دراسة إحصائية لقياس العوامل المسببة للأمراض المزمنة في فلسطين باستخدام (التحليل التمييزي والشبكات العصبية)

أشرف إسماعيل محمود أبو سمرة

باحث إحصائي - فلسطين

Ashrafsamra80@hotmail.com

كريم خلف عزر

جامعة سومر- قسم الاحصاء

Kareemalataby28@gmail.com

الملخص:

تعد الامراض المزمنة (ضغط الدم، السكر) مشكلة صحية كبيرة وعالمية. حيث تهدف هذه الدراسة لاختيار أفضل نموذج إحصائي للعوامل المؤثرة على الامراض المزمنة في (فلسطين)، من خلال المقارنة بين الشبكات العصبية، ونموذج التحليل التمييزي 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 ٩٣.١% ومعدل الخطأ ٦.٩%. وهذا يعود إلى أن الشبكات العصبية تقدم أفضل نموذج يقترب من البيانات المتاحة.

الكلمات المفتاحية: (الامراض المزمنة، الشبكات العصبية، التحليل التمييزي، منحنى ROC، Cross- validation، Bootstrapping، مصفوفة التصنيف).

A statistical study to measure the factors causing chronic disease in Palestine using (Discriminative Analysis and neural networks)

Ashraf Ismail Mahmoud Abu Samra

Statistical researcher – Palestine

Karim Khalaf Motif

Sumer University – Department of Statistics

Abstract:

Diabetes is a chronic, global health problem. Where this study aims to choose the best statistical model for the factors affecting chronic disease in the Gaza Strip (Palestine), by comparing neural networks, and the logistic regression model 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, logistic regression, neural networks, Cross- validation with half of the observations, leave-one-out cross-validation, Bootstrapping, confusion matrix, ROC curve).

أولاً: مقدمة

تودي الأمراض المزمنة بحياة نحو ٤١ مليون شخص سنوياً، بما يعادل ٧١% من إجمالي جميع الوفيات في العالم، حيث يتوفى ١٥ مليون سنوياً بسبب أحد هذه الامراض وتقع ٨٥% من هذه الامراض في البلدان المنخفضة الدخل والمتوسطة الدخل.

وتتزايد الاعداد بزيادة العمر وتغير اسلوب المعيشة. ووفقا لتقرير وزارة الصحة ٢٠٢١ تبين أن عدد المصابين بأحد الامراض المزمنة (السكر وضغط الدم) يزيد على 165,213 مريض بمعدل انتشار ١٠٠/٣.١ نسمة و٦.٢% في الفئة العمرية أكبر من ١٨ 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 212,736 زيارة.

ثانياً: مشكلة الدراسة

تعتبر الأمراض المزمنة من أكثر التحديات الصحية التي تواجه عالمنا في القرن الواحد والعشرين حاملاً معه امراضاً أكثر خطورة منه كالمضاعفات القلبية والكلى والقدم السكرية وارتفاع معدل الدهون وغيرها من الامراض، وتدخل عوامل بشرية كثيرة بانتشاره وزيادته، فهناك عوامل وراثية وهناك عوامل غذائية وسلوكية خاطئة أضافة لضغوط الحياة المتزايدة. لذلك تأتي الدراسة لتحديد أهم عوامل الخطر للإصابة بالأمراض المزمنة من خلال أفضل نموذج إحصائي. وكيفية التحقق من جدوى وفعالية نماذج تسجيل عوامل الخطر للمرضى باستخدام التحليل التمييزي والشبكات العصبية.

ثالثا: أهمية الدراسة

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

رابعاً: أهداف الدراسة

التحقق من جدوى وفعالية نماذج التصنيف (التحليل التمييزي والشبكات العصبية) وإجراء مقارنة بين أداء هذه النماذج في تصنيف المرضى الذين يعانون من الامراض المزمنة. وتوصيف أفضل نموذج إحصائي يساعد على تحديد احتمالية الاصابة بالأمراض المزمنة. وتحديد أكثر المتغيرات تأثيراً على الامراض المزمنة 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

١. يوجد تأثير جوهري للعوامل التالية (العمر ومؤشر كتلة الجسم والهيموجلوبين السكري والكوليسترول الكلي وتحليل الدهون الثلاثي، النوع والتدخين والتاريخ العائلي للمرض والعمل والمنطقة) القدرة التنبؤية في التمييز والتشخيص لمصاب وغير مصاب بالأمراض المزمنة.

٢. أن النماذج المقترحة التالية (التحليل التمييزي والشبكات العصبية) ملائمة لتوفيق بيانات الامراض المزمنة.

٣. يوجد اختلاف بين طرق التقييم المستخدمة في تقدير دقة النموذج لصالح طريقة البوتستراب Bootstrapping.

سادسا: حدود الدراسة

الحدود المكانية (الجغرافية): عيادات الرعاية الأولية بدولة فلسطين.

الحدود الزمانية: تشمل المرضى الزائرين الذين قاموا بمراجعة العيادات الأولية خلال سنة ٢٠٢١.

سابعاً: مجتمع وعينة الدراسة

يتكون مجتمع الدراسة من الأشخاص المرضى الزائرين لعيادات مراكز الرعاية الاولية بقطاع غزة والبالغ عدد الزيارات لهؤلاء الأشخاص حوالي ١٢٢,١٦٦ زيارة، واستخدم الباحث أسلوب العينة العشوائية البسيطة بحجم ٣٨٤ مشاهدة.

ثامناً: منهجية الدراسة

١. التحليل التمييزي: Discriminate Analysis

يعد التحليل التمييزي أحد أساليب التحليل المتعدد، ويتم تحليل المتغيرات الداخلة في النموذج بطريقة مترابطة مع الأخذ في الحسبان العلاقات المتداخلة بين هذه المتغيرات، كما أنه يسعى إلى تكوين نموذج إحصائي يصور العلاقة المتبادلة بين المتغيرات المختلفة، وتعود أهميته بصفة أساسية إلى فاعليته في التمييز بين المشاهدات باستخدامه العديد من المتغيرات، وذلك من خلال إيجاد تركيبات خطية Linear Combination لمجموعة من المتغيرات يطلق عليها متغيرات التمايز.

١.١ النموذج التمييزي في حالة مجموعتين:

إن النموذج التمييزي هو نموذج يمكن صياغته اعتمادا على مؤشرات العينة التي اختيرت مفرداتها بشكل عشوائي ووضعت في مجموعتين مختلفتين، بواسطة هذا النموذج نستطيع أن نختبر المفردة ونحدد انتمائها إلى أي مجموعة، وتكون صيغ النموذج التمييزي للمجموعتين كالتالي:

$$Y_1 = \alpha_1 X_{1i1} + \alpha_2 X_{1i2} + \ldots + \alpha_P X_{1ip} i = 1,2\ldots, n_1 \ldots\ldots\ldots\ldots\ldots (1)$$

$$Y_1 = \alpha_1 X_{2j1} + \alpha_2 X_{2j2} + \ldots + \alpha_P X_{2jp} j = 1,2\ldots, n_2 \ldots . (2)$$

حيث أن n_1, n_2 المشاهدات للمجموعتين. حيث تمثل Y تركيبة من المتغيرات التوضيحية تسمى بالنموذج التمييزي، حيث ان المعاملات α_p ,...., α_2 , α_1, α_0، تقدر بحيث تجعل النموذج التمييزي يعطي أفضل تميز بين المجموعتين، وفي حال وجود مجموعتين يكون لدينا نموذج تمييزي واحد، وفي حالة وجود ثلاث مجموعات يكون لدينا نموذجين تمييزيين. وبعد استخراج المعاملات α، تصنف المشاهدة الى احدى المجموعتين بالاعتماد على نقطة وسط المجموعتين (L) التي تجعل احتمال التصنيف أقل ما يمكن. حيث أن:

$$\boldsymbol{L = \frac{\bar{Y_1} + \bar{Y_2}}{2}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

تصنف المشاهدة الى المجموعة الأولى إذا كانت $\hat{Y} \succ L$

تصنف المشاهدة الى المجموعة الأولى إذا كانت $\hat{Y} \prec L$

تصنف المشاهدة عشوائيا الى المجموعة الأولى أو الثانية إذا كانت $\hat{Y} = L$

٢.١ اختبارات التحليل التمييزي:

بعد تقدير النموذج التمييزي لا بد من اختبار معنويته، وهذه الاختبارات هي:

أولاً: اختبار باستخدام إحصائية Hotelling T^2

ثانياً: اختبار ويلكس لامبدا Wilkes's Lambda Test:

ثالثا: اختبار F test

٣.١ المعاملات التمييزية المعيارية وغير المعيارية

تتمثل المعاملات المعيارية بقيم ($\hat{\alpha}_n$) الظاهرة في المعادلة التالية

$$\hat{Y} = \hat{\alpha}_1 X_1 + \hat{\alpha}_2 X_2 + \hat{\alpha} X_3 \ldots + \hat{\alpha}_n X_n \quad \ldots\ldots\ldots(4)$$

حيث ان:

$\hat{Y}$: القيمة التمييزية المعيارية، X_n: المتغير التمييزي المعياري، $\hat{\alpha}_n$: المعامل التمييزي المعياري ، n: عدد المتغيرات التمييزية المعيارية في المعادلة التمييزية ويساوي (عدد المتغيرات المؤثرة −١) وتستخدم معاملات المعادلة التمييزية المعيارية في تحديد أهمية المتغيرات، وتعني إشارة المعامل التمييزي المعياري ان مساهمة النسبة في التمييز هي مساهمة موجبة او سالبة، ويتم باستخدام المعادلة التمييزية المعيارية تحديد الحد الفاصل بين المعاملات التمييزية المعيارية للمجاميع. وتتمثل المعاملات غير المعيارية بقيم (b) حسب المعادلة التالية:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_2 X_3 \ldots + b_n X_n \quad \ldots\ldots\ldots(5)$$

$\hat{Y}$: القيمة التمييزية غير المعيارية، b_0 : الثابت، X_n: المتغيرات التمييزية غير المعيارية، b_n: المعاملات التمييزية غير المعيارية.

٤.١ ملائمة جودة النموذج التمييزي

أن مقياس جودة التوفيق للنموذج تعني مدى اقتراب القيم المشاهدة من خط التقدير، والملائمة تعني هل أن النموذج الإحصائي ملائم لبيانات عينة الدراسة، وجودة الملائمة تقيس التقارب بين القيم المشاهدة والقيم المتوقعة للنموذج. ومن الاختبارات المستخدمة في جودة الملائمة هي:

٥.١ اختبار مربع كاي χ^2

يرمز له بالرمز χ^2 ومعرف بالصيغة الاتية

$$\chi^2 = \sum_{i=1}^{n} \frac{(X_i - Y_i)^2}{Y_i} \dots\dots\dots\dots\dots\dots\dots\dots (6)$$

حيث أن: X_i التكرار المشاهد، Y_i : تمثل التكرار المتوقع، n: عدد المشاهدات في الجدول وأن توقع الاحصاءة χ^2 يقترب من توزيع χ^2 بدرجة حرية = عدد الخلايا – عدد المعالم في النموذج ويكشف اختبار مربع كاي χ^2 فيما إذا كانت هناك علاقة ذات دلالة بين متغيرين فئويتين.

٦.١ معامل الارتباط القانوني

يقيس معامل الارتباط القانوني جودة التوفيق لنموذج التحليل التمييزي، حيث أن القيمة المرتفعة تكون مؤشرا على جودة توفيق عالية للنموذج التمييزي، ويحسب معامل الارتباط التجميعي بقسمة مربعات التباينات بين المجموعات على الجذر التربيعي لمجموع مربعات البيانات الكلي.

٧.١ معامل التحديد للنموذج التمييزي

يمكن استخدام إحصاءه (R^2) لاختبار قوة النموذج التمييزي، أي نسبة مساهمة العوامل المؤثرة التي يتضمنها النموذج المقدر على المتغير التابع من خلال قيمة الجذر الكامن وهي نسبة التباين المفسر بين المجموعات والتي تعود الى الفروق بينها في النموذج التمييزي، ومعامل الارتباط القانوني الذي يقيس الارتباط بين النموذج التمييزي والعوامل المؤثرة التي تمثل تمييز الإصابة، وبتربيع معامل الارتباط القانوني نحصل على قيمة معامل التحديد.

٢.الشبكة العصبية الإصطناعية 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

$$n = b + \sum_{j=1}^{R} w_j P_j \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (7)$$

n : ناتج مدخلات التركيبة الخطية

b : الحد المطلق (الثابت) bias

w_j : الأوزان المرتبطة بالمدخلات وهي تقابل المعلمات في نموذج الإنحدار

P_j : متغيرات الأدخال

وتعرف أيضاً بدالة التحويل (Transfer Function) لأنها تقوم بتحويل المدخلات عِبر 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.

١.٢ 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 رئيسين:

١- شبكات وحيدة المستوى أو الطبقة (Single-level or layer) وهي لا تمتلك مستوى (طبقة) مخفية.

٢- شبكات متعددة المستويات (الطبقات) (Multi-level or layers) لها مستوى (طبقة) مخفي واحد أو أكثر وهي نوعان أيضا 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:

١- مستوى (طبقة) الإدخال ((Input level (layer)): هو المستوى الأول يحتوي على عدد من العقد تمثل عدد المتغيرات المستقلة (المدخلات).

٢- المستوى (الطبقة) المخفي (Hidden Level): وهو المستوى الأوسط الذي يقع بين المستوى الأول (الإدخال) والمستوى الأخير (الإخراج).

٣- مستوى (طبقة) الإخراج (Output Level): وهو المستوى الأخير الذي يمثل اخراجات 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 فلكل عقدة (خلية) وزن يربطها مع المستوى السابق، ووزن يربطها مع المستوى اللاحق.

٢.٢ 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 المكيفة (Adaptive N. Nets) والتي لها القابلية على تغيير أوزانها. ويقصد بمعالجة المعلومات في الشبكات العصبية مرور البيانات في الشبكات العصبية المكيفة بمرحلتين أساسيتين هما: مرحلة التدريب أو (التعلم) ومرحلة العمل الاسترجاع.

٣.٢ 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 أيضا بدالة العتبة threshold function، لأنها تناسب الاستخدامات التصنيفية والتمييزية ولأنها تعطي نتيجتين للمخرج (١،٠) كما في الصيغة التالية:

$$f(x) = \begin{cases} 0, if\ x < 0 \\ 1, if\ x \geq 0 \end{cases} \dots\dots\dots\dots\dots\dots\dots\dots (8)$$

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

$$f(x) = \frac{1}{1 + \theta^{-s}} \quad \dots\dots\dots\dots\dots\dots\dots\dots (9)$$

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

$$S = \sum_{i=1}^{n} w_i x_i + \theta, \quad \dots\dots\dots\dots\dots\dots (10)$$

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، وبالاعتماد على المدخل فاذا كان المخرج ٠ هذا يعني ان المشاهدة تتبع لتصنيف المجموعة الأولى، اما إذا كان المخرج ١ فان المشاهدة تتبع لتصنيف المجموعة الثانية.

٣. طرق التحقق من صلاحية النماذج الإحصائية

١.٣ البوتستراب Bootstrap

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

يوصف التوزيع (التوزيع التجريبي) بانه توزيع عينات البوتستراب التي تم اخذها من خلال أي مجموعة حقيقية من البيانات، أي انه اذا كانت $X = x_1, x_2, x_3, ..., x_n$ فان احتمال أخذ العينات لاي بيانات متقطعة هو $\frac{1}{n}$. العينة العشوائية بحجم n من مجموعة البيانات الاصلية يشار لها $X^* = x^*_1, x^*_2, x^*_3, ..., x^*_n$ حيث تشير النجمة فوق الرمز على ان X* ليست البيانات الحقيقة للمجموعة X، لكنها عبارة عن مجموعة اخذت كعينة من البيانات الاصلية.

إن العينة المأخوذة من البيانات الحقيقية تكون كعملية وصف للتوزيع التجريبي $\hat{F} = (X^* = x^*_1, x^*_2, x^*_3, ..., x^*_n)$ حيث أن عملية المعاينة تكون مع التكرار، وعمليات التكرار في البوتستراب لاي إحصاء ممكن حسابه من المعادلة $\hat{\theta} = S(X^*)$ حيث ان $S(X^*)$هو مقدر احصائي في عملية التكرار للبوتستراب من البيانات الاصلية. وهذا الإحصاء قد يكون المتوسط او الانحراف المعياري أو أي احصاء آخر من عينة البوتستراب التي قد تكون تولدت لإيجاد تقديرات للإحصاء حيث $\hat{\theta} = S(X^{*b})$, b=1,2,...B وتقديرات هذا الإحصاء يمكن استخدامها لبناء توزيع معاينة .

٢.٣ 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.

٣.٣ 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.

٤.٣ 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.

٥.٣ 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 (١): مصفوفة التشوش لفصلين (إيجابي وسلبي)

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

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

TN عدد العينات التي صنفت سالبة (لا تمتلك الصفة) وهي في الحقيقة سالبة

TPعدد العينات التي صنفت موجبة (تمتلك الصفة) وهي في الحقيقة موجبة

N عدد العينات الكلي

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

FN عدد العينات التي صنفت سالبة وهي في الحقيقة موجبة.

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

FP عدد العينات التي صنفت موجبة وهي في الحقيقة سالبة

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

معيار: $\text{Precision.} = \frac{TP}{TP + FP}$

٦.٣ منحني 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 ٠.٥، بينما يحتوي المصنف المثالي على ١. لذا يجب أن تكون المصنفات المستخدمة عمليًا في مكان ما بينهما، ويفضل أن تكون قريبة من ١، يمكن انشاء 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:

$$A_{ROC} = \int_0^1 \frac{TP}{P} d\,\frac{FP}{N} = \frac{1}{PN}\int_0^N TP * \mathrm{dFP} \ldots (11)$$

تاسعاً: الدراسة التطبيقية للنماذج الإحصائية المستخدمة

١. متغيرات الدراسة وتعريفاتها الإجرائية

اعتمدت الدراسة على ١٠ متغيرات مستقلة، منهم ٥ متغيرات تصنيفية وهم (الجنس، التدخين والتاريخ العائلي للمرض، المنطقة والعمل)، و٥ مؤشرات رقمية (العمر، مؤشر كتلة الجسم، تحليل الدهون الثلاثية، تحليل الكوليسترول الكلي، الهيموجلوبين السكرى).

٢. تحليل البيانات:

تم استخدام البرنامج الإحصائي R وبرنامج SPSS V.28 لتصنيف النماذج وتقييم دقتها، وأجرينا محاكاة باستخدام ١٠٠٠ عينة عشوائية واستخدمنا خوارزميات R المناسبة في ملاءمة النماذج وتقييم دقتها. ويتضح من جدول (٣) الاحصائيات الوصفية للمتغيرات الكمية وجود فروق في المتوسطات بين المصابين غير المصابين حيث بلغ معدل الاعمار ٥٠.٠٨ سنة وكان معدل الاعمار للمرضى اعلى منه لغير المرضى حيث بلغ ٥٤.٣٦ سنة، و بلغ معدل مؤشر كتلة الجسم (BMI) ٣٠.٤ هو مؤشر يعبر عن وجود زيادة في الوزن وسمنة، ويشير معدل قياس الهيموغلوبين السكري(HBA1C)٦.٠٤% لمقدمات الاصابة بأحد الامراض المزمنة عند الزائرين، ويشير معدل تحليل الدهون الثلاثية (T. G) ١٥٤.٤٦ ملغم/ديسيلتر لارتفاع بنسبة الدهون الثلاثية وهو مؤشر لبداية الإصابة بأحد الامراض المزمنة ، ويشير معدل تحليل الكوليسترول الكلي (T.CH)١٥٩.٠٥ لبداية ارتفاع في مستوى الكوليسترول السيء في الدم مما يدلل على بدايات الاصابة بأحد الامراض المزمنة.

جدول (٢) الاحصائيات الوصفية لمتغيرات الدراسة

	غير مريض		مريض		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 (٣) التوزيع التكراري للمتغيرات الفئوية، حيث ٥٩.٤% من افراد عينة الدراسة غير مصابين ، ٤٠.٦% مصابين بأحد الامراض المزمنة ، ويتضح أن ٥٣.٦% من افراد العينة هم من الذكور، ٤٦.٤% من الاناث، بينما ٢٦.٣% من الزائرين للعيادات الصحية من المدخنين، ٧٣.٧%% غير مدخنين، ويتضح أن ٥٢.١% من أفراد العينة هم من سكان المخيمات، ٤٧.٩% سكان المدن، ويتضح أن ٧٣.٤% من افراد العينة ليس لديهم تاريخ عائلي بأحد الامراض المزمنة ، ويتضح ٥٤.٧% لا يعملون.

جدول (٣) جدول تكراري للمتغيرات التصنيفية الفئوية

		غير مريض		مريض		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	58.٧%	81	٤١.٣%	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%

٣- تقدير البيانات باستخدام التحليل التمييزي

من جدول (٤) يتضح وجود فروق بين متوسطات العوامل المؤثرة بين افراد العينة (مصابين /غير مصابين) بأحد الامراض المزمنة، فمن خلال متوسطات عوامل المصابين يتضح انها اعلى من غير المصابين، وهذا طبيعي لان كلما زادت متوسطات Smoking, Region (Family , History, Age, T.G, HBA1C, Work, Bmi , T.CH) زادت احتمالية الإصابة بأحد الامراض المزمنة وكذلك متغيري (Sex, Region)

بلغت قيم معامل الارتباط القانوني فى جدول (٥) حوالى ٠.٧٤٧ مما يدل على ٧٤.٧% من التباين الى الفروق في نموذج التمييز بين المجموعتين (مصابين وغير مصابين) وهذا يعني أن العوامل المؤثرة ساهمت ب ٧٤.٧% من التباين الذي يحصل في تمييز الإصابة بأحد الامراض المزمنة.

٤- اختبار فرضية تساوي المجموعتين:

من خلال (٤) يتضح قيمة (Chi-square=٣٠٨.٧٩٢) وأنها معنوية، بمعنى عدم تساوي متوسطات المجموعتين، أي انه يوجد تمييز بين المجموعتين.

جدول (٤) ويلس لامبدا (Wilks' Lambda) في التحليل التمييزي

Wilks' Lambda	Chi-square	df	Sig.
0.442	308.792	7	0.000

من خلال جدول (٥) يتضح ان قيمة الجذر الكامن (Eigenvalue =1.261) وهي توضح نسبة التباين المفسر بين مجموعتي الإصابة بأحد الامراض المزمنة، قيمة معامل الارتباط القانوني إلى ٠.٧٤٧ ويشير إلى الارتباط بين النموذج التمييزي والعوامل المؤثرة وكان مربع الارتباط القانوني ٠.٥٥٨، بمعنى ان العوامل المؤثرة ساهمت في تفسير الإصابة بنسبة ٥٥.٨% وهذا يدل على جودة توفيق الدالة التمييزية.

جدول (٥) الجذر الكامن (Eigenvalue) في التحليل التمييزي

Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1.261	100.0	100.0	0.747

٥- اختبار معنوية المتغيرات في النموذج التمييزي

من خلال جدول (٦) يتضح أن العوامل المؤثرة(المستقلة) التي دخلت النموذج التمييزي، فمن خلال قيمة ('Wilks Lambda) والمعنوية يتضح أن العوامل (Work، Smoking ,Bmi ,History, Family, Age, HBA1C, T.CH,) كانت معنوية في النموذج وفي التأثير على متغير الإصابة بأحد الامراض المزمنة وبقيت في النموذج حسب التحليل التمييزي التدريجي.

جدول (٦) اختبار F لكل العوامل المؤثرة في النموذج التمييزي

Var.	Wilks' Lambda	F	df1	df2	Sig.
HBA1C	.560	195.415	2	381.000	.000
Age	.494	137.709	3	380.000	.000
Family history	.479	108.840	4	379.000	.000
Work	.465	90.027	5	378.000	.000
BMI	.456	76.898	6	377.000	.000
T.CH	.450	67.737	7	376.000	.000
Smoking	.442	195.415	2	381.000	.000

٦- تقدير النموذج التمييزي:

باستخدام المعاملات التمييزية غير المعيارية تتكون معادلة تحليل التمايز كالتالي :

$$\boldsymbol{DA} = -11.058 + 0.526(\text{Smoking}) + 0.464(\text{Family}) + 0.052(\text{Age}) + 0.038(\text{BMI}) + 1.082(\text{HBA1C}) + 0.005(\text{T.CH}) - 0.626(\text{Work})$$

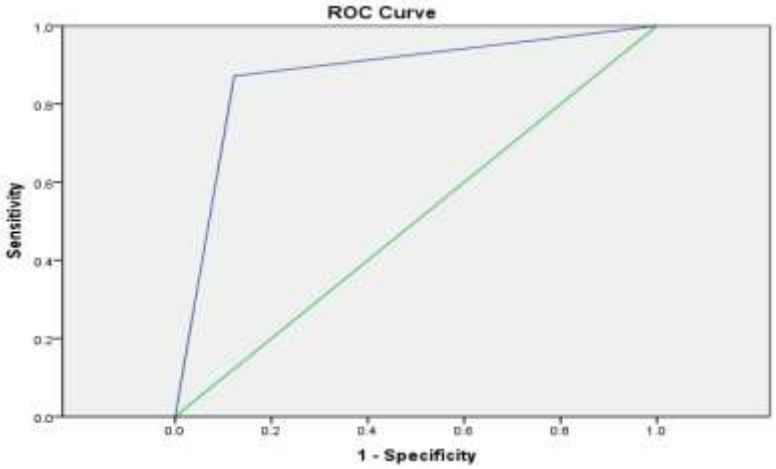
٧- جودة التصنيف بالنموذج التمييزي:

يشير جدول (٨) الى مصفوفة التمييز بعد ادخال المتغيرات المؤثرة لنموذج التحليل التمييزي حيث بلغت دقة النموذج التمييزي ٨٧.٥%، و من جدول (٨) يتضح ملخص نتائج جودة تصنيف النموذج التمييزي، بلغت المساحة تحت

منحنى ROC للنموذج بلغت ٨٧.٤% وهذا يوضح أن النموذج يساعد على التنبؤ بتصنيف عوامل المتغير التابع (الاصابة بأحد الامراض المزمنة) أكثر مما تعمله الصدفة، وبلغت قيمة Precision ٨٢.٩%

جدول (٨) التصنيف لنموذج التحليل التمييزي

Observed			Predicted		
			Diabetes		
			not diabetes	diabetes	Percentage Correct
	Diabetes	not diabetes	200	28	87.7%
		diabetes	20	136	87.2%
	Overall Percentage		237	147	87.5%
DA	Sensitivity	Specificity	Accuracy	Precision	Error rate
	87.2%	87.7%	87.5%	82.9%	12.5%
	Area ROC	87.4%			



شكل (٢) منحنى ROC للنموذج التمييزي

٣. تقدير البيانات باستخدام الشبكات العصبية (ANN)

١.٤ 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 (١،٠) كما في الصيغة التالية:

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

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

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

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

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

٤. تطبيق وتحليل البيانات باستخدام الشبكات العصبية الاصطناعية:

تم تدريب الشبكة باستخدام ١٠ متغيرات مستقلة لها تأثير على الإصابة ، حيث أجريت مرحلة التدريب على ٢٧٦ مشاهدة بنسبة ٧١.٩%، ومرحلة الاختبار أجريت على باقي المشاهدات ١٠٨ بنسبة ٢٨.١% لاختبار صلاحية الشبكة من خلال متغير التقسيم. و يتضح أنه في مرحلة الادخال تم ادخال ١٠ متغيرات وهي (zone،sex ، Smoking، bmi, age, family history ,work، T.G، HBA1C, T.CH) كما يتضح ان بوحدة الادخال ٥ وحدات، اما الطبقة الخفية فيوجد طبقة واحدة، كما يوجد ٥ وحدات في الطبقة الخفية، وان الدالة المستخدمة بدالة التنشيط (Activation Function) هي 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 ٣.٩% وهي متقاربة، وهذا يعبر على ان الشبكة تدربت جيدا.

٥. جودة التصنيف باستخدام الشبكة العصبية:

من جدول (٧) يتضح أن معدل دقة التنبؤ الكلي للنموذج في عينة التدريب بـ ٨٩.٩٪، مما يؤكد قوة ودقة التصنيف باستخدام الشبكة العصبية. وبلغ معدل الخطأ في التصنيف ١٠.١٪. بينما يبلغ التصنيف الصحيح لعينة التدريب لغير المصابين ٩٢.٣%، يبلغ التصنيف الصحيح للمصابين في عينة التدريب للمصابين ٨٥.٩%. بينما في عينة الاختبار معدل دقة التنبؤ الكلي لنموذج دقة (التصنيف الصحيح) في عينة الاختبار بـ ٩٣.١٪. معدل الخطأ في التصنيف

٦.٩٪. بينما يبلغ التصنيف الصحيح لعينة الاختبار لغير المصابين ٩٨.٣%، ويبلغ التصنيف الصحيح للمصابين ٨٧.٧%، وبلغت المساحة أسفل منحنى ROC ٩٨.٢%.

جدول (٧) تصنيف الشبكة العصبية للبيانات

		Predicted			
	Observed	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%	
Testing	Sensitivity	Specificity	Accuracy	Precision	Error rate
	87.7%	98.3%	93.1%	98.0%	6.9%
Area ROC		98.2%			

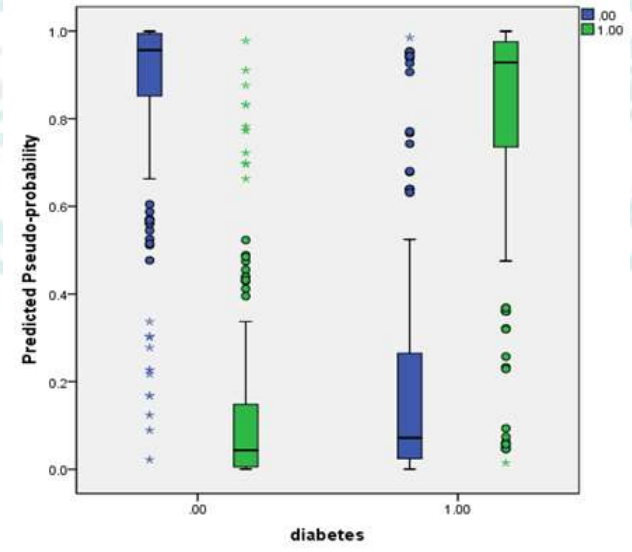
من خلال جدول (٨) يتضح أهمية المتغيرات في التصنيف باستخدام الشبكات العصبية، حيث أن العامل HBA1C كان الأكثر تأثيرا في التصنيف بالإصابة بأحد الامراض المزمنة ٠.٢٩ يليه عامل العمر بنسبة ٠.١٧، يليه عامل التاريخ العائلي للمرض ١٤.٦%، يليه عامل تحليل الدهون الثلاثية T.G بنسبة ٠.٠٩٥، يليه عامل كتلة الجسم BMI بنسبة ٠.٠٨١، يليه عامل التدخين smoking ٠.٠٦٧ ، يليه عامل تحليل الكوليسترول الكلي T.CH ٠.٠٨٧ ، يليه عامل النوع الاجتماعي بنسبة ١.٩%، وأخيرا عامل المنطقة بنسبة ٠.٠١٣ علما ان الأهمية النسبية للعوامل تمثل الاوزان للمتغيرات.

جدول (٨) 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%

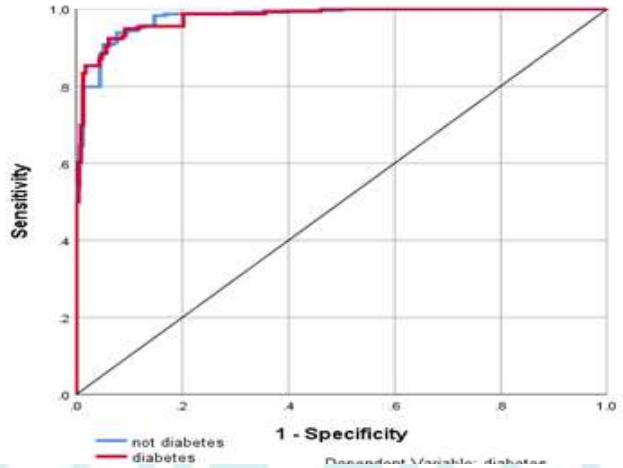
HBA1C	.290	100.0%
T. G	.095	32.7%
T.CH	.087	30.1%
work	.029	9.8%
Region	.013	4.6%
smoking	.067	23.1%

يوضح الشكل (٢) تقدير احتمالات تجمعات التنبؤ الصحيح للشبكة العصبية، حيث يمثل المحور الافقي (استجابة المشاهدات للإصابة بأحد الامراض المزمنة)، أما المحور الراسي فيمثل احتمال التنبؤ الصحيح، حيث يوجد أربع مستطيلات تمتل كل حالات التنبؤ، حيث المستطيل الأول على اليسار فيمثل المشاهدات التي تحقق عدم الاصابة P(٠/٠)، وهو يقع في الجهة العليا بين ٠.٨، ١ صحيح مما يدل على قدرة النموذج على التنبؤ، ويمثل المستطيل الثاني للمشاهدات التي حققت اصابة وتنبأ بها بعدم الاصابة P(٠/١) وهو خطا من النوع الثاني (β)، مما يدل على انخفاض احتمال الخطأ في التصنيف، ويقع في الأسفل اقل من ٠.٢، 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 ٠.١، ٠ P(١/٠) وهو يمثل الخطأ النوع الأول (α)،، بينما المستطيل الرابع فيمثل التصنيف بالإصابة وهو يقع ما بين ٠.٨، ١ P(١/١) ، مما يدل على قدرة النموذج على التنبؤ الصحيح.



شكل (٢) تقدير احتمالات التنبؤ الصحيح للشبكة العصبية

يوضح الشكل (٣) المساحة تحت منحنى ROCللنموذج بلغت ٩٨.٦% وهذا يوضح أن النموذج يساعد على التنبؤ بتصنيف عوامل المتغير التابع (الاصابة بأحد الامراض المزمنة) أكثر مما تعمله الصدفة.



شكل (٣) منحنى ROC الشبكات العصبية

عاشراً: مناقشة ومقارنة النتائج:

١. المقارنة في حالة النموذج العام وباستخدام دقة التصنيف للنماذج

يتضح من خلال جدول (٩) أن تصنيف البيانات بطريقة النموذج التمييزي (DA) حيث بلغ معدل دقة التصنيف الكلي للنموذج ٨٧.٥%، وبمعدل خطأ بلغ ١٢.٥%، بينما الشبكات العصبية الاصطناعية (ANN) كان الافضل حيث بلغت دقة التصنيف للنموذج الكلي ٩٣.١%، وبلغ معدل الخطأ ٦.٩%. وهذه النتيجة تتفق مع كل الدراسات حيث يتمتع نموذج الشبكات العصبية (ANN) بكفاءة تصنيف عالية وبأقل خطأ ممكن.

جدول (٩) معايير دقة تصنيف النماذج

	Sensitivity	Specificity	Accuracy	Precision	Error rate
DA	87.2%	87.7%	87.٥ %	82.9%	12.5%
ANN	87.7%	98.3%	93.1%	98.0%	6.9%

٢. المقارنة بين الأساليب الإحصائية باستخدام Area ROC

كانت نسبة المساحة المحصورة أسفل منحنى ROC للنموذج التمييزي (DA) بنسبة بلغت ٨٧.٤%، بينما لتقنية الشبكات العصبية بلغت ٩٨.٢% حيث انها تمثل المساحة الأكبر مما يدلل على مدى دقة الشبكات العصبية بتصنيف المشاهدات بدرجة قوية وكبيرة.

٣.المقارنة بين الأساليب الإحصائية في حالة تقنية Bootstrap

يتضح من خلال جدول (١٠) أن تصنيف البيانات بطريقة الشبكات العصبية الاصطناعية (ANN) كان الافضل حيث بلغت دقة التصنيف للنموذج الكلي ٩٢.٢%، وبلغ معدل الخطأ ٧.٨%، بينما التحليل التمييزي (DA) بلغ معدل دقة التصنيف الكلي للنموذج 88.3%، وبمعدل خطأ بلغ 11.7% حسب بيانات الدراسة حيث بلغ دقة التصنيف الكلي للنموذج 86.7%.

جدول (١٠) معايير دقة النموذج التمييزي والشبكات العصبية باستخدام طريقة bootstrap

	Sensitivity	Specificity	Accuracy	Precision	Error rate
DA	86.2%	89.8%	88.3%	86.7%	11.7%
ANN	86.8%	96.8%	92.2%	95.8%	7.8%

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

يتضح من خلال جدول (١١) أن تصنيف البيانات في حالة Two-fold Cross- validation حسب طريقة الشبكات العصبية الاصطناعية (ANN) حيث بلغت دقة التصنيف للنموذج الكلي ٨٨.٥%، وبلغ معدل الخطأ ١١.٥%، النموذج التمييزي (DA) حيث بلغ معدل دقة التصنيف الكلي للنموذج 88.4% وبلغ معدل الخطأ 11.6%.

جدول (١١) معايير دقة النموذج التمييزي والشبكات العصبية باستخدام طريقة Two-fold Cross- validation

	Sensitivity	Specificity	Accuracy	Precision	Error rate
DA	86.2%	90.2%	88.4%	86.2%	11.6%
ANN	87.0%	89.5%	88.5%	84.8%	11.5%

٥. المقارنة بين الأساليب الإحصائية في حالة تقنية Leave-one-out cross-validation

يتضح من خلال جدول (١٢) أن تصنيف البيانات بطريقة الشبكات العصبية الاصطناعية (ANN) كان الافضل حيث بلغت دقة التصنيف للنموذج الكلي 93.0%، وبلغ معدل الخطأ 7.0%، نموذج النموذج التمييزي (DA) دقة التصنيف الكلي للنموذج 87.2% وبمعدل خطأ 12.8%.

جدول (١٢) معايير دقة النموذج التمييزي والشبكات العصبية باستخدام طريقة Leave-one-out cross-validation

	Sensitivity	Specificity	Accuracy	Precision	Error rate
DA	٨٢.٢%	٨٧.٣%	87.2%	٨٢.٤%	١٢.٨%
ANN	87.0%	98.1%	93.0%	97.6%	7.0%

من مناقشة النتائج السابقة يتضح:

وجود فروق بين متوسطات العوامل المؤثرة بين افراد العينة (مصابين /غير مصابين) بأحد الامراض المزمنة، فمن خلال متوسطات عوامل المصابين يتضح انها اعلى من غير المصابين، وهذا طبيعي لان كلما زادت متوسطات (Work, BMI, T.CH Age,T.G,HBA1C, ,Family, History , Region, Smoking) زادت احتمالية الإصابة بأحد الامراض المزمنة وكذلك متغيري (Sex, Region) فاحتمالية إصابة الذكور اعلى منها عند النساء كما أوضحت الدراسات ، وكذلك منطقة السكن حيث الكثافة السكنية والضغط النفسي العالي جدا في المخيمات.

بلغت قيم معامل الارتباط القانوني ٠.٧٤٧ مما يدل على ٧٤.٧% من التباين الى الفروق في نموذج التمييز بين المجموعتين (مصابين وغير مصابين) وهذا يعني أن العوامل المؤثرة ساهمت ب ٧٤.٧% من التباين الذي يحصل في تمييز الإصابة بأحد الامراض المزمنة.

يتضح ان قيمة الجذر الكامن (Eigenvalue=1.261) وهي توضح نسبة التباين المفسر بين مجموعتي الإصابة بأحد الامراض المزمنة ، قيمة معامل الارتباط القانوني إلى ٠.٧٤٧ ويشير إلى الارتباط بين النموذج التمييزي والعوامل المؤثرة وكان مربع الارتباط القانوني ٠.٥٥٨، بمعنى ان العوامل المؤثرة ساهمت في تفسير الإصابة بنسبة ٥٥.٨% وهذا يدل على جودة توفيق الدالة التمييزية.

أظهرت النتائج أن العوامل المؤثرة التي دخلت النموذج التمييزي، ولها تأثير على متغير الإصابة بأحد الامراض المزمنة وبقيت في النموذج حسب التحليل التمييزي التدريجي هي حسب معادلة العوامل المؤثرة للبيانات

كما أظهرت نتائج الشبكة العصبية: أهمية المتغيرات في التصنيف باستخدام الشبكات العصبية، حيث أن العامل HBA1C كان الأكثر تأثيرا في التصنيف بالإصابة بأحد الامراض المزمنة29 %، يليه عامل العمر بنسبة ١٧.٢% ، يليه عامل التاريخ العائلي للمرض ١٤.٦%، يليه عامل تحليل الدهون الثلاثية T.G بنسبة ٩.٥%، يليه عامل كتلة الجسم BMI بنسبة ٨.١%، يليه عامل التدخين smoking ٦.٧% ، يليه عامل تحليل الكوليسترول الكلي T.CH ٨.٧% ،يليه عامل النوع الاجتماعي بنسبة ١.٩%، وأخيرا عامل المنطقة بنسبة ١.٣%، علما ان الأهمية النسبية للعوامل تمثل الاوزان للمتغيرات.

التوصيات

١. نوصى الباحثين باستخدام نموذج الشبكات العصبية في تقدير العوامل المؤثرة على الامراض المزمنة حيث أظهرت النتائج فعاليته على نموذج التحليل التمييزي.
٢. نوصى الباحثين والدارسين بإجراء المزيد من الابحاث حول الامراض المزمنة للمساعدة في تحديد العوامل الخطرة والمسببة للأمراض المزمنة واستخدام طرق احصائية حديثة تعتمد على التعلم اخرى غير المستخدمة في هذا البحث وكذلك تضمين متغيرات اخرى لها علاقة بالأمراض المزمنة.

٣. اجراء فحص مستوى الهيموغلوبين السكري HBA1C في الدم لأنه احد اهم الأسباب التي من خلالها يتم التشخيص، 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.

٤. البطالة أحد الأسباب المؤثرة بالإصابة بأحد الامراض المزمنة، لذلك نوصي بعدم التكاسل وبدل المزيد من العمل والحركة لما لها مردود إيجابي على الشخص.

٥. نوصي وزارة الصحة بتوفير قاعدة بيانات الكترونية جيدة لجميع المرضى والمصابين بالأمراض وكذلك المراجعين حيث يمكن استخدامها في مجال البحث.

٦. تعميم فكرة استخدام الاساليب الاحصائية للتمييز والتصنيف في المجالات الاجتماعية والاقتصادية وعدم تركيزها على المجالات الطبية فقط.

المراجع:

١. رولا رضا شريقي، فاعلية برنامج إرشادي لرفع مستوى الرضا عن الحياة لدى مرضى السكري، رسالة دكتورة – جامعة دمشق –٢٠١٤

٢. زياد عبد الكريم القاضي(٢٠٠٠م)، مقدمة في الذكاء الاصطناعي، الطبعة الأولى، مكتبة المجتمع العربي للنشر والتوزيع، عمان

٣. علي ابشر فضل المولى سليمان، ٢٠١٥، المقارن بين التحليل التمييزي والنموذج اللوجستي الثنائي ونماذج الشبكات العصبية في تصنيف المشاهدات–دراسة على العوامل المؤثرة على دخل الاسرة – جامعة السودان

٤. مخللاتي، جلال،1984، التغذية وصحة الإنسان، الجامعة الإسلامية، غزة.

٥. وزارة الصحة، ٢٠٢١، التقرير السنوي الامراض المزمنة بقطاع غزة –فلسطين.

6. 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.
7. David W. Hosmer. Jr.,(2000) Applied logistic regression, Stanley Lemeshow. 2 nd ed, Jone Wiley & Sons, Inc
8. Gokhan Zorluoglu, Mustafa Agaoglu, Diagnosis of Breast Cancer Using Ensemble of Data Mining Classification Methods, Goztepe Campus 34722 / Kadikoy - Istanbul TURKEY
9. Hosmer, D. W. and Lemeshow, Stanley. (2000), "Applied logistic regression", 2nd Edition. Published by Johan Wiley and Sons, Wiley, New York.

10. Johnson, R.A. and Wichern, D.W. (2007) Applied Multivariate Statistical Analysis., Scientific Research an Academic Publisher Journal, 6th Edition.
11. Jnan Roman Rabunal and Julin Dorrod, (2006) , Artificial Neural Network in real-life applications Idea Group Publishing , USA.
12. Ferrer, Alvaro J. Arce and Wang, Lin (1999). Comparing the Classification Accuracy among Nonparametric, Parametric Discriminant Analysis and Logistic Regression Methods. Paper Presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23,1999).
13. 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
14. Nettina، s(1996): Manual of Nursing practice،6th edition،Lippincott company، New York.
15. 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.
16. 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