانات من خلال تقدير الدقة ومعدل الخطأ لكل نموذج. وقد بينت نتائج هذه المقارنات أن الشبكات العصبية الاصطناعية هي الأفضل من حيث الدقة ومعدل الخطأ، حيث بلغت درجة الدقة 93.1% ومعدل الخطأ 6.9%. وهذا يعود إلى أن الشبكات العصبية تقدم أفضل نموذج يقترب من البيانات المتاحة.

الكلمات المفتاحية: مرض السكر، الشبكات العصبية، نموذج آلة المتجه، منحنى ROC، Cross-validation، Bootstrapping، مصفوفة التصنيف.

A Comparative Study Between Neural Networks Analysis And Support Vector Machine Model Applied To Diabetics In Palestine

Dr. Ashraf Ismail Mahmoud Abu Samra

Statistical researcher, Gaza - Palestine

Abstract

Diabetes is a chronic, global health problem. Where this study aims to choose the best statistical model for the factors affecting diabetes in the Gaza Strip (Palestine), by comparing neural networks, and the Support Vector Machine on real data for patients at health clinics. In this study, a comparison was made between statistical models using four different methods (Cross-validation with half of the observations, leave-one-out cross-validation, Bootstrapping, and ROC curve) to reach the best data model by estimating the accuracy and error rate of each model. The results of these comparisons showed that the artificial neural networks are the best in terms of accuracy and error rate, with an accuracy of 93.1% and an error rate of 6.9%. This is because neural networks provide the best model approximation of the available data.

Keywords: Diabetes, Support Vector Machine, neural networks, Cross-validation with half of the observations, leave-one-out cross-validation, Bootstrapping, confusion matrix, ROC curve.

أولاً: مقدمة

يُعاني من مرض السكر ما لا يقل عن 537 مليون نسمة، وتتزايد الاعداد بزيادة العمر وتغير اسلوب المعيشة. ووفقاً لتقرير وزارة الصحة 2021 تبين أن عدد المصابين بمرض السكري في قطاع غزة يزيد على 66,209 مريض بمعدل انتشار 100/3.1 نسمة و 6.2% في الفئة العمرية أكبر من 18 سنة. منهم



من يتلقى العلاج في عيادات الرعاية الأولية لوزارة الصحة، وآخرون يتلقونه في العيادات التابعة لوكالة الغوث (أونروا). ويبلغ عدد الزيارات لعيادات الأمراض المزمنة (السكر) 122,166 مريض سنوياً.
ثانياً: مشكلة الدراسة

يُعتبر مرض السكر من أكثر التحديات الصحية التي تواجه عالمنا في القرن الواحد والعشرين حاملاً معه أمراضاً أكثر خطورة منه كالمضاعفات القلبية والكلية والقدم السكرية وغيرها من الأمراض، وتهتم الدراسة بالنوع الثاني من مرض السكر نظراً لوجود عوامل بشرية كثيرة تدخل بانتشاره وزيادته، فهناك عوامل وراثية وهناك عوامل غذائية وسلوكية خاطئة أضافت لضغوط الحياة المتزايدة. لذلك تأتي الدراسة لتحديد أهم عوامل الخطر للإصابة بمرض السكر من خلال أفضل نموذج إحصائي. وكيفية التحقق من جدوى وفعالية نماذج تسجيل عوامل الخطر لمرضى السكر باستخدام الشبكات العصبية و نموذج آلة المتجه.

ثالثاً: أهمية الدراسة

تأتي أهمية الدراسة كونها تدرس العلاقة بين العوامل المؤثرة على الإصابة بمرض السكر الثاني بفلسطين، وكونها تستخدم أساليب إحصائية متعددة ومختلفة لتصنيف الزائرين لعيادات الصحة بقطاع غزة بفلسطين كذلك تأتي أهمية الدراسة كونها ستقوم بمقارنة النماذج (نموذج آلة المتجه والشبكات العصبية) من حيث كفاءتها وقدرتها على تصنيف المشاهدات. كذلك بناء نموذج إحصائي تنبؤي لأفضل وأكثر العوامل تأثيراً على مرض السكر، مما يؤدي إلى تقليل الإصابة بمرض السكر الثاني، ودراسة ما هي الطريقة الأفضل لمقارنة دقة تقييم البيانات بالطرق (Leave-one-out cross-validation، Bootstrap Method، K-fold cross-validation، ROC curve).

رابعاً: أهداف الدراسة

التحقق من جدوى وفعالية نماذج التصنيف (نموذج آلة المتجه والشبكات العصبية) وإجراء مقارنة بين أداء هذه النماذج في تصنيف المرضى الذين يعانون من مرض السكر في فلسطين. وتوصيف أفضل نموذج إحصائي يساعد على تحديد احتمالية الإصابة بمرض السكر. وتحديد أكثر المتغيرات تأثيراً على مرض السكر النوع الثاني في فلسطين. وذلك من خلال المقارنة بين طرق التقييم (Bootstrap Method، K-fold cross-validation، ROC curve)، للوصول إلى أفضل نموذج للبيانات من خلال تقدير الدقة ومعدل الخطأ لكل نموذج.

خامساً: فروض الدراسة

1. يوجد تأثير جوهري للعوامل التالية (العمر ومؤشر كتلة الجسم والهييموجلوبين السكري والكوليسترول الكلي وتحليل الدهون الثلاثي، النوع والتدخين والتاريخ العائلي للمرض والعمل والمنطقة) القدرة التنبؤية في التمييز والتشخيص لمصاب وغير مصاب بمرض السكر.
2. أن النماذج المقترحة التالية (نموذج آلة المتجه والشبكات العصبية) ملائمة لتوفيق بيانات مرضى السكر من النوع الثاني.
3. يوجد اختلاف بين طرق التقييم المستخدمة في تقدير دقة النموذج لصالح طريقة البوتستراپ Bootstrapping.

سادساً: حدود الدراسة

الحدود المكانية (الجغرافية): عيادات الرعاية الأولية بقطاع غزة بدولة فلسطين.
الحدود الزمانية: تشمل المرضى الزائرين الذين قاموا بمراجعة العيادات الأولية خلال سنة 2021.
سابعاً: مجتمع وعينة الدراسة



يتكون مجتمع الدراسة من الأشخاص المرضى الزائرين لعيادات السكر لمراكز الرعاية الأولية بقطاع غزة والبالغ عدد الزيارات لهؤلاء الأشخاص حوالي 122,166 زيارة، واستخدم الباحث أسلوب العينة الطبقية العشوائية البسيطة بحجم 384 مشاهدة.

ثامناً: منهجية الدراسة

تمهيد:

يُعد فرط الجلوكوز في الدم، الذي يعرف أيضاً بارتفاع مستوى السكر في الدم، من النتائج الشائعة الدالة على خلل في ضبط مستوى السكر في الدم، ويؤدي مع الوقت إلى الإضرار الخطيرة بالعديد من أجهزة الجسم، ولاسيما الأعصاب والأوعية الدموية. ويمكن باتباع نظام غذائي صحي بالتزامن مع نشاط بدني جيد مع الامتناع عن التدخين منع مرض السكري من النوع الثاني أو تأخير الإصابة به. ويمكن تجنب عواقبه أو تأخير ظهوره من خلال الأدوية والفحص المنتظم وعلاج أية مضاعفات.

1. آلة المتجه الداعمة (Support Vector Machine)

لقد أثبتت آلات المتجهات الداعمة (SVM)، التي قدمتها شركة Vapnik في أوائل التسعينيات، أنها أداة حسابية فعالة للغاية في التعلم الآلي. لقد تفوق SVM بالفعل على معظم منهجيات الذكاء الحسابي الأخرى أساساً لأنها تستند إلى مبادئ رياضية سليمة لنظرية التعلم الإحصائي ونظرية التحسين. إنها خوارزمية تعلم خاضعة للإشراف، وتم تطبيقها بنجاح على مجموعة واسعة من المجالات المختلفة. ويقسم أسلوب أو تقنية آلة المتجه الداعم إلى قسمين والقسم الأول يتحدث عن آلة المتجه الداعمة للتصنيف Support vector machine for classification، وهو موضوع هذا البحث، والقسم الثاني هو آلة المتجه الداعم للانحدار Support Vector Machine For Regression.

1.1 آلة المتجه الداعمة للتصنيف

هي إحدى أساليب تعلم الآلة Machine learning قدمها في العام 1992 الباحث Vapnik وهي عبارة عن خوارزمية تعلم عن طريق مشرف أو موجه Supervised، وتستند في عملها إلى نظرية التعلم الإحصائية Statistical Learning Theory.

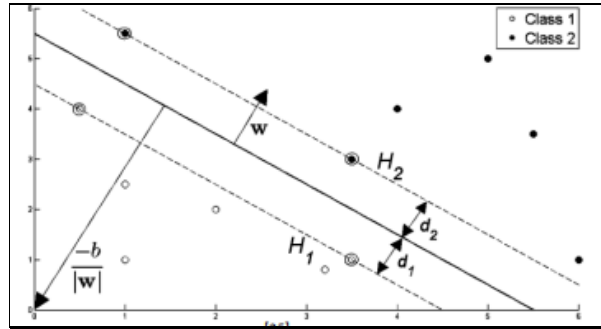
لقد كان اكتشاف تقنية آلة المتجه الداعم في الأصل لحل مسائل تمييز الأنماط Pattern Recognition عن طريق تحديد المستوى الفاصل للبيانات، إذ أن الهدف الأساسي من هذه التقنية هو إيجاد أفضل مستوى فاصل للبيانات المراد فصلها وتصنيفها إلى صنفين. ويمكن إستخدامها في مسائل التصنيف الخطية وغير الخطية إذ يمكنها التصنيف بالاعتماد على مصنف خطي Linear Classifier ومصنف غير خطي Nonlinear Classifier، علماً أن المصنف غير الخطي أتى من بعض مسائل التصنيف التي لا يكون لديها مستوى فاصل بسيط لكي يستخدم كمعيار فاصل، حيث يتم إيجاده عبر استخدام مفهوم الانوية. إذا فرضنا أنه كان لدينا L من النقاط بحيث أن كل قيمة x_i مدخله لها D Dimensionality إذ أن قيم y_i تساوي 1 أو -1 فتكون البيانات المدربة (training data) بالشكل التالي:

$$\{x_i, y_i\} \text{ where } i = 1, \dots, L, y_i \in \{-1, +1\}, x \in R^D \dots (1)$$

وعلى إفتراض ان البيانات مفصولة خطياً فبالإمكان رسم خط x_1 ضد x_2 يقوم بفصل البيانات إلى صنفين أو مجموعتين إذا كانت $D=2$ ويكون المستوى الفاصل على الرسم ل $X_1, \dots, X_D$ في حالة $D > 2$ ، ويمكن التعبير عن المستوى الفاصل (Hyperplane) حسب الصيغة التالية:

$$w'x_i + b = 0$$

حيث أن: $X_1, \dots, X_D$: متغيرات توضيحية تمثل خصائص المشاهدات W : متجه الاوزان ويتم تطبيقها (Normalization) للمستوى الفاصل (Hyperplane)، b : يمثل حد القطع أو التحيز (bias) وتكون النقاط الأقرب إلى المستوى الفاصل (Hyperplane) عبارة عن المتجهات الداعمة (support vector) وتكون قيمتها إما 1+ للمجموعة الاولى و -1 للمجموعة الثانية ويمكن ملاحظتها من خلال الرسم التالي:



شكل (1) البيانات المفصولة بواسطة مصنف خطي

H_1 : يمثل المستوى الثاني الأول والذي تقع عليه متجهات الداعمة للنوع الأول عندما $w'x_i + b = 1$
 H_2 : يمثل المستوى الثاني الثاني والذي تقع عليه المتجهات الداعمة للنوع الثاني عندما $w'x_i + b = -1$

d_1 : أقصر مسافة إلى أقرب نقطة من النوع الأول

d_2 : أقصر مسافة إلى أقرب نقطة من النوع الثاني

$\frac{-b}{||w||}$: تمثل المسافة العمودية من المستوى الفاصل (Hyperplane) إلى نقطة الأصل والمستوى الذي

بين المستويين H_1 و H_2 هو المستوى الفاصل الرئيسي (Hyperplane) حيث أن الهدف من آلة المتجهات الداعمة (Support Vectors Machine (SVM) هي جعل المستوى الفاصل (Hyperplane) أبعد ما يمكن من نقاط الصنفين.

وإن عملية تصنيف المشاهدات تتم على وفق الصيغتين الآتيتين:

$$w'x_i + b \geq +1 \quad \text{for } y_i = +1 \dots \dots \dots (2)$$

$$w'x_i + b \leq -1 \quad \text{for } y_i = -1 \dots \dots \dots (3)$$

فإذا كانت قيمة y أكبر من $+1$ يعني أن المشاهدة تنتمي للمجموعة الأولى وإذا كانت قيمة y أقل من -1 فإن المشاهدات تنتمي للمجموعة الثانية وإذا كانت $+1$ أو -1 فتكون عبارة عن ال (Support Machine Vectors) المتجهات الداعمة وهي النقاط التي تكون أقرب نقاط المجموعتين إلى المستوى الفاصل Hyperplane

علما ان الصيغتين المذكورتين سابقا تم الحصول عليهما من الصيغة الآتية:

$$y_i (w'x_i + b) \geq 1$$

$$y_i (w'x_i + b) - 1 \geq 0 \dots \dots \dots (4)$$

تعرف المسافة بين المتجهات الداعمة والمستوى الفاصل Hyperplane بهامش المتجهات الداعمة (SVM Margin) ولجعل المستوى الفاصل Hyperplane بعيد قدر الامكان عن المتجهات الداعمة (Support Vectors) تحتاج الى تعظيم الهامش والمسافة بين المستويين H_1 و H_2 تكون مساوية

$$\frac{2}{||w||} = \frac{2}{\sqrt{w'w}} \quad \text{ويتم تعظيمها.}$$

2. الشبكة العصبية الاصطناعية Artificial Neural Network

تتكون الخلية العصبية الاصطناعية من عناصر مناظرة للخلية البيولوجية وإن أهمها وحدة عناصر المعالجة التي تحتوي على جزئين هما:

دالة الجمع وهي تحدد طريقة وصيغة إدخال المعلومات إلى الشبكة العصبية والتي تعرف بالمدخلات وغالباً ما تكون عبارة عن تركيبة خطية بدلالة الأوزان وتوصف كما يلي

$$n = b + \sum_{j=1}^R w_j P_j \dots \dots \dots (5)$$

n : ناتج مدخلات التركيبة الخطية



b : الحد المطلق (الثابت) bias

w_j : الأوزان المرتبطة بالمدخلات وهي تقابل المعلومات في نموذج الانحدار

P_j : متغيرات الإدخال

وتعرف أيضاً بدالة التحويل (Transfer Function) لأنها تقوم بتحويل المدخلات عبر تفاعلها مع الأوزان من صيغة رياضية إلى أخرى وهي تكون على عدة أنواع منها الخطية وغير الخطية وتمتاز بتحديد نوع العلاقة ما بين المدخلات والمخرجات عند كل عقدة من عقد الشبكة. أما في عقدة طبقة الإخراج فإنه لا يوجد تحديد لدالة التنشيط المستعملة فقد تكون الدالة خطية أو غير خطية فإذا تضمن التدريب على التنبؤ فإن الدالة تكون خطية أما إذا تضمنت عملية التدريب على تصنيف البيانات فإن دالة الإخراج تكون غير خطية.

1.2 معمارية (هيكلية) الشبكة العصبية الاصطناعية

يقصد بمعمارية الشبكة العصبية الاصطناعية ترتيب العقد في المستويات أو الطبقات وشكل الترابط ضمن المستويات (الطبقات) أو بينها، فهي من أهم خصائص الشبكة العصبية والتي توصف على أساسها الشبكة. كما تصنف الشبكات بحسب عدد مستوياتها (طبقاتها) إلى صنفين رئيسيين:

- 1- شبكات وحيدة المستوى أو الطبقة (Single-level or layer) وهي لا تمتلك مستوى (طبقة) مخفية.
- 2- شبكات متعددة المستويات (الطبقات) (Multi-level or layers) لها مستوى (طبقة) مخفي واحد أو أكثر وهي نوعان أيضاً شبكة أمامية التغذية (Feed Forward Network) وشبكة عكسية التغذية ((Feed Backward Network)). وبشكل عام، فإن معمارية الشبكة العصبية الاصطناعية النموذجية مكونة من ثلاث مستويات أو طبقات هي:

- 1- مستوى (طبقة) الإدخال (Input level (layer)): هو المستوى الأول يحتوي على عدد من العقد تمثل عدد المتغيرات المستقلة (المدخلات).
- 2- المستوى (الطبقة) المخفي (Hidden Level): وهو المستوى الأوسط الذي يقع بين المستوى الأول (الإدخال) والمستوى الأخير (الإخراج).
- 3- مستوى (طبقة) الإخراج (Output Level): وهو المستوى الأخير الذي يمثل إخراجات الشبكة العصبية. ويتكون كل مستوى من المستويات الثلاثة أعلاه من: العقد أو الخلايا (Nodes)، والمستوى (Level) والأوزان (Weights) وهي تشير إلى مدى قوة الارتباط العصبي بين مستويات (طبقات) الشبكة العصبية فكل عقدة (خلية) وزن يربطها مع المستوى السابق، ووزن يربطها مع المستوى اللاحق.

2.2 معالجة المعلومات في الشبكة العصبية (التدريب والتعلم)

إن الشبكات العصبية الاصطناعية نوعان هما الشبكات الثابتة (Fixed N Nets.) وهي التي لا تتغير أوزانها عند التدريب أو التعلم، والشبكات المكيّفة (Adaptive N. Nets) والتي لها القابلية على تغيير أوزانها. ويقصد بمعالجة المعلومات في الشبكات العصبية مرور البيانات في الشبكات العصبية المكيّفة بمرحلتين أساسيتين هما: مرحلة التدريب أو (التعلم) ومرحلة العمل الاسترجاع.

3.2 استخدام الشبكات العصبية في التصنيف:

تستخدم الشبكات العصبية الاصطناعية ANN كأساس للتصنيف، من خلال دالة تنشيط Activation Function خاصة بهذا الغرض، وهناك العديد من دوال التنشيط التي قدمت من قبل الباحثين والتي تختلف باختلاف المخرجات وباختلاف الهدف المراد تحقيقه، وسوف نستخدم دالة الخطوة Step Function وتسمى أيضاً بدالة العتبة threshold function، لأنها تناسب الاستخدامات التصنيفية والتمييزية ولأنها تعطي نتيجتين للمخرج (1,0) كما في الصيغة التالية:

$$f(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1, & \text{if } x \geq 0 \end{cases} \dots \dots \dots (6)$$



حيث تستخدم هذه الدالة في طبقة المخرجات، بينما تستمد الدالة اللوجستية σ في خلايا الطبقة الخفية، والتي تعطى بالصيغة التالية:

$$f(x) = \frac{1}{1 + \theta^{-s}} \dots\dots\dots (7)$$

حيث s تمثل المجموع الموزون للمدخلات مضافا اليه حد التحيز (Bias) الذي نرمز له بالرمز θ ، أي أن مجموع المجموع الموزون يعطى بالصيغة:

$$S = \sum_{i=1}^n w_i x_i + \theta, \dots\dots\dots (8)$$

حيث: w_i تمثل الاوزان، وتحسب قيمة S في وحدة الطبقة الخفية اعتمادا على الدالة اللوجستية، بينما في طبقة المخرج تحسب اعتمادا على دالة الخطوة، وباعتماد على المدخل فاذا كان المخرج 0 هذا يعني ان المشاهدة تتبع لتصنيف المجموعة الأولى، اما إذا كان المخرج 1 فان المشاهدة تتبع لتصنيف المجموعة الثانية.

3. طرق التحقق من صلاحية النماذج الإحصائية

1.3 البوتستراب Bootstrap

تعد هذه الطريقة إحدى طرائق التقدير التي تعتمد على مبدأ المعاينة مع الارجاع، وهي اعادة عينة تشتمل على (n) من العناصر المسحوبة بالارجاع بشكل عشوائي من (N) من البيانات الاصلية، وهي تقنية تعتمد على الحاسوب.

يوصف التوزيع (التوزيع التجريبي) بأنه توزيع عينات البوتستراب التي تم اخذها من خلال أي مجموعة حقيقية من البيانات، أي انه اذا كانت $X = x_1, x_2, x_3, \dots, x_n$ فان احتمال أخذ العينات لاي بيانات منقطعة هو $\frac{1}{n}$. العينة العشوائية بحجم n من مجموعة البيانات الاصلية يشار لها X^* $x_1^*, x_2^*, x_3^*, \dots, x_n^*$ حيث تشير النجمة فوق الرمز على ان X^* ليست البيانات الحقيقية للمجموعة X ، لكنها عبارة عن مجموعة اخذت كعينة من البيانات الاصلية.

إن العينة المأخوذة من البيانات الحقيقية تكون كعملية وصف للتوزيع التجريبي $\hat{F} = (X^* = x_1^*, x_2^*, x_3^*, \dots, x_n^*)$ حيث أن عملية المعاينة تكون مع التكرار، وعمليات التكرار

في البوتستراب لاي إحصاء ممكن حسابه من المعادلة $\theta = S(X^*)$ حيث ان $S(X^*)$ هو مقدر احصائي في عملية التكرار للبوتستراب من البيانات الاصلية. وهذا الإحصاء قد يكون المتوسط او الانحراف المعياري أو أي إحصاء آخر من عينة البوتستراب التي قد تكون تولدت لإيجاد تقديرات للإحصاء حيث $\hat{\theta} = S(X^{*b}), b=1,2,\dots,B$ وتقديرات هذا الإحصاء يمكن استخدامها لبناء توزيع معاينة .

2.3 طريقة التحقق باستخدام Cross-Validation Method

يعد التحقق إجراء عاما يستخدم في إنشاء النماذج الإحصائية. يمكن استخدامها لاتخاذ قرار بشأن ترتيب نموذج إحصائي. وهي طريقة إحصائية لتقويم ومقارنة خوارزميات التعلم عن طريق تقسيم البيانات إلى جزأين: أحدهما يستخدم لتعلم نموذج أو تدريبه والآخر يستخدم للتحقق من صحة النموذج. وعند الانتهاء من التدريب، يمكن استخدام البيانات التي تمت إزالتها لاختبار أداء النموذج الذي تم تعلمه على بيانات "جديدة". هذه هي الفكرة الأساسية لفئة كاملة من طرائق التقويم النموذجية تسمى التحقق من الصحة.

وفي هذه الطريقة، $\binom{k}{n}$ يتم تصميم المصنفين. يتم تصميم كل مصنف عن طريق اختيار k من n ملاحظات كمجموعة تدريب، ويتم تقدير معدل الخطأ الخاص به باستخدام الملاحظات المتبقية $(K-n)$. يتم تكرار هذه العملية لكل الخيارات المميزة للأنماط ويتم حساب متوسط معدلات الخطأ. لذلك فإن متوسط معدل الخطأ لكل مجموعة فرعية هو تقدير لمعدل الخطأ للمصنف.

3.3 طريقة التحقق باستخدام K-fold cross validation



تستخدم عملية التحقق من الصحة عبر الطي k جزءا من البيانات المتوفرة لملاءمة النموذج، وجزءا مختلفا لاختبارها. يتم تقليل تباين التقدير الناتج عند زيادة k. العيب في هذه الطريقة هو أنه يجب إعادة تشغيل خوارزمية التدريب من وقت الصفر، مما يعني أن إجراء التقييم يستغرق عدة مرات. ومن بين أشكال هذه الطريقة تقسيم البيانات عشوائيا إلى مجموعة اختبار وتدريب k مرة مختلفة. ميزة القيام بذلك هي أنه يمكن أن تختار بشكل مستقل حجم كل مجموعة اختبار وعدد المحاولات في المتوسط.

4.3 طريقة التحقق باستخدام Leave-one-out cross-validation

التحقق من الصحة من خلال ترك واحد للخارج (LOOCV) هو حالة خاصة للتحقق من الصحة عبر الطي k. وبعبارة أخرى، في كل تكرار تقريبا تستخدم جميع البيانات باستثناء ملاحظة واحدة للتدريب، ويجري اختبار النموذج المقدر على هذه الملاحظة الواحدة. من المعروف أن دقة النموذج المقدر باستخدام LOOCV غير متحيزة تقريبا، لكنها تختلف بشكل كبير، مما يؤدي إلى تقديرات لا يمكن الاعتماد عليها. ويتضمن إثبات صحة النتيجة من خلال ترك العمل لمرة واحدة (LOOCV) استخدام ملاحظة واحدة من العينة الأصلية كبيانات التحقق من الصحة والملاحظات المتبقية كبيانات التدريب. ويتم تكرار ذلك بحيث يتم استخدام كل ملاحظة في العينة مرة واحدة كبيانات التحقق من الصحة.

5.3 مصفوفة التصنيف (التشويش) Confusion Matrix

تعتبر مصفوفة التصنيف مؤشر احصائي على مدى ملائمة النموذج ومن ثم مطابقته للبيانات، حيث يعمل على تصنيف الظواهر ثنائية الحدث عن طريق استخدام مصفوفة الخلط (التشويش) Confusion matrix، التي تظهر الانتماء الفعلي مقابل الانتماء المتنبأ به لكل مجموعة، اما الشكل العام لجدول التصنيف فهو التالي:

جدول (1): مصفوفة التشويش لفصلين (إيجابي وسلبي)

التنبؤ		Confusion Matrix	
إيجابي	سالب	سالب	المشاهدات
FP	TN	إيجابي	
TP	FN		

الدقة: $Accuracy = \frac{TN+TP}{N}$ حيث يمثل:

TN عدد العينات التي صنفنا سالبة (لا تمتلك الصفة) وهي في الحقيقة سالبة
TP عدد العينات التي صنفنا موجبة (تمتلك الصفة) وهي في الحقيقة موجبة
N عدد العينات الكلي

حساسية النموذج: $Sensitivity = \frac{TP}{TP+FN}$ حيث يمثل

FN عدد العينات التي صنفنا سالبة وهي في الحقيقة موجبة.

خصوصية النموذج: $Specificity = \frac{TN}{TN+FP}$ حيث تمثل

FP عدد العينات التي صنفنا موجبة وهي في الحقيقة سالبة

معدل الخطأ $Error\ rate = \frac{FP+FN}{N}$

معيار: $Precision = \frac{TP}{TP+FP}$

6.3 منحنى ROC

يعتبر أداة مرنة لإنشاء منحنيات أداء ثنائية الأبعاد محددة المعالم، وهي طريقة للكشف عن العلاقة بين الحساسية (Sensitivity) والخصوصية (Specificity). يحتوي المصنف العشوائي على منطقة أسفل المنحنى 0.5، بينما يحتوي المصنف المثالي على 1. لذا يجب أن تكون المصنفات المستخدمة عملياً في مكان ما بينهما، ويفضل أن تكون قريبة من 1، يمكن إنشاء منحنى ROC برسم دالة التوزيع التراكمي



(المساحة الواقعة تحت التوزيع الاحتمالي من $-\infty$ الى ∞ ، وتستخدم المنطقة تحت المنحنى ROC عادة كمقياس لجودة التصنيف الاحتمالي، ويتم احتساب المساحة تحت منحنى ROC باستخدام الصيغة التالية:

$$A_{ROC} = \int_0^1 \frac{TP}{P} d \frac{FP}{N} = \frac{1}{PN} \int_0^N TP * dFP, \dots\dots\dots (9)$$

تاسعاً: الدراسة التطبيقية للنماذج الإحصائية المستخدمة

1. متغيرات الدراسة وتعريفاتها الإجرائية

اعتمدت الدراسة على 10 متغيرات مستقلة، منهم 5 متغيرات تصنيفية وهم (الجنس، التدخين والتاريخ العائلي للمرض، المنطقة والعمل)، و5 مؤشرات رقمية (العمر، مؤشر كتلة الجسم، تحليل الدهون الثلاثية، تحليل الكوليسترول الكلي، الهيموجلوبين السكري).

2. تحليل البيانات:

تم استخدام البرنامج الإحصائي R وبرنامج SPSS V.28 لتصنيف النماذج وتقييم دقتها، وأجرينا محاكاة باستخدام 1000 عينة عشوائية واستخدمنا خوارزميات R المناسبة في ملائمة النماذج وتقييم دقتها. ويتضح من جدول (3) الاحصائيات الوصفية للمتغيرات الكمية وجود فروق في المتوسطات بين المصابين بمرض السكر وبين غير المصابين حيث بلغ معدل الاعمار 50.08 سنة وكان معدل الاعمار لمرضى السكر اعلى منه لغير مرضى السكر حيث بلغ 54.36 سنة، وبلغ معدل مؤشر كتلة الجسم (BMI) 30.4 هو مؤشر يعبر عن وجود زيادة في الوزن وسمنة، ويشير معدل قياس الهيموغلوبين السكري (HBA1C) 6.04% لمقدمات مرض سكر عند الزائرين، ويشير معدل تحليل الدهون الثلاثية (T. G) 154.46 ملغم/ديسيلتر لارتفاع بنسبة الدهون الثلاثية وهو مؤشر لبداية الإصابة بمرض السكر، ويشير معدل تحليل الكوليسترول الكلي (T.CH) 159.05 لبداية ارتفاع في مستوى الكوليسترول السيء في الدم مما يدل على بدايات الاصابة بمرض السكر.

جدول (2) الاحصائيات الوصفية لمتغيرات الدراسة

	غير مريض		مريض سكر		الاجمالي	
	الوسط	الانحراف المعياري	الوسط	الانحراف المعياري	الوسط	الانحراف المعياري
AGE	47.15	9.51	54.36	8.149	50.08	9.64
BMI	29.93	5.22	31.09	5.326	30.40	5.29
HBA1C	5.48	0.72	6.87	0.836	6.04	1.03
T. G	144.23	39.38	169.39	82.569	154.46	61.89
T.CH	146.89	38.89	176.82	45.294	159.05	44.08

يتضح من خلال جدول (3) التوزيع التكراري للمتغيرات الفئوية، حيث 59.4% من افراد عينة الدراسة غير مصابين بالسكر النوع الثاني، 40.6% مصابين بالسكر النوع الثاني، ويتضح أن 53.6% من افراد العينة هم من الذكور، 46.4% من الاناث، بينما 26.3% من الزائرين للعيادات الصحية من المدخنين، 73.7% غير مدخنين، ويتضح أن 52.1% من أفراد العينة هم من سكان المخيمات، 47.9% سكان المدن، ويتضح أن 73.4% من افراد العينة ليس لديهم تاريخ عائلي لمرض السكر، ويتضح 54.7% لا يعملون.

جدول (3) جدول تكراري للمتغيرات التصنيفية الفئوية

	غير مريض سكر		مريض سكر		الاجمالي	%
	العدد	النسبة	العدد	النسبة	العدد	
Sample	228	59.4%	156	40.6%	384	100%



Sex	female	112	62.9%	66	37.1%	178	46.4%
	male	116	56.3%	90	43.7%	206	53.6%
Smoking	yes	31	30.7%	70	69.3%	101	26.3%
	no	197	60.9%	86	39.1%	283	73.7%
Region	Camp	117	60.0%	75	40.0%	200	52.1%
	City	111	58.7%	81	41.3%	184	47.9%
Family history	no	193	68.0%	91	32.0%	282	73.4%
	yes	35	35.0%	65	65.0%	102	26.6%
Work	no	135	64.3%	75	35.7%	210	54.7%
	yes	93	53.4%	81	46.6%	174	45.3%

3. تقدير النموذج باستخدام آلة المتجه الداعمة (SVM)

أن الهدف من آلة المتجه الداعم هو جعل المستوى الفاصل أبعد ما يمكن عن نقاط التصنيفين، وبعد ذلك تحديد دالة النواه التي تستخدم في بناء النموذج، ومن خلال الصيغة التالية يمكننا ان نصنف أي نقطة جديدة يراد اختبارها لأي مجموعة تصنيف.

$$f(x) = \left(\sum_{i=1}^L \alpha_i y_i \cdot x_i + b \right)$$

تم بناء وتدريب نموذج SVM باستخدام 10 متغيرات مستقلة لها تأثير على الإصابة بمرض السكر النوع الثاني، ومن خلال جدول (9) يتضح ان هناك مرحلتين تمت الأولى مرحلة التدريب وأجريت على 269 مشاهدة بنسبة 70%، ومرحلة الاختبار أجريت على باقي المشاهدات 115 بنسبة 30% لاختبار صلاحية (SVM).

جدول (9) ملخص لتدريب واختبار بيانات آلة المتجه الداعمة

		N	Percent
Sample	Training	269	70%
	Testing	115	30%

9- جودة التصنيف باستخدام نموذج (SVM):

من خلال جدول (10) يتضح أن معدل دقة التنبؤ الكلي لنموذج دقة (التصنيف الصحيح) في عينة التدريب ب 87.3%، مما يؤكد قوة ودقة التصنيف باستخدام (SVM). وبلغ معدل الخطأ في التصنيف 12.7%. بينما يبلغ التصنيف الصحيح لعينة التدريب للمصابين 78.9%. بينما في عين الاختبار معدل دقة التنبؤ الكلي لنموذج دقة (التصنيف الصحيح) في عينة الاختبار ب 91.3%. وبلغ معدل الخطأ في التصنيف 8.7%. بينما يبلغ التصنيف الصحيح لعينة الاختبار للمصابين 95.6%، يبلغ التصنيف الصحيح للمصابين بالسكر 85.1%، وبلغت المساحة أسفل منحنى ROC 93.5%.

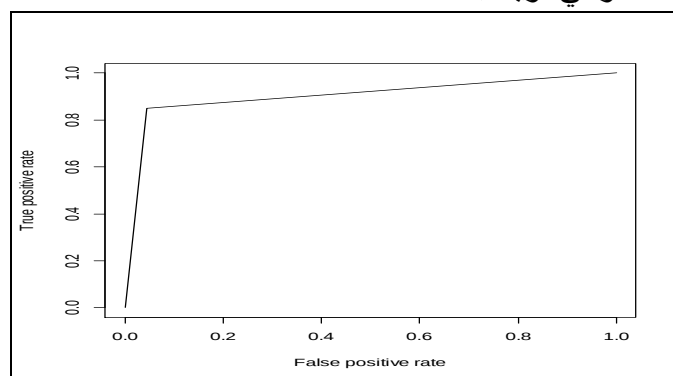
جدول (10) تصنيف نموذج (SVM) للبيانات في حالتي التدريب والاختبار

Sample	Observed	Predicted		
		not diabetes	diabetes	Percent Correct
Trainin g	not diabetes	149	11	93.1%
	diabetes	23	86	78.9%



	Overall Percent	63.9%	36.1%	87.3%	
Testing	not diabetes	65	3	95.6%	
	diabetes	7	40	85.1%	
	Overall Percent	62.6%	37.4%	91.3%	
	Sensitivity	Specificity	Accuracy	Precision	Error rate
SVM (Testing)	85.1%	95.6%	91.3%	93.0%	8.7%
	Area ROC	93.5%			

يوضح الشكل (3) منحنى ROC بين حيث بلغت المساحة الواقعة أسفل منحنى ROC في آلة المتجه الداعمة لحالتنا هي 93.5 % وهي قوية جدا.



شكل (3) منحنى ROC بين آلة المتجه الداعمة

تقدير البيانات باستخدام الشبكات العصبية (ANN)

1.4 توصيف نموذج الشبكات العصبية:

سوف نستخدم دالة الخطوة Step Function وتسمى أيضا بدالة العتبة Threshold function، لأنها تناسب الاستخدامات التصنيفية والتمييزية ولأنها تعطي نتيجتين للمخرج (1,0) كما في الصيغة التالية:

$$f(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1, & \text{if } x \geq 0 \end{cases} \dots\dots\dots 2.3$$

حيث تستخدم هذه الدالة في طبقة المخرجات، بينما تستخدم الدالة اللوجستية sigmoid في خلايا الطبقة الخفية، والتي تعطى بالصيغة التالية:

$$f(x) = \frac{1}{1 + \theta^{-s}} \dots\dots\dots (2.3)$$

حيث s تمثل المجموع الموزون للمدخلات مضافا اليه حد التحيز (Bias) الذي نرمز له بالرمز θ ، أي أن مجموع المجموع الموزون يعطى بالصيغة:

$$S = \sum_{i=1}^n w_i x_i + \theta, \dots\dots\dots (2.4)$$

3. تطبيق وتحليل البيانات باستخدام الشبكات العصبية الاصطناعية:

تم تدريب الشبكة باستخدام 10 متغيرات مستقلة لها تأثير على الإصابة بمرض السكر النوع الثاني، حيث أجريت مرحلة التدريب على 276 مشاهدة بنسبة 71.9 %، ومرحلة الاختبار أجريت على باقي المشاهدات 108 بنسبة 28.1 % لاختبار صلاحية الشبكة من خلال متغير التقسيم. ويتضح أنه في مرحلة



الادخال تم ادخال 10 متغيرات وهي (work, family history age, , Smoking, zone, sex) كما يتضح ان بوحدة الادخال 5 وحدات، اما الطبقة الخفية فيوجد طبقة واحدة، كما يوجد 5 وحدات في الطبقة الخفية، وان الدالة المستخدمة بدالة التنشيط (T.CH ,HBA1C, T.G ,bmi Activation) هي (Function Hyperbolic tangent) ، كما يتضح ان هناك متغير تابع واحد (diabetes) ، وان دالة التنشيط المستخدمة (SoftMax) وتعرف أيضا بالدالة اللوجستية او دالة sigmoid. ويتضح ان معدل التصنيف الخاطئ في عينة التدريب في التحليل 7.8%، بينما نسبة التصنيف الخاطئ في عينة الاختبار 3.9% وهي مقاربة، وهذا يعبر على ان الشبكة تدرت جيدا.

4. جودة التصنيف باستخدام الشبكة العصبية:

من جدول (7) يتضح أن معدل دقة التنبؤ الكلي للنموذج في عينة التدريب بـ 89.9٪، مما يؤكد قوة ودقة التصنيف باستخدام الشبكة العصبية. وبلغ معدل الخطأ في التصنيف 10.1٪. بينما يبلغ التصنيف الصحيح لعينة التدريب لغير المصابين 92.3٪، يبلغ التصنيف الصحيح للمصابين في عينة التدريب للمصابين 85.9٪. بينما في عينة الاختبار معدل دقة التنبؤ الكلي لنموذج دقة (التصنيف الصحيح) في عينة الاختبار بـ 93.1٪. معدل الخطأ في التصنيف 6.9٪. بينما يبلغ التصنيف الصحيح لعينة الاختبار لغير المصابين 98.3٪، ويبلغ التصنيف الصحيح للمصابين 87.7٪، وبلغت المساحة أسفل منحني ROC 98.2٪.

جدول (7) تصنيف الشبكة العصبية للبيانات

	Observed	Predicted			
		not diabetes	diabetes	Percent Correct	
Training	not diabetes	156	13	92.3%	
	diabetes	14	85	85.9%	
	Overall %	61.3%	38.7%	89.9%	
Testing	not diabetes	58	1	98.3%	
	diabetes	7	50	87.7%	
	Overall %	56.0%	44.0%	93.1%	
Testin	Sensitivity	Specificity	Accura cy	Precisio n	Error rate
	87.7%	98.3%	93.1%	98.0%	6.9%
Area ROC		98.2%			

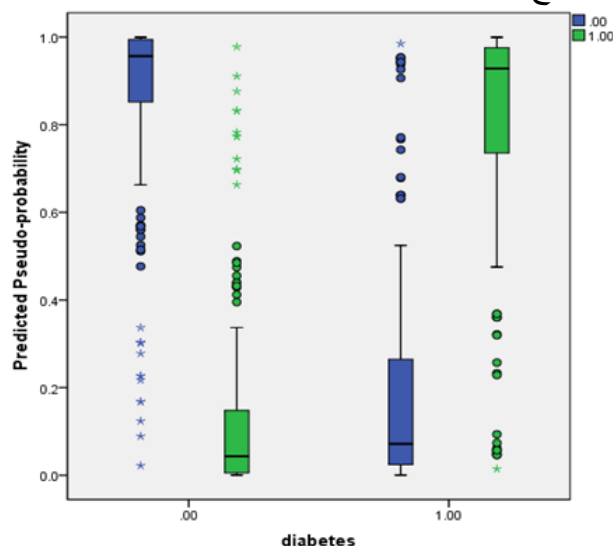
من خلال جدول (8) يتضح أهمية المتغيرات في التصنيف باستخدام الشبكات العصبية، حيث أن العامل HBA1C كان الأكثر تأثيراً في التصنيف بالإصابة بمرض السكر 0.29 يليه عامل العمر بنسبة 0.17، يليه عامل التاريخ العائلي للمرض 14.6٪، يليه عامل تحليل الدهون الثلاثية T.G بنسبة 0.095، يليه عامل كتلة الجسم BMI بنسبة 0.081، يليه عامل التدخين smoking 0.067 ، يليه عامل تحليل الكوليسترول الكلي T.CH 0.087 ، يليه عامل النوع الاجتماعي بنسبة 1.9٪، وأخيراً عامل المنطقة بنسبة 0.013 علماً ان الأهمية النسبية للعوامل تمثل الاوزان للمتغيرات.

جدول (8) المتغيرات المؤثرة حسب الأهمية في التأثير على المتغير التابع



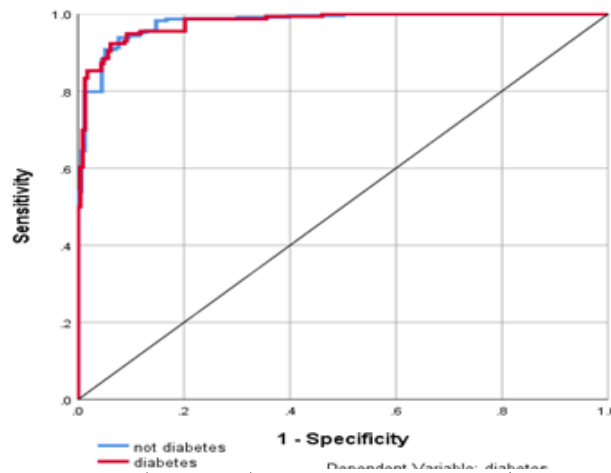
	Importance	Normalized Importance
sex	.019	6.6%
family history	.146	50.2%
age	.172	59.4%
bmi	.081	27.9%
HBA1C	.290	100.0%
T. G	.095	32.7%
T.CH	.087	30.1%
work	.029	9.8%
Region	.013	4.6%
smoking	.067	23.1%

يوضح الشكل (2) تقدير احتمالات تجمعات التنبؤ الصحيح للشبكة العصبية، حيث يمثل المحور الأفقي (استجابة المشاهدات للإصابة بالسكر)، أما المحور الراسي فيمثل احتمال التنبؤ الصحيح، حيث يوجد أربع مستطيلات تمثل كل حالات التنبؤ، حيث المستطيل الأول على اليسار فيمثل المشاهدات التي تحقق عدم الإصابة $P(0/0)$ ، وهو يقع في الجهة العليا بين 0.8، 1 صحيح مما يدل على قدرة النموذج على التنبؤ، ويمثل المستطيل الثاني للمشاهدات التي حققت إصابة وتنبأ بها بعدم الإصابة $P(0/1)$ وهو خطأ من النوع الثاني (β) ، مما يدل على انخفاض احتمال الخطأ في التصنيف، ويقع في الأسفل أقل من 0.2، مما يدل على انخفاض احتمال الخطأ في التصنيف، بينما المستطيل الثالث فيمثل عدم الإصابة وتم التنبؤ بها بشكل صحيح، وتم تصنيفه بالإصابة ويقع ما بين 0.1، 0 $P(1/0)$ وهو يمثل الخطأ النوع الأول (α) ، بينما المستطيل الرابع فيمثل التصنيف بالإصابة وهو يقع ما بين 0.8، 1 $P(1/1)$ ، مما يدل على قدرة النموذج على التنبؤ الصحيح.



شكل (2) تقدير احتمالات التنبؤ الصحيح للشبكة العصبية

يوضح الشكل (3) المساحة تحت منحنى ROC للنموذج بلغت 98.6% وهذا يوضح أن النموذج يساعد على التنبؤ بتصنيف عوامل المتغير التابع (الإصابة بمرض السكر النوع الثاني) أكثر مما تعمله الصدفة.



شكل (3) منحني ROC الشبكات العصبية

عاشراً: مناقشة ومقارنة النتائج:

1. المقارنة في حالة النموذج العام وباستخدام دقة التصنيف للنماذج

يتضح من خلال جدول (9) أن تصنيف البيانات بطريقة الشبكات العصبية الاصطناعية (ANN) كان الأفضل حيث بلغت دقة التصنيف للنموذج الكلي 93.1%، وبلغ معدل الخطأ 6.9%، أظهرت النتائج أن تصنيف البيانات بطريقة نموذج آلة المتجه الداعمة (SVM) حيث بلغ دقة التصنيف الكلي للنموذج 91.3% وبمعدل خطأ 8.7%. وهذه النتيجة تتفق مع كل الدراسات حيث يتمتع نموذج الشبكات العصبية (ANN) بكفاءة تصنيف عالية وبأقل خطأ ممكن.

جدول (9) معايير دقة تصنيف النموذج

	Sensitivity	Specificity	Accuracy	Precision	Error rate
SVM	85.1%	95.6%	91.3%	93%	8.7%
ANN	87.7%	98.3%	93.1%	98.0%	6.9%

2. المقارنة بين الأساليب الإحصائية باستخدام Area ROC

كانت نسبة المساحة المحصورة أسفل منحني ROC لتقنية الشبكات العصبية المساحة الأكبر حيث بلغت 98.2% مما يدل على مدى دقة الشبكات العصبية بتصنيف المشاهدات بدرجة قوية وكبيرة، توضح نسبة منحني ROC للأساليب حيث كانت نسبة المساحة المحصورة أسفل منحني ROC لتقنية نموذج SVM المساحة الأكبر حيث بلغت 91.3%.

3. المقارنة بين الأساليب الإحصائية في حالة تقنية Bootstrap

يتضح من خلال جدول (10) أن تصنيف البيانات بطريقة الشبكات العصبية الاصطناعية (ANN) كان الأفضل حيث بلغت دقة التصنيف للنموذج الكلي 92.2%، وبلغ معدل الخطأ 7.8%، نموذج آلة المتجه الداعمة (SVM) حسب بيانات الدراسة لمرضى السكر حيث بلغ دقة التصنيف الكلي للنموذج 85.2% وبمعدل خطأ 14.8%.

جدول (10) معايير دقة النماذج باستخدام طريقة bootstrap

	Sensitivity	Specificity	Accuracy	Precision	Error rate
SVM	80.2%	88.2%	85.2%	82.6%	14.8%
ANN	86.8%	96.8%	92.2%	95.8%	7.8%

4. المقارنة بين الأساليب الإحصائية في حالة تقنية Two-fold Cross- validation

يتضح من خلال جدول (11) أن تصنيف البيانات في حالة Two-fold Cross- validation حسب طريقة الشبكات العصبية الاصطناعية (ANN) حيث بلغت دقة التصنيف للنموذج الكلي 88.5%، وبلغ



معدل الخطأ 11.5%، نموذج آلة المتجه الداعمة (SVM) بلغ دقة التصنيف الكلي للنموذج 85.3% وبمعدل خطأ 14.7%.

جدول (11) معايير دقة النماذج باستخدام طريقة Two-fold Cross- validation

	Sensitivity	Specificity	Accuracy	Precision	Error rate
SVM	80.5%	88.6%	85.3%	82.7%	14.7%
ANN	87.0%	89.5%	88.5%	84.8%	11.5%

5. المقارنة بين الأساليب الإحصائية في حالة تقنية Leave-one-out cross-validation يتضح من خلال جدول (12) أن تصنيف البيانات بطريقة الشبكات العصبية الاصطناعية (ANN) كان الأفضل حيث بلغت دقة التصنيف للنموذج الكلي 93.0%، وبلغ معدل الخطأ 7.0%، نموذج آلة المتجه الداعمة (SVM) دقة التصنيف الكلي للنموذج 83.0% وبمعدل خطأ 17%.

جدول (12) معايير دقة النماذج باستخدام طريقة Leave-one-out cross-validation

	Sensitivity	Specificity	Accuracy	Precision	Error rate
SVM	80.4%	85.2%	83.0%	82.2%	17%
ANN	87.0%	98.1%	93.0%	97.6%	7.0%

من مناقشة النتائج السابقة يتضح:

ملائمة نموذج نموذج آلة المتجه للبيانات حيث أظهرت النتائج بان قيمة (Cox & Snell R Square) تساوي (0.509) وهي تشير الى ان (50.9%) من التغير الحاصل في المتغير التابع يتم تفسيره من خلال انموذج نموذج آلة المتجه ، وكذلك قيمة ($R^2_{Nagelkerke}$) تساوي (0.687) مما يعني ان 68.7% من أسباب الإصابة بمرض السكر سببها المتغيرات (BMI ، HBA1C،work ،age، Smoking، family history)، ومن خلال اختبار (Hosmer and Lemeshow Test) يتضح جودة توفيق النموذج للبيانات وصلاحيته للتنبؤ.

وأظهرت النتائج أن العوامل (التدخين، التاريخ العائلي، العمر، الهيموغلوبين السكري، العمل) لها تأثير معنوي بالإصابة بمرض السكر النوع الثاني.

كما أظهرت نتائج الشبكة العصبية: أهمية المتغيرات في التصنيف باستخدام الشبكات العصبية، حيث أن العامل HBA1C كان الأكثر تأثيراً في التصنيف بالإصابة بمرض السكر 29%، يليه عامل العمر بنسبة 17.2% ، يليه عامل التاريخ العائلي للمرض 14.6%، يليه عامل تحليل الدهون الثلاثية T.G بنسبة 9.5%، يليه عامل كتلة الجسم BMI بنسبة 8.1%، يليه عامل التدخين smoking 6.7% ، يليه عامل تحليل الكوليسترول الكلي T.CH 8.7% ، يليه عامل النوع الاجتماعي بنسبة 1.9%، وأخيراً عامل المنطقة بنسبة 1.3%، علماً ان الأهمية النسبية للعوامل تمثل الاوزان للمتغيرات.

التوصيات

1. نوصي الباحثين باستخدام نموذج الشبكات العصبية في تقدير العوامل المؤثرة على مرض السكر حيث أظهرت النتائج فعاليتها على نموذج آلة المتجه .
2. نوصي الباحثين والدارسين بإجراء المزيد من الابحاث حول مرض السكر للمساعدة في تحديد العوامل الخطرة والمسببة لمرض السكر واستخدام طرق احصائية حديثة تعتمد على التعلم اخرى غير المستخدمة في هذا البحث وكذلك تضمين متغيرات اخرى لها علاقة بمرض السكر.
3. اجراء فحص مستوى الهيموغلوبين السكري HBA1C في الدم لأنه احد اهم الأسباب التي من خلالها يتم تشخيص مرض السكر، وهو يوضح متوسط كمية الجلوكوز المرتبط بالهيموغلوبين خلال الأشهر الثلاثة الماضية، ويعود السبب في ذلك أن حياة خلايا الدم الحمراء في مجرى الدم عادةً ما تكون ثلاثة أشهر فقط.



4. البطالة أحد الأسباب المؤثرة بزيادة مرض السكر، لذلك نوصي بعدم التكاسل وبدل المزيد من العمل والحركة لما لها مردود إيجابي على الشخص.
5. ننصح مرضى السكر بوضع خطة غذائية مناسبة لما لذلك من آثار ايجابية على الحد من الإصابة بمرض السكر.
6. نوصي بتخفيف الوزن واجراء تمارين رياضية باستمرار، ذلك ان بتقليل كتلة الجسم (الوزن /مربع الطول) نحافظ على أنفسنا من الإصابة بالأمراض عامة ومرض السكر النوع الثاني خاصة.
7. نوصي وزارة الصحة بتوفير قاعدة بيانات الكترونية جيدة لجميع المرضى والمصابين بالأمراض وكذلك المراجعين حيث يمكن استخدامها في مجال البحث، وإعطاء أهمية كبيرة لمرض السكر.
8. تعميم فكرة استخدام الاساليب الاحصائية للتمييز والتصنيف في المجالات الاجتماعية والاقتصادية وعدم تركيزها على المجالات الطبية فقط.

المراجع:

1. أحمد عبد الصمد الجبوري، 2018، استخدام أنموذجي اله المتجه الداعم SVM والانحدار اللوجستي LR في تصنيف البيانات مع تطبيق عملي على مرضى داء السكري في مستشفى الموانئ العام بالبصرة – رسالة دكتوراة، بغداد
2. رولا رضا شريقي، فاعلية برنامج إرشادي لرفع مستوى الرضا عن الحياة لدى مرضى السكري، رسالة دكتوراة -جامعة دمشق -2014
3. زياد عبد الكريم القاضي(2000م)، مقدمة في الذكاء الاصطناعي، الطبعة الأولى، مكتبة المجتمع العربي للنشر والتوزيع، عمان
4. مخللاتي، جلال، 1984، التغذية وصحة الإنسان، الجامعة الإسلامية، غزة.
5. وزارة الصحة، 2021، التقرير السنوي الامراض المزمنة بقطاع غزة -فلسطين.
1. Agresti, A. (2007). "Building and applying logistic regression models", An Introduction to Categorical Data Analysis. Hoboken, New Jersey: Wiley.
2. Algamal , Zakariya Y. , And , Resheed , Khairy B. , " Re - Sampling in Linear Regression Model Using " Jackknife and Bootstrap " , Research Published In The Iraqi Journal Statistical Science , 2010.
3. Ivanciuc, O. (2007). Applications of support vector machines in chemistry. Reviews in computational chemistry, 23,p 291
4. Gokhan Zorluoglu, Mustafa Agaoglu, Diagnosis of Breast Cancer Using Ensemble of Data Mining Classification Methods, Goztepe Campus 34722 / Kadikoy - Istanbul TURKEY
5. Johnson, R.A. and Wichern, D.W. (2007) Applied Multivariate Statistical Analysis., Scientific Research an Academic Publisher Journal, 6th Edition.
6. Jnan Roman Rabunal and Julin Dorrod, (2006) , Artificial Neural Network in real- life applications Idea Group Publishing , USA.
7. K. Saravananathan and T. Velmurugan, Analyzing Diabetic Data using Classification Algorithms in Data Mining, Indian Journal of Science and Technology, Vol 9(43), DOI: 10.17485/ijst/2016/v9i43/93874, November 2016
8. Nettina s(1996): Manual of Nursing practice,6th edition,Lippincott company, New York.



9. Nichols, Jerry L.; Obrenovac, Paul M.; Ingold, Scott et al (1998). Using Logistic Regression to Identify New "At-Risk" Freshmen. Journal of Marketing for Higher Education, Vol a (1) 1998. The Haworth Press, Inc. PP. 25-37.
10. Fine, T.L. 1999. Feedforward Neural Network Methodology, 3rd ed. New York: Springer-Verlag.
 1. <https://www.pcbs.gov.ps/>
 2. <https://www.who.int/ar/campaigns/world-diabetes-day/2021>
 3. <https://www.moh.gov.ps>