استخدام أساليب التعلم الآلي للتنبؤ بمرض السكري

Using Machine Learning Techniques For Prediction Diabetes

تحت اشراف

أ.د/ فاطمة علي محمد عبد العاطي

أستاذ الإحصاء التطبيقي

كلية التجارة-جامعة المنصورة

أ.د/ هشام محمد رجب المنجي

أستاذ الإحصاء المساعد

كلية التجارة-جامعة المنصورة

إعداد

ساره شعبان المتولي

المستخلص

يعتبر مرض السكري من الأمراض غير المعدية المعروفة في العالم، قدر على أنه السبب السابع للوفاة، يتسبب مرض السكري في وفاة عدد كبير من الأشخاص كل عام ولا يدرك عدد كبير من الأشخاص المصابين بالمرض حالته الصحية مبكرا بما فيه الكفاية Zhu et al (2019). يتم في هذا البحث اقتراح نموذج للتنبؤ بمرض السكري في مرحلة مبكرة باستخدام أساليب التعلم الآلي، يتكون النموذج من استخدام نموذج الانحدار اللوجستي (تعلم آلي باشراف) ويتم أيضا تحسين دقته عن طريق استخدام تحليل المكونات الرئيسية (تعلم آلي بدون اشرف) لتقليل الأبعاد بدون فقد الكثير من المعلومات قبل استخدام نموذج الانحدار اللوجستي.

الكلمات الافتتاحية

التعلم الآلي، التعلم بإشراف، التعلم بدون اشراف، الانحدار اللوجستي، تحليل المكونات الرئيسية.

Abstract

Diabetes is well-known non-transmittable diseases in the world. It is assessed to be the seventh leading cause for death, Diabetes causes a large number of deaths each year and a large number of people living with the disease do not realize their health condition early enough Zhu et al (2019). in this paper, a model for predicting diabetes at an early stage using machine learning methods is proposed, The model consists of using the Logistic Regression Model (supervised machine learning) and it is also improved accuracy by using principle component analysis (un supervised machine learning) to reduce the dimensions without losing a lot of information before using the logistic regression model.

Key words

Machine learning, supervised learning, unsupervised learning, logistic regression, principle component analysis

المقدمة

التعلم الآلي هو عبارة عن مجال علمي يتعامل مع الطرق التي تتعلم بها الآلات من التجربة Joshi & Chawan (2018) وهو عبارة عن نوع من أنواع الذكاء الإصطناعي Artificial intelligence يمكن النظام من التعلم بنفسه ويطور نماذج المعرفة Knowledge models لاتخاذ قرار بالتنبؤ ببيانات غير معروفة أو تصنيف بيانات معينة Singh et al (2017). يعد التعلم الآلى أحد التقنيات سريعة النمو في آخر 15 عام، له العديد من التطبيقات في مختلف المجالات مثل الرؤية الحاسوبية computer vision، المعلومات الحيوية bioinformatics، تحليل الأعمال business analytics، الرعاية الصحية healthcare، القطاع المصرفي banking sector، الكشف عن الاحتيال fraud detection، التنبؤ بالاتجاهات prediction of trends وغير ذلك من المجالات Reddy et al (2020).

الانحدار اللوجستي هو عبارة عن خوارزمية تعلم آلي بإشراف Zhu et al(2 019) هذه الخوارزمية شائعة للتصنيف الثنائي De Cock et al (2021) وهو عبارة عن نموذج انحدار يكون فيه المتغير التابع فئوي بمعني أنه يأخذ قيمتين فقط 0 و 1 والتي تمثل النتائج مثل النجاح والفشل, الفوز والخسارة, على قيد الحياة وميت, معافى ومريض Joshi & Chawan (2018) والعوده والشفاء على عكس المتغير التابع يمكن تطبيق المتغيرات المستقلة في كل من الأشكال الفئوية والمستمرة وذلك بافتراض عدم وجود توزيعات طبيعية Seo et al (2020). الانحدار اللوجستي مهم من الناحية النظرية والتطبيق للتعلم الآلي الحديث، يتم استخدامه لمهام مثل تصنيف الفئات category classification، توقى نسبة النقر إلى الظهور click-through-rate prediction وتقييم الخطر risk assessment. يتكون النموذج من مجموعة من المتغيرات والتي تؤثر معلماتها تأثير على بعض النتائج Jacquet et al (2021).

أثناء تحليل البيانات يكون من الصعب جدا الحصول على جميع العلاقات بين البيانات، يسمح تحليل المكونات الرئيسية بتحويل كمية كبيرة من المعلومات في البيانات المترابطة الأولية إلى مجموعة جديدة من المركبات المتعامدة، مما يحل من مشكلة الارتباط التي تجعل من الصعب على خوارزمية التصنيف أن تجد العلاقات بين البيانات فيساعد تطبيق تحليل المكونات الرئيسية على فلترة المتغيرات غير المهمة فيؤدي ذلك إلى تخفيض وقت التدريب، التكلفة ويؤدي إلى زيادة أداء النموذج Zhu et al (2019).

تحليل المكونات الرئيسية في مجال التعلم الآلي وعلوم البيانات هو عبارة عن أسلوب تعلم غير خاضع للإشراف ، يتم استخدامه لتقليل الأبعاد حيث أنه يستخدم التحويل المتعامد ويسمح بتحديد الأنماط والارتباطات في مجموعة البيانات للتحويل إلى مجموعة بيانات ذات بعد أقل بدون فقد أي معلومات مهمة Sarker et al (2020)، يتم تعريفه رياضيا على أنه تحويل خطي متعامد حيث يحول البيانات إلى نظام إحداثي جديد بحيث يقع اسقاط البيانات على الإحداثيات الجديدة التي تسمى بالمكون الرئيسي، فيسمى أكبر تباين باسقاط البيانات على الإحداثي الأول المكون الرئيسي الأول، ثاني أكبر تباين على الإحداثي الثاني بالمكون الرئيسي الثاني وهكذا Ji et al (2020) وهو أيضا عبارة عن إجراء إحصائي يستخدم التحويل المتعامد حيث أنه يقوم بتحويل مجموعة من المتغيرات المترابطة إلى مجموعة من المتغيرات غير المترابطة، يستخدم في تحليل البيانات الاستكشافية exploratory data analysis، فحص العلاقات بين مجموعة من المتغيرات وأيضا يمكن استخدامه لتقليل الأبعاد Reddy et al (2020).

(1) الاستعراض المرجعي

استخدمت دراسة **Jhaldiyal & Mishra (2014)** أسلوبين مختلفين مع استخدام تحليل المكونات الرئيسية الأسلوب الأول هو استخدام آلة المتجهات الداعمة SVM والأسلوب الثاني هو استخدام PEP و توصلت هذه الدراسة إلى نسبة 93.66% مع الأسلوب الأول و نسبة 79.93% مع الأسلوب الثاني. استخدم Mahajan et al (2017) أسلوب الشبكات العصبية بمفردة واستخدم تحليل المكونات الرئيسية لتقليل الأبعاد قبل استخدام أسلوب الشبكات العصبية، توصلت هذه الدراسة تحسين الدقة باستخدام أسلوب المكونات الرئيسية من 72.9% إلى 92.2%.تناولت دراسة **Islam et al (2020)** استخدام خوارزمية نايف بايز وخوارزمية الانحدار اللوجستي وخوارزمية الغابة العشوائية وتم بعد ذلك تطبيق أساليب التحقق المتبادل ذات العشرة أضعاف و percentage split technique وكانت أفضل دقة تم الحصول عليها كانت باستخدام خوارزمية الغابة العشوائية حيث تم تصنيف 97.4% من الحالات بشكل صحيح باستخدام التحقق المتبادل ذات العشرة أضعاف وتم تصنيف 99% من الحالات بشكل صحيح باستخدام percentage split technique. استخدمت دراسة Zhu et al (2019) استخدام أسلوب المكونات الرئيسية، وذلك بهدف تحسين نتائج k-means clustering ودقة الانحدار

اللوجستي ، بلغ تطبيق k-means clustering قبل الانحدار اللوجستي 82% أما عند استخدام الأسلوبين معا بلغت الدقة 97%.

(2) النماذج محل الدراسة

أولا: نموذج الانحدار اللوجستي Logistic regression model

يستند نموذج الانحدار اللوجستي على نموذج الانحدار الخطي المعطى من العلاقة

$$y = h_\theta(x) = \theta^t x \qquad (1)$$

ولكن هذه المعادلة غير مفيدة بشكل كبير للتنبؤ بالقيم الثانية $\left(y^{(i)}\varepsilon\{0,1\}\right)$ لذلك فإنه يتم تقديم دالة للتنبؤ باحتمال أن مريض معين ينتمي إلى الفئة 1 (إيجابي) مقابل احتمالية أن ينتمي هذا المريض إلى الفئة 0 (سلبي).

$$\mathrm{P}(y = 1|x) = h_\theta(x) = \frac{1}{1+exp(-\theta^t x)} = \sigma(\theta^t x) \qquad (2)$$

$$\mathrm{P}(y = 0|x) = 1 - \rho(y = 1|x) = 1 - h_\theta(x) \qquad (3)$$

وبتطبيق المعادلة

$$\sigma(\mathrm{t}) = \frac{1}{(1+\mathrm{e}^{-1})} \qquad (4)$$

التي تعرف باسم الدالة السينية sigmoid function تظل القيمة $\boldsymbol{\theta^t x}$ داخل المجال [0,1] ثم يتم البحث عن قيمة θ بحيث أن احتمالية أن $\mathrm{P}(y = 1|x) = h_\theta(x)$ تكون كبيرة عندما تنتمي x إلى الفئة 1 وصغيرة عندما تنتمي إلى الفئة صفر أي أن $\mathrm{P}(y = 0|x)$ صغيرة.

ثانيا: تحليل المكونات الرئيسية Principal component analysis

الهدف من استخدام تحليل المكونات الرئيسية هو تحويل مجموعة البيانات X ذات البعد P إلى مجموعة جديدة Y ذات بعد أقل L $(L < P)$ حيث أن

$$Y = PC(x) \qquad (5)$$

يتم ذلك باتباع الخطواط التالية

1. حساب المتوسط

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (6)$$

2. حساب التباين

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)} \qquad (7)$$

3. حساب التغاير

$$X^{n \times n} = (x_{ij}, x_{ij}) = cov(D_i m_i, D_i m_j) \qquad (8)$$

حيث أن

- $X^{n \times n}$ هي مصفوفة البيانات تحتوي على n من الصفوف و n من الأعمدة.
- $D_i m_j$ هي عبارة عن البعد.

4. حساب المتجهات المميزة والجذور المميزة.

إذا كانت A هي مصفوفة $n \times n$ فإن X هي متجه غير صفري تسمى متجه مميز.

إذا كانت AX هي عدد متعدد في x هذا يعني أن

$$AX = \lambda x \qquad (9)$$

بحيث أن

- λ تسمى بالقيم المميزة.

- x تسمى بالجذور المميزة المقابلة ل λ.

نظرا لأن المتجهات المميزه تكون مقابلة للجذور المميزة للمصفوفة A فإن المتجه غير الصفري يفي بالمعادلة

$$(\lambda \mathrm{I} - A)x = 0 \qquad (10)$$

5. يتم بعد ذلك تحديد المجموعة E لكل المتجهات X التي تفي بالمعادلة السابقة

$$E = \{x: (A - \lambda \mathrm{I})x\} = 0 \qquad (11)$$

6. يتم بعد ذلك الحصول على المتجهات المميزة بواسطة القيم المميزه الاعلى ثم الأقل، يؤدي ذلك إلى إزالة المكون الأقل أهمية وترك المكون الرئيسي الذي يوفر تقريب للبيانات الأصلية **Zhu et al (2019)**.

(3) تقييم النماذج

الدقة

هي عبارة عن النسبة المئوية للتنبؤات الصحيحة التي قام بها المصنف عند مقارنتها بالقيم الفعلية للتصنيف في مرحلة الاختبار.

$$Accuracy = (TN + TP)/(TN + TP + FN + FP) \qquad (12)$$

حيث أن **TP** (true positive) هي إيجابي حقيقي، **TN** (true negative) هي سلبي حقيقي، **FP** (false positive) هي إيجابي زائف، **FN** (false negative) هي سلبي كاذب Reddy et al (2020).

(4) الجانب التطبيقي:

يتناول الجانب التطبيقي بيانات تم الحصول عليها من مستشفى الباطنة التخصصي بجامعة المنصورة، تم جمع هذه البيانات عن طريق استخدام استبيان مباشر من الأشخاص الذين أصيبوا مؤخرا بمرض السكري أو الغير مصابين بالمرض ولكن لديهم الكثير أو القليل من الأعراض، حيث حصلت الباحثة على (203) مريض

تم حذف المرضى الذين تحتوي البيانات الخاصة على فقد بالنسبة لمتغير الاستجابة فأصبح عدد المرضى (185) مريض وتم التعويض عن باقي القيم المفقودة باستخدام المتوسط الأقرب حيث كانت نسبة المفقود أقل من 1%.

وصف المتغيرات المستخدمة في البحث:

متغير الاستجابة

متغير الاستجابة الذي تم اكتشافه في هذه الدراسة هو الإصابة بمرض السكري (نعم ، لا)، يرمز للمصاب في هذه الدراسة بالرمز 1 أما الغير مصاب فيرمز له بالرمز0.

المتغير المستقل

المتغيرات المستقلة في هذه الدراسة هي عبارة عن مجموعة من المتغيرات تتعلق بأعراض مرتبطة بمرض السكري بمعني أنها ربما تسبب الإصابة بمرض السكري وبعض التحاليل الخاصة به وأيضا بعض المتغيرات الاقتصادية والاجتماعية والديموغرافية فيتم أخذ قياسات 35 متغير لكل مريض ويتم التعريف بهذه المتغيرات من خلال الجدول التالي:

جدول (1) التعريف بالمتغيرات المستقلة

المتغير	الاسم باللغة الانجليزية	الاسم باللغة العربية	التعريف
X_1	Gender	النوع	يأخذ القيمة 1 إذا كانت أنثى والقيمة صفر إذا كان ذكر.
X_2	Age	العمر	يمثل السن لكل مريض.
X_3	Country	البلد	تأخذ القيمة 1 إذا كانت المنطقة قرية والقيمة 2 إذا كانت المنطقة مدينة.

تابع جدول (1) التعريف بالمتغيرات المستقلة

X_4	Marital Status	الحالة الاجتماعية	تأخذ ثلاث مستويات أعزب يرمز له بالرمز 1، متزوج يرمز له بالرمز 2، مطلق يرمز له بالرمز 3 وأرمل يرمز له بالرمز 4.
X_5	Numper of time pregnant	عدد مرات الحمل	تمثل عدد مرات الحمل لكل أنثى.
X_6	Employment status	الوظيفة	تأخذ 3 مستويات فيرمز للايعمل بالرمز 1، يعمل بالرمز 2 وكان يعمل بالرمز 3.
X_7	Smoking habit	التدخين	تأخذ 3 مستويات فيرمز للايدخن بالرمز 1، يدخن بالرمز 2 ومدخن سابق بالرمز 3.
X_8	Snakes consumption	استهلاك الوجبات الخفيفة	تأخذ 3 مستويات فيرمز للايستهلك بالرمز 1، يستهلك بالرمز 2 ومتوقف عن الاستهلاك بالرمز 3.
X_9	Increased urination	كثرة التبول	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{10}	Increased thrist	كثرة العطش	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{11}	Increased appetite	زيادة الشهية	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{12}	Obesity	السمنه	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{13}	Weight loss	فقدان الوزن فجأة	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{14}	Itching	حكة في الجلد	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{15}	Delayed healing	تأخر في التئام الجروح	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.

تابع جدول (1) التعريف بالمتغيرات المستقلة

X_{16}	Genital thrush	فطريات في الأماكن الحساسة	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{17}	Visual blurring	زغللة في العين	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{18}	Eating pattern	نظام غذائي	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{19}	Exercise	التمارين الرياضية	يأخذ 3 مستويات فيرمز للا يلعب بالرمز 1، يلعب بالرمز 2 و لسبق له اللعب بالرمز 3.
X_{20}	Alopecia	الصلع	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{21}	Family history	التاريخ العائلي	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{22}	Fatigue	الشعور بالإرهاق	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{23}	Abdominal pain	آلام في البطن	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{24}	Nausea	الشعور بالغثيان	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{25}	Ketone small	رائحة النفس الكريهة	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{26}	Tachycardia	زيادة معدل سرعة التنفس	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{27}	Confusion	الارتباك	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{28}	Numbness	ألم وتنميل في الأطراف	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{29}	Seizure	نوبات تشنجية	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{30}	Mood swings	تقلبات في المزاج	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{31}	Dyspnia at night	صعوبة في النفس أثناء الليل	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.

تابع جدول (1) التعريف بالمتغيرات المستقلة

X_{32}	Swelling in the foot and ankle	حرارة في القدمين	تأخذ القيمة 1 وتعني نعم والقيمة 0 وتعني لا.
X_{33}	Diastolic pressure	الضغط الإنقباضي	يمثل الضغط الإنقباضي للمريض وقت الإختبار.
X_{34}	Systolic pressure	الضغط الانبساطي	يمثل الضغط الانبساطي للمريض وقت الاختبار.
X_{35}	Fasting blood glucose	السكر الصايم	يمثل السكر الصايم للمريض 7 صباحا.

تم في هذه الدراسة تطبيق كل من:

- نموذج الانحدار اللوجستي.
- تحليل المكونات الرئيسية قبل نموذج الانحدار اللوجستي.

أولا تطبيق نموذج الانحدار اللوجستي:

عند تطبيق نموذج الانحدار اللوجستي يتم تقسيم البيانات إلى مرحلتين مرحلة التدريب ومرحلة الإختبار والسبب في ذلك ذلك أنه يمثل نوع من أنواع التعلم الآلي بدون إشراف، تحتوي مرحلة التدريب في هذه الدراسة على نسبة 80% من البيانات الخاصة بهذه الدراسة (148) أما مرحلة الإختبار فهي تحتوي على نسبة 20% من البيانات (37)، في مرحلة التدريب يتم تدريب النموذج عن طريق إدخال المدخلات والمخرجات معا فالمدخلات تكون عبارة عن جميع البيانات الخاصة بالمريض وهي في هذه الدراسة عبارة عن 35 متغير أما المخرجات هي عبارة عن المعلومة التي تبين هل هو مريض بالسكري أم لا بحيث عند إدخال بيانات جديدة خاصة بالمريض يستطيع النموذج التنبؤ هل هو مريض سكري أم لا، فيوضح الجدول التالي مصفوفة الإرتباك الخاصة بمرحلة الإختبار وهي عبارة عن تقنية لقياس الأداء الخاص بتصنيف التعليم الآلي وهذه المصفوفة توضح الفئات التي تم تصنيفها بشكل صحيح والفئات التي تم تصنيفها بشكل خطأ.

جدول(2) مصفوفة الارتباك الخاصة بمرحلة الاختبار عند تطبيق نموذج الانحدار اللوجستي

	0	1
0	17	4
1	5	11

يتضح من الجدول السابق أن هناك 11 مريض تم تصنيفهم علي أنهم مرضى سكري وهم بالفعل مصابين مرض السكري ويطلق عليهم إيجابي حقيقي وهناك 17 مريض تم تصنيفهم علي أنهم غير مصابين بالسكري وهم بالفعل غير مصابين ويطلق عليهم سلبي حقيقي ويوضح أيضا أن هناك 5 مرضى تم تصنيفهم على أنهم ليسو مصابين بالسكري وهم في الحقيقة مصابين بالمرض ويطلق عليهم سلبي كاذب وهناك 4 مرضى تم تصنيفهم على أنهم مصابين بالسكري ولكنهم في الحقيقة غير مصابين بالسكري ويطلق عليهم إيجابي زائف.

$$Accuracy = (TN + TP)/(TN + TP + FN + FP)$$

$$Accuracy = (17 + 11)/(17 + 11 + 5 + 4) = 75\%$$

بلغت الدقة التي تم الحصول عليها باستخدام نموذج الانحدار اللوجستي 75%.

ثانيا: تحليل المكونات الرئيسية قبل نموذج الانحدار اللوجستي:

يتم تطبيق تحليل المركبات الرئيسية قبل نموذج الانحدار اللوجستي حيث أنه يقوم بتقليل الأبعاد بدون فقد الكثير من المعلومات فيتم في هذا الأسلوب ادخال المدخلات فقط وليس المدخلات والمخرجات كما تم في مرحلة تطبيق الانحدار اللوجستي بمفردة والسبب في ذلك هو أن أسلوب المركبات الرئيسية هو عبارة عن نوع من أنواع التعلم الآلى بدون اشراف فيقوم الأسلوب باستخراج المركبات التي تساهم في تفسير التباين الحادث بين المتغيرات ويتم ذلك من خلال عدة خطوات:

تم تقسيم متغيرات هذه الدراسة الوصفية إلى ثلاث أقسام:

- القسم الأول: يتضمن المعلومات الخاصة بالمريض.
- القسم الثاني يتضمن 11 متغير من الأعراض التي من الممكن أن يعاني منها المصاب بمرض السكري.

- القسم الثالث: يتضمن 10 متغيرات من الأعراض التي من الممكن أن يعاني منها المصاب بمرض السكري.

بالنسبة للقسم الأول:

يحتوي على عدد من المتغيرات التي هي عبارة عن المعلومات الخاصة بالمريض مثل النوع (X_1)، البلد (X_3)، الحالة الإجتماعية (X_4)، الوظيفة (X_6)، التدخين (X_7)، استهلاك الوجبات الخفيفة (X_8)، نظام غذائي (X_{18}) وممارسة التمارين الرياضية (X_{19})

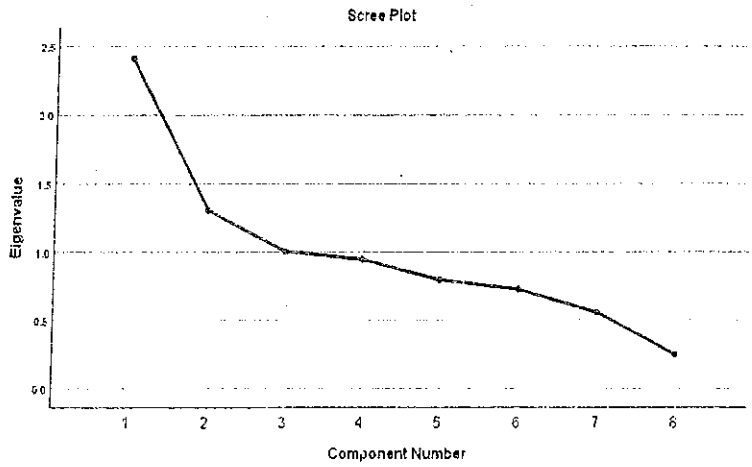
جدول (3) الجذر الكامن للمركبات الخاصة بالقسم الأول

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.412	30.146	30.146	2.412	30.146	30.146
2	1.307	16.333	46.479	1.307	16.333	46.479
3	1.003	12.540	59.019	1.003	12.540	59.019
4	0.947	11.835	70.855			
5	0.797	9.961	80.816			
6	0.726	9.081	89.897			
7	0.557	6.964	96.861			
8	0.251	3.139	100.000			

فيوضح الجدول (3) ما يلي:

1. استخراج 3 مكونات، بقيم الجذر الكامن لها أكبر من الواحد الصحيح حيث أن قيمة الجذر الكامن الأول 2.412، قيمة الجذر الكامن الثاني 1.307 وقيمة الجذر الكامن الثالث 1.003.
2. نسبة تفسير للتباين للمكون الأول تساوي 30.164%، للمكون الأول والثاني 46.479% والمكون الأول والثاني والثالث 59.019% أي أنهم يفسروا 59.019% من الدلالة العامة للمتغيرات بالنسبة لهذا القسم.
3. المكون الأول الذي قيمتة 2.412 قد فسر 30.146 من التباين وهي أعلى نسبة حيث تعد قيم المكونات معيار لكل مكون لما يستطيع أن يفسره من تباين، المكون الثاني الذي قيمته 1.307 قد فسر 16.333 من نسبة التباين وهي أقل من النسبة التي فسرها المكون الأول أما المكون الثالث الذي قيمته 1.003 قد فسر 12.540 من نسبة التباين وهذه النسبة أقل من النسبة التي فسرها كل من المكون الأول والثاني.

يوضح الرسم البياني رقم (1) قيم الجذور الكامنة لكل عامل على المحور الصادي ورقم المكون على المحور السيني



شكل رقم (1) المركبات الرئيسية التي تم استخراجها

يتضح من الرسم البياني رقم (1) أن هناك ثلاث عوامل أكبر من الواحد الصحيح وهي التي تم استخراجها أما بقية المركبات فهي أقل من الواحد الصحيح.

بالنسبة للقسم الثاني:

يحتوي على 11 متغير وتمثل هذه المتغيرات الأعراض التي من الممكن أن يصاب بها مريض السكري وهذه الأعراض عبارة عن كثرة التبول (X_9)، كثرة العطش (X_{10})، زيادة الشهية (X_{11})، السمنة (X_{12})، فقدان الوزن فجأة (X_{13})، حكة في الجلد (X_{14})، تأخر في التئام الجروح (X_{15})، فطريات في الأماكن الحساسة (X_{16})، زغللة في العين (X_{17})، الصلع (X_{20})، التاريخ العائلي (X_{21}) والشعور بالإرهاق (X_{22}).

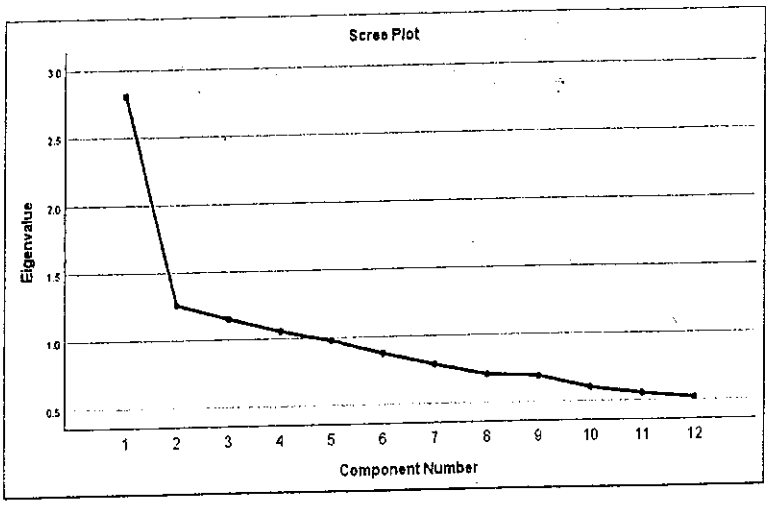
جدول (4) الجذر الكامن للمركبات الخاصة بالقسم الثاني

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.806	23.386	23.386	2.806	23.386	23.386
2	1.260	10.500	33.885	1.260	10.500	33.885
3	1.153	9.606	43.491	1.153	9.606	43.491
4	1.053	8.778	52.269	1.053	8.778	52.269
5	0.980	8.170	60.439			
6	0.879	7.326	67.765			
7	0.788	6.569	74.334			
8	0.709	5.905	80.239			
9	0.692	5.766	86.005			
10	0.603	5.021	91.026			
11	0.555	4.627	95.653			
12	0.522	4.347	100.000			

يوضح الجدول (4) ما يلي:

1. استخراج 4 مكونات بقيم الجذر الكامن لها أكبر من الواحد الصحيح حيث أن قيمة الجذر الكامن الأول 2.806، قيمة الجذر الكامن الثاني 1.260، قيمة الجذر الكامن الثالث 1.153 وقيمة الجذر الكامن الرابع 1.053.
2. نسبة تفسير التباين للمكون الأول 23.386%، نسبة تفسير التباين للمكون الأول والثاني معا 33.885%، نسبة تفسير التباين للمكون الأول والثاني والثالث معا 43.491% ونسبة تفسير التباين للمكونات الأول والثاني والثالث والرابع 52.269% أي أنهم يفسروا 52.269% من الدلالة العامة للمتغيرات بالنسبة لهذا القسم.
3. المكون الأول الذي قيمته 2.806 قد فسر 23.386% من نسبة التباين وهي أعلى نسبة حيث تعد قيم المكونات معيار لكل مكون لما يستطيع أن يفسره من تباين، المكون الثاني الذي قيمته 1.260 قد فسر 10.500% من نسبة التباين وهي أقل من النسبة التي فسرها المكون الأول، المكون الثالث الذي قيمته 1.153 قد فسر 9.606% من نسبة التباين وهي أقل من النسبة التي فسرها المكون الأول والنسبة التي فسرها المكون الثاني والمكون الرابع الذي قيمته 1.053 فسر 8.778% من نسبة التباين وهي أقل من القيمة التي فسرها المكون الأول والقيمة التي فسرها المكون الثاني والقيمة التي فسرها المكون الثالث.

يوضح الرسم البياني رقم (2) قيم الجذور الكامنة لكل عامل على المحوز الصادي ورقم المكون على المحور السيني



شكل رقم (2) المركبات الرئيسية التي تم استخراجها

يوضح الرسم البياني رقم (2) أن هناك 4 مكونات أكبر من الواحد الصحيح وهذه المكونات هي التي تم استخراجها أما بقية المكونات فهي أقل من الواحد الصحيح.

بالنسبة للقسم الثالث:

المتغيرات التي يحتوي عليها هذه القسم هي عبارة عن آلام في البطن X_{23}، الشعور بالغثيان أو التقيؤ X_{24}، رائحة النفس الكريهة X_{25}، زيادة معدل سرعة التنفس X_{26}، الارتباك X_{27}، ألم وتنميل في الأطراف X_{28}، نوبات تشنجية X_{29}، تقلبات في المزاج X_{30}، صعوبة في النفس أثناء الليل X_{31} وحرارة في القدمين.

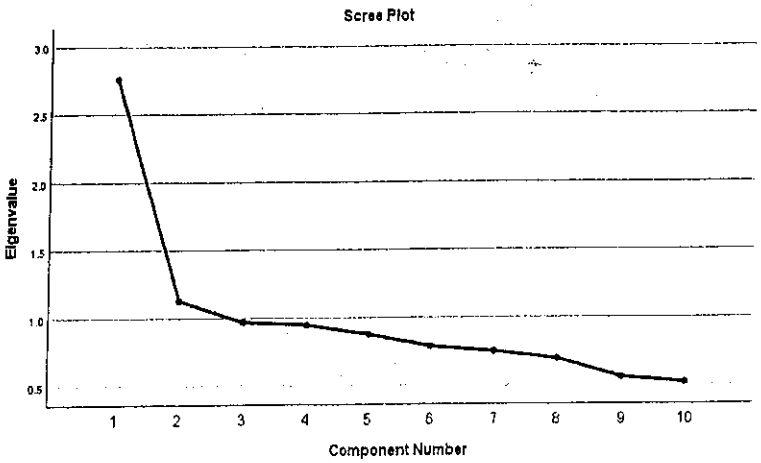
جدول (5) الجذر الكامن للمركبات الخاصة بالقسم الثالث

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.759	27.591	27.591	2.759	27.591	27.591
2	1.131	11.307	38.897	1.131	11.307	38.897
3	0.970	9.696	48.594			
4	0.949	9.491	58.085			
5	0.879	8.793	66.877			
6	0.788	7.880	74.757			
7	0.750	7.501	82.258			
8	0.695	6.951	89.209			
9	0.560	5.598	94.807			
10	0.519	5.193	100.000			

يبين الجدول (5) ما يلي:

1. تم استخراج مكونين بقيم الجذر الكامن لهم أكبر من الواحد الصحيح حيث أن قيمة الجذر الكامن الأول 2.795 و قيمة الجذر الكامن الثاني 1.131.
2. نسبة تفسير التباين للمكون الأول 27.591% ونسبة تفسير المكونين الأول والثاني 38.879% أي أنهم يفسروا 38.897% من الدلالة العامة للمتغيرات بالنبة لهذا القسم.
3. المكون الأول الذي قيمته 2.759 قد فسر 27.591% وهي أعلى من قيمة المكون الثاني حيث تعد قيم المكونات معيار لكل مكون لما يستطيع أن يفسره من تباين والمكون الثاني الذي قيمته 1.131 قد فسر 11.307% من نسبة التباين.

يوضح الرسم البياني رقم (3) قيم الجذور الكامنة لكل عامل على المحور الصادي ورقم المكون على المحور السيني



شكل رقم (3) المركبات الرئيسية التي تم استخراجها

يوضح الرسم البياني رقم (3) أن هناك مكونين أكبر من الواحد الصحيح وهذه المكونات هي التي تم استخراجها أما بقية المكونات فهي أقل من الواحد الصحيح .

بعد ذلك يتم أخذ المتغيرات التي تم استخراجها من تحليل المكونات الرئيسية من الأقسام الثلاثة وعددهم 9 متغيرات بالإضافة إلى المتغيرات الكمية التالية: السن، عدد مرات الحمل، السكر، الضغط الانقباضي، الضغط الانبساطي ليصبح عدد المتغيرات 14 متغيرا فيتم تطبيق أسلوب الانحدار اللوجستي وتقسيم البيانات بعدد المتغيرات 18 إلي مرحلة التدريب ومرحلة الاختبار.

يوضح الجدول (6) مصفوفة الارتباك الخاصة لمرحلة الاختبار عند تطبيق تحليل المركبات الرئيسية قبل تطبيق نموذج الانحدار اللوجستي

جدول (6) مصفوفة الارتباك الخاصة بمرحلة الاختبار عند تطبيق تحليل المركبات الرئيسية قبل تطبيق نموذج الانحدار اللوجستي

	0	1
0	22	2
1	2	11

فيوضح الجدول (6):

تم تصنيف 11 مريض على أنهم مصابين بالسكري وهم بالفعل مصابين ويطلق على هذا التصنيف إيجابي حقيقي وتم تصنيف 22 مريض علي أنهم غير مصابين بالسكري وهم بالفعل غير مصابين ويطلق على هذا التصنيف سلبي حقيقي لكنها توضح أن تم تصنيف مريضين علي أنهم مصابين بالمرض وهم في الحقيقة غير مصابين بالسكري ويطلق علي هذا التصنيف سلبي كاذب وتم تصنيف مريضين على أنهم غير مصابين بالسكري وهم في الحقيقة مصابين بالسكري ويطلق على هذا التصنيف إيجابي زائف.

$$Accuracy = (TN + TP)/(TN + TP + FN + FP)$$

$$Accuracy = (22 + 11)/(22 + 11 + 2 + 2) = 89\%$$

بلغت الدقة التي تم الحصول عليها باستخدام تحليل المكونات الرئيسية قبل نموذج الانحدار اللوجستي 89%.

(5) البرامج المستخدمة:

1. برنامج SPSS.
2. برنامج R.

(6) النتائج:

يمكن ايجاز النقاط التي تم التوصل إليها في هذا البحث في كل من:

1. بلغت دقة تطبيق نموذج الانحدار اللوجستي بمفرده للتنبؤ بمرض السكري 75%.
2. بلغت دقة استخدام تحليل المكونات الرئيسية لتقليل الأبعاد قبل استخدام نموذج الانحدار اللوجستي للتنبؤ 89% مما يدل ذلك على أن استخدام تحليل المكونات الرئيسية قبل استخدام نموذج الانحدار اللوجستي قد زاد من دقة التنبؤ بمرض السكري من 75% إلى 89%.

(7) التوصيات:

بناء على ما تم التوصل إليه البحث من نتائج، يوصى بالآتي:

استخدام أساليب أخرى للتنبؤ بمرض السكري مثل الشبكات العصبية، شجرة القرار، نايف بايز والغابة العشوائية وغير ذلك من الأساليب التي يمكن استخدام للتنبؤ.

(8) المراجع

De Cock, M., Dowsley, R., Nascimento, A. C., Railsback, D., Shen, J., & Todoki, A. (2021). High performance logistic regression for privacy-preserving genome analysis. *BMC Medical Genomics*, *14*(1), 1-18.

Islam, M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis* (113-125). Springer, Singapore.

Jacquet, P., Shamir, G., & Szpankowski, W. (2021, March). Precise Minimax Regret for Logistic Regression with Categorical Feature Values. In *Algorithmic Learning Theory* (755-771). PMLR.

Jhaldiyal, T., & Mishra, P. K. (2014). Analysis and prediction of diabetes mellitus using PCA, REP and SVM. *International Journal of Engineering and Technical Research (IJETR)*, *2*(8), 164-166.

Ji, K., Wen, R., Ren, Y., & Dhakal, Y. P. (2020). Nonlinear seismic site response classification using K-means clustering algorithm: Case study of the September 6, 2018 Mw6. 6 Hokkaido Iburi-Tobu earthquake, Japan. *Soil Dynamics and Earthquake Engineering*, *128*, 105907.

Joshi, T. N., & Chawan, P. P. M. (2018). Diabetes prediction using machine learning techniques. *Ijera*, *8*(1), 9-13.

Mahajan, A., Kumar, S., & Bansal, R. (2017, May). Diagnosis of diabetes mellitus using PCA and genetically optimized neural network. In *2017*

International Conference on Computing, Communication and Automation (ICCCA) (pp. 334-338). IEEE.

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *IEEE Access, 8*, 54776-54788.

Sarker, I. H., Abushark, Y. B., & Khan, A. I. (2020). Contextpca: Predicting context-aware smartphone apps usage based on machine learning techniques. *Symmetry, 12*(4), 499.

Seo, J. H., Kim, H. J., & Lee, J. Y. (2020). Nomogram construction to predict dyslipidemia based on a logistic regression analysis. *Journal of Applied Statistics, 47*(5), 914-926.

Singh, D. A. A. G., Leavline, E. J., & Baig, B. S. (2017). Diabetes prediction using medical data. *Journal of Computational Intelligence in Bioinformatics, 10*(1), 1-8.

Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked, 17*, 100179.