استخدام طرق التصنيف لتحديد أهم عوامل الخطر على مرضى السكري في فلسطين -قطاع غزة

هـارون موسى بهار
قسم الإحصاء، جامعة الاسراء
غزة، دولة فلسطين

عبد الهادي خليل أبوسعده
قسم الإحصاء، جامعة فلسطين
غزة، دولة فلسطين

تاريخ استلام البحث: 2021/06/13
تاريخ قبول البحوث: 2021/06/30
نشر البحث في العدد الثاني عشر: يوليو 2021

رمز التصنيف ديوي / النسخة الالكتروني (Online) 2522-64X/519.5;616.4
رمز التصنيف ديوي / النسخة الورقية (Print) 2519-948X/519.5;616.4

استخدام طرق التصنيف لتحديد أهم عوامل الخطر على مرضى السكري في فلسطين -قطاع غزة

هـارون موسى بهار
قسم الإحصاء، جامعة الاسراء
غزة، دولة فلسطين

عبدالهادي خليل ابوسعده
قسم الإحصاء، جامعة فلسطين
غزة، دولة فلسطين

المستخلص

مرض السكر يُعتبر من أكثر الامراض انتشاراً، وهناك العديد من الدراسات التي تناولت هذا الموضوع، لكن تأتي أهمية هذا البحث من انه تمكن من بناء نموذج إحصائي يستطيع التنبؤ بحالات الاصابة بمرض السكر والذي من خلاله نستطيع تحديد أهم عوامل الخطر التي إذا ما توفرت لدى اي شخص كان احتمال اصابته بهذا المرض أكبر بكثير. حيث تم استخدام ثلاث نماذج إحصائية (التحليل اللوجستي المتعدد ، تحليل التمايز ، تحليل الشبكات العصبية) للحصول على أفضل نموذج يستطيع تصنيف مرض السكر باعتباره متغير تابع وأهم عوامل الخطر التي تساعد على وجود هذا المرض، وكانت خلاصة التحليل أن أفضل نموذج حصلنا عليه كان باستخدام (نموذج تحليل الشبكات العصبية) حيث بلغت درجة دقة التصنيف والتنبؤ (95.7%)؛ وأهم عوامل الخطر التي ظهرت في النموذج هي: الضغط النفسي المرتفع، كمية الفواكه والخضروات التي يتناولها أسبوعيا، كمية اللحوم المستهلكة أسبوعا، المؤهل العلمي. وفي هذه الدراسة تم استخدام بعض أدوات التقييم الإحصائي لتقييم النموذج الذي تم الحصول علية ولمعرفة مدى جودة التصنيف التي يحققها هذا النموذج، وهي: (leave-one-out cross validation, classification table and ROC curve). وقد تم تحليل البيانات باستخدام البرمجيات R ، SPSS.

الكلمات المفتاحية:

التحليل اللوجستي، التحليل التميزي، الشبكات العصبية، جدول التصنيف، منحنى ROC، نسبة الأرجحية، مرض السكري.

The Use of Classification Methods to Identify the Most Important Risk Factors for Diabetics Patients in Palestine- Gaza Strip

Abstract:

Diabetes considers one of the most prevalent diseases; there are many research dealt with this topic. but this research plays an important role as it was able to build a statistical model which can predict the cases of diabetes, that we can identify the most important risk factors which the more appear on a person, the far more likely to be infected with the disease.

Three statistical models (Multiple Logistics Analysis, discriminant analysis, neural network analysis) were tested to obtain the best model can classify diabetes as a dependent variable and the most important risk factors that help the presence of this disease.

The conclusion was that the best model we got is by using (neural network analysis) as the accuracy of classification and prediction (95.7%).

The most important risk factors were: (1) high psychological stress (2), quantity of fruits and vegetables consumed per week, (3) amount of meat consumed per week, (4) scientific qualification.

In this research, the researchers used some statistical methods for model evaluation and for evaluating the quality of the classification achieved by this model: (leave-one-out cross validation, classification table and ROC curve). Data in this research has been analyzed using software R, SPSS.

Keywords: Discriminant Analysis, Logistic regression, Neural network, classification table, ROC curve, odds ratio, diabetic patients.

مقدمة:

يعتبر مرض السكر أحد الامراض المزمنة والتي تحدث نتيجة عجز غدّة البنكرياس عن إنتاج الكمية الكافية من الإنسولين أو عندما يعجز الجسم عن استخدام الإنسولين الذي أنتجه بالشكل المطلوب، ويذكر تقرير منظمة الصحة العالمية حول مرض السكر: أنه ارتفع عدد الأشخاص المصابين بالسكري من 108 ملايين شخص في عام 1980 إلى 422 مليون شخص في عام 2014 ، كما ارتفع معدل انتشار السكري على الصعيد العالمي لدى البالغين الذين تزيد أعمارهم على 18 سنة من 4.7% في عام 1980 إلى 8.5% في عام 2014 ، وسجل معدل انتشار السكري ارتفاعاً أسرع في البلدان ذات الدخل المتوسط والمنخفض(WHO). وتأتي أهمية هذا البحث من انه تمكن من بناء نموذج إحصائي يستطيع التصنيف والتنبؤ بحالات مرض السكر باستخدام بيانات تم جمعها من داخل المراكز الصحية، وكانت عينة الدراسة عبارة 232 شخص مقسمين الى 172مريض سكري و60 شخص غير مريض بالسكري. من كلا الجنسين والتي تزيد أعمارهم عن 25سنة.

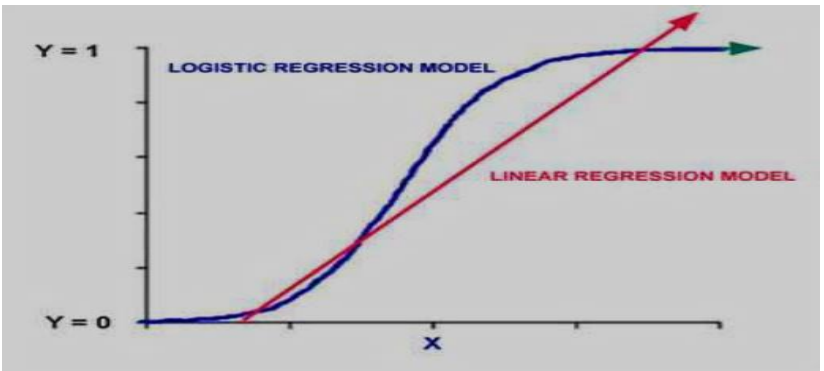
البحوث العلمية لا تهدف إلى وصف الموقف الحالي لمختلف الظواهر فحسب، بل تهدف ايضا إلى توضيح العلاقات بين السبب والنتيجة، ويمكننا القول أن البحث العلمي يحاول تقديم العلاقات بين السبب والنتيجة في صورة نموذج. بحيث تصف هذه النماذج العلاقة بين النتيجة على أنها (متغير تابع) ومجموعة الأسباب التي تؤثر على النتيجة على أنها (متغير مستقل أو متغيرات مستقلة). نماذج الانحدار الخطي المختلفة هي نماذج شائعة الاستخدام في مختلف العلوم، لكن هذه النماذج تتطلب أن يكون المتغير التابع متغير كمي(عددي). علاوة على ذلك؛ فإنها تتطلب أيضًا افتراضات التوزيع الطبيعي للبيانات. لكن عادة ما يكون المتغير التابع الذي يتم الحصول عليه في العلوم الاجتماعية وغيره من العلوم متغير وصفي وبالتالي لا يمكن تطبيق نماذج الانحدار الخطي. في هذه الحالة علينا استخدام الطرق الاحصائية التي تناسب هذه الحالة -عندما يكون المتغير التابع وصفي مثل: الانحدار اللوجستي ، التحليل التمييزي والشبكات العصبية ، التي سنطبقها في هذه الدراسة، سنناقش أيضًا بعض تقنيات التقييم لطرق التصنيف أعلاه.

1. الانحدار اللوجستي الثنائي (*Logistic Regression*):

الانحدار اللوجستي (LR) هو اسلوب نمذجة إحصائية لتحليل البيانات الفئوية ، وهو يستخدم في الكثير من المجالات؛ من البحوث الطبية الحيوية إلى مجالات الأعمال التجارية، المالية، الهندسة، التسويق، الاقتصاد، السياسة، الصحية... زاد توافر البرامج الإحصائية المتطورة والحواسيب عالية السرعة من زيادة فائدة الانحدار اللوجستي كأداة إحصائية مهمة. (Shmueli et al. 2010).

1.1. مفهوم الانحدار اللوجستي:

يعرف الانحدار اللوجستي بشكل عام بانه التحليل الذي يختص بدراسة العلاقة بين متغير واحد يعرف بالمتغير التابع (في حال كان المتغير التابع نوعيا مكون من فئتين Dichotomous – في هذه الحالة يسمى الانحدار اللوجستي الثنائي) ومتغير واحد او أكثر يعرف بالمتغير المستقل او المتغيرات المستقلة (المفسرة) وذلك بغرض التقدير أو التنبؤ. ويقوم نموذج الانحدار اللوجستي على فرض أساسي وهو أن المتغير التابع Y والذي نهتم بدراسته وهو متغير ثنائي يأخذ القيمة (1) باحتمال (P) والقيمة (0) باحتمال (1-P). في هذه الحالة وعند القيام برسم افضل خط مستقيم لتوفيق البيانات سيكون غير ملائم والسبب هو ان الخط المستقيم سوف ياخذ قيم اكبر من الواحد الصحيح واقل من الصفر، الا اذا كان الميل صفر. ويرى (Pample,2000) بان احد الحلول لهذه المشكلة هي اعتماد صيغة القمة والقاع ووفقا لهذا المبدأ فان هناك حدود للقيم المتنبأ بها بحيث الا تتجاوز الواحد الصحيح ولا تقل عن الصفر كما في الشكل:



شكل رقم (1) يوضح العلاقة غير الخطية بين المتغير التابع والمتغيرات المستقلة

وبناء على ذلك فان توفيق البيانات في حالة المتغير التابع الثنائي لن يكون من خلال استخدام افضل خط مستقيم ولكن من خلال المنحنى اللوجستي والذي تقع قيمه بين الصفر والواحد والذي يأخذ شكل حرف ال S هو الانسب لتوفيق هذه البيانات (Walker, 1998)، بواسطة (بابطين:1429هـ). وكذلك قال (Schmidt,2000) أن العلاقة الغير خطية الاكثر ملائمة هي المشابهة لحرف S، بحيث تكون مستويان المنحنى محصورة بين الصفر والواحد. الانحدار اللوجستي مناسب بشكل خاص لتقدير المتغيرات الفئوية (ثنائية التفرع أو المتجانسة) باستخدام إجراء تقدير الاحتمال الاعظم (Maximum Likelihood Estimation (MLE)). تستخدم نماذج الانحدار اللوجستي MLE كمعيار للتقارب. يسمح الانحدار اللوجستي للتنبؤ بنتيجة ثنائية مثل الوجود / الغياب ، النجاح / الفشل ، الشراء / عدم الشراء ، البقاء على قيد الحياة /

الموت. قد تكون المتغيرات المستقلة قاطعة أو مستمرة أو مزيج من الاثنين معا. يمكننا التفكير في المتغير التابع على أنه يقسم الملاحظات إلى عدة فئات. على سبيل المثال، إذا كانت Y كمتغير تابع تشير إلى توصية بقبول أو رفض الصفقة، فعندئذ يكون لدينا متغير تابع مع فئتين. وبالتالي فإن الصفقة سوف تنتمي إلى واحدة من فئتين، "قبول" الصفقة أو "رفض" الصفقة. بالطبع، هذا يعتمد على البيانات المتاحة (المتغيرات المستقلة). (Shmueli et al. 2010). بشكل عام، نموذج الانحدار اللوجستي يكون كالتالي:

$$\text{Log}\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n = x\beta \qquad \text{--------(1)}$$

حيث أن p هو احتمال الحصول على النتيجة P(Y=1) =p ، β_0 هو الحد الثابت، , i=1,2,...n β_i هي معاملات المتغيرات المستقلة (التوضيحية) x_i,

$$x_i = (1, x_1, x_2, \ldots, x_n), \text{ and } \beta_i = (\beta_0, \beta_1, \ldots, \beta_n)$$

ويمكن التعبير على احتمال الحصول على النتيجة، p ، كدالة غير خطية كما يلي :

$$p = \frac{1}{1+ e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}} \qquad \text{--------} (2)$$

مع العلم أن القيمة $\frac{p}{1-p}$ تعرف ب معامل الارجيحة (odds) ويمكن كتابتها كما يلي :

$$p = \frac{odds}{1+odds} \qquad \text{---------}(3)$$

وللتخلص من log في المعادلة (2-1) نقوم بحساب اللوغارتيم للطرفين فنحصل على:

$$\frac{p}{1-p} = e^{x\beta} \Rightarrow p = e^{x\beta} - pe^{x\beta} \Rightarrow p(1 + e^{x\beta}) = e^{x\beta}$$

$$p = \frac{e^{x\beta}}{1+ e^{x\beta}} \qquad \text{----------} (4)$$

نسبة الارجحية (odds ratio) تساوي exp(p), واحيانا تكتب على الصيغة e^x.

لا تفترض الصيغة أعلاه أن المتغيرات التوضيحية او التفسيرية موزعة بشكل طبيعي أو لها مصفوفات تغاير متساوية. بل يجب أن تكون مستقلة ويمكن أن تكون ثنائية أو متصلة . وهذا يجعل الانحدار اللوجستي قويًا نسبيًا مقارنة بالانحدار الخطي. أيضا، نماذج الانحدار اللوجستي لا تفترض التماثل او التجانس (homoscedasticity) بين المتغيرات التابعة والمستقلة. تم استخدام الانحدار اللوجستي على نطاق واسع في

العديد من الدراسات والعلوم الاجتماعية، والأعمال التجارية والمجالات ذات الصلة. (Hosmer & Lemeshow, 2000) & (Pai., 2009).

2.1. افتراضات الانحدار اللوجستي:

يفترض الانحدار اللوجستي وجود علاقة خطية بين لوغاريتم (logit) المتغيرات المستقلة والمتغير التابع. ولا يفترض وجود علاقة خطية بين المتغيرات المستقلة والمتغير التابع . تنخفض موثوقية التقدير عندما يكون هناك عدد قليل من الحالات في كل فئة. التوزيع الطبيعي للمتغير التابع ليس ضروريًا، تماثل (Homoscedasticity) ليست ضروري لكل مستوى من مستويات المتغيرات المستقلة، ليس مطلوب أن تكون الأخطاء (errors) تتبع التوزيع الطبيعي .(Schüppert, 2009)

اختبار جودة الملاءمة للانحدار اللوجستي: تقيس جودة الملاءمة للنموذج مدى جودة وصف النموذج لمتغير الاستجابة. يتضمن تقييم جودة الملاءمة دراسة مدى قرب القيم التي تنبأ بها النموذج من القيم الملاحظة. تقوم إحصائيات *Hosmer-Lemshow* بتقييم مدى الملاءمة من خلال إنشاء 10 مجموعات مرتبة ، ثم تقارن العدد الفعلي في كل مجموعة (يتم ملاحظتها) مع القيم التي يتم التنبؤ بها بواسطة نموذج الانحدار اللوجستي (المتوقع). وبالتالي، ويتم استخدام اختبار كاي تربع، للتأكد من أن نتائج تنبؤ النموذج لا يختلف اختلافًا كبيرًا عن الملاحظة المرصودة. إذا كان النموذج جيدًا، فسيتم تصنيف معظم الملاحظات بشكل صحيح *Hosmer & Lemeshow, 2000*).

2. التحليل التمييزي (*Discriminant Analysis -DA*):

الهدف من التحليل التمييزي (DA) هو تصنيف الكائنات، باستخدام مجموعة من المتغيرات المستقلة؛ إلى واحدة من فئتين أو أكثر من الفئات -بحسب فئات المتغير التابع-. يعتمد نموذج التحليل التمييزي على الوصول إلى دالة التمايز (Discriminant Function) التي تعمل على تعظيم الفروق بين متوسط المجموعات وتقليل التشابه في أخطاء التصنيف في الوقت ذاته، وذلك من خلال إيجاد تجميعات خطية لمجموعة من المتغيرات (Johnson & Wichern, 2007). على سبيل المثال، على أساس عمر مقدم الطلب وطوله ودخله ومستواه التعليمي؛ وما إلى ذلك، يمكن لنا تصنيف الأشخاص على أنهم مقبولون أم لا ضمن مسابقة معينة. وتكون هناك دالة تمييزية لكل مجموعة او فئة، وتكون معادلة التمييز على الشكل التالي:

$$Y_i = b_0 + b_1 \quad x_{1i} + b_2 \quad x_{2i} + \ldots + b_n \, x_{ni} \qquad -------(5)$$

إجراءات التصنيف هي كما يلي:

يتم تصنيف الفرد على انه ضمن المجموعة الاولى $Y_i > Y_{\text{critical}}$

يتم تصنيف الفرد على انه ضمن المجموعة الثانية $Y_i < Y_{\text{critical}}$

حيث أن : Y_{critical} هي القيمة الحرجة

1.2. افتراضات تحليل التمايز:

1- **حجم العينة:** يجب أن يكون أصغر حجم للعينة 20 على الأقل لكل فئة من فئات المتغير التابع ، ومن الافضل بشكل عام أن تكون عدد المشاهدات 4 أو 5 أضعاف عدد المتغيرات المستقلة.(Tabachnick and Fidell, 1996).

2 - **التوزيع الطبيعي:** من الافضل أن جميع المتغيرات تتبع التوزيع الطبيعي. ولكن حتى لو لم تتبع البيانات التوزيع الطبيعي فيمكن الاعتماد على النتائج وتكون نتائج موثوقة (Tabachnick and Fidell, 1996).

3- **تجانس التباين / التغايرات:** تحليل التمايز حساس للغاية لعدم تجانس مصفوفات التباين والمتغاير. قبل قبول الاستنتاجات النهائية، من المستحسن مراجعة الفروق داخل المجموعات ومصفوفات الارتباط.

4 - **القيم المتطرفة:** تحليل التمايز حساس للغاية للقيم المتطرفة. إذا احتوت إحدى المجموعات في الدراسة على القيم المتطرفة التي تؤثر على الوسط ، فسوف تزيد أيضًا من التباين.

5 - **الترابط الداخلي:** إذا كان أحد المتغيرات المستقلة يرتبط ارتباطًا كبيرًا بمتغير آخر، فهذا سوف يؤثر بشكل كبير نتائج دالة التمييز .(Poulsen and French, 1999)

- تصنيف الحالات. بمجرد احتساب درجات التصنيف لحالة ما ، يكون من السهل تحديد كيفية تصنيف الحالة. بشكل عام ، نقوم بتصنيف الحالة على أنها تنتمي إلى المجموعة التي حصلت على أعلى درجة تصنيف لها. ولفهم كيفية حصول ذلك علينا أن ننظر أولاً في ما يسمى بمسافة (Mahalanobis)ماهالانوبيس، وهي عبارة عن مقياس للمسافة بين نقطتين في الفضاء المحدد بواسطة اثنين أو أكثر من المتغيرات. لكل حالة يمكننا بعد ذلك حساب المسافات (Mahalanobis) (لكل حالة على حدة). ويتم تصنيف الحالة على أنها تنتمي إلى المجموعة التي هي الأقرب إليها ، أي حيث تكون مسافة (Mahalanobis) أصغر. (StatSoft, Inc., 1984-2000).

3. الشبكات العصبية (*Neural network*):

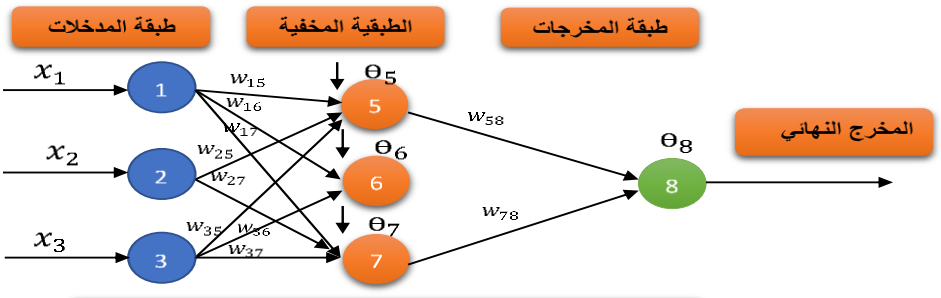
الشبكات العصبية هي الأدوات المفضلة للعديد من تطبيقات التنقيب عن البيانات بسبب قوتها ومرونتها. الشبكات العصبية التنبؤية مفيدة بشكل خاص في التطبيقات التي تكون فيها العملية الأساسية معقدة ، مثل:

- التنبؤ بطلب المستهلكين لتبسيط تكاليف الإنتاج والتسليم.
- سجل المتقدمين لتحديد مخاطر تقديم الائتمان.
- الكشف عن المعاملات الاحتيالية في قاعدة بيانات مطالبات التأمين.

الشبكات العصبية المستخدمة في التطبيقات التنبؤية ، مثل الشبكات متعددة الطبقات *Perceptron (MLP)*، هي تقنيات خاضعة للإشراف بمعنى أن النتائج المتوقعة للنموذج يمكن مقارنتها بالقيم المعروفة للمتغيرات المستهدفة. (شركة *IBM*، 2011). تسمى الشبكات العصبية أيضًا الشبكة العصبية الاصطناعية (*NN*). بدأت الدراسات الأولى على *NN* مع نمذجة الخلايا العصبية التي تشكل الدماغ ، وتطبيق هذه النماذج في أنظمة الكمبيوتر. بعد ذلك ، أصبح من الشائع في العديد من المجالات بالتوازي مع التطور في أنظمة الكمبيوتر. *NN* له استخدام فعال في العديد من المجالات مثل الطب والصناعة والبيولوجيا والأنظمة الإلكترونية ، والعلوم الاجتماعية (*Golden*، 1996).

1.3. إعداد الشبكات العصبية:

تأتي أهمية الشبكات العصبية من قدرتها على اكتشاف مشاكل العلاقات غير الخطية وحلها. علاوة على ذلك، تشير الأبحاث على مدى العقدين الماضيين إلى أن الشبكات العصبية قد تحقق تصنيفًا وتوقعًا أفضل مقارنة بالطرق الإحصائية القياسية الأخرى. (*Pai, 2009*) &(*Sharda, 1994*). وقد تم إثبات ذلك من خلال عدد من تطبيقات *NN* الناجحة مثل التنبؤ بالإفلاس، وتوقع فشل البنك ، وتجزئة السوق على سبيل المثال لا الحصر. يلتقط هيكل الشبكات العصبية العلاقات المعقدة بين متغيرات التوقع ومتغير الاستجابة من خلال طبقة من الخلايا العصبية. لدى بعضها شبكات عصبية أحادية الطبقة (*SLNN*) وبعضها لديها شبكات عصبية متعددة الطبقات (*MLNN*). أنجح التطبيقات في التصنيف والتنبؤ هي شبكات التغذية متعددة الطبقات. الطبقة التي يتم فيها تطبيق أنماط الإدخال هي طبقة الإدخال. الطبقة التي نحصل منها المخرجات هي طبقة المخرجات. في حالة وجود نتيجة ثنائية يكون لدى الشبكة عقدة إخراج واحدة فقط. تُعرف الطبقات بين طبقات المدخلات والمخرجات باسم الطبقات المخفية أو طبقات النقل، لأن مخرجاتها لا يمكن ملاحظتها بسهولة. .(Pai. , 2009) يوضح الشكل التالي شبكة بسيطة متعددة الطبقات تتكون من العقد والسهام. تمثل العقد في الشبكة الخلايا العصبية بينما تشير الأسهم إلى مسار الاتصال المرتبط. تربط الأسهم الخلايا العصبية من طبقة واحدة إلى الطبقة التالية. مخرجات العقدة في الطبقة هي مدخلات للعُقد في الطبقة التالية. تسمى القيم على أسهم الاتصال الأوزان ، ويرمز لها بالرمز ($w_{(i,j)}$).



شكل رقم (2) رسم توضيحي للشبكة العصبية

2.3. تدريب النموذج:

يتم تدريب نموذج الشبكة العصبية أولاً على مجموعة من المدخلات والمخرجات باستخدام بيانات مجموعة التدريب. تدريب النموذج يعني تقدير الأوزان التي تؤدي إلى أفضل النتائج التنبؤية. الطريقة الأكثر شيوعًا المستخدمة في الضبط هي خوارزمية تسمى الانتشار الخلفي. في هذه الطريقة، يتم ضبط الأوزان لتقليل الفرق التربيعي بين مخرجات النموذج والمخرجات المرغوب فيها. تستند التعديلات إلى خوارزمية النسب التدرج (*Shmueli et al. 2010*) & (*Pai,2009*).

على الرغم من العديد من المزايا مثل الأداء التنبئي الجيد، وقدرتها على حل مشاكل العلاقات المعقدة. فإن الشبكات العصبية لديها بعض العيوب. أولاً: لا توفر الشبكة العصبية نظرة جيدة حول بنية العلاقة بين متغيرات التنبؤ والاستجابة. ثانياً: ليس للشبكة العصبية آلية اختيار متغيرة مضمنة، وبالتالي هناك حاجة لتقييم أهمية إضافة متغيرات التوقع إلى النموذج باستخدام طرق إحصائية أخرى. ثالثًا: تعتمد الشبكة العصبية اعتمادًا كبيرًا على وجود بيانات كافية لأغراض التدريب، وإلا فإن أداء النموذج سيكون سيئ. تستغرق الشبكة العصبية وقتًا حسابيًا أعلى نسبيًا (*Pai,2009*) & (*Shmueli et al.,2009*).

4. طرق تقيم النموذج:

1.4. جدول التصنيف *Classification Tables*:

إن استخدام جدول التصنيف يعتبر احدى طرق فحص جودة مطابقة النموذج للبيانات ، وتعتمد هذه الطريقة على انشاء جدول يوضح عدد الحالات التي تمتلك الصفة المرغوب فيها أو الحالات التي لا تمتلك الصفة المرغوب فيها والتي تم تصنيفها بطريقة صحيحة أو بطريقة خاطئة (Soderstrom & Leitner, 1997) ، وتتطلب هذه الطريقة الحصول على متغير تابع مشتق من النموذج من خلال تحديد نقطة قطع C، ثم مقارنة الاحتمالات المتوقعة بتلك النقطة بحيث اذا تجاوزت الاحتمالات المتوقعة نقطة القطع C أعطيت تلك الحالة تصنيفا متوقعا يساوي واحد، وما عدا ذلك فان

الحالة يعطي لها تصنيف متوقع يساوي الصفر، علما بأنه غالبا ما تكون نقطة القطع C تساوي 0.5 (Frass & Newman, 2003) وتعتمد فكرة استخدام هذا التحليل على أن النموذج اذا قام بتوقع تصنيف الحالات بشكل صحيح اعتماد على معيار ما ، فان ذلك يعطي برهانا بأن النموذج يطابق البيانات المشاهدة (Ferrer & Wang, 1999).

جدول(1) جدول التصنيف

التصنيف		التوقع		المجموع
		الموجب	السالب	
المشاهد	الموجب P	الموجب الصحيح TP	السالب الخاطئ FN	P= TP+FN
	السالب N	الموجب الخاطئ FP	السالب الصحيح TN	P'= FP+TN
المجموع		Q= TP+FP	Q'= FN+TN	N=TP+FN+FP+TN

الحساسية Sensitivity: ويرمز لها بالمز SE ، وتعرف بانها قيمة الاحتمال بأن يكون التصنيف المتوقع موجبا للحالة التي تكون فعلا موجبة وتحسب حسب المعادلة :

$$SE = \frac{TP}{TP + FN} = \frac{TP}{P}$$

الدقة Specifity: ويرمز له بالمز SP، وتعرف بانها قيمة احتمال أن يكون التصنيف المتوقع سالبا للحالة التي تكون فعلا سالبة، وتعطى حسب المعادلة

$$SP = \frac{TN}{FP + TN} = \frac{TN}{P'}$$

الدقة Accuracy: تعني قرب القيمة المقاسة من القيمة القياسية أو القيمة الفعلية، على سبيل المثال ، إذا كنت في المختبر تحصل على قياس للوزن يبلغ 3.2 كغ لمادة معينة ، لكن الوزن الفعلي أو المعروف هو 10 كغم ، فإن قياسك غير دقيق (not accurate). في هذه الحالة، لا يكون قياسك بالقرب من القيمة المعروفة. ويمكن قياس الدقة (accuracy) من خلال العلاقة التالية:

$$\text{Accuracy} = \frac{TN+TP}{N}$$

الإحكام أو الضبط Precision: تشير إلى قرب قياسات اثنين أو أكثر من بعضها البعض. باستخدام المثال أعلاه، إذا كنت تزن مادة معينة خمس مرات، وتحصل على 3.2 كجم في كل مرة، يكون القياس محكم أو مضبوط للغاية (very precise). الضبط أو الإحكام(Precision) مستقلة عن الدقة (accuracy). يمكنك أن تكون محكم ومضبوط جدًا (very precise) ولكن غير دقيق (accuracy)، كما هو موضح أعلاه.

يمكنك أيضًا أن تكون دقيقًا (accuracy) ولكن غير محكم (دقيق)(imprecise). والدقة أو الاحكام (Precision) يتم حسابها كالتالي:

$$\text{Precision}=\frac{TP}{TP+FP}$$

يتم احتساب معدل الخطأ (error rate) لكل التصنيفات حسب الصيغة التالية :

$$\text{Error Rate} = \frac{FN+FP}{N}$$

معدل الخطأ الموجب (The false positive rate) يتم حسابة كم خلال الصيغة:

$$\text{The false positive rate}=\frac{FP}{FP+TP}$$

نسبة التصنيف الصحيح Hit Ratio: وتعرف بأنها قيمة احتمال التصنيف الصحيح، كما أنها تعرف؟ أيضا بنسبة الكفاءة، واذا كانت الكفاءة Efficiency ، والتي يرمز لها بالمز EF وتعرف بانها: $EF = \text{TP} + \text{TN}$ فان نسبة التصنيف الصحيح أو ما يعرف بنسبة الكفاءة تساوي

$$HitRatio = \frac{EF}{Total} = \frac{(TP+TN)}{P+P'} = \frac{(TP+TN)}{(Q+Q')}$$

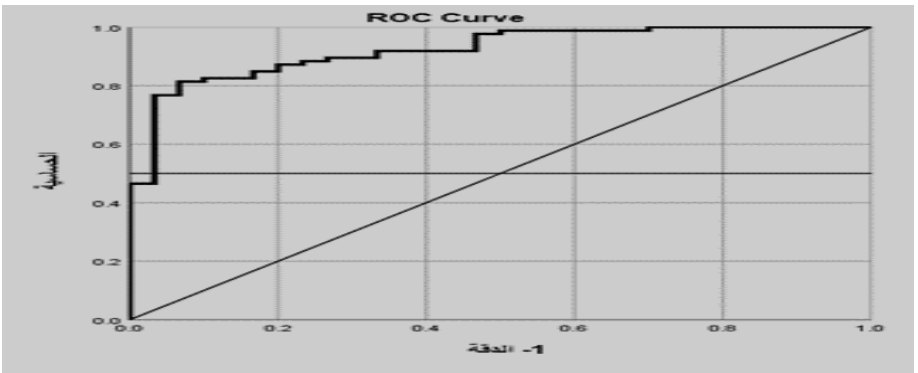
علما بأن جميع هذه المقاييس تتطلب قاعدة للتقرير واتخاذ القرار(أو ما يسمى ب threshold) لتصنيف نتائج الاختبار إما موجبا أو سالبا (Ferrer& Wang,1999) & (Obuchowski ,2005).

لكن كما يقول (Hosmer & Lemshow, 2000, p. 156) بأنه لسوء الحظ فان هذا الوضع وهو استخدام هذا التحليل مبرهناً على مطابقة النموذج للبيانات قد لا يعمل دائما، حيث من السهل تصميم وضعية يكون فيها تحليل الانحدار اللوجستي صحيحا ويطابق البيانات المشاهدة ، ومع ذلك يعطي تحليل جداول التصنيف نتائج سيئة وضعيفة ، إن سبب ذلك يعود الى أن صحة التصنيف تعتمد على التوزيع الاصلي للمتغير التابع وحساسية التحليل لنسب احجام مجموعتي العينة، حيث دائما ما يفضل التحليل تصنيف الحالات لصالح المجموعة الاكبر حجما وهي حقيقة مستقلة عن وجود مطابقة النموذج، بمعنى أن دقة التصنيف أو عدمها لا يعكسان المعيار المفترض لجودة المطابقة ، وهو اقتراب او بعد المسافات بين القيم المشاهدة والمتوقعة للنموذج كما يلاحظ أنه اذا كانت هناك حالات كثيرة لها احتمالات تقترب من نقطة القطع فان من المتوقع ان يكون مقدار سوء التصنيف كبير(Nichols, et al., 1998) & (Soderstrom & Leitner:1998)

ويرى (Ferrer & Wang: 1999) أن احدى مميزات هذه الطريقة أنها تمكن الباحث من مقارنة نتائج التحليل لأسلوبين إحصائيين مختلفين تماما كما هو الحال في مقارنة نتائج تحليل الانحدار اللوجستي مع نتائج التحليل التمييزي ، وذلك لان كلا الاسلوبين يعيطان جدول التصنيف نفسه والذي يمكن من خلالهما اجراء المقارنة.

2.4. تحليل منحنى ROC :

ان تمثيل (1-الدقة) في مقابل (الحساسية) لجميع نقاط القطع فأن ذلك يعطي شكلا في غاية الأهمية في تقيم النموذج وهو ما يسمى منحنى خاصية تشغيل المستقبل Receiver Operation Characteristic والمعروف اختصارا بمنحنى ROC كما في الشكل التالي (Chatellier,1998) & (Westin,2005)



شكل رقم (3) يمثل منحنى ROC للبيانات المتوقعة للنموذج اللوجستي

لقد بدأ استخدام منحنى ROC خلال الحرب العالمية الثانية اعتمادا على نظرية التقاط الإشارات، والتي توضح كيفية التقاط المشغل المستقبل إشارات الرادار عند وجود التشويشات، وهي ترسم احتمال التعرف على الإشارة الصحيحة (sensitivity) والاشارة الخاطئة (1- specificity) على المدى الكلي لنقاط القطع الممكنة(Fawcett,2005) ويبدأ منحنى ROC بالإحداثي(0,0) المقابل لنقطة القطع للحالات السالبة، اما الطرف الاخر فإن الاحداثي (1,1) يقابل نقطة القطع للقرار بأن جميع الحالات موجبة. أما الخط الذي يصل بين النقطتين (0,0) و(1,1) فيسمى بقطر الصدفة Chance Diagonal، وهو يمثل منحنى ROC لاختبار التصنيف الذي ليس له قدرة على التمييز بين الحالات الموجبة والحالات السالبة. ولكن عندما يكون المنحنى واقعا أعلى من قطر الصدفة فإن هذا يعني ان النموذج له قدرة تصنيفية وتميزية بين الحالات الموجبة والسالبة، وكلما ابتعد المنحنى عن قطر الصدفة نحو الجهة اليسرى والى الأعلى كلما ذادت قدرة النموذج على التصنيف (Obuchowski,2005). وتعطى المساحة تحت المنحنى ROC والتي تتراوح ما بين الصفر والواحد صحيح مقياسا لمدى قدرة النموذج على التصنيف بين الحالات التي تمتلك السمة موضع الدراسة او الفحص والحالات التي لا تمتلك هذه السمة، وهي أي المساحة تحت المنحنى ROCتعتبر من افضل مقاييس دقة التصنيف (Hosmer&Lemshow,2000). وتكون المساحة تحت قطر الصدفة تساوي 0.5

وكلما زادت القدرة التمييزية للنموذج كلما ابتعد المنحنى عن قطر الصدفة متجها الى اعلى نحو اليسار الامر الذي يترتب علية زيادة المساحة تحت المنحنى حتى تصل الى الواحد صحيح والتي تعني التمييز التام للحالات. طبعا في الواقع من النادر إيجاد مساحة تحت المنحنى اكبر من 0.95. ممكن نجد مساحات تحت منحنى ROC اقل من 0.5 والتي تفسر عندها على ان النموذج له قدرة تنبئية أسواء من الصدفة (Obuchowski,2005). ويرى هوزمر وليمشو أن قيمة المساحة تحت المنحنى ROC يمكن أن تفسر على النحو التالي (Hosmer&Lemshow,2000).

ROC=0.5 النموذج ليس له قدرة تمييزية تختلف عن الصدفة
$0.7 \le ROC \le 0.8$ قدرة تميزية مقبولة
$0.8 \le ROC \le 0.9$ قدرة تميزية ممتازة
$0.9 \le ROC$ قدرة تميزية خارقة

5. تقنيات التحقق المتقاطع (Cross Validation Techniques):

يعد التحقق من المتقاطع (Cross Validation) أحد الإجراءات العامة المستخدمة في بناء النماذج الإحصائية. ويمكن استخدامه لاتخاذ قرار بشأن النموذج الإحصائي وذلك يشمل نماذج السلاسل الزمنية ونماذج الانحدار ونماذج توزيع المزيج ونماذج التميز (Chernick,2008).
تعد عملية التحقق المتقاطع طريقة إحصائية للتقييم والمقارنة، وذلك من خلال تقسيم البيانات إلى جزأين: يتم استخدام أحدهما لتقدير المعلمات أو تدريب النموذج والقسم الآخر يستخدم للتحقق من صحة النموذج واختباره. يتم إجراء التحقق من الصحة بطرق مختلفة.

في هذا البحث سوف نقوم بتطبيق (Leave-One-Out Cross-Validation)

التحقق من صحة التدفقات المتقاطعة (Leave-One-Out Cross-Validation) (LOOCV): يعتبر التحقق من صحة التدفقات المتقاطعة (LOOCV) حالةً خاصة من التحقق المتقاطع k-fold حيث k يساوي عدد أجزاء البيانات، وبعبارة أخرى يتم استخدام جميع البيانات تقريبًا باستثناء جزء واحد للتدريب ويتم اختبار النموذج المقدر على هذا الجزء من البيانات المتبقي والذي لم يدخل في بناء النموذج. من المعروف أن دقة النموذج المقدر باستخدام LOOCV غير متحيزة تقريبًا، ولكنها ذات تباين عالٍ، مما يؤدي إلى تقديرات غير موثوقة (Efron, 1983). وهذه الطريقة لا تزال تستخدم على نطاق واسع عندما تكون البيانات المتوفرة نادرة للغاية، خاصة في مجال المعلوماتية الحيوية حيث تتوفر عشرات البيانات فقط. يمكن تطبيق التحقق المتقاطع لملائمة النموذج عدد n من المرات، في كل مرة يتم استبعاد بعض المشاهدات وبعد ذلك يتم اختبار النموذج لتقدير معالم النموذج أو التنبؤ بالمشاهدات التي تم استبعادها في كل مرة، وهذا يوفر اختبارًا عادلاً من خلال اختبار مشاهدات لا تستخدم في ملائمة النموذج، كما أنه فعّال في استخدام البيانات لملائمة النموذج نظرًا لأنه في كل مرة يتم استخدام (n - 1) من المشاهدات في الملاءمة النموذج. نسبة النجاح (Hit ratio) هي النسبة

المئوية للحالات (الأفراد، المستجيبون، الشركات، إلخ) والتي تم تصنيفها بشكل صحيح حسب النموذج. يتم حسابه على أنه عدد الكائنات في قطر مصفوفة التصنيف مقسوماً على العدد الكلي (n). تُعرف نسبة النجاح أيضًا باسم النسبة المئوية للأجسام المصنفة بشكل صحيح (Hair et al. 2009).

1.5. بيانات الدراسة:

تم جمع البيانات من الفئة المستهدفة وهم الأشخاص الذين يراجعون العيادات الطبية أو المراكز الصحية سواء كانت هذه المراكز حكومية أو خاصة أو حتى تتبع لوكالة غوث وتشغيل اللاجئين الفلسطينيين، في فلسطين - قطاع غزة، والذين تزيد أعمارهم عن 25سنة وذلك خلال عام 2017. كانت العينة مكونة من232شخص مقسمين الى 172مريض سكري و60غير مريض بالسكري حيث تم جمع هذه البيانات داخل المراكز الصحية والعيادات وتم تقسيم الاشخاص إلى مجموعتين (مريض/غير مريض) وفقا لنتيجة اختبار فحص السكر في الدم، وكانت العينة موزعة على الجنسين.

2.5. البرمجيات المستخدمة:

استخدامت كل من برمجيات R وSPSS في تحليل البيانات وتطبيق الطرق الاحصائية الثلاثة.

3.5. التحليل الإحصائي للبيانات/ التحليل الوصفي للبيانات:

مقدمة: وقد تم اختيار مجموعة من الأشخاص الذين يراجعون العيادات الطبية، تم الحصول على بعض نتائج التحليل الطبية التي أجراها الأشخاص المراجعين لهذه العيادات وتم تعيين استبانة ملحقة بهذه البيانات لنفس الأشخاص.

جدول (2) يحتوي على بعض الإحصاءات الوصفية للمتغيرات الكمية في الدراسة، حيث كان الوسط creatinine للمرضى يساوي 0.91 ولغير المرضى 0.80، ومتوسط Uric Acid للمرضى يساوي4.55 ولغير المرضى يساوي 3.81، ومتوسط كرات الدم الحمراء HGB يساوي 13.4 ولغير المرضى 16.0، بينما متوسط مؤشر كتلة الجسم BMI للمرضى يساوي 33.12 ولغير المرضى 29.85، ومتوسط urea للمرضى يساوي34.1 ولغير المرضى 28.33، وأن متوسط العمر للمرضى 56.9 سنة ولغير المرضى 47.6 سنة، ومتوسط الوزن للمرضى يساوي 84.8 كجم ولغير المرضى 77.9 كجم، ومتوسط HDL للمرضى يساوي 55.7 ولغير المرضى 44.4، ومتوسط LDL للمرضى يساوي 117.16، ولغير المرضى 111.3، كما نلاحظ أن متوسط الطول للمرضى يساوي 162.2سم، ولغير المرضى 164.4 سم. كما نجد أن متوسط Cholesterol للمرضى يساوي 199.6 ولغير المرضى 189.6، الجدول التالي يوضح هذه النتائج:

جدول (2) يوضح الإحصاءات الوصفية للمتغيرات الكمية للبيانات

المتغير	فئات المتغير	N	Mini	Max	Mean	Std. Deviation
Creatinine	مريض	172	0.68	3.10	0.91	0.31
	غير مريض	60	0.71	1.07	0.80	0.09
Uric acid	مريض	172	2.90	12.00	4.55	1.49
	غير مريض	60	1.70	6.40	3.81	0.98
HGB	مريض	172	9.01	16.70	13.41	1.67
	غير مريض	16.00	16.0	16.00	16.00	16.00
BMI	مريض	172	16.6	73.60	33.12	10.70
	غير مريض	60	21.2	48.50	29.85	6.37
Urea	مريض	172	21	170	34.10	18.99
	غير مريض	60	21	43	28.33	5.10
HDL	مريض	172	36	185	55.69	24.44
	غير مريض	60	42	54	49.40	3.13
العمر	مريض	172	31	78	56.93	8.72
	غير مريض	60	35	72	47.57	9.53
الوزن	مريض	172	50	125	84.84	16.79
	غير مريض	60	50	95	77.90	12.12
LDL	مريض	172	1	252	117.16	39.41
	غير مريض	60	71	198	111.33	30.20
الطول	مريض	172	120	195	162.22	14.00
	غير مريض	60	135	185	164.40	11.72
Cholesterol	مريض	172	141	330	199.63	36.98
	غير مريض	60	147	261	189.60	27.90

الجدول رقم(3) يوضح بعض الإحصاءات الوصفية للمتغيرات الفئوية التي تم جمعها من أفراد الدراسة، حيث يوضح الجدول أن 44.2% من المرضى كن من الاناث و55.8% كانوا من الذكور، وأن 83.7% من المرضى كانوا يعانوا من الضغط النفسي المرتفع، وأن 16.3% لم يعانوا من الضغط النفسي المرتفع، 75.6% من المرضى لا يوجد لديهم خطة غذائية ، وأن 24.4% كان لديهم خطة غذائية، كما نجد أيضا أن 56.4% من المرضى يستهلكون اللحوم مرة في الأسبوع، وأن 43.6% من المرضى يستهلكون اللحوم أكثر من مرة خلال الأسبوع، 27.9% من المرضى يستهلكون الخضروات والفواكه مرة خلال الأسبوع وأن 72.1% يستهلكون الخضروات والفواكه أكثر من مرة خلال الأسبوع، 91.9% من المرضى لا يعانوا من الفشل الكلوي، وأن 8.1% كانوا يعانوا من الفشل الكلوي، 91.9% من المرضى لا يعانون من ارتفاع ضغط الدم ، وأن 8.1% كانوا يعانون

من ارتفاع ضغط الدم، 75.6% من المرض ليسوا مدخنين وأن 24.4% كانوا مدخنين، 84.9% من المرضى لا يحملون مؤهل جامعي و15.1% يحملون مؤهل جامعي. 2.3% من المرضى كانوا من محافظة خانيونس 1.2% كانوا من الوسطى، وأن 46.5% كانوا من محافظة غزة، وأن 50% كانوا من محافظة الشمال، والجدول التالي يوضح النتائج السابقة.

جدول (3) يوضح الإحصاءات الوصفية للمتغيرات الكمية للبيانات

الاسم المتغير	الفئة	مريض n=172	غير مريض n=60
الجنس	انثى	%44.2	%53.3
	ذكر	%55.8	%46.7
الضغط النفسي المرتفع	لا يوجد	%16.3	%56.7
	يوجد	%83.7	%43.3
يوجد لديه خطه غذائيه	لا يوجد	%75.6	%46.7
	يوجد	%24.4	%53.3
استهلاك اللحوم فى الأسبوع	مرة واحدة	%56.4	%66.7
	اكثر من مرة	%43.6	%33.3
كمية استهلاك الخضروات والفواكه أسبوعيا	مرة واحدة	%27.9	%3.3
	اكثر من مرة	%72.1	%96.7
فشل كلوي	لا يوجد	%91.9	%96.7
	يوجد	%8.1	%3.3
ارتفاع ضغط الدم	لا يوجد	%91.9	%96.7
	يوجد	%8.1	3.3%
مد خن	لا	%75.6	%80.0
	نعم	%24.4	%20.0
التعليم الجامعي	لا	%84.9	%60.0
	نعم	%15.1	%40.0
المحافظة	رفح	%0	%0
	خانيونس	%2.3	%43.3
	الوسطى	%1.2	%23.4
	غزة	%46.5	%13.3
	الشمال	%50	%20

4.5. الطرق الاحصائية لتحليل البيانات:

1.4.5 الانحدار اللوجستي: في القسم التالي سوف نتطرق الى الانحدار اللوجيستي لإيجاد النموذج الاحصائي الخاص بيانات مرضى السكري، ثم نقوم بعد ذلك بتقييم دقة النموذج الذي تم الحصول عليه باستخدام ثلاث طرق للتقييم وهي جدول التصنيف، ومنحنى ROC وكذلك تقنيات التحقق المتقاطع Cross Validation Techniques .

نموذج الانحدار اللوجيستي الثنائي لبيانات مرضى السكري:

- أهم المتغيرات في نموذج الانحدار اللوجيستي:

كان المتغير التابع لهذه الدراسة (Y) ويأخذ القيم (0، 1)، مريض بالسكري، وغير مريض. كما أحتوى النموذج اللوجستي على عدد من المتغيرات المستقلة والتي تعتبر أهم عوامل الخطر والتي تساهم بالإصابة بمرض السكر وهي موضحة في المعادلة التالية:

$$\log it\left[p(Y=1)\right]=\ln\left[\frac{p(Y=1)}{1-P(Y=1)}\right]=-27.7+0.262X_1+2.945X_2-0.425X_3+0.167X_4-2.226X_5+0.705X_6+1.705X_7$$

جدول (4) المتغيرات الهامة في النموذج

الرمز	اسم المتغير	الرمز	اسم المتغير
Y	المتغير التابع مرض السكري (مريض/غير مريض)	X4	العمر
X1	ضغط الدم المنخفض	X5	المؤهل العلمي
X2	الضغط النفسي المرتفع	X6	استهلاك اللحوم في الأسبوع
X3	كمية استهلاك الخضروات والفواكه أسبوعيا	X7	وجود خطة غذائية

- عوامل الخطر على مرضى السكري:

الجدول التالي يوضح معاملات النموذج مع الخطأ المعياري ونسب الأرجحية لعوامل الخطر التي تؤدي الى الاصابة بمرضى السكري.

جدول (5) معاملات النموذج المقدر والخطأ المعياري ونسبة الارجحية

	B	STD	Z-VALUE	SIG	EXP(B)
الثابت	-27.67	5.252	-5.269	0.00	0
ضغط الدم المنخفض	0.245	0.054	4.865	0.00	1.3
الضغط النفسي المرتفع	2.942	0.631	4.665	0.00	19.1
كمية استهلاك الخضروات والفواكه أسبوعيا	-0.425	0.114	-3.716	0.00	0.65
العمر	0.167	0.034	4.884	0.00	1.18
المؤهل	-2.226	0.643	-3.461	0.001	0.107
استهلاك اللحوم في الأسبوع	0.704	0.326	2.157	0.03	2.02
يوجد لديه خطة غذائية	1.705	0.591	-2.883	0.003	1.82

يوضح الجدول السابق، نتائج التحليل اللوجيستي الثنائي حيث احتوى النموذج على أهم عوامل الخطر، على مرضى السكري، كذلك يوضح الجدول السابق (Odds Ratio) لكل عامل من عوامل الخطر لمرضى السكري، حيث نلاحظ أنه كلما ابتعد ت قيمة (OR) عن الواحد الصحيح سواء بالزيادة أو النقصان كان هناك تأثير واضح لهذه المتغيرات المستقلة على المتغير التابع، حيث أظهر الجدول السابق أن متغير (الضغط النفسي المرتفع) كان له تأثير قوي جداً على مرضى السكري حيث يعتبر أهم عوامل الخطر على مرضى السكري حيث كانت قيمة (sig=0.0) وقيمة (OR=19.1)، وهذا يعني أن الشخص الذي يتعرض للضغط النفسي المرتفع، معرض للإصابة بمرض السكري 19 ضعفا مقارنة بالشخص الذي لا يعاني من هذا الضغط النفسي، في حين أن العامل الثاني كان (التعلم الجامعي) حيث كانت قيمة (sig=0.001) وقيمة (OR=0.107) وهذا يعني أن الشخص الذي ليس لديه مؤهلاً جامعياً معرض للإصابة بمرض السكري (9.3) مرة عن الشخص الذي يحمل مؤهل جامعي. في حين يوضح الجدول السابق أيضا أن العامل الثالث من حيث الأهمية هو (استهلاك اللحوم في الأسبوع) حيث كانت قيمة sig=0.000 وقيمة OR=2.02 وهذا يعني أن الشخص الذي يستهلك اللحوم أكثر من مرة في الأسبوع، معرض للإصابة ضعف الشخص الذي يستهلك اللحوم مرة واحدة خلال الأسبوع. في حين أن العامل الرابع كان (يوجد لديه خطة غذائية) حيث كانت قيمة sig=0.003 وقيمة OR=1.82 بمعنى أن الشخص الذي لا يوجد لديه خطة غذائية معرض للإصابة بمرض السكري بمقدار 1.8 عن الشخص الذي لديه خطة غذائية. في حين أن العامل الخامس كان (كمية استهلاك الخضروات والفواكه أسبوعيا) حيث نجد أن قيمته المعنوية sig=0.00، وقيمة OR= 0.65 بمعنى أن الشخص الذي يستهلك الخضروات والفواكه مرة واحدة في الأسبوع معرض للإصابة بمرض السكري بمقدار(1.5) عن الشخص الذي يستهلك الخضروات والفواكه أكثر من

مرة أسبوعياً. كذلك يوضح الجدول السابق أن ضغط الدم المنخفض كان العامل السادس من عوامل الخطر حيث كانت قيمة sig=0.00 قيمة OR=1.3 . في حين أن العمر كان كذلك عامل الخطر السابع، والذي يؤثر على مرضى السكري حيث كانت قيمة sig=0.00وقيمة OR=1.18 وهو أقل عوامل الخطر تأثيراً على مرضى السكري. وبالتالي يمكن ترتيب عوامل الخطر كالتالي من حيث الأهمية كالتالي: (الضغط النفسي المرتفع، التعليم الجامعي، استهلاك اللحوم في الأسبوع، يوجد لديه خطة غذائية، كمية استهلاك الخضروات والفواكه أسبوعيا، ضغط الدم المنخفض، العمر).

طرق تقييم النموذج:

جدول التصنيف لنموذج الانحدار اللوجستي (Logistic Regression- LR):

يوضح الجدول رقم (6) التالي نتائج التصنيف لنموذج الانحدار اللوجستي:

جدول رقم (6) التصنيف للنموذج النهائي الذي يحتوي على المتغيرات المستقلة

التصنيف	التوقع			النسبة المئوية للتصنيف الصحيح
	مريض	غير مريض	المجموع	
مريض	162	10	172	94.50%
غير مريض	12	48	60	80%
المجموع	174	58	232	90.50%

حيث نلاحظ أن نسبة التصنيف الصحيح للمرضى بلغت 94.5%؛ وكذلك نسبة التصنيف الصحيح لغير المرضى 80%؛ وبالتالي كانت نتائج التصنيف للنموذج ككل 90.5%.

ومن النتائج التي حصلنا عليها في جدول التصنيف رقم (6) يمكن الحصول على الجدول التالي:

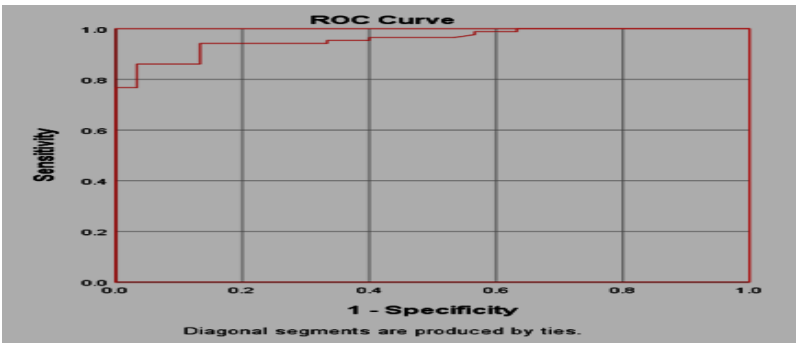
جدول رقم (7) يوضح الحساسية، معدل الدقة ومعدل الخطأ وكذلك المساحة تحت المنحنى

النموذج	Sensitivity	Specifity	Accuracy	Precision	False Positive	Error Rate	Area Roc
الانحدار اللوجيستي	0.945	0.80	0.905	0.942	0.058	0.095	0.957

نلاحظ من الجدول السابق أن دقة تصنيف النموذج للأشخاص المرضى كانت 94.5%؛ في حين دقة تصنيف الاشخاص غير المرضى 80%؛ بينما كانت دقة النموذج في التصنيف الصحيح 90.5%. مما يؤكد أن النموذج يتمتع بدقة عالية ويمكن الاعتماد عليها في

التصنيف. كما ونلاحظ أن تقارب القيم المقدرة والتي نحصل عليها من النموذج كانت عالية وبلغت 94.2%. وفي هذه الحالة عندما يكون قياس القيم المقدرة دقيقة ومحكمة(**precision**)، وكذلك دقة التصنيف(**Accuracy**) عالية فهذا مؤشر قوى على قوة النموذج وجودته. أيضا نلاحظ أن معدل الخطأ في التصنيف كان بنسبة 9.5%. في حين كان معدل الخطأ الإيجابي 5.8% بمعنى أن يكون الشخص غير مريض بالسكري ويصنف أنه مريض. كما يمكن ملاحظة أن المساحة تحت منحنىRoc بلغت حوالي 95.7% وهي نسبة جيد جداً وتعني أن النموذج استطاع أن يصنف الحالات المرضية والغير المرضية بشكل جيد ويمكن الاعتماد عليه في التصنيف.

3.2.2) منحنى Roc Curve لتقييم النموذج: الشكل التالي يوضح المساحة تحت منحنى ROC.



الشكل رقم (4) يوضح منحنىROC لنتائج الانحدار اللوجستي

يوضح الشكل السابق أن المساحة تحت منحنى ROC للنموذج تساوي (95.7%) عند مستوى دلالة (0.05) وهذا يعني أ ن النموذج يساعد على التنبؤ بتصنيف حالات المتغير التابع أكثر مما تفعله الصدفة بشكل كبير. خلاصة ما سبق: نجد أن نمذجة المتغيرات الخاصة بمرضى السكر وتصنيفهم إلى مرضى وغير مرضى باستخدام الانحدار اللوجستي أعطي نموذجاً جيداً وملائماً ويمكن الاعتماد عليه في عملية التصنيف والتنبؤ بأن يكون الشخص مصاب أو غير مصاب بمرض السكري.

2.4.5. طريقة leave-one-out cross-validation:

وفقا لهذه التقنية تم توليد 1000 عينة عشوائية بدون استبدال أو إرجاع . وقد أدرجت 7متغيرات في النموذج حيث كان النموذج مناسب لتصنيف ويمكن الاعتماد عليه في تصنيف الحالات المرضية، المصابة بمرضى السكري وفي النهاية تم تقييم النموذج وتم الحصول على الجدول التالي الذي يوضح نسبة التصنيف الصحيح وكذلك الدقة.

جدول (8) التصنيف للنموذج اللوجستي
باستخدام leave-one-out cross-validation

التصنيف	التوقع			النسبة المئوية للتصنيف الصحيح
	مريض	غير مريض	المجموع	
مريض	711	41	752	%90.1
غير مريض	78	170	248	%80.6
المجموع	789	211	1000	88.10%

يمكن أن نلاحظ من الجدول (8) أن الدقة المقدرة لنموذج الانحدار اللوجيستي لمن يعانون من مرض السكري هو 90.1%؛ في حين أن أولئك الذين لا يعانوا من مرض السكري كانت نسبة الدقة في تصنيفهم 80.6%؛ دقة التصنيف للنموذج بشكل عام كانت 88.1%. مما يؤكد دقة التصنيف باستخدام نموذج الانحدار اللوجيستي جيدة.
الجدول التالي يوضح بعض الإحصاءات الخاصة بأداء النموذج، مثل الدقة والحساسية والتي حصلنا عليها من الجدول السابق:

جدول (9) يوضح الحساسية، معدل الدقة ومعدل الخطأ للنموذج اللوجستي
باستخدام cross-validation

	sensitivity	Specificity	Accuracy	precision	false positive rate	error rate
LR	**0.901**	**0.806**	**0.881**	**0.945**	**0.054**	**0.119**

يمكن أن نلاحظ من خلال الجدول السابق أن حساسية النموذج بلغت 90.1% وهي تعبر عن نسبة التصنيف الصحيح للمرضى، بمعنى أن النموذج يستطيع تصنيف المرضى على أنهم مرضي بشكل صحيح بنسبة 90.1%. كما نلاحظ ان دقة تصنيف الحالات غير المرضى بشكل صحيح بنسبة 80.6%. ودقة تصنيف النموذج بشكل عام بلغت 88.1%. ونلاحظ أن تقارب القيم المقدرة والتي نحصل عليها من النموذج كانت عالية وبلغت 94.5%. وفي هذه الحالة عندما يكون قياس القيم المقدرة دقيقة ومحكمة (**precision**)، وكذلك دقة التصنيف (**Accuracy**) عالية فهذا مؤشر قوى على قوة النموذج وجودته.

كذلك نلاحظ أن معدل الخطأ الإيجابي 5.4% بمعنى أن يكون الشخص غير مريض بالسكري ويصنف أنه مريض. واخيرا كانت نسبة التصنيف الخاطئ للنموذج بشكل عام تساوي 11.9%.

3.4.5 التحليل التميزي الخطي: (Linear Discriminant Analysis- LDA)

تقدير النموذج التميزي: يبين الجدول التالي المعاملات غير المعيارية للارتباط بين كل متغير من المتغيرات المستقلة الداخلة في التحليل التميزي وبين المتغير التابع والذي يمثل مرض السكري (مصاب / غير مصاب).

جدول رقم (10) يوضح معاملات الدالة التميزية غير المعيارية

اسم المتغير	المتغير	المعاملات	اسم المتغير	المتغير	المعاملات
الثابت	A	-9.179	يوجد لديه خطة غذائية	X4	-0.614
العمر	X1	0.057	كمية استهلاك الخضروات والفواكه اسبوعيا	X5	-0.126
ضغط الدم المنخفض	X2	0.072	استهلاك اللحوم في الأسبوع	X6	0.233
المؤهل	X3	-0.815	الضغط النفسي المرتفع	X7	1.129

ونستخرج من الجدول أعلاه قيمة معاملات المتغيرات في الدالة التميزية، ويمكن كتابة معادلة التمييز الخطية من الجدول السابق في الصورة التالية:

$$D = -9.179 + 0.57X_1 + 0.072X_2 - 0.815X_3 - 0.614X_4 - 0.126X_5 + 0.233X_6 + 1.129X_7$$

حيث يمثل D المتغير التابع (مريض/ غير مريض).

مقاييس جودة النموذج:

جدول رقم(11) يوضح قيم الجذر الكامن

الدالة	الجزر الكامن	التباين المفسر%	التباين التجميعي%	الارتباط التجميعي
1	0.952[a]	100	100	0.698

يوضح الجدول السابق قيمة الجذر الكامن لدالة التميزية حيث كانت قيمة الجذر الكامن تساوي 0.952 بنسبة تباين مفسر يساوي 100%، في حين أن الارتباط التجمعي فقد بلغ 0.698 وهذا يدل على جودة توفيق دالة التمايز الخطية.

جدول رقم (12) يوضح اختبار Wilks' Lambda

الدالة	Wilks' Lambda	Chi-square	درجة الحرية	المعنوية
1	0.512	151.496	7	0.00

من خلال الجدول السابق نلاحظ أن قيمة Wilks' Lambda تساوي 0.512 وهي صغيرة، كذلك نلاحظ أن اختبار Chi-square يساوي 151.5 وقيمة المعنوية sig=0.0 وهذا يؤكد على جودة النموذج الذي تم الحصول عليه باستخدام التحليل التميزي. بمعنى أن النموذج جيد في فرز وتصنيف البيانات (الحالات المرضية والغير مرضية الخاصة بمرضى السكري) وتصنيفها حسب النموذج الذي حصلنا عليه من خلال التحليل التميزي.

طرق تقييم النموذج: جدول التصنيف للنموذج النهائي:

يوضح الجدول التالي التصنيف للنموذج النهائي، حيث أن هذا النموذج أظهر أن نسبة التصنيف الصحيح بلغت 87.9%. حيث يتضح من الجدول أن 93.02% من الحالات تم تصنيفهم مرضى وهم مرضى، وأن 73.3% من الحالات تم تصنيفهم غير مرضى وهم غير مرضى، في حين كانت النسبة الإجمالية للتصنيف الصحيح حوالي 87.9% وهذا يعني أن النموذج الذي حصلنا عليه جيد في التصنيف والتنبؤ بحالات مرضى السكري.

جدول رقم (13) التصنيف للنموذج النهائي الذي يحتوي على المتغيرات المستقلة

التصنيف	التوقع			النسبة المئوية للتصنيف الصحيح
	مريض	غير مريض	المجموع	
مريض	160	12	172	93.02 %
غير مريض	16	44	60	73.3%
المجموع	176	56	232	87.9%

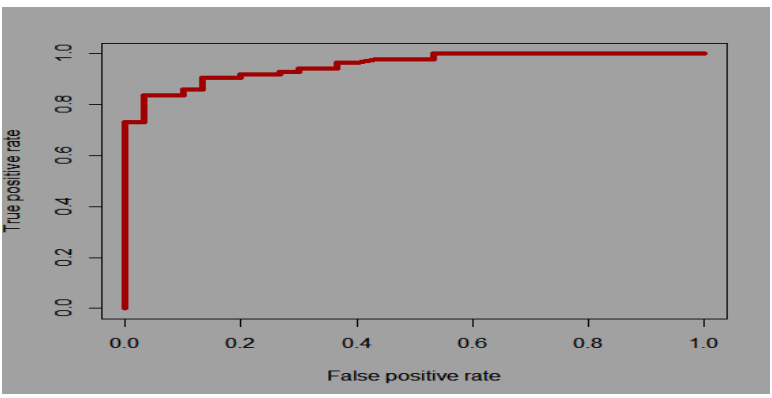
ومن النتائج التي حصلنا عليها في جدول التصنيف رقم (13) يمكن الحصول على الجدول التالي:

جدول رقم (14) يوضح الحساسية، معدل الدقة ومعدل الخطأ وكذلك المساحة تحت المنحنى

النموذج	sensitivity	Specificity	Accuracy	precision	False positive rate	error rate	Area Roc
LDA	0.909	0.733	0.879	0.930	0.070	0.121	0.953

يتضح من الجدول (14) السابق أن النموذج نجح في التصنيف بشكل صحيح لمن يعانون من مرض السكري بنسبة 90.9%؛ في حين أن أولئك الذين لا يعانون من مرض السكري كانت بنسبة 73.3%؛ بينما كان المعدل الإجمالي للتنبؤ الصحيح للنموذج، (التصنيف الصحيح) 87.9%. ونلاحظ أن تقارب القيم المقدرة والتي نحصل عليها من النموذج كانت عالية وبلغت 93%. وفي هذه الحالة عندما يكون قياس القيم المقدرة دقيقة ومحكمة (**precision**)، وكذلك دقة التصنيف (**Accuracy**) عالية فهذا مؤشر قوى على قوة النموذج وجودته. أيضا نلاحظ أن معدل الخطأ في التصنيف كان بنسبة 12.1%. في حين كان معدل الخطأ الإيجابي 7% بمعنى أن يكون الشخص غير مريض بالسكري ويصنف أنه مريض. كما يمكن ملاحظة أن المساحة تحت منحنى Roc بلغت حوالي 95.3% وهي نسبة جيد جداً وتعني أن النموذج استطاع أن يصنف الحالات المرضية والغير المرضية بشكل جيد ويمكن الاعتماد عليه في التصنيف.

منحنى Roc Curve لتقييم النموذج: الشكل التالي يوضح المساحة تحت منحنى ROC اذ انه يوضح الشكل السابق أن المساحة تحت منحنى ROC للنموذج تساوي (95.3%) عند مستوى دلالة (0.05) وهذا يعني أن النموذج يساعد على تصنيف حالات المتغير التابع بشكل جيد.



الشكل رقم (5) يوضح منحنىROC لنتائج الانحدار التميزي

الخلاصة: نجد أن نمذجة المتغيرات الخاصة بمرضى السكر وتصنيفهم إلى مرضى وغير مرضى باستخدام تحليل التمايز أعطي نموذجاً جيداً وملائماً ويمكن الاعتماد عليه في عملية التصنيف والتنبؤ بأن يكون الشخص مصاب أو غير مصاب بمرض السكر.

4.4.5. طريقة leave-one-out cross-validation

وفقا لهذه التقنية تم توليد 1000 عينة عشوائية بدون استبدال أو إرجاع . وقد أدرجت 7متغيرات في النموذج حيث كان النموذج مناسب لتصنيف ويمكن الاعتماد عليه في تصنيف الحالات المصابة بمرضى السكري وفي النهاية تم تقييم النموذج وتم الحصول على الجدول التالي الذي يوضح معدل التصنيف الصحيح للمشاهدات وكذلك الدقة.

جدول (15) التصنيف لنموذج تحليل التمايز باستخدام تقنية leave-one-out cross-validation

التصنيف	التوقع			النسبة المئوية للتصنيف الصحيح
	مريض	غير مريض	المجموع	
مريض	705	47	752	93.75%
غير مريض	81	167	248	67.34%
المجموع	786	214	1000	87.2%

يمكن أن نلاحظ من الجدول (15) أن الدقة المقدرة لنموذج تحليل التمايز لمن يعانون من مرض السكري هو 93.75%؛ في حين أن أولئك الذين لا يعانوا من مرض السكري كانت نسبة الدقة لتصنيفهم 67.34%؛ وأن التنبؤ الصحيح للتصنيف بشكل عام كان 87.2% . مما يؤكد دقة التصنيف باستخدام نموذج تحليل التمايز.

الجدول التالي يوضح بعض الإحصاءات الخاصة بأداء النموذج، مثل الدقة والحساسية:

جدول (16) يوضح الحساسية، معدل الدقة ومعدل الخطأ للنموذج االتميزي باستخدام cross-validation

	Sensitivity (The true positive rat)	Specificity	Accuracy	precision	false positive rate	error rate
LDA	**0.897**	**0.673**	**0.872**	**0.937**	**0.063**	**0.128**

يمكن أن نلاحظ من خلال الجدول السابق أن حساسية النموذج بلغت 89.7% وهي تعبر عن نسبة التصنيف الصحيح للمرضى، بمعنى أن النموذج يستطيع تصنيف المرضى على أنهم مرضي بشكل صحيح بنسبة 89.7%. كما نلاحظ ان دقة تصنيف الحالات غير المرضى بشكل صحيح بنسبة 67.3%. ودقة تصنيف النموذج بشكل عام بلغت 87.2%. ونلاحظ أن تقارب القيم المقدرة والتي نحصل عليها من النموذج كانت عالية وبلغت 93.7%. وفي هذه الحالة عندما يكون قياس القيم المقدرة دقيقة ومحكمة (**precision**)، وكذلك دقة التصنيف (**Accuracy**) عالية فهذا مؤشر قوى على قوة النموذج وجودته. كما نلاحظ أن معدل الخطأ الإيجابي 6.3% بمعنى أن يكون الشخص غير مريض بالسكري ويصنف أنه مريض. واخيرا كانت نسبة التصنيف الخاطئ للنموذج بشكل عام تساوي 12.8%.

5.4.5. تحليل الشبكات العصبية (Artificial neural networks-ANN).

تقدير نموذج الشبكات العصبية: من خلال تطبيق نموذج الشبكات العصبية الاصطناعية تم ادخال المتغيرات المستقلة التالية في النموذج وهي (**العمر، ضغط الدم المنخفض، المؤهل العلمي، يوجد لديه خطة غذائية، كمية استهلاك الخضروات والفواكه أسبوعيا، استهلاك اللحوم في الأسبوع، الضغط النفسي المرتفع**) حيث كانت نتائج التصنيف عبارة عن ثلاث مستويات من الطبقات وهي كالتالي 11طبقة ادخال و8 طبقات مخفية وطبقة واحدة للمخرجات حيث كانت الأوزان التي حصلنا عليها من خلال نموذج الشبكات العصبية الاصطناعية هي على النحو التالي:

```
summary(aa.nnet)
a 11-8-1 network with 105 weights
options were - decay=5e-04
  b->h1  i1->h1  i2->h1  i3->h1  i4->h1  i5->h1  i6->h1  i7->h1  i8->h1  i9->h1 i10->h1 i11->h1
  -0.72   -0.30   -0.55   -0.09   -0.21    0.27    0.46   -3.97   -2.02    2.26    5.93    3.99
  b->h2  i1->h2  i2->h2  i3->h2  i4->h2  i5->h2  i6->h2  i7->h2  i8->h2  i9->h2 i10->h2 i11->h2
  -0.54    9.81   -5.87    1.44    3.11   -5.14   -4.08   -0.26   -2.19    5.47    0.58    1.38
  b->h3  i1->h3  i2->h3  i3->h3  i4->h3  i5->h3  i6->h3  i7->h3  i8->h3  i9->h3 i10->h3 i11->h3
  -0.07   -2.59   -0.06    0.56   -1.82    1.59    0.36    1.86   -0.23   -1.30   -0.87   -0.84
  b->h4  i1->h4  i2->h4  i3->h4  i4->h4  i5->h4  i6->h4  i7->h4  i8->h4  i9->h4 i10->h4 i11->h4
  -0.29    2.64    2.68    5.19   -0.26   -3.09   -1.03   -0.91    1.97    3.57    6.95    3.84
  b->h5  i1->h5  i2->h5  i3->h5  i4->h5  i5->h5  i6->h5  i7->h5  i8->h5  i9->h5 i10->h5 i11->h5
   0.00   -0.01    0.00   -0.01   -0.01   -0.02    0.00    0.00    0.00    0.00    0.00    0.00
  b->h6  i1->h6  i2->h6  i3->h6  i4->h6  i5->h6  i6->h6  i7->h6  i8->h6  i9->h6 i10->h6 i11->h6
   0.00   -0.01    0.00   -0.01   -0.01   -0.03    0.00    0.00    0.00    0.00    0.00    0.00
  b->h7  i1->h7  i2->h7  i3->h7  i4->h7  i5->h7  i6->h7  i7->h7  i8->h7  i9->h7 i10->h7 i11->h7
   0.00    0.02    0.00    0.02    0.02    0.05    0.00    0.00    0.00    0.00    0.00    0.00
  b->h8  i1->h8  i2->h8  i3->h8  i4->h8  i5->h8  i6->h8  i7->h8  i8->h8  i9->h8 i10->h8 i11->h8
   0.00   -0.01    0.00   -0.02   -0.01   -0.03    0.00    0.00    0.00    0.00    0.00    0.00
 b->o h1->o h2->o h3->o h4->o h5->o h6->o h7->o h8->o
-2.61 11.64  5.52 -7.35  6.79 -0.03  0.00 -2.61  0.00
```

طرق تقييم النموذج: جدول التصنيف للنموذج النهائي باستخدام الشبكات العصبية:

يوضح الجدول التالي التصنيف للنموذج النهائي باستخدام الشبكات العصبية الاصطناعية (ANN)، اذ إن هذا النموذج أظهر أن نسبة التصنيف الصحيح بلغت 95.7%. كما هو موضح في الجدول التالي :

جدول رقم (16) التصنيف للنموذج النهائي الذي يحتوي على المتغيرات المستقلة

التصنيف	التوقع			النسبة المئوية للتصنيف الصحيح
	مريض	غير مريض	المجموع	
مريض	168	4	172	97.6 %
غير مريض	6	54	60	90%
المجموع	174	58	232	95.7%

من الجدول السابق نلاحظ أن دقة التصنيف الصحيح للحالة الاولى مريض بلغت 97.6% بمعنى تصيف المريض بشكل صحيح أنه مريض؛ وكذلك تصنيف الفئة الثانية غير مريض بشكل صحيح بلغت 90%، بمعنى تصنيف غير المريض بشكل صحيح أنه غير مريض؛ في حين كانت النسبة الإجمالية للتصنيف الصحيح للنموذج 95.7% وهذا يعني أن النموذج الذي حصلنا عليه جيد في التصنيف والتنبؤ بحالات مرضى السكري، والجدول التالي يوضح جدول بعض خصائص النموذج. ومن النتائج التي حصلنا عليها في جدول التصنيف رقم (16) يمكن الحصول على الجدول التالي:

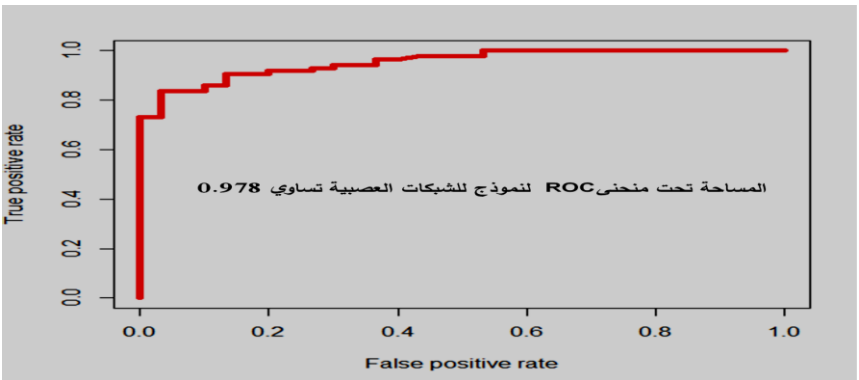
جدول رقم (17) يوضح الحساسية، معدل الدقة ومعدل الخطأ وكذلك المساحة تحت المنحنى

النموذج	Sensitivity	Specificity	Accuracy	precision	False positive rate	error rate	Area Roc
ANN	0.966	0.90	0.957	0.977	0.023	0.043	0.978

يتضح من الجدول (17) السابق أن النموذج نجح في التنبؤ بشكل صحيح بمن يعانون من مرض السكري بنسبة 96.6%؛ في حين أن أولئك الذين لا يعانون من مرض السكري كانت دقة تصنيفهم 90%؛ بينما كان المعدل الإجمالي للتنبؤ الصحيح في النموذج (التصنيف الصحيح) 95.7%؛ ونلاحظ أن تقارب القيم المقدرة والتي نحصل عليها من النموذج كانت عالية وبلغت 97.7%. وفي هذه الحالة عندما يكون قياس القيم المقدرة دقيقة ومحكمة (**precision**)، وكذلك دقة التصنيف (**Accuracy**) عالية بهذا مؤشر قوى على قوة النموذج وجودته. في حين كان معدل الخطأ الإيجابي 2.3% بمعنى أن يكون

الشخص غير مريض بالسكري ويصنف أنه مريض؛ واخيرا كانت نسبة التصنيف الخاطئ للنموذج بشكل عام تساوي 4.3%. وكانت المساحة تحت منحنى ROC تساوى 97.8% وهي قيمة عالية جداً.

منحنى Roc Curve لتقييم النموذج: الشكل التالي يوضح المساحة تحت منحنى ROCوالذي تم باستخدام نموذج الشبكات العصبية:



شكل رقم (6) يوضح منحنىROC باستخدام نموذج الشبكات العصبية

حيث يوضح الشكل السابق أن المساحة تحت منحنى ROC للنموذج تساوي (97.8%) عند مستوى دلالة (0.05) وهذا يعني أ ن النموذج يساعد على التنبؤ بتصنيف حالات المتغير التابع أكثر مما تفعله الصدفة بشكل كبير جداً.

وخلاصة ما سبق: أن نمذجة المتغيرات الخاصة بمرضى السكر وتصنيفهم إلى مرضى وغير مرضى باستخدام تحليل الشبكات العصبية الاصطناعية أعطي نموذجاً ملائماً لتصنيف البيانات. وبالتالي يمكن الاعتماد على هذا النموذج من أجل التصنيف والتعرف على احتمال أن يكون الشخص مصاب أو غير مصاب بمرض السكري.

6.4.5. طريقة leave-one-out cross-validation

وفقا لهذه التقنية تم توليد 1000 عينة عشوائية بدون استبدال أو إرجاع وقد أدرجت 7متغيرات في النموذج حيث كان النموذج مناسب لتصنيف ويمكن الاعتماد عليه في تصنيف الحالات أما مصابة او غير مصابة بمرضى السكري وفي النهاية تم تقييم المشاهدات المتوقعة وتم الحصول على معدل التصنيف الصحيح للمشاهدات المتوقعة وثم الحصول على جدول تصنيف النتائج وكذلك الدقة التصنيف باستخدام هذه التقنية:

جدول (18) التصنيف للنموذج باستخدام الشبكات العصبية باستخدام تقنية leave- one-out cross-validation

التصنيف	التوقع			النسبة المئوية للتصنيف الصحيح
	مريض	غير مريض	المجموع	
مريض	699	43	742	94.2%
غير مريض	45	213	258	82.55%
المجموع	744	256	1000	91.2%

أن نستنتج من الجدول (18) أن الدقة المقدرة باستخدام نموذج الشبكات العصبية الاصطناعية (ANN) لمن يعانون من مرض السكري بشكل صحيح هو 94.2%؛ في حين أن أولئك الذين لم يعانوا من مرض السكري كانت نسبة الدقة في تصنيفهم 82.55%؛ أما التصنيف الصحيح للنموذج ككل بلغت 91.2%. هذا يؤكد دقة النموذج في التصنيف. كما يمكن أن استخلاص بعض خصائص النموذج من الجدول السابق وهي موضحة في الجدول التالي:

جدول(19) يوضح الحساسية، معدل الدقة ومعدل الخطأ للنموذج اللوجستي باستخدام cross-validation

النموذج	Sensitivity	Specificity	Accuracy	Precision	False Positive Rate	Error rate
ANN	0.940	0.826	0.912	0.942	0.058	0.088

من الجدول السابق نلاحظ أن حساسية النموذج 94% وهي تُعبر عن نسبة التصنيف الصحيح للمرضى، بمعنى أن النموذج يستطيع تصنيف المرضى على أنهم مرضي بشكل صحيح بنسبة 94%. كما نلاحظ ان دقة تصنيف الحالات غير المرضى بشكل صحيح بنسبة 82.6%. ودقة تصنيف النموذج بشكل عام بلغت 91.2%. ونلاحظ أن تقارب القيم المقدرة والتي نحصل عليها من النموذج كانت عالية وبلغت 94.2%. وفي هذه الحالة عندما يكون قياس القيم المقدرة دقيقة ومحكمة (precision)، وكذلك دقة التصنيف (Accuracy) عالية بهذا مؤشر قوى على قوة النموذج وجودته. في حين كان معدل الخطأ الإيجابي 5.8% بمعنى أن يكون الشخص غير مريض بالسكري ويصنف أنه مريض.

6. المقارنة بين الطرق الاحصائية الثلاثة (ANN, LDA, LR):

من خلال تطبيق الطرق الإحصائية الثلاثة وهي الانحدار اللوجيستي الثنائي، وتحليل التمايز الخطي، تحليل الشبكات العصبية الاصطناعية على بيانات مرضى السكري في قطاع غزة حصلنا على عدة نتائج يمكن الاعتماد عليها. الان سوف نقوم بالمقارنة بين هذه الطرق الثلاثة من خلال دقة التصنيف والمساحة تحت منحنى Roc بالإضافة الى نتائج تطبيق(Leave-One-Out Cross-Validation).

1.6. المقارنة بين الطرق الإحصائية باستخدام دقة التصنيف:

الجدول (20) يوضح نتائج دقة التصنيف للطرق الاحصائية الثلاثة

النموذج	sensitivity	Specificity	Accuracy	Precision	False positive rate	error rate
LDA	0.909	0.733	0.879	0.930	0.070	0.121
LR	0.931	0.800	0.905	0.942	0.058	0.095
ANN	0.966	0.900	0.957	0.977	0.023	0.043

يوضح الجدول السابق دقة التصنيف الصحيح لنموذج (**Accuracy**) التحليل التميزي كانت الاقل حيث بلغت 87.9%؛ وجاء الانحدار اللوجيستي في المرتبة الثانية حيث بلغت دقة تصنيفه 90.5%؛ وكان نموذج الشبكات العصبية هو الافضل حيث بلغت دقة النموذج في التصيف 95.7%. وبالتالي نستنتج أن نموذج الشبكات العصبية هو النموذج الافضل في تصنيف بيانات المتوفرة لدينا (بيانات مرض السكري في قطاع غزة) حيث حصل على أعلى دقة وأقل معدل خطأ.

2.6. المقارنة حسب المساحة تحت منحى ROC:

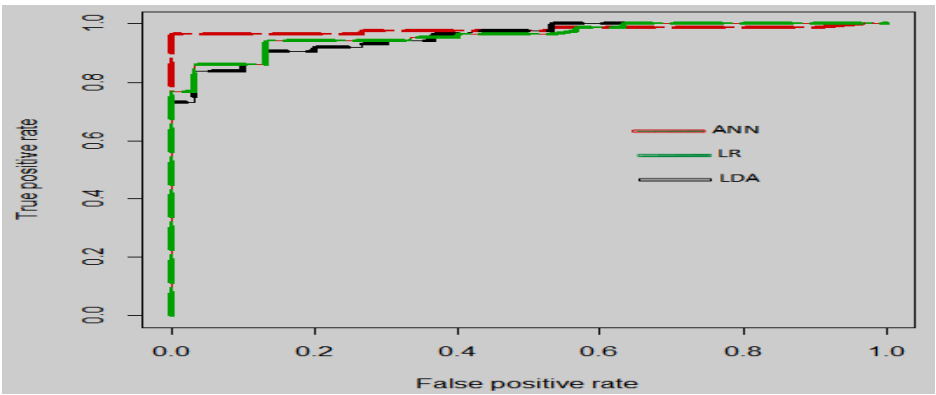
الجدول (21) يوضح المساحة تحت منحنى ROC للطرق الاحصائية الثلاثة:

النموذج	LDA	LR	ANN
المساحة تحت المنحنى	0.953	0.957	0.978

نلاحظ أن المساحة تحت منحنى Roc في نموذج التحليل التميزي تساوي 95.3%؛ أما في نموذج الانحدار اللوجيستي تساوي 95.7%؛ كما نلاحظ أن المساحة تحت منحنى

Roc في نموذج الشبكات العصبية الاصطناعية تساوي 97.8%، وهي الاعلى. بمعنى أن نموذج الشبكات العصبية هو الافضل في تصيف البيانات المتوفرة لدينا.

والشكل التالي يوضح المساحة تحت المنحنى للنماذج الثلاثة:



الشكل رقم (7) يوضح المساحة تحت منحنى ROC
للطرق الاحصائية الثلاثة ANN و LDA و LR

3.6. المقارنة باستخدام تقنية (Leave-One-Out Cross-Validation): وتعتبر تقنية Leave-One-Out Cross-Validation احدى التقنيات التي يعتمد عليها في تقيم النماذج الإحصائية حيث يمكن الحصول على بعض المؤشرات التي يمكن من خلالها تقييم هذه النماذج الإحصائية، حيث تلاحظ من الجدول رقم (22) أن نسبة الخطأ في التصنيف لنموذج التحليل التميزي 12.8%؛ ونسبة الخطأ في التصنيف لنموذج الانحدار اللوجيستي يساوي 11.9% بينما كانت نسبة الخطأ في التصنيف لنموذج الشبكات العصبية يساوي 8.8%. كذلك نلاحظ أن نسبة التصنيف الصحيح لنموذج التحليل التميزي 87.2%، وللنموذج اللوجيستي يساوي 88.1% ولنموذج الشبكات العصبية يساوي 91.2%.

جدول (22): يوضح الحساسية، والدقة، معدل الخطأ، ومعدل والخطأ الإيجابي، باستخدام: Leave-One-Out Cross-Validation

	Sensitivity	Specificity	Accuracy	Precision	False positive rate	error rate
LDA	**0.897**	**0.673**	**0.872**	**0.938**	**0.063**	**0.128**
LR	**0.901**	**0.685**	**0.881**	**0.945**	**0.055**	**0.119**
ANN	**0.940**	**0.826**	**0.912**	**0.942**	**0.058**	**0.088**

كما نلاحظ أن تقارب القيم المقدرة والتي نحصل عليها من نموذج التحليل التميزي بلغت 93.8%، ومن النموذج اللوجستي بلغت 94.5%؛ ومن نموذج الشبكات العصبية 94.2%. نلاحظ هنا أنه ووفق (**precision**) كانت نتائج الانحدار اللوجستى هو الافضل لكن بشكل عام ووفق طرق المقارنة المختلفة كان نموذج الشبكات العصبية هو الافضل لتصنيف البيانات المتوفرة لدينا وهي (بيانات مرضى السكر في قطاع غزة).

النتائج والتوصيات:

أهم النتائج:

تبين من خلال تحليل البيانات الخاصة بمرضى السكري أن أهم عوامل الخطر على مرضى السكري كانت مرتبة حسب درجة تأثيرها على مرضى السكري وهي كالتالية مرتبة حسب الأهمية: الضغط النفسي المرتفع وحصل على المرتبة الأولى من حيث التأثير على مرضى السكري، التعلم الجامعي وحصل على المرتبة الثانية من حيث درجة التأثير، في حين كمية استهلاك اللحوم في الأسبوع حصل على المرتبة الثالثة، يليه هل يوجد خطة غذائية لدى الشخص وحصل هذا المتغير على المرتبة الرابعة، كذلك كمية استهلاك الخضروات والفواكه المستهلكة أسبوعيا حصلت على المرتبة الخامسة من ناحية التأثير على مرض السكري، في حين أن ضغط الدم المنخفض كان له أثر واضح على مرضى السكري، حيث حصل على المرتبة السادسة من حيث الأهمية، وأخيراً كان متغير العمر أحد عوامل الخطر التي تكون لها أثار على مرضى السكري. وفي المجمل

استطعنا الحصول على أهم العوامل الخطر على مرضى السكري وذلك من خلال استخدام عدة نماذج إحصائية وهي (الانحدار اللوجيستي الثنائي، وتحليل التمايز، والشبكات العصبية الاصطناعية) حيث أعطيت جميع هذا النماذج نتائج جيد جداً في التنبؤ بحالة المريض المصاب بمرض السكري الا ان نموذج الشبكات العصبية الاصطناعية كان الأفضل مقارنة بالنماذج الاخرى، كما تم استخدام ثلاث طرق إحصائية متطورة للتأكد من ملائمة هذه النماذج لتصنيف البيانات الخاصة بمرض السكري. وكانت النتائج قوية وتدعم استخدام نموذج الشبكات العصبية على اعتبار انه حصل على افضل نموذج تصنيفي في تحليل هذه البيانات والحصول على نتائج ذات مدلولات قوية. حيث كانت المساحة تحت منحنى ROC تساوي(97.8%) وفي حين أن جدول التصنيف كانت نسبة التصنيف الصحية تساوي (95.7%)

في حين أن جدول Cross validation أعطى نسبة تصنيف صحيحة تساوي (91.2%) وهذه المؤشرات تدعم الرأي القائل بأن نموذج الشبكات العصبية الاصطناعية جيد في تصنيف الحالات المرضية والغير مرضية، الخاصة ببيانات مرضى السكري في قطاع غزة. والنموذج النهائي الذي حصلنا عليه كان حسب المعادلة التالية.

التوصيات:

من النتائج السابقة يمكن أن نسرد بعض التوصيات الخاصة بمرضى السكري:

1. تقليل الضغط النفسي المرتفع، حيث أثبتت الدراسة أنه أحد أهم الأسباب التي تؤدي إلى مرض السكري، وقد يصنف الضغط النفسي المرتفع في المستقبل إلى السبب الرئيسي الغير مباشر للوفاة في فلسطين، نتيجة ما يترتب عليه من أعراض أو أمراض كثيرة، تؤدي في نهاية المطاف إلى الوفاة.
2. نوصي وزارة الصحة والمراكز الصحية بالاهتمام بالجانب النفسي لجميع المرضى وخصوصاً مرضى السكري لما لهذا المرض من أثر فتاك وقاتل والذي قد يعتبر القاتل الرئيس في المستقبل.
3. تنظم ضغط الدم، اتخاذ كافة السبل للحيلولة دون عدم انتظامه وتحديداً ضغط الدم المنخفض لما له من أثر كبير على زيادة نسبة الإصابة بمرض السكري.
4. حث مرضى السكري بزيادة استهلاك الخضروات والفواكه حيث أثبتت هذه الدراسة أن الزيادة من كمية استهلاك الفواكه والخضروات تؤثر بشكل إيجابي على تقليل نسبة الإصابة بمرض السكري.
5. نوصي مرضى السكري في الحد من استهلاك اللحوم حيث أن زيادة استهلاك اللحوم يؤثر سلباً على زيادة معدل الإصابة بمرض السكري.
6. ننصح مرضى السكري بوضع خطة غذائية مناسبة لما لذلك من أثار إيجابية على الحد من الإصابة بمرضى السكري.
7. ننصح الشباب بمتابعة دراستها الجامعية حيث أن ذلك يزيد من توعيتهم بمخاطر مرض السكري وبالتالي المحافظة على صحتهم الأمر الذي ينعكس إيجابياً على الحد من الإصابة بهذا المرض.
8. نوصي وزارة الصحة بتوفير قاعدة بيانات الكترونية جيدة لجميع المرضى والمصابين بالأمراض وكذلك المراجعين حيث يمكن استخدامها في مجال البحث، وإعطاء أهمية كبيرة لمرض السكري.
9. نوصي وزارة الصحة بتأسيس مستشفى للطبيب النفسي والاكلينيكي والتدخل المبكر لمعالجة المرضى الذين من المتوقع إصابتهم بأمراض مختلفة نتيجة الضغط النفسي لتلافي ذلك في المستقبل.
10. نوصي الباحثين بإجراء المزيد من الأبحاث حول مرض السكر للمساعدة في تحديد كل العوامل الخطورة المسببة لمرض السكر واستخدام طرق إحصائية أخرى غير المستخدمة في هذا البحث وكذلك تضمين متغيرات أخرى لها علاقة بمرض السكر.

المراجع

1) بابطين، عادل بن احمد بن حسن(1429 هـ): الانحدار اللوجستي وكيفية استخدامه في بناء نماذج التنبؤ للبيانات ذات المتغيرات التابعة الثنائية القيمة. بحيث مقدم للحصول على درجة الدكتوارة – تخصص احصاء وبحوث .
2) Chatellier, Gilles (1998). Logistic Regression Model: Conditions Required for Stability of Prediction. Medical Informatics Department, Broussais Hospital, Paris, France.
3) Chernick, M. R. (2008), "Bootstrap Methods: A Guide for Practitioners and Researchers", Second Edition, Wiley, Inc., New York.
4) Efron B. (1983), "Estimating the error rate of a prediction rule: improvement on cross-validation". J. Am. Stat. Assoc., 78:316–331.
5) Electronic text book, StatSoft , Inc., 1984- 2000, "Discriminant Function Analysis" , http://www.uta.edu/faculty/sawasthi/Statistics/stdiscan.html#index
6) Fraas, Jojan W. & Newman Isadore (2003). Ordinary Least Squares Regression, Discriminant Analysis, and Logistic Regression: Question Researchers and Practitioners Should address When Sclecting an Analytic Technique. Paper Presented at the Annual Meeting of the Education Research Association (Hilton Heal Island, GA, February 26-March 1, 2003).
7) Fawcett, Tom (2005). An Introduction to ROC Analysis. Pattern Recognition Letters. No.27,2006,861-874.
8) Ferrer, Alvaro J. Arce & 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).
9) Hair, J. F., Anderson, R. E., Babin, B. J., & Black, W. C. (2009), "Multivariate Data Analysis", 7th Edition, Maxwell Macmillan, International, New York.
10) Hosmer, D. W. and Lemeshow, Stanley. (2000), "Applied logistic regression", 2nd Edition. Published by Johan Wiley and Sons, Wiley, New York.
11) Golden, R. M. (1996), "Mathematical methods for neural network analysis and design. USA: Massachusetts Institute of Technology", From ozean journal of applied science Volume 3, Issue 2.

12) Johnson, R.A. and Wichern, D.W. (2007) Applied Multivariate Statistical Analysis., Scientific Research an Academic Publisher Journal, 6th Edition.
13) 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.
14) Obuchowski, Nancy A. (2005). Fundamentals of Clinical Research for Radiologists ROC Analysis. American Roeutgen Ray Society. No.184, February 2005, 364-372.
15) Pai., Dinesh R., (2009), "Determining the Efficacy of Mathematical Programming Approaches for Multi-Group Classification", A Ph. D. Dissertation, The Graduate School-Newark Rutgers, The State University of New Jersey.
16) Pample, Fred C. (2000). Logistic Regression Aprimer. Sage University Paper series on
17) Poulsen, John and French, Aaron. (1999), "Discriminant Function Analysis ". online: http://userwww.sfsu.edu/efc/classes/biol710/discrim/discrim.pdf
18) Quantitative Applications in the Social Sciences. No.07-132. Beverly Halls, CA: Sage.
19) Shmueli, Galit and Nitin, Patel R. and Bruc, C. Peter. (2010), "Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner". 2nd Edition, published by John Wiley and Sons,Inc., Hoboken N.J.
20) Schmidt, Amy Elizabeth (2000). An Approximation of a Hierarchical Logistic Regression Model Used Establish the Predictive Validity of Scores on A Nursing Licensure Exam. Educational and Psychological Measurement, Vol.60, No.3, June2000, 463-478.
21) Shmueli, Galit and Nitin, Patel R. and Bruc, C. Peter. (2010), "Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XL Miner". 2nd Edition, published by John Wiley and Sons, Inc., Hoboken N.J.
22) Schüppert, Anja (2009), "Binomial (or Binary) Logistic Regression" online:

http://www.let.rug.nl/~nerbonne/teach/rema-stats-meth-seminar/presentations/Binary-Logistic-Regression-Schueppert-2009.pdf

23) Sharda, R. (1994), "Neural networks for the MS/OR analyst: An application bibliography". Interfaces. 24(2) 116-130.
24) Soderstrom, Irina R. & Leitner, Dennis W. (1997). The Effects of Base Rate, Selection Ratio, Sample Size, and Reliability of Predictors on Predictive Efficiency Indicies Associated with Logistic Regression Models. Paper Presented at the Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL, October 15-18, 1997).
25) Soderstrom, Irina R. & Leitner, Dennis W. (1997). The Effects of Base Rate, Selection Ratio, Sample Size, and Reliability of Predictors on Predictive Efficiency Indicies Associated with Logistic Regression Models. Paper Presented at the Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL, October 15-18, 1997).
26) Tabachnick, Barbara G. and Fidell, Fidell, Linda S. (1996)," Using multivariate statistics", 3rd edition, Publisher: Harpercollins College Publishers (New York, Ny).
27) Walker,Marilyn D.(1998).Discriminant Function Analysis. Lesson8.
28) Westin, Lena Kallin (2005). Receiver Operating Characteristic (ROC) Analysis Evaluating Discriminance EffortsAmong Decision Support Systems. ISSN-0348-0542.
29) (WHO) World Health Organization, Geneva, 2016, Global report on diabetes.