استخدام التحليل التمييزي في التصنيف والتنبؤ

(دراسة تطبيقية)

مقدمه

الدكتور/ عبد الرحيم عوض عبد الخالق بسيوني

dr-abdelreheembassuny@outlook.com

https://caf.journals.ekb.eg/

الملخص:

يهدف البحث إلى تحديد العوامل المؤثرة في الإصابة بمرض السُكري وإنشاء دالة تميزية تقوم بالتمييز والفصل بين الأشخاص إلى مجموعتين هُما (مصاب، غير مصاب) وذلك من خلال مجموعة من العوامل المؤثرة وهي الوراثة، الوزن، ضغط الدم، السن، النوع، التدخين، ممارسة الرياضة، مرض النقرس، الكوليسترول، الحالة الاجتماعية، وأمراض القلب والكلى. وتم التطبيق على عينة من ٣٥٠ شخص منهم ١٦٨ مُصاب و١٨٢ غير مصاب وبعد التأكد من ملائمة التحليل التمييزي للبيانات وتكوين دالة تمييزية للفصل والتمييز تبين أن أكثر العوامل المؤثرة في الإصابة بمرض السُكري الوزن، السن، ضغط الدم، التدخين، ممارسة الرياضة، الوراثة، الكوليسترول، النوع، وتم استبعاد باقي المُتغيرات لعدم معنوياتها. أما عن أكثر المُتغيرات مُساهمة في التمييز والفصل بين المجموعات هو الوزن بنسبة (٧٩,٤%) يليه ضغط الدم بنسبة (٣٣,٢%) يليه الوراثة بنسبة (٣١%) ثم السن بنسبة (٢٨,٩%) والكوليسترول بنسبة (٢٤,٢%) والتدخين بنسبة (٢٠,٨%) ثم ممارسة الرياضة بنسبة (١٩,٨%) وأخيرًا النوع بنسبة (١٠%) كما تبين كفاءة الدالة التمييزية في التصنيف بنسبة (٩٠,٦%) بنسبة خطأ (٩,٤%) وحساسية (٩٢,٩%) ونوعية بنسبة (٨٨,١%).

Abstract:

The research aims to determine the factors that affect the incidence of diabetes and establish discriminant function that separates people from with and without disease groups. this is done through a set of factors, genetics, weight, blood pressure, age, gender, smoking and exercise, gout and cholesterol, social status, and heart and kidney disease. After checking the fit of the discriminant analysis to the data and creating the discriminant function for separating and distinguish, it was concluded that the most important factors affecting the incidence of diabetes are weight, age, blood pressure, smoking, exercise, genetics, cholesterol and gender the remaining variables were excluded because they were not significant. the variables that contribute most to segregation are weight (79.4%), blood pressure (33.2%), genetic (31%), age (28.9%) cholesterol (24.2%), smoking (20.8%), exercise (19.8%) and gender (10%) it also shows that the discriminative function is (90.6%) efficient, and attributed if wrong (9.4%), sensitivity (92.9%), qualitative (88.1%).

مقدمة:

شهدت السنوات الأخيرة زيادة ملحوظة في اعداد المصابين بمرض السكري حتى اننا لا نكاد نرى بيت بدون مريض السكري حتى أصبح غالبية المصريين يتعايشون مع هذا المرض على أنه صديق وفي ولكنه غير وفي لأنه في أي لحظة قد يفتك بحياة المريض وخاصة ان لمرض السكري مضاعفات خطيرة تصيب القلب والشرايين والاعصاب والكلى والعين فقد كان ومازال مرض السكري وراء ارتفاع نسبة الوفيات سواء كان بشكل مباشر أو غير مباشر، ويرجع الأطباء السبب إلى عدم مقدرة البنكرياس على فرز كمية الانسولين المطلوبة أو الانسولين المنتج غير فعال مما يؤدي إلى اضطرابات ومنها ارتفاع نسبة السكر في الدم مما يضر الكلى والأوعية الدموية والقلب والعين وغيرها ويرجع مرض السكري إلى مجموعة من العوامل أهمها عامل الوراثة وضغط الدم والوزن الزائد والنوع وممارسة الرياضة والتدخين والكوليسترول وأمراض القلب والكلى.

ويعتبر أسلوب تحليل التمايز أهم الأساليب الإحصائية متعددة المتغيرات التي تستخدم في معالجة البيانات الوصفية ويعتمد على بناء دالة تسمى دالة التمايز وهي عبارة عن توليفة خطية لمجموعة من المتغيرات المستقلة وهذه الدالة تعمل على تقليل التشابه في أخطاء التصنيف ويهدف التحليل التميزي إلى تصنيف المشاهدات إلى مجموعاتها الصحيحة بأقل خطأ تصنيف ممكن، ويختلف أسلوب التمايز مع كُل من تحليل التباين وتحليل الانحدار حيث المتغير التابع نوعي، بينما في الأسلوبين الأخيرين يكون مُتغير كمي، كما يتشابه تحليل التمايز مع الانحدار اللوجستي إذا يفسر كُل منهما مُتغير وصفي ألا إنَ الانحدار اللوجستي لا يتطلب أن تكون المُتغيرات المُستقلة تتبع التوزيع الطبيعي.

مُشكلة البحث:

يُعد الاهتمام بالقطاع الصحي وصحة المواطن من أولويات أي حكومة حتى يواكب أي تقدم كما يعد مرض السكري من الامراض الأصيلة في الشعب المصري في مختلف الاعمار ومن هذا المنطلق تبدأ مشكلة البحث في الظهور كما أن معظم الدراسات اهتمت بالتحليل احادي المتغير أو ثنائي المتغير ولم يتطرق الا القليل إلى التحليل متعدد المتغيرات ومن هنا لزم البحث في الأساليب الإحصائية التي تهتم بدراسة متغيرات متعددة او مجموعة من المتغيرات في آن واحد حتى يمكن الوقوف على أهم العوامل المؤثرة على الإصابة بمرض السكري ومحاولة الحد منها.

أهمية البحث:

يستمد البحث أهميته من خطورة مرض السكري وما ينتج عنه من مضاعفات خطيرة كأمراض القلب والكلى والعين وغيرها مما يؤدي إلى ارتفاع نسبة الوفيات واهدار الكثير من الأموال على القطاع الصحي دون جدوى، كما نوضح الدور الذي يقوم به الاحصائيين جنبًا إلى جنب الأطباء وتوفير نموذج احصائي أو دالة تمييزية لها القدرة على التمييز والفصل بين الأشخاص إلى مجموعتين أحدهما المصابة

والأخرى غير المصابة بمرض السكري ثم تصنيف المشاهدات الجديدة وتوزيعها على إحدى المجموعتين وبالتالي التشخيص المبكر والحفاظ على أرواح البشر والأموال الطائلة التي تنفق عليها.

أهداف البحث:

للبحث أهداف عديدة أهمها: -

١- القاء الضوء على أسلوب التحليل التمييزي كأحد أساليب التحليل الاحصائي متعدد المتغيرات.

٢- تحديد العوامل المؤثرة في الإصابة بمرض السكري.

٣- انشاء دالة تمييزية تصنف الأشخاص إلى مجموعتين (مصاب وغير مصاب).

٤- تحديد الأهمية النسبية للعوامل المؤثرة في الإصابة بمرض السكري ومدى مساهمة كل عامل في التمييز والتصنيف.

٥- تصنيف الأشخاص أو المفردات الجديدة وتوزيعها على إحدى المجموعتين.

٦- التنبؤ باحتمالية إصابة أو عدم إصابة الشخص بمرض السكري بناءً على مجموعة العوامل مما يؤدي للتشخيص المبكر وتفادي تدهور الحالة الصحية للمريض.

٧- توضيح دور الإحصائي في المساهمة والمساعدة في المجالات الطبية.

مُتغيرات البحث: -

١- المتغير التابع: (متغير نوعي) (مصاب وغير مصاب).

٢- العوامل المؤثرة (المتغيرات المستقلة): -

٣-

- الوراثة.
- النوع.
- الكوليسترول.
- الوزن.
- التدخين.
- الحالة الاجتماعية.
- ضغط الدم.
- ممارسة الرياضة.
- أمراض القلب والكلى.
- السن.
- مرض النقرس.

في حالة توافر الخاصية يأخذ المتغير (١) وعدم توافرها (٠).

مصادر البيانات:

أخذت البيانات من عينة من ٣٥٠ شخص بمستشفى كفر الشيخ العام ومستشفى جامعة كفر الشيخ وإحدى المستشفيات الخاصة، تنقسم العينة إلى ١٦٨ مصاب و١٨٢ غير مصاب باستخدام أحد البرامج الإحصائية "SPSSV23".

الدراسات السابقة:

١- دراسة Pohar and Blas (2004):

هدفت الدراسة إلى المقارنة بين الانحدار اللوجستي وتحليل التميز الخطي، دراسة محاكاة وتوصلت هذه الدراسة إلى أن تحليل التمايز الخطي يستخدم إذا كانت المتغيرات تتبع التوزيع الطبيعي وان الانحدار

اللوجستي يستخدم في حالة العينات الصغيرة حيث ان لا يشترط ان تتبع المتغيرات التوزيع الطبيعي كما توصلت الدراسة إلى أن نتائج الطريقتين كانت متقاربة عندما كان حجم العينة كبير.

٢- دراسة عبد الكريم (٢٠٠٦): -

هدفت الدراسة إلى استخدام الطرق التمييزية الإحصائية لتشخيص بعض أمراض القلب حيث تناول البحث أسلوب التحليل التمييزي والنموذج اللوجستي وتم التطبيق على عينة من ٢٠٦ مريضًا جمعت من ثلاث مستشفيات مختلفة وكانت المتغيرات المستقلة هي العمر والوزن والطول وضغط الدم ونسبة الكوليسترول في الدم والنوع والمتغير التابع نوع المريض (تصلب الشرايين = ١، جلطة قلبية = ٢) وتم التوصل إلى النموذج اللوجستي أعطى نسبة خطأ تصنيف أقل من النموذج التمييزي.

٣- دراسة الجاعوني، غانم (٢٠٠٧): -

تَضَمن هذا البحث دراسة أحد أساليب التحليل الإحصائي متعدد المتغيرات وهو أسلوب التحليل التمييزي الذي يعد من الأساليب الإحصائية المتقدمة التي تستخدم في توصيف وتوزيع الاسر داخل الهيكل الاقتصادي والاجتماعي للمجتمع ويساعد في رسم خطط التنمية الاقتصادية والاجتماعية التي تهدف إليها الدولة والوقوف إلى أنسب الطرق في حيث عدالة توزيع الدخل والعِبْء الضريبي والإعانات الحكومية لأسر المجتمع بصورة أكثر واقعية.

٤- دراسة الشمراني (٢٠٠٨): -

هدفت الدراسة إلى التعرف على كيفية استخدام التحليل التمييزي وكذلك استخدام تحليل التباين متعدد المتغيرات وذلك في حالة عامل واحد أو عاملين وكذلك مدى إمكانية تقييم كفاءة النموذج التمييزي واختيار القدرة التمييزية للنموذج ومقارنة جوانب الشبه والاختلاف بينها، توصلت الدراسة إلى ان تشابه افتراضات تحليل التباين المتعدد والتحليل التمييزي وفي حالة وجود عاملين يعد استخدام تحليل التباين المتعدد امرًا ضروريًا عن وجود تفاعل أم لا وبعده يأتي دور التحليل التمييزي لتحديد الدوال التمييزية لكل مجموعة.

٥- دراسة Roush and Kelly (2009):

هدفت هذه الدراسة إلى عمل مقارنة بين تحليل التمايز الخطي Linear discriminant analysis (LDA) وتحليل التمايز اللوجستي Linear logistic discrimination analysis (LLD) وتحليل التمايز الخطي باستخدام الرتب LDA based on ranks وتحليل التمايز المختلط Mixture discriminant analysis (MDA)بالاعتماد على دراسة محاكاة مونت كارلو وتوصلت الدراسة إلى أن كل من تحليل التمايز الخطي وتحليل التمايز اللوجستي لهم نفس الدقة في التصنيف كما أشارت النتائج إلى تحليل التمايز المختلط بشكل عام هو أكثر النماذج قابلية للتطبيق وأكثرهم دقة للتصنيف وخاصة إذا كانت البيانات لا تتبع التوزيع الطبيعي كما أنَ تحليل التمايز بالاعتماد على الرتب أكثر دقة للتصنيف عن كل من تحليل التمايز الخطي وتحليل التمايز اللوجستي.

٦- دراسة الجزار (٢٠١٢): -

هدفت الدراسة إلى المقارنة بين انسب أساليب التصنيف والتنبؤ وهي التحليل التمييزي الخطي وأسلوب الانحدار اللوجستي المتعدد وكان المعيار المستخدم للمقارنة بينهما هو دقة التصنيف والمساحة تحت المنحنى Area under the roc curve (AUC) لتحليل الـ Receiver operating characteristic curve (ROC) على بيانات مولدة بالحاسب بهدف مقارنة قدرة كلا النموذجين على التصنيف والتنبؤ تحت تأثير الاختلاف في حجم البيانات وعدد فئات المتغير التابع والمسافة بين متوسطات المجموعات التي تحتاج إلى تصنيف وتبين تشابه كبير في المعاملات التي تم تقديرها وكان الأسلوب اللوجستي أعلى بقليل من التحليل التمييزي في دقة التصنيف إلا أنه عند أخذ معيار Sensitivity , Specificity والمساحة تحت المنحنى AUC لتحليل ROC فقد وجد ان الفارق بين النموذجين ضئيل جدًا.

٧- دراسة هاشم (٢٠١٤): -

سعت الدراسة إلى استخدام التحليل التمييزي المتعدد لتصنيف مراحل الإصابة بمرض الفشل الكلوي المزمن وتضمنت الدراسة عينة من ٣٢٢ وتم إيجاد الدالة التمييزية وتصنيف مراحل الإصابة بمرض الفشل الكلوي المزمن على أساس مجموعة من المتغيرات وهي العمر والحالة الاجتماعية والمهنة والسكن وعدد الأطفال ومستوى الدخل وتبين معنوية متغير المهنة في التمييز والتصنيف كما أن للدالة قدرة عالية على التمييز وتصنيف الأشخاص.

٨- دراسة سُليمان (٢٠١٥): -

هدفت الدراسة إلى المقارنة بين التحليل التمييزي والنموذج اللوجستي ونماذج الشبكات العصبية في تصنيف المشاهدات وتم ذلك بالتطبيق على العوامل المؤثرة على كفاية دخل الأسرة وهي حجم الأسرة وطبيعة ملكية السكن ووجود طلبة يدرسون بالجامعات وتوصل التحليل التمييزي إلى الدالة التمييزية ومعنوية تأثير متغيرين فقط وهما حجم الأسرة وملكية السكن وعدم معنوية وجود طلبة يدرسون في الجامعة كما تبين ان الشبكات العصبية أفضل من النموذج اللوجستي أفضل من التحليل التمييزي.

٩- دراسة خوالدي (٢٠١٧): -

أوضحت الدراسة دور التحليل التمييزي في التنبؤ بالفشل المالي للمؤسسات الاقتصادية الصغيرة والمتوسطة لولاية أم البواقي بالاستعانة بالنسب المالية المحسوبة من القوائم المالية بالإضافة إلى التواصل إلى أفضل النسب تمييزًا لوضعية المؤسسات سواء كانت ناجحة أو فاشلة وذلك من خلال اتباع خطوات التحليل التمييزي باستخدام برنامج "SPSS" حيث تم استخدام عينة مكونة من ٣٠ مؤسسة منها ١٧ مؤسسة ناجحة و١٣ مؤسسة فاشلة واظهرت النتائج كفاءة النموذج المستخدم الذي يُمكن من التنبؤ بالفشل المالي للمؤسسات الصغيرة والمتوسطة والذي يتكون من نسبتين ماليتين (معدل دوران الأصول المتداولة ومعدل دوران إجمالي الأصول) من اصل تسع نسب لهم القدرة على التمييز بين المؤسسات الناجحة والمؤسسات الفاشلة.

١٠- دراسة النويري (٢٠١٨): -

أوضحت الدراسة أهم العوامل التي لها دور في تمييز مرض السكري المصابين من غير المصابين بالفشل الكلوي واستخدام التحليل التمييزي للتوصل لنموذج رياضي يُمَكِن من تصنيف مرض السكري المصابين وغير المصابين بالفشل الكلوي وذلك بالاعتماد على متغيرات (نسبة السكر في الدم – اليوريا – الكرياتينين – العُمر) لمرضى السكري لمعرفة مدى أهمية المتغيرات في التمييز وتم جمع عينة من ٢٠٠ مريض سكري منهم ١٠٠ مصاب و١٠٠ غير مصاب بالفشل الكلوي وباستخدام الدالة التمييزية تم التوصل إلى: –

- هناك فروق معنوية بين متوسطات المتغيرات للمجموعتين باستخدام F مما يعني قدرة الدالة على التصنيف.
- هناك متغيرات لها الأثر الأكبر في التمييز وهما اليوريا والكرياتينين.
- النموذج التمييزي له دقة تصنيف عالية ٩١% وخطأ ٩%.

١١- دراسة Abdul Hussein (2019):

سعت هذه الدراسة إلى التمييز بين مجموعتين (المصابين وغير المصابين) بمرض القلب باستخدام إحصائية Mahalnobis (D^2) واشتق الباحث قاعدة للتمييز مشتقة من تلكَ الإحصائية أسمها R_{D^2} وهي تتبع توزيع F واستخدامها في بناء دالة التمييز الخطية وطبقت هذه الدراسة على ٤٠ مريض تم تقسيمهم إلى مجموعتين الأولى وعددها ١٦ مصاب والثانية وعددها ٢٤ من الأشخاص غير المصابين بمرض القلب وتوصلت الدراسة إلى أن إحصائية Mahalnobis مُهمة لإنشاء قاعدة تمييز بين مجموعتين كذلك بالاعتماد على هذه الإحصائية أمكننا إيجاد متجه معاملات التمايز بسهولة من مُعاملات الانحدار.

١٢- دراسة بغرش (٢٠٢٠): –

هدفت الدراسة إلى استخدام التحليل التمييزي كأسلوب يُستخدم للتنبؤ بمتغيرات اسمية تابعة بناءً على علاقتها بمتغيرات كمية وتستخدم هذه الطريقة في البنوك لتصنيف المقرضين إلى جيدين أو سيئين بحسب الاخلال بالسداد، تم التطبيق من مجموعة من المشروعات التي استفاد أصحابها من القروض من طرف الوكالة الوطنية لتسيير القرض المصغر خلال الفترة من ٢٠٠٤ إلى ٢٠٠٦.

المبحث الأول

التحليل التمييزي The Discriminate ate analysis

مقدمة:

يلعب التحليل الإحصائي دورًا هامًا في تحليل وتفسير الظواهر الاجتماعية والطبيعية في المجتمع ويُعد التحليل الإحصائي أحد طرق البحث العلمي الذي يستخدم عند دراسة المشاكل الاجتماعية والصحية والاقتصادية وتم تقسيم التحليل الإحصائي إلى التحليل احادي وثنائي المتغيرات ويبنى على حزمة من المتغيرات أو العوامل من الأساليب الإحصائية للتحليل متعدد المتغيرات هو أسلوب التحليل التمييزي والذي يشاع استخدامه في المجالات الطبية حيث يهتم التحليل التمييزي بكيفية التمييز بين مجموعتين أو أكثر من الافراد أو الأشياء وتصنيف المفردات الجديدة على المجموعات التي سبق تعريفها ويعتمد أسلوب تحليل التمايز على الوصول إلى دالة تسمى دالة التمايز تعمل على زيادة الفروق بين متوسط المجموعات حيث كُلما كان هُناك تباعد بين متوسط المجموعات كلما كان التمييز كفء وبالتالي يقل خطأ التصنيف ويعتبر التحليل التميزي بين مجموعتين أو أكثر من الأفراد أو الأشياء وتصنيف المفردات الجديدة على المجموعات التي سبق تعريفها ويعتبر التحليل التمييزي استكشافيًا بطبيعته حيث يكتشف أسباب الاختلاف المشاهدة عندما لا تستطيع فهم العلاقات السببية بدرجة كافية الدقة. (الجاعوني، غانم: ٢٠٠٧)

أهمية التحليل التمييزي:

ترجع أهمية التحليل التمييزي كأحد أساليب التحليل مُتعدد المُتغيرات إلى مقدرته في التمييز بين مجموعتين أو أكثر من خلال مجموعة من المتغيرات ويتم ذلك بإنشاء دوال تمايز "Discriminate Function" تعمل على تعظيم الاختلاف أو الفروق بين المجموعات بأقل خطأ للتصنيف.

أنواع التحليل التمييزي:

هناك ثلاث أنواع من التحليل التمييزي تتمثل في: (النويري: ٢٠١٣)

١- التحليل التمييزي المُباشر Direct discriminate analysis: حيث تدخل المتغيرات إلى التحليل دفعة واحدة دون إعطاء أي أهمية لأي مُتغير.

٢- التحليل التميزي الهرمي Hierarchical discriminate analysis: يتم فيها ادخال المتغيرات حسب رؤية الباحث.

٣- التحليل التميزي المتدرج Stepwise discriminate analysis: يتم ادخال المُتغيرات للتحليل حسب معيار إحصائي يُحدد أولوية إدخال المُتغيرات إلى النموذج حيث يتم إضافة المُتغيرات إلى الدوال التميزية واحد تلو الآخر حتى نجد أن إضافة مُتغيرات لا يُعطي تمييزًا أفضل.

أهداف التحليل التمييزي: -

هُناك عدة أهداف للتحليل التمييزي أهمها: -

- انشاء دوال تميزية للفصل أو التمييز بين فئات المتغير التابع.

- تعمل هذه الدوال على تعظيم الفروق بين المجموعات (فئات المتغير التابع).
- ترتيب المتغيرات التي تسهم بقدر كبير في التمييز أو توضيح الاختلافات بين المجموعات (فئات المتغير التابع).
- تصنيف المشاهدات الجديدة وتوزيعها على المجموعات (فئات المتغير التابع).
- الوصول إلى أقل نسبة خطأ للتوصيف – تقييم دقة التصنيف كنسبة مئوية.

شروط التحليل التمييزي:

١- عدم تساوي متوسطات المجموعات (فئات المُتغير التابع).

٢- تساوي مصفوفة التباين والتغاير بين المجموعتين.

٣- ان تكون المجموعات منفصلة وقابلة للتحديد.

٤- ان تتوزع المتغيرات التابعة والكمية توزيعًا طبيعيًا.

٥- العينة تختار عشوائيًا.

٦- استقلال المشاهدات؛ أي عدم وجود ارتباط بين المتغيرات المستخدمة في الدراسة أو ما يعرف بمشكلة Multicollinearityحيث كلما كان هُناك ارتباط بين المتغيرات كلما كان هناك صعوبة في تفسير نتائج تحليل التمايز وذلك صعوبة في تحديد المساهمة النسبية لكل متغير على حدة.

٧- عدم وجود قيمة متطرفة حيث أن تحليل التمايز أكثر حساسية وتأثرًا بالقيم الشاذة ووجودها يبعد توزيع البيانات عن التوزيع الطبيعي.

الدالة التميزية Discriminate Function:

تقوم الدالة التمييزية على فكرة أساسية وهي تقسيم الأشخاص إلى مجموعتين هُما (مصاب أو غير مُصاب) وذلك بالاعتماد على مجموعة من المُتغيرات أو العوامل وتعمل الدالة على زيادة درجة التجانس بين مفردات المجموعة الواحدة وتقليل درجة التجانس بين المجموعتين وبالتالي تسهيل إمكانية تصنيف أي مُشاهدة جديدة إلى إحدى المجموعتين بأقل خطأ للتصنيف كما تعمل الدالة على استبعاد المُتغيرات التي ليس لها تأثير معنوي في التمييز والفصل بين المجموعتين.

ويتم حساب الدالة التمييزية كالتالي: –

في حالة تعدد المجموعات تتعدد الدوال التمييزية ولكننا سنقتصر على الدالة التمييزية بين مجموعتين فقط.

أولاً: – حساب متوسطات المُتغيرات في كل مجموعة وإيجاد الفرق بين متوسط: –

$$\bar{x}_{i\,(1)} = \begin{bmatrix} \bar{x}_{1\,(1)} \\ \bar{x}_{2\,(1)} \\ \vdots \\ \vdots \\ \bar{x}_{k\,(1)} \end{bmatrix}$$

متوسطات المُتغيرات في المجموعة الثانية: –

$$\bar{x}_{i\,(2)} = \begin{bmatrix} \bar{x}_{1\,(2)} \\ \bar{x}_{2\,(2)} \\ \vdots \\ \vdots \\ \bar{x}_{k\,(2)} \end{bmatrix}$$

k عدد المتغيرات المستقلة

الفرق بين متوسط المتغير في المجموعتين:

$$d_i = \bar{x}_{i\,(1)} - \bar{x}_{i\,(2)} = \begin{bmatrix} \bar{x}_{11} - \bar{x}_{12} \\ \bar{x}_{21} - \bar{x}_{22} \\ \vdots \quad \vdots \\ \vdots \quad - \quad \vdots \\ \bar{x}_{k\,(1)} \quad \bar{x}_{k\,(2)} \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ \vdots \\ d_k \end{bmatrix} \text{ (المسافة)}$$

ثانيًا: إيجاد التباين والتغاير المشترك بين المجموعتين: –

$$S_{ii} = \sum x_i^2 - \frac{(\sum xi)^2}{n}$$

$$S_{ij} = \sum x_i x_j - \frac{\sum x_i \sum x_i}{n}$$

∴ التباين المشترك

$$V_{ii} = \frac{S_{ii} + S_{ii^{(2)}}}{n_1 + n_2 - 2}$$

∴ التغاير المشترك

$$V_{ij} = \frac{S_{ij\,(1)} + S_{ij^{(2)}}}{n_1 + n_2 - 2}$$

مصفوفة التباين والتغاير المشترك بين المجموعتين.

$$\mathrm{v} = \begin{bmatrix} \mathrm{v}_{11}\mathrm{v}_{12}\mathrm{v}_{13} \dots \dots \dots \mathrm{v}_{1k} \\ \mathrm{v}_{21}\mathrm{v}_{22}\mathrm{v}_{23} \qquad\qquad \mathrm{v}_{2k} \\ \vdots \quad \vdots \quad \vdots \qquad\qquad \vdots \\ \mathrm{v}_{k1}\mathrm{v}_{k2}\mathrm{v}_{k3} \dots \dots \dots \mathrm{v}_{kk} \end{bmatrix}$$

وهي عبارة عن مصفوفة مربعة ومتماثلة والقطر الرئيسي لها يُمثل التباين المُشترك وباقي العناصر التغاير المشترك.

بناء الدالة التمييزية:

تأخذ الدالة التمييزية بمعاملات معيارية الشكل التالي:

$$\hat{L} = \hat{\propto}_1 \; x_1 + \propto_2 \; x_2 + \qquad\qquad + \propto_k \; x_k$$

حيث

$$\hat{\propto} = \mathrm{v}^{-1}\, d$$

-١

$$\begin{bmatrix} \propto_1 \\ \propto_2 \\ \vdots \\ \propto_k \end{bmatrix} = \begin{bmatrix} \mathrm{v}_{11}\,\mathrm{v}_{12} \cdots\cdots\cdots \mathrm{v}_{1\mathrm{k}} \\ \mathrm{v}_{21}\,\mathrm{v}_{22} \cdots\cdots\cdots \mathrm{v}_{2\mathrm{k}} \\ \vdots \quad \vdots \qquad\qquad \vdots \\ \mathrm{v}k_1\mathrm{v}k_2 \cdots\cdots\cdots \mathrm{v}_{kk} \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_k \end{bmatrix}$$

حيث

$\hat{\alpha}$ معاملات الدالة التمييزية المعيارية.

v^{-1} : معكوس مصفوفة التباين والتغاير المشترك.

d_i : مصفوفة المسافة بين متوسط المُتغيرات في كلا المجموعتين.

الأهمية النسبية للعوامل المؤثرة (المُتغيرات المستقلة): –

بعد قيام التحليل التمييزي بإنشاء وتكوين الدوال التمييزية تظهر له ميزة إضافية وهي تحديد الأهمية النسبية للمتغيرات المستقلة والمؤثرة في عملية التمييز والفصل بين المجموعات وترتيبها ويتم ذلك من خلال استبعاد إشارات المعاملات المعيارية لدالة التمييز وصاحب أعلى قيمة هو الأكثر أهمية أما عن نسبة المُساهمة في عملية التمييز تحدد من خلال مُعامل الارتباط القانوني "Canonical correlation"

اختبارات الدالة التمييزية:

لاختبار قدرة الدالة على التمييز والفصل بين المجموعات تستخدم الاختبارات الآتية: –

١– اختبار F (F test)

وذلك لاختبار قدرة الدالة على التمييز وعن طريق الفرضية التي تنص على ان الدالة ليس لديها القدرة على التمييز (H_0) ضد الدالة لديها القدرة على التمييز (H_1) ويعتمد هذا الاختبار على قياس الاختلافات بين المجموعات وداخل المجموعات بين المفردات ويتم ذلك من خلال تكوين جدول تحليل التباين التالي: –

Source	SS	Df	Ms	F
بين المجموعات Between x's	SSB	k−1	M_{SB}	M_{SB}
الخطأ Within x's	SSE	n−k	M_{SE}	M_{SE}
الكلي Total	SST	n−1		

حيث ان: –

١– مجموع مربعات الأخطاء يحسب كالتالي: –

$$SSE = D^2 = \hat{\alpha}_1\ d_1 + \hat{\alpha}_2\ d_2 + \cdots\cdots\cdots\cdots\cdots\cdots + \hat{\alpha}_k\ d_k$$

٢– مجموع مربعات بين المتغيرات: –

$$SSB = \frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 2)} \times (D^2)^2$$

٣– مجموع مربعات الكلى: –

$$SST = SSB + SSE$$

ويتم الاختيار كالتالي:

١- صياغة الفروض:

H_0 : الدالة ليس لها قدرة على التمييز

H_1 : الدالة لها القدرة على التمييز

٢- القيمة المحسوبة:

$$F = \frac{M_{SB}}{M_{SE}}$$

٣- القيمة الجدولية:

$$F\ (k-1, n\ -k)$$

٤- القرار:

إذا كانت F المحسوبة أكبر من F الجدولية نرفض الفرض العدمي ونقبل بالفرض البديل ويكون للدالة قدرة عالية على التمييز والعكس صحيح.

٢- اختبار ويلكس لمدا (A) Wil'ks Lambda:

تأخذ الفروض الشكل الآتي:

الدالة ليس لها مقدرة على التمييز

H_0: $\mu_1 = \mu_2$

الدالة لها القدرة على التمييز H_1: $\mu_1 \neq \mu_2$

القيمة المحسوبة

$$\Lambda = \prod_{i=1}^{k} \frac{1}{1+\lambda_i}$$

λ_i الجذر الكامن (eigenvalues) لكل المُتغيرات.

k عدد المُتغيرات

القرار: تنحصر قيمة

$$o \leq \Lambda \leq 1$$

إذا كان

$\Lambda = 1$ معناها تساوي متوسطات المجموعتين وبالتالي عدم مقدره الدالة على التمييز والفصل.

$\Lambda = 0$ معناها عدم تساوي متوسطات المجموعتين والدالة لها القدرة عالية على التمييز.

إذا اقتربت قيمة Λ من الواحد دليل على عدم مقدرة الدالة على التمييز وإذا اقتربت من الصفر دليل على قدرة الدالة على التمييز.

وتستخدم إحصائية "ويلكس لمدا" لاختبار معنوية المتغيرات الداخلة في النموذج حيث يتم الإبقاء على المُتغيرات لها أدنى قيمة لإحصائية Wilk's Lambda وأعلى قيمة لـ F.

٣- اختبار هوتلنج (T^2) Hotelling – Lawely test

إحصاء هوتلنج تأخذ الشكل الآتي: -

$$T^2 = \sum_{i=1}^{s} \lambda_i$$

حيث ان

λ_i eigenvalues : الجذور المميزة للمُتَغيرات

s : عدد المتغيرات

وتعادل إحصائية هوتلنج قيمة F من جدول تحليل التباين ويمكن تحويله إلى قيمة لها توزيع F تقريبي صيغته كالتالي:

$$F = \frac{n_1 + n_2 - k - 1}{(n_1 + n_2 - 2)k} * T^2$$

والقيمة الجدولية:

$$F_{\propto}(k-1, n_1 + n_2 - k - 1)$$

إذا كانت F المحسوبة أكبر من F الجدولية رفض الفرض العدمي وقبول البديل بإن للدالة قدرة عالية على التمييز.

نقطة الفصل (القطع) Cut Of Point:

بعد تكوين الدالة التمييزية واختبار قدرتها على التمييز والفصل بين المجموعتين يبدأ الاستخدام الثاني لها وهو كيفية تصنيف المشاهدة الجديدة إلى أي المجموعتين تنتمي ويتم ذلك من خلال الخطوات الآتية: -

١- تحديد نقطة الفصل وهي تمثل متوسط المتوسطين:

$$\bar{\bar{L}} = \frac{\bar{L}_{(1)} + \bar{L}_{(2)}}{2}$$

حيث ان

$\bar{\bar{L}}$: نقطة الفصل.

$\bar{L}_{(1)}$ متوسط القيم التمييزية للمجموعة الأولى.

$\bar{L}_{(2)}$ متوسط القيم التمييزية للمجموعة الثانية.

قاعدة التصنيف Classification Role:

من خلال هذه القاعدة يُمكن تصنيف أو التنبؤ بانتماء مفردة جديدة لإحدى المجموعتين بأقل خطأ تصنيف على النحو التالي:

١) إذا كان $\bar{L}_{(1)} > \bar{L}_{(2)}$

وإذا كانت القيمة التمييزية للمفردة الجديدة أكبر من نقطة الفصل تصنف ضمن المجموعة الأولى وإذا كانت القيمة التمييزية للمفردة الجديدة أقل من نقطة الفصل تصنف ضمن المجموعة الثانية وإذا ساوت نقطة الفصل تصنف عشوائيًا ضمن أي مجموعة من المجموعتين.

٢) إذا كان $\bar{L}_{(1)} < \bar{L}_{(2)}$

وإذا كانت القيمة التمييزية للمفردة الجديدة أعلى من نقطة الفصل تصنف ضمن المجموع الثانية وإذا كانت أقل تصنف ضمن المجموعة الأولى وإذا تساوت معها تصنف عشوائيًا ضمن أي مجموعة في المجموعتين.

أخطاء التصنيف:

يقصد بأخطاء التصنيف وضع المفردة في مجموعة غير مناسبة لها أي وضع مفردة في مجموعة ما ولكن هي تنتمي لمجموعة أخرى ويعتبر خطأ التصنيف عامل مُهم عند الحكم على كفاءة الدالة التمييزية.

هناك نوعان من أخطاء التصنيف هُما:

١- خطأ التصنيف الظاهري.

ويحسب من جدول التصنيف التالي.

مجموع	تابع المجموعة الثانية (٢)	تابع المجموعة الأولى (١)	المجموعة
n_1	n_{12}	n_{11}	الأولى (١)
n_2	n_{22}	n_{21}	الثانية (٢)

n_{11} : عدد المفردات من المجموعة الأولى والتي تم تصنيفها في نفس المجموعة وبالتالي هي صنفت بطريقة صحيحة.

n_{12}: عدد المفردات من المجموعة الأولى والتي تم تصنيفها خطأ في المجموعة الثانية.

n_{21}: عدد المفردات التي تنتمي بالأصل إلى المجموعة الثانية وتم تصنيفها خطأ في المجموعة الأولى.

n_{22}: عدد المفردات في المجموعة الثانية التي تم تصنيفها في نفس المجموعة وبالتاي هي صنفت بطريقة صحيحة.

ويُحسب الخطأ الظاهري كما يلي:

$$P_{12} = \frac{n_{12}}{n_1}$$

P_{12} نسبة المفردات التي تنتمي للمجموعة الأولى وصنفت خطأ للثانية.

$$P_{21} = \frac{n_{21}}{n_2}$$

P_{21} نسبة المفردات التي تنتمي للمجموعة الثانية وصنفت خطأ في الأولى. ويمكن حساب معدل الخطأ الظاهري باستخدام المعادلة

$$\frac{n_{12} + n_{21}}{n_1 + n_2}$$

٢- الخطأ الحقيقي: يمثل نسبة التصنيف الخاطئ في المجتمع:

$$P_{12} = P_{21} = F\left[\frac{-\sqrt{D^2}}{2}\right]$$

حيث F دالة التوزيع الطبيعي المعياري، D إحصائية Mahalnobis. تحسب القيمة بين القوسين ويحسب الاحتمال المقابل لها من جدول التوزيع الطبيعي المعياري وكلما اقترب الاحتمال من الصفر دل على صنف وانخفاض خطأ التوصيف وبالتالي قدرة الدالة على التمييز والتصنيف اما إذا كان الاحتمال قريب من الواحد يدل على ارتفاع خطأ التوصيف وانخفاض قدره الدالة على التمييز والتصنيف.

الدالة التمييزية بمعاملات غير معيارية: -

تأخذ الشكل التالي:

$$y = b_o + b_1 x_1 + b_2 x_2 \ \cdots\cdots\cdots\cdots\cdots\ b_k x_k$$

y: الدالة التمييزية غير المعيارية.

b_o: ثابت التمايز.

b_n's: معاملات التمييز غير المعيارية.

x_n's: المتغيرات غير المعيارية.

وللحكم على جودة النموذج التمييزي من خلال مُعامل الارتباط القانوني Canonical correlation حيث ان القيم المرتفعة لمعامل الارتباط القانوني تكون مؤشر لجودة التوفيق العالي للنموذج التمييزي وبتربيع قيمة مُعامل الارتباط القانوني تحصل على قيمة معامل التحديد "R^2" الذي يحدد نسبة مُساهمة المُتغيرات المستقلة في التمييز والتصنيف.

المبحث الثاني

الجانب التطبيقي

تمهيد:

بالرغم من الدور الذي يلعبه التحليل احادي المتغير أو ثنائي المتغير في تفسير وتحليل كثير من الظواهر الاقتصادية والاجتماعية والطبية إلا أنهُ عندما يتعلق الأمر بعدد كبير من المُتغيرات فلابد من اللجوء إلى التحليل مُتعدد المتغيرات ومن أهم أساليب التحليل متعدد المتغيرات والشائع استخدامه في المجالات الطبية هو أسلوب التحليل التمييزي والذي يقوم بدوره بالتمييز وفصل الأشخاص إلى مجموعتين رئيسيتين هُما أما (مُصاب أو غير مُصاب) بمرض السكري وذلك على عينة من ٣٥٠ شخص منهم ١٦٨ مصاب و١٨٢ غر مصاب بهدف الوصول إلى دالة تمييزية من خلالها يتم تصنيف الأشخاص أو المشاهدات الجديدة على احدى المجموعتين بناءً على فرضيات معينة.

١- مُتغيرات البحث:

تتمثل مُتغيرات البحث في مُتغير تابع نوعي ثنائي القيمة (y) غير مُصاب (٠) ومُصاب (١) ومجموعة من العوامل المؤثرة (المُتغيرات المستقلة) وهي.

١. الوراثة (x_1)تأخذ (٠ لا يوجد، ١ يوجد).
٢. الوزن.(x_2)
٣. ضغط الدم (x_3) تأخذ (٠ طبيعي، ١ مُرتفع).
٤. العُمر(x_4) .
٥. النوع (x_5) تأخذ (٠ أنثى، ١ ذكر).
٦. التدخين (x_6) تأخذ (٠ لا يُدخن، ١ مُدخن).
٧. ممارسة الرياضة (x_7) تأخذ (٠ لا يُمارس، يُمارس رياضة).
٨. مرض النُقرس (x_8) تأخذ (٠ لا يوجد، ١ يوجد).
٩. الكوليسترول (x_9) تأخذ (٠ لا يوجد، ١ يوجد).
١٠. الحالة الاجتماعية (x_{10}) تأخذ (٠ أعزب، ١ متزوج).
١١. أمراض القلب والكلى(x_{11}) تأخذ (٠ لا يوجد، ١ يوجد).

ولاستخدام تحليل التمييزي مجموعة من الافتراضات لابد من توافرها وهي:

١- اختبار التوزيع الطبيعي للبيانات: -

نظرًا ان حجم العينة يزيد عن ٣٠ مُفردة طبقًا لنظرية النهاية المركزية فإن البيانات تتبع التوزيع الطبيعي ولا داعي لإجراء اختبار الطبيعية.

٢- اختبار تساوي متوسطي المجموعتين.

بالنظر إلى الجدول رقم (١) التالي:

المتغير	المجموعة		Sig
	الأولى (المصابين) (١)	الثاني (غير المصابين) (١)	
	المتوسط	المتوسط	
x_1	0.7381	0.3352	000
x_2	100.3274	73.7692	000
x_3	0.7202	0.3187	000
x_4	48.333	38.2363	000
x_5	0.6667	0.5385	0.014
x_6	0.3512	0.6154	000
x_7	0.3512	0.6044	000
x_8	0.5476	0.5549	0.891
x_9	0.7143	0.4121	000
x_{10}	0.7381	0.6099	0.011
x_{11}	0.5179	0.2802	000

من المُلاحظ من خلال جدول (١) ان قيمة Sig أقل من ٠,٠٥ وبالتالي معنوية الفرق بين متوسطي كل متغير في المجموعتين أو خلال فئات المُتغير التابع فيما عدا المُتغير الثامن.

كما نُلاحظ أن المتوسطات الأعلى في المجموعة الأولى للمصابين في كلاً من x_2، x_4، x_1، x_9، x_3 وهُما (الوزن والعُمر والوراثة والكوليسترول وضغط الدم وذلك أمر بديهي فإن المصابين بمرض السكري نسبة الضغط العالي والوزن الزائد والعُمر المتقدم أما المتوسطات الأعلى في المجموعة الثانية وهي مجموعة غير المصابين تتمثل في x_2، x_4، x_6.

ومن خلال إحصائية ويلكس لامدا Wilks's lambda distribution التالي:

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	Df	Sig.
1	.365	346.661	8	.000

حيث ان الفروض تصاغ:

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

حيث Sig (000) أقل من ٠,٠٥ رفض الفرض العدمي وقبول الفرض البديل وبالتالي هناك اختلاف بين متوسطي المجموعتين كما ان إحصاء ويكلس لامدا تساوي ٠,٣٦٥ وهي تقترب من الصفر دليل على وجود اختلاف بين متوسطي المجموعتين وهذا يعني أن الدالة التميزية لديها القدرة على التمييز وتصنيف المشاهدات إلى مجتمعها الحقيقي.

٣- اختبار فرضية تجانس التباين بين المجموعتين:

حيث تصاغ الفروض الإحصائية كالتالي:

$$H_0\ \Sigma_1 = \Sigma_2$$
$$H_1\ \Sigma_1 \neq \Sigma_2$$

وباستخدام اختبار Box's M كانت النتائج كالتالي:

Log Determinants

Y	Rank	Log Determinant
N	8	1.230
Y	8	.976
Pooled within-groups	8	1.244

Test Results

Box's M		47.341
F	Approx.	1.283
	df1	36
	df2	402161.071
	Sig.	0.119

ومن الملاحظ أن قيمة (0.119) Sig أكبر من ٠,٠٥ وبالتالي قبول الفرض العدمي بتساوي مصفوفة التباين والتغاير للمجموعتين وبالتالي تحقق افتراض تجانس التباين بين المجموعتين.

٤- اختبار معنوية العوامل المؤثرة (المُتغيرات المستقلة) في النموذج التمييزي:

تم اختبار معنوية جميع العوامل المؤثرة في النموذج التمييزي لمعرفة أهمية كل مُتغير ومدى إسهامه في عملية التمييز والتصنيف وكانت كالتالي **Tests of Equality Group of Means**

	Wilks' Lambda	F	df1	df2	Sig.
x1	.857	58.046	1	348	.000
x2	.477	381.840	1	348	.000
x3	.839	66.816	1	348	.000
x4	.873	50.479	1	348	.000
x5	.983	6.052	1	348	.014
x6	.930	26.082	1	348	.000
x7	.936	23.829	1	348	.000
x8	1.000	.019	1	348	.891
x9	.908	35.422	1	348	.000
x10	.981	6.594	1	348	.011
x11	.941	21.827	1	348	.000

ومن المُلاحظ أن جميع المُتغيرات تتمتع بمعنوية عالية حيث ان (000) Sig أقل من (0.05) لجميع المتغيرات ماعدا المتغير الثامن وذلك يدل على المتغيرات لها تأثير معنوي كبير في عملية التمييز بين المجموعتين ومن ثم توصلنا إلى النموذج التحليل التميزي مناسب لبيانات مرضى السكري.

١) تكوين الدالة التمييزية:

لإنشاء الدالة التمييزية تحدد أولاً المُتغيرات الداخلة في تكوين الدالة التمييزية حيث ان اختزال عدد المُتغيرات في نموذج التمييز يفيد في قياس المتغيرات ذات العلاقة المعنوية وذات الصلة الأكبر بالموضوع محل الدراسة وللتعرف على المتغيرات ذات القوة التمييزية المعنوية والتي تغطي أقل خطأ تصنيف (James, 1985). وهُناك عدة معايير إحصائية للإدخال والحذف وهي الإبقاء على التمييز صاحب القيمة الأكبر لـ F وأقل قيمة لإحصائية ويلكس لامدا Wilks's lambdaكما في الجدول الآتي:

Variables Entered/Removed[a,b,c,d]

Step	Entered	Wilks' Lambda							
						Exact F			
		Statistic	df1	df2	df3	Statistic	df1	df2	Sig.
1	x2	.477	1	1	348.000	381.840	1	348.000	.000
2	x4	.441	2	1	348.000	219.946	2	347.000	.000
3	x3	.417	3	1	348.000	160.993	3	346.000	.000
4	x1	.400	4	1	348.000	129.221	4	345.000	.000
5	x7	.390	5	1	348.000	107.676	5	344.000	.000
6	x5	.380	6	1	348.000	93.114	6	343.000	.000
7	x6	.370	7	1	348.000	83.038	7	342.000	.000
8	x9	.365	8	1	348.000	74.142	8	341.000	.000

المصدر: SPSS V23

ونُلاحظ من الجدول أنه تم استبعاد ثلاثة مُتغيرات وتم الإبقاء على ثمان مُتغيرات التي لها قدرة أعلى في التمييز والفصل بين المجموعتين المصابين وغير المصابين والتي لها اعلى قيمة F وأقل قيمة Wilks's lambda وتم الاختيار بناءً على الاختيار التدريجي على ٨ خطوات وبالتالي المتغيرات الداخلة للنموذج هي x_2، x_4، x_3، x_1، x_7، x_5، x_6، x_9.

إيجاد الدالة التمييزية: –

أولاً: من جدول رقم (١) تم إيجاد متوسطات المُتغيرات في كلا المجموعتين.

ثانيًا: إيجاد الفرق بين متوسط كل مُتغير في كلا المجموعتين.

$$d = \bar{x}_{i\,(1)} - \bar{x}_{i\,(2)} = \begin{bmatrix} \bar{x}_{11} - \bar{x}_{12} \\ \bar{x}_{21} - \bar{x}_{22} \\ \vdots \\ \bar{x}_{9\,(1)} - \bar{x}_{9\,(2)} \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ | \\ d_9 \end{bmatrix}$$

ثالثًا: مصفوفة التباين والتغاير بين المجموعتين.

$$V = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_9 \end{array} \begin{array}{c} \begin{array}{cccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_9 \end{array} \\ \begin{bmatrix} 0.284 & 2.969 & 0.50 & 1.446 & 0.049 & 0-0.41 & -0.41 & 0.57 \\ 2.969 & 337.465 & 3.136 & 62.014 & 0.293 & 0-1.640 & -1.528 & 2.410 \\ 0.050 & 3.136 & 0.251 & 0.960 & -.007 & -.047 & -.024 & 0.035 \\ 1.446 & 62.014 & 201.452 & 201.45 & -.348 & -1.571 & -.814 & 0.916 \\ 0.049 & 0.293 & -.007 & -.348 & 0.241 & 0.053 & -.013 & -.014 \\ -.041 & -1.640 & -.047 & -1.571 & 0.053 & 0.251 & 0.016 & -.041 \\ -.041 & -1.528 & -.024 & -.814 & -.013 & 0.016 & 0.250 & -.035 \\ 0.057 & 2.410 & 0.035 & 0.914 & -.014 & -.041 & -.035 & 0.247 \end{bmatrix} \end{array}$$

وبذلك الدالة التمييزية بمعاملات معيارية: –

$$L = \hat{\propto}_1 x_1 + \hat{\propto}_2 x_2 + \hat{\propto}_3 x_3 + \quad + \hat{\propto}_9 x_9$$

حيث ان

$$\hat{\propto} = v^1 d$$

$$\hat{\propto} - \begin{bmatrix} 0.160 \\ 0.797 \\ 0.267 \\ 0.295 \\ 0.268 \\ -0.203 \\ -0.187 \\ 0.154 \end{bmatrix}$$

$$\hat{L} = 0.160 x_1 + 0.797 x_2 + 0.267 x_3 + 0.295 x_4 + 0.268 x_5 - 0.203 x_6 \; 0.187 x_7 +$$

$$0.154 x_9$$

ولتحديد الأهمية النسبية للعوامل المؤثرة ونسبة المساهمة في التمييز والتنبؤ في النموذج التمييزي:

لتحديد أكثر العوامل أثر على مستوى الإصابة ومُساهمة العامل في التمييز والتصنيف مكان كالتالي:

المُتغير	المُعامل	الأهمية النسبية (مُعامل الارتباط القانوني التمييزي)
x_2	0.797	0.794
x_4	0.295	0.289
x_5	0.268	0.10
x_3	0.267	0.332
x_6	−0.203	−0.208
x_7	−0.187	−0.198
x_1	0.16	0.310
x_9	0.154	0.242

ولمعرفة أهم العوامل المؤثرة تنظر لعمود المعاملات المعيارية ($\propto_i$)حيث انه القيمة المطلقة الكبيرة يقابلها العامل الأكثر أهمية في التأثير على الإصابة وتكون هذه الأهمية موجبة أو سالب.

نُلاحظ أن أكثر المُتغيرات أهمية (x2) الوزن ثم (x4) العُمر ثم (x5) النوع ثم (x3) ضغط الدم ثم (x6) التدخين ثُم (x9) ممارسة رياضة ثم (x1) الوراثة ثم (x9) الكوليسترول.

أما الأهمية النسبية فإن (x2) الوزن يساهم بنسبة أكبر في عملية تمييز المجموعتين ٧٩,٤% يليه ضغط الدم ٣٣,٢% والوراثة ٣١% العُمر ٢٨,٩% وكوليسترول ٢٤,٢% والتدخين ٢٠,٨% وممارسة الرياضية ١٩,٨% وأخيرًا النوع ١٠%.

ولاختبار قدرة الدالة على التمييز:

١) باستخدام جدول تحليل التباين واختبار F.

- ومن خلال الفروض الآتية:

H_0 الدالة ليس لها قدرة على التمييز

H_1 الدالة لها القدرة على التمييز

- إيجاد قيمة مجموع مُربعات الأخطاء (داخل المجموعات).

$$SSE = D^2 = \propto_1 d_1 + \propto_2 d_2 + \propto_3 d_3 + \propto_4 d_4 + \propto_5 d_5 + \propto_6 d_6 + \propto_7 d_7 + \propto_9 d_9$$

$$= [0.160 \quad 0.797 \quad 0.267 \quad 0.295 \quad 0.268 \quad -0.203 \quad -0187 \quad 0.154] \begin{bmatrix} 0.4029 \\ 26.55 \\ 0.4015 \\ 10.09 \\ 0.1282 \\ -0.2642 \\ -0.2532 \\ 0.3022 \end{bmatrix}$$

$$SSE = 24.50$$

مجموع مُربعات بين المتغيرات

$$SSB = \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2)} \times (D^2)^2 = \frac{168\times182}{(168+182)(168+182-2)} * 24.5^2 = 150.68$$

مجموع المُربعات الكلية:

$$SST = SSB + SSE = 150.68 + 24.5 = 175.18$$

جدول تحليل التباين

Source	Ss	Df	*Ms*	F
بين المجموعات	150.68	k−1 7	٢١,٥٢٥	
داخل المجموعات	24.5	n−k 342	٠,٠٧٢	٣٠٠,٥
الكلي	175.18	n−1 349		

القيمة المحسوبة:

$$F = 300.5$$

القيمة الجدولية:

$$F0.05\ (7,\ 342) =\ 1.40$$

F القيمة المحسوبة أكبر F الجدولية ∴ رفض الفرض العدمي وقبول الفرض البديل فإن للدالة قدرة عالية على التمييز والفعل بين المجموعتين.

٢) اختبار ويلكس لامدا Wilks' lambda:

تُصاغ الفروض كالتالي:

$H_0\ \mu_1 = \mu_2$ الدالة ليس لها قدرة على التمييز

$H_1\ \mu_1 \neq \mu_2$ الدالة لها القدرة على التمييز

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	Df	Sig.
1	.365	346.661	8	.000

أولاً: قيمة إحصائية ويلكس لامدا ٠,٣٦٥ وهي أقرب للصفر وذلك دليل على القدرة العالية للدالة على التمييز. وكما نُلاحظ أن Sig (00) أقل من ٠,٠٥ وبالتالي رفض الفرض العدمي وقبول الفرض البديل بإن للدالة قدرة على التمييز والفصل بين المجموعتين.

كما أن في جدول:

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.739[a]	100.0	100.0	.797

ومن المُلاحظ قيمة الجذر الكامن ١,٧٣٩ = ⋋ وتُشير إلى أن نسبة التباين المُفسر بين مجموعتي المصابين وغير المصابين والتي تعود للفروق بينها في النموذج التمييزي الوحيد وجمع **قيمة مُعامل الارتباط القانوني ٠,٧٩٧ مُعامل الارتباط بين مجموعة العوامل المؤثرة ونموذج التمييز الوحيد وبتربيع هذه القيمة تحصل على ٦٣,٥% وهذا يعني نسبة مُساهمة العوامل المؤثرة في التباين والاختلاف في التمييز بين المجموعتين.**

للاستخدام الثاني للنموذج التمييزي وهو التصنيف فكانت النتاج كالتالي:

	غير مُصاب (٠)	مصاب (١)	مجموع
غير مُصاب (٠)	١٦٩	١٣	١٨٢
مُصاب (١)	٢٠	١٤٨	١٦٨
٩٠,٦٠% النسبة الإجمالية.			

حيث ان النموذج التمييزي الذي يتكون من ثمان مُتغيرات هُما x_1، x_2، x_3، x_4، x_5، x_6، x_7، x_9، قام بالتصنيف الصحيح ١٦٩ مفردة من غير المُصابين بمرض السكري وتصنيف غير صحيح ١٣ أي نسبة تصنيف صحيح ٩٢,٩% في الحالات بشكل صحيح. أما بالنسبة للمصابين قام تصنيف الصحيح ١٤٨ مُفردة وغير الصحيح ٢٠ مُفردة بنسبة إجمالية صحيحة ٨٨,١% وإن الدقة الإجمالية للتصنيف هي ٩٠,٦% بخطأ ٨,٤%.

نسبة الخطأ الظاهري:

للمجموعة الأولى:

$$p_{12} = \frac{n_{12}}{n_1} = \frac{13}{182} = 0.0714$$

للمجموعة الثانية:

$$p_{21} = \frac{n_{21}}{n_1} = \frac{20}{168} = 0.119$$

تقدير النموذج التمييزي بمُعاملات غير معيارية: -

يتم تقدير النموذج التمييزي بمعاملات غير معيارية كما يلي:

المُتغير	b
x1	0.324
x2	0.063
x3	0.581
x4	0.022
x5	0.551
x6	−.420
x7	−.386
x9	0.325
(Constant)	−6.974

$$\hat{y} = -6.974 + 0.324\,x_1 + 0.063\,x_2 + 0.581\,x_3\ 0.022\,x_4$$
$$+0551\,x_5 - 0420\,x_6 - 0386\,x_7 + 0.325\,x_9$$

وهذا النموذج فعال وقادر على التصنيف الصحيح للمرضى بنسبة ٩٠,٦% وأقل خطأ تصنيف ٩,٤%.

تصنيف المُشاهدات الجديدة:

١- إيجاد نقطة الفصل:

حساب متوسط المجموعة الأولى (المصابين) $\bar{y}_1 = 1.369$

حساب متوسط المجموعة الثانية (غير المصابين) $\bar{y}_2 = 1.263$

نقطة الفصل $\bar{\bar{y}} = \frac{1.369-1.263}{2} = 0.052$

$$\bar{y}_1 > \bar{y}_2 \because$$

إذا كانت المُفردة الجديدة أكبر في النقطة الفاصلة تصنف المُفردة إلى المجموعة الأولى وإذا كانت أقل في النقطة الفاصلة تصنف إلى المجموعة الثانية.

فمثلاً: – إذا كانت هُناك سيدة ($x_5 = 0$) لديها مرض السكري وراثة ($x_1 = 1$) والوزن ($x_2 = 100$) والعُمر (x_4=70) ولديها ضغط الدم ($x_3 = 1$) ولا تُدخن ($x_6 = 0$) ولا تُمارس رياضة ($x_7 = 0$) وليس لديها كوليسترول ($x_9 = 0$).

$$\hat{y} = -6.974 + 0.324\,(1) + 0.063\,(100) + 0.581\,(1) + 0.022\,(70) + 0.551\,(0) - 0.420\,(0) - 0.386\,(0) + 0.325\,(0) = 1.771$$

∴ القيمة التمييزية للمُشاهدة الجديدة أكبر من نقطة الفصل بالتالي تصنف ضمن المجموعة الأولى (المُصابين) لمرض السُكري.

وإذا كان هناك رجل ($x_5 = 1$) ليس لديه مرض وراثي ($x_1 = 0$) وزنه ($x_2 = 65$) وضغط الدم عادي ($x_3 = 0$) وعُمره ($x_4 = 60$) يُدخن ($x_6 = 1$) ولا يُمارس رياضة ($x_7 = 0$) وكوليسترول عالي ($x_9 = 1$) القيمة التمييزية له.

$$\hat{y} = -6.974 + 0.324\,(0) + 0.063\,(65) + 0.581\,(0) + 0.022\,(60) + 0.551\,(1) - 0.420\,(1) - 0.386\,(0) + 0.325\,(1) = -1.103$$

القيمة التمييزية للمشاهدة أقل من نقطة الفصل وبالتالي تصنف ضمن المجموعة الثانية (غير المُصابين). وبالتالي فالنموذج التمييزي المقدر بنسبة مُساهمة العوامل المؤثرة فيه ٦٣,٥% وكفاءة النموذج في النصف ٩٠,٦% أما الحساسين أي تصنف غير المصاب على أنهُ غير مصاب تمثل ٩٢,٩% أما النوعية تصنف المصاب على أنهُ مصاب بنسبة ٨٨,١% ونسبة خطأ تصنيف ٩,٤%.

النتائج والتوصيات:

يهدف البحث إلى استخدام التحليل التمييزي كأحد أساليب التحليل مُتعدد الحدود لتحديد أهم العوامل المؤثرة فالإصابة بمرض السُكري وذلك من خلال مُتغير تابع نوعي (مُصاب أو غير مُصاب) مجموعة من المُتغيرات (العوامل) المُستقلة وهي عامل الوراثة والوزن وضغط الدم والعُمر والنوع والتدخين وممارسة رياضة ومرض النقرس والكوليسترول والحالة الاجتماعية وأمراض القلب والكلى.

وتم التوصل إلى النتائج التالية:

١- بعد التأكد من توافر افتراضات أسلوب تحليل التمييزي وهي شرط طبيعة البيانات وشرط عدم تساوي متوسطات المجموعتين وتساوي مصفوفة التباين والتغاير بين المجموعتين ومعنوية غالبية العوامل المؤثرة توصلت إلى ملائمة أسلوب التحليل التمييزي لبيانات مرضى السكري أي يُمكن استخدامه في تمييز وتصنيف المفردات الجديدة إلى مصابين أو غير مُصابين وفقًا لمجموعة العوامل المُستقلة.

٢- باختبار معنوية العوامل المؤثرة تم استبعاد ثلاث مُعاملات (متغيرات) هي أمراض النقرس، الحالة الاجتماعية وامراض القلب اما باقي المُتغيرات لها معنوية عالية في أسلوب تحليل التمييزي.

٣- وبالتالي فإن الدالة التمييزية للفصل والتمييز بين المجموعتين بمعاملات معيارية هي.

$$\hat{L} = 0.160\,x_1\; 0.797\,x_2 + 0.267\,x_3 + 0.295\,x_4 + 0.268\,x_5 - 0.203\,x_6 - 0.187\,x_7 + 0.154\,x_9$$

٤- أكثر العوامل المؤثرة وأهمها على الإصابة بمرض السُكري هو الوزن ثم العُمر ثم ضغط الدم ثم التدخين وممارسة الرياضة والوراثة والكوليسترول.

٥- أكثر العوامل مُساهمة في التمييز بين المجموعتين هو الوزن لنسبة مُساهمة ٧٩,٤% يليه ضغط الدم بنسبة ٣٣,٢% يليه الوراثة ٣١% ثم العُمر ٢٨,٩% والكوليسترول بنسبة ٢٤,٢% ثم التدخين بنسبة ٢٠,٨% وممارسة الرياضة ١٩,٨% وأخيرًا النوع بنسبة ١٠%.

٦- العوامل المؤثرة في الإصابة بمرض السُكري تساهم بنسبة ٦٣,٥% من التمييز والتصنيف بين المجموعتين.

٧- النموذج التمييزي ذو كفاءة عالية التصنيف بنسبة ٩٠,٦% وحساسية ٩٢,٩% ونوعية ٨٨,١%.

٨- نسبة خطأ التنصيف صغيرة ٨,٤%.

التوصيات:

يوصي الباحث بـ: -

١- التوسع في استخدام التحليل التمييزي كأحد أساليب التحليل مُتعدد المُتغيرات في المجالات الاقتصادية والاجتماعية.

٢- استخدام التحليل التمييزي لتحديد العوامل المؤثرة في الإصابة بمرض السُكري مع إضافة مُتغيرات أخرى كالنظام الغذائي وتناول الكحوليات وغيرها.

٣- استخدام نموذج الدالة التمييزية في التشخيص المُبكر.

٤- الوصول للوزن المثالي تفاديًا من الإصابة بمرض السُكري.

٥- الاهتمام بالتحليل الإحصائي وابراز الدور الهام له في الجانب الطبي.

٦- استخدام أساليب تحليل مُتعدد المعدات لوس كالانحدار اللوجستي والانحدار المُتعدد وتحليل التباين في تحديد العوامل المؤثرة في الإصابة بمرض السُكري ومقارنة النتائج بالتحليل التمييزي.

المراجع العربية:

١- الجاعوني، غانم (٢٠٠٧) "التحليل الإحصائي مُتعدد المتغيرات (التحليل التمييزي) في توصيف وتوزيع الاسر داخل الهيكل الاقتصادي الاجتماعي في المجتمع"، ورقة بحثية منشورة، مجلة جامعة دمشق للعلوم الاقتصادية والقانونية المُجلد (٢٣) العدد الثاني، سوريا.

٢- الجزار، ماجد عطية (٢٠١٢) "دراسة مُقارنة بين التحليل التمييزي الخطي والانحدار اللوجستي المُتعدد في التصنيف والتنبؤ" جامعة الأزهر، كلية الاقتصاد والعلوم الإدارية، قسم الإحصاء التطبيقي، غزة.

٣- الشمراني، مُحمد بن موسى الشمراني (٢٠٠٨) "دراسة مُقارنة بين التحليل التمييزي وتحليل التباين في تحليل البيانات مُتعددة المُتغيرات" رسالة دكتوراه، كلية التربية، جامعة أم القُرى، المملكة العربية السعودية.

٤- النويري، فريال محمد (٢٠١٣) "استخدام الدالة التمييزية الخطية لتمييز مرضى السكري المصابين من غير المصابين بالفشل الكلوي" رسالة ماجستير، جامعة الجزيرة، كلية الاقتصاد والتنمية الريفية، قسم الإحصاء التطبيقي. السودان.

٥- بغرش، سعيدة، (٢٠٢٠) "استخدام التحليل التمييزي في تقدير خطر عدم تسديد القرض من طرق الوكالة الوطنية لتسيير القرض المُصغر" مجلة نماء للاقتصاد والتجارة – المُجلد (٤)، العدد (١) الجزائر.

٦- خوالدي، سليمة (٢٠١٧) "دور التحليل التمييزي في التنبؤ بالفشل المالي لعينة من المؤسسات الصغيرة والمتوسطة بولاية أم البواقي لفترة (٢٠١٤-٢٠١٦)" رسالة ماجستير غير منشورة جامعة العربي بن مهيدي، كلية العلوم الاقتصادية والتجارية، أم البواقي.

٧- سليمان، على ابشر فضل (٢٠١٥) "المقارنة بين التحليل التمييزي والنموذج اللوجستي ونماذج الشبكات العصبية في تصنيف المشاهدات" رسالة دكتوراه غير منشورة، جامعة السودان للعلوم والتكنولوجيا، كلية الدراسات العليا.

٨- عبد الكريم، أنوار ضياء (٢٠٠٦) "استخدام الطرائق التمييزية الإحصائية لتشخيص بعض أمراض القلب "ورقة بحثية منشورة، مجلة جامعة كركوك – الدراسات العلمية، مجلد (١) العدد (٢)، العراق.

٩- هاشم، غراء (٢٠١٤) "استخدام التحليل التمييزي المُتعدد لتصنيف مراحل الإصابة بمرض الفشل الكلوي

١٠-المُزمن"، جامعة السودان للعلوم والتكنولوجيا – كلية العلوم، قسم الإحصاء التطبيقي

المراجع الأجنبية:

1- Abdul Hussein, S. F., (2019), The use of mahalnobis statistic in the linear discriminant analysis between two groups, Al – Mustans-yriah university, the journal of administration and economics, vol-119, pp. 59-66.

2- Ahmed, M. S., (1998) "A Companion of The Discriminant and Logistic Regression Approach" ph.p thesis, ISSR, Cairo University.

3- Anderson, T. W., (1984). An introduction to Multivariate statistical analysis. and Edition, John and Sons, New York, USA.

4- Chandran, R.K., (2009). The Effectiveness of stepwise Discriminant Analysis. Antimicrob Agents Chemother. 2009 July; 53 (7): 2887-2891.

5- Ferrer, A. J. A. and Wang, L. (1999). Comparing The Classification Accuracy among Non-parametric, parametric at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).

6- Geofry J. M., (1992), Discriminant Analysi and statistical pattern Rocognitien The university of Queen slan, Wiley.

7- Hosmer .D and Lemeshow S., (2000). Appleid Logestic Rogression. 2nd Edtion wiley: New York. In CBSU Librany Athired edtion is due to be published in 2013.

8- Kandil, A, M, (1992) "Discriminant with mixture of continuous, discrete and nominal variables", The Egyptian Journal, ISSR, Cairo Univ, Vol36 No1, 102, 117.

9- Krznowski, W, T. (1995). "Multivariate statistics Classification covariance structure", John Wiley Stens, Inc, New York.

10- Pohar, M., and Blas, M., (2004), Comparison of logistic regression and Linear discriminat analysis: Asimulation study, Metodoloski Zvzki, Vol-1, pp, 143-161.

11- Rausch, j.R., and Kelley, k. (2009). A Comparison of linear and mixture models for discriminat analysis under Non normality, Behavior research methods, Vol.41, pp. 85-98.

12- Wang .y., (2008). Comparing Linear Discriminainat Analysis With Classi fcatim trees using forest Landwenr Survey Data a case study. M. S Thesis. The Universty of tennesses. knoxville.