طريقة هجينة للتصنيف بالاعتماد على طريقة تحليل التمايز الخطي لفيشر وخوارزمية آلية متجه الدعم (Fisher-LDA & SVM)

هبة حباق*، خضر الكريدي، محمد طاهر عنان****

* طالبة دراسات عليا(ماجستير)، قسم الإحصاء الرياضي، كلية العلوم، جامعة حلب

** قسم الإحصاء الرياضي، كلية العلوم، جامعة حلب

الملخص

تمّ في هذا البحث اقتراح منهجية جديدة للتصنيف الثنائي تقوم على بناء قاعدة تصنيف بالاعتماد على تحليل التمايز الخطي لفيشر وآلية متجه الدعم، الهدف منها التخلص من الشروط المفروضة على تحليل التمايز والاستفادة من دالة التمايز الخطية في استخلاص المميزات من مجموعة المتغيرات الأصلية واعتبارها قاعدة أولية لآلية متجه الدعم.

طُبّقت الطريقة الجديدة في التصنيف على بيانات تمّ توليدها عشوائياً وعلى بيانات حقيقية لأنواع مرض السكري (النمط الأول والنمط الثاني)، حيث تمّ بناء قاعدة تصنيف يمكن استخدامها في التنبؤ لشخص ما غير مصاب في المستقبل بأن يصاب بأحد أنماط السكري إمّا النمط الأول أو النمط الثاني، حيث أعطت الطريقة الهجينة في التطبيقين تصنيفاً صحيحاً بمقدار 100%.

الكلمات المفتاحية: تحليل التمايز الخطي، دالة التمايز، آلية متجه الدعم، الهامش الأعظمي.

ورد البحث للمجلة بتاريخ 14/ 4/ 2019

قبل للنشر بتاريخ 1 / 7 / 2019

A Hybrid Method of Classification Based on the Fisher Linear Discriminant Analysis Method and the Support Vector Machine Algorithm (Fisher-LDA & SVM)

Hiba Habbak*, Khoder Alkrede, Mohammad Taher Anan****

*Postgraduate Student (MSc), Dept. of Statistics, Faculty of Science, University of Aleppo

** Dept. of Statistics, Faculty of Science, University of Aleppo

Abstract

This research proposed a new methodology for binary classification based on classification base building which is based on Fisher linear discriminant analysis and the support vector machine. The main objective is to get rid of the conditions imposed on differentiation analysis and make use of linear discriminant function to draw the features from the Original variables group and consider it as an initial base of support vector machine.

The new method was applied to randomly generated data classification and real data types for diabetes (type 1 and type 2). A base of classification was built to be used in predicting if someone could be injured in the future with diabetes type 1 or type 2. The hybrid method in both applications was rated true by 100%.

Keywords: linear discriminant analysis, discriminant function, maximum margin support vector mechanism.

Received 14 /4 /2019
Accepted 1 /7 /2019

مقدمة (Introduction):

يعتمد أسلوب التحليل الإحصائي على نوع المشكلة محل القياس والمعالجة ونوع البيانات، ويعتبر التصنيف أحد أساليب التحليل الإحصائي متعدد المتغيرات، حيث يعتمد على وصف وتحليل الظواهر ذات الأبعاد والمتغيرات المتعددة، فإذا كانت المتغيرات تشترك فيما بينها بمجموعة من الخصائص والصفات بدرجات متفاوتة وتؤثر على متغير ما، فإنّ التحليل الإحصائي يتناول دراسة بيانات تلك المتغيرات والتعبير عنها بشكل رياضي من خلال أكثر المتغيرات تأثيراً في الظاهرة محل الدراسة. الجوهر الأساسي لبحثنا كان التركيز على طريقة تحليل التمايز الخطي لفيشر واعتماد دالة التمايز الناتجة عنها كقاعدة أساسية لاستخدامها في آلية متجه الدعم في التصنيف الثنائي وذلك للحصول على دقة تصنيف أفضل للمتغيرات المدروسة.

أهمية البحث (The Importance of Research):

تكمن أهمية البحث في إيجاد قاعدة تصنيف بالاعتماد على خوارزمية آلية متجه الدعم واعتماد دالة التمايز الخطية لفيشر كقاعدة لبنائها، حيث يمكن اعتماد قاعدة التصنيف الناتجة عن الطريقة المقترحة كقاعدة أساسية في التنبؤ لقيم جديدة غير معروفة عائديتها إلى أي مجموعة.

هدف البحث (The Aim Of Research):

1- التعرّف على خوارزمية آلية متجه الدعم في التصنيف الثنائي لما لها من أهمية في تحليل البيانات متعددة المتغيرات.

2- اقتراح منهجية جديدة تعتمد على طريقة تحليل التمايز الخطي لفيشر وخوارزمية آلية متجه الدعم في التصنيف الثنائي.

3- المقارنة بين الطريقة المقترحة مع طريقة تحليل التمايز الخطي لفيشر.

منهجية البحث (Methodology):

اعتمد منهج التحليل الاستدلالي للوصول إلى النتائج العلمية والمنطقية، بالإضافة لاستخدام البرامج الحاسوبية المناسبة EXCEL، R بما يخدم أغراض البحث.

الدراسة النظرية (Theoretical Study):

1- طريقة تحليل التمايز الخطي (Linear Discriminant Analysis):

يعتبر تحليل التمايز من أساليب التحليل متعدد المتغيرات التي تهتم بفصل مجموعات مختلفة من المفردات (المشاهدات)، إنّ عملية التصنيف هي العملية اللاحقة لعملية تكوين الدالة المميزة التي هي عبارة عن تركيب خطي للمتغيرات الأصلية، يتم الاعتماد على هذه الدالة لتصنيف المفردة الجديدة لإحدى المجموعات قيد الدراسة بأقل خطأ تصنيف ممكن.[1]

التمايز الخطي هو إحدى حالات التمييز التي يتطلب فيها شرط تساوي التباينات للمجموعات المدروسة، هناك تمايز خطي في حالة مجموعتين، وتمايز خطي في حالة أكثر من مجموعتين. أمّا التمايز غير الخطي فيستخدم في حالة عدم تساوي التباينات.

تُستخدم دالة التمايز الخطية عندما تكون المجتمعات المدروسة: 1) ذات توزيع طبيعي متعدد المتغيرات 2) متجهات متوسطات مختلفة 3) مصفوفات تباين وتباين مشترك متساوية.[2][1]

ففي بحثنا هذا سنهتم بدراسة التمايز الخطي في حال مجموعتين.

تحليل التمايز الخطي لفيشر (Fisher linear Discriminant Analysis): [3]

إذا كانت لدينا المتغيرات $(X_1, X_2, \dots, X_p)$ فالدالة الخطية يمكن أن تأخذ الصيغة التالية:

$$Y = b_1X_1 + b_2X_2 + \cdots + b_pX_p \qquad (1,1)$$

حيث Y تركيبة خطية للمتغيرات الأصلية تدعى بدالة التمايز وتكتب بالصيغة التالية:

$$Y = b'X \qquad (1,2)$$

اقترح فيشر لتقدير معاملات التصنيف $(b_1, b_2, \dots, b_p)'$ بحيث تجعل الدالة تعطي أفضل تمييز بين المجموعتين وذلك بجعل مربع الفرق بين متوسطي المجموعتين إلى التباين (مجموع المربعات داخل المجموعات) أكبر ما يمكن.

$$J(b) = \frac{Between\ Group}{Within\ Group} \qquad (1,3)$$

ويعبر عن تابع فيشر بالشكل التالي:

$$J(b) = \frac{b' S_B b}{b' S_W b} \qquad (1,4)$$

حيث أنّ: S_B: (التباين بين المجموعات) وتعرّف كما يلي:

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)' \qquad (1,5)$$

$J(b)$: تابع فيشر.

μ_1 : متجه المتوسط للمجموعة الأولى و μ_2 متجه المتوسط للمجموعة الثانية.

S_W :(التباين ضمن المجموعات) وتعطى بالعلاقة التالية:

$$S_W = S_1^2 + S_2^2 \qquad (1,6)$$

$$S_i^2 = \sum_{x \in \Pi_i} (x - \mu_i)(x - \mu_i)' \quad ; i = 1,2$$

حيث S_i^2 مصفوفة التباين للمجموعة i.

بأخذ القيمة العظمى لتابع فيشر نحصل على متجه "معاملات تابع التصنيف" من المرتبة $(p \times 1)$

$$b = arc\ max \left[\frac{b' S_B b}{b' S_W b}\right] = S_w^{-1}(\mu_1 - \mu_2) \qquad (1,7)$$

نلاحظ أنّ $Y = b'X$ عبارة عن توليفة خطية ناتجة عن القيمة العظمى لنسبة التباين بين المجموعات إلى داخل المجموعات، وتمثل المستقيم الفاصل بين المجموعتين. في حالة وجود مجموعتين يكون لدينا دالة تمايز واحدة فقط، وفي حالة وجود ثلاث مجموعات يكون لدينا دالتي تمايز وهكذا...[4]

نصنّف المشاهدات الجديدة الناتجة عن Y إلى إحدى المجموعتين بالاعتماد على نقطة وسط المجموعتين (L) التي تجعل احتمال التصنيف الخاطئ أقل ما يمكن.

$$L = \frac{\bar{Y}_1 + \bar{Y}_2}{2} \qquad (1,8)$$

حيث أنّ: $\bar{Y}_1$ ، $\bar{Y}_2$ تمثل متوسطي القيم التمييزية للمجموعتين الأولى والثانية على الترتيب.

1) إذا كانت $Y > L$ نصنّف المشاهدة إلى المجموعة الأولى G_1.
2) إذا كانت $Y < L$ نصنّف المشاهدة إلى المجموعة الثانية G_2.
3) أما إذا كانت $Y = L$ تصنّف المشاهدة عشوائياً إلى المجموعة الأولى أو الثانية.

2- خوارزمية آلية متجه الدعم (Support Vector Machine):[5]

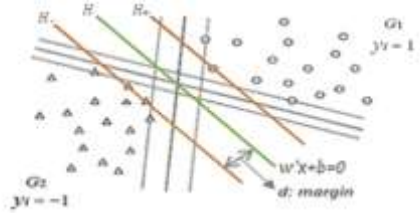
قُدّمت هذه التقنية في عام (1992) من قبل العالم فابنك (Vapnik) وهي عبارة عن خوارزمية تعلّم عن طريق مشرف أو موجّه تستعمل للتصنيف مستندة إلى نظرية التعلّم الإحصائية (Statistical Learning Theory)، واعتمدت مبدأ المخاطرة البنيوية ((SRM) Structural Risk Minimization) في تخفيض الحد الأعلى للمخاطرة المتوقعة.

تَستخدم هذه التقنية مجموعة جزئية من أمثلة التدريب تعرف بمتجهات الدعم (Support Vector) لتمثيل حد القرار (Decision Boundary) من خلال البحث عن أكبر هامش تبتعد عنه البيانات من الطرفين، وهو مستو فائق ذو هوامش أعظمية (Maximum Margin Hyperplane).[6]

ولتكن $X = (x_1, x_2, \dots, x_p)'$ مجموعة من السمات لبيانات تصنيف ثنائية ويراد تصنيفها في مجموعتين G_1 , G_2 ، وبأخذ مجموعة تدريب $x_1, x_2, \dots, x_n$ من مجموعة السمات p ، الهدف هو العثور على مستو فائق يفصل بيانات التدريب فصلاً تاماً إلى مجموعتين G_1 ،G_2، فإذا لوحظ أنّ النقطة x_i تنتمي إلى المجموعة G_1 نضع $y_i = 1$ وإذا كانت تنتمي للمجموعة G_2 وضعنا $y_i = -1$، يمكننا تمثيل n من بيانات التدريب على النحو التالي:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n); y_i = \begin{cases} 1 & if \quad x_i \in G_1 \\ -1 & if \quad x_i \in G_2 \end{cases} \qquad (2,1)$$

منهجية التصنيف لـ SVM Classification Methodology:



الشكل (1) آلية التصنيف وفق الـ SVM

نقوم برسم مستقيمين متوازيين H_+ , H_- من خلال نقاط الحافة لكل مجموعة، نرسم حد القرار H هو منتصف المسافة بين المستقيمين المرسومين عند نقاط الحافة،

يمكن ملاحظة وجود أكثر من مستو لفصل البيانات إلا أنّنا نبحث عن المستو الذي تبتعد أو تقع عليه متجهات الدعم، بحيث يكون أبعد ما يمكن عن نقاط المجموعتين.

من المفترض أنّ كل بيانات التدريب يجب أن تصنّف في مجموعتين G_2، G_1بالشكل:

$$x_i \in G_1 \Leftrightarrow w'x_i + b \geq 1 \qquad (2,2)$$
$$x_i \in G_2 \Leftrightarrow w'x_i + b \leq -1$$

حيث: w الأوزان، b الانحياز.

وتعطى المسافة بين بيانات التدريب وحد القرار بالشكل التالي:

$$d = \frac{|w'x + b|}{\|w\|} \; ; \; \|w\| = \sqrt{w_1^2 + w_2^2 + \cdots + w_p^2} \qquad (2,3)$$

- **<u>تعلّم آلية متجه الدعم</u>**:[8][7][4]

تهدف عملية التعلّم الوصول لأكبر هامش تبتعد عنه البيانات من الطرفين، والتي تؤول إلى مسألة برمجة غير خطية مقيّدة متمثلة بدالة الهدف:

$$f(w) = \min_w \frac{\|w\|^2}{2} \qquad (2,4)$$
$$Subject\ to \quad y_i(w'x_i + b) \geq 1 \quad ; \quad i = 1,2,\dots,N$$

تُحل هذه المسألة باستخدام طريقة لاغرانج، وهي طريقة لإيجاد قيم النهايات الحدية المحلية لدالة لعدة متحولات خاضعة لواحد أو أكثر من القيود، نشكّل دالة لاغرانج من أجل مسألة الأمثلية:

$$L(w, b, \lambda_1, \lambda_2, \dots \lambda_n) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \lambda_i (y_i(w'x_i + b) - 1) \qquad (2,5)$$

حيث λ_i هي مضاريب لاغرانج.

لجعل دالة لاغرانج أصغرية نعطي المشتقات الجزئية لـ L بالنسبة لـ w,b القيمة صفر:

$$\frac{\partial L(w, b, \lambda_1, \lambda_2, \dots \lambda_n)}{\partial \mathrm{w}} = 0 \Rightarrow w = \sum_{i=1}^{n} \lambda_i y_i x_i \qquad (2,6)$$

$$\frac{\partial L(w, b, \lambda_1, \lambda_2, \dots \lambda_n)}{\partial \mathrm{b}} = 0 \Rightarrow \sum_{i=1}^{n} \lambda_i y_i = 0 \qquad (2,7)$$

نعوض في العلاقة (2,5) فنجد:

$$L_D(\lambda_1, \lambda_2, \dots, \lambda_n) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i' x_j \qquad (2,8)$$

وبالتالي نكون قد حصلنا على مسألة مزدوجة بين w و λ_i ، تابع الهدف للمسألة المزدوجة يحتاج لإيجاد الحد الأعظمي:

$$Max\ L_D(\lambda_1, \lambda_2, \dots, \lambda_n) = \max \left\{ \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j x_i' x_j \right\}$$

$$Subject\ to \quad \lambda_i \geq 0 \quad ; i = 1,2, \dots, n \qquad (2,9)$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0$$

وهي مسألة برمجة تربيعية ((QP) Quadratic Programming) بحلها نحصل على:

متجه الأوزان $\widehat{w} = \sum_{i=1}^{n} \hat{\lambda}_i y_i x_i$ والانحياز $\hat{b} = -\frac{1}{2}(\widehat{w}' x_+ + \widehat{w}' x_-)$

الأمثليين وحد قرار مع هوامش أعظمية $\widehat{w}' x + \hat{b} = \sum_{i=1}^{n} \hat{\lambda}_i y_i x_i' x + \hat{b} = 0$

وفقاً لذلك فإنّ قاعدة التصنيف للمجوعتين G_1،G_2 تكون كالتالي:

$$\widehat{w}' x + \hat{b} = \begin{cases} \sum_{i=1}^{n} y_i \hat{\lambda}_i x_i' + \hat{b} \geq 0 \Rightarrow G_1 \\ \sum_{i=1}^{n} y_i \hat{\lambda}_i x_i' + \hat{b} < 0 \Rightarrow G_2 \end{cases} \qquad (2,10)$$

3- الطريقة المقترحة:

تتلخص الطريقة الجديدة المقترحة في التصنيف بالخطوات التالية:

* الخطوة الأولى: في البداية نطبّق تحليل التمايز الخطي لفيشر لإيجاد دالة التمايز الخطية $Y = b'X$
* الخطوة الثانية: تعويض كل متجه من البيانات في دالة التمايز السابقة لنحصل على متغير جديد z_1 .

$$z_{i1} = b_1\, x_{i1} + b_2\, x_{i2} + \cdots + b_p\, x_{ip}\ ; i = 1,2, \dots, N \qquad (3,1)$$

* الخطـوة الثالثـة: إيجـاد انحرافـات القـيم x_{ij} حيـث أنّ $(i = 1,2, \dots, N\ , j = 1,2, \dots, p)$ عـن متوسـطاتها لكـل مجموعـة وتعويضـها فـي دالـة التمـايز لنحصـل أيضاً على متغير جديد z_2 .

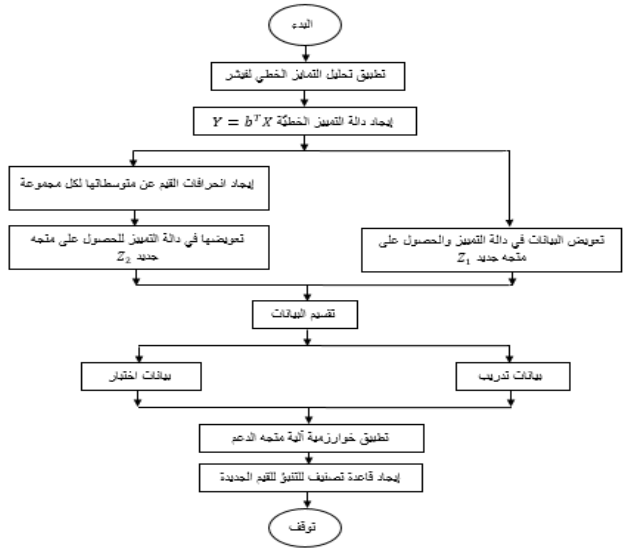
$$z_{i2} = b_1 m_{i1} + b_2 m_{i2} + \cdots + b_p m_{ip} \qquad (3,2)$$

حيث انحرافات القيم عن متوسطاتها تعطى بالعلاقة:

$$m_{ij} = x_{ij} - \bar{x}_{kj}\ ; i = 1,2, \dots, N\ , j = 1,2, \dots, p\ , k = 1,2$$

$\bar{x}_{kj}$: متوسط المتغير j للمجموعة k.

* الخطـوة الرابعـة: نعتبـر المتغيـرين السـابقين $Z_1\ , Z_2$ متغيـرات جديـدة ونطبّـق خوارزميـة آليـة متجـه الـدعم عليهـا لنحصـل علـى قاعـدة تصـنيف يمكـن الاعتمـاد عليها للتنبؤ في عائدية المفردات الجديدة إلى المجموعتين $G_1\ , G_2$.



الشكل (2) المخطط النهجي للطريقة المقترحة

الدراسة التطبيقية:

➢ التطبيق الأول:

جُمعت البيانـات مـن خـلال عيّنـة أُخذت مـن البرنـامج الـوطني للسـكري فـي محافظة حلب في عام 2018، يعتبر داء السكري متلازمة تتصف باضطراب وارتفاع شاذ في تركيز سكر الدم ناتج: إمّا عن عوز هرمون الأنسولين أو مقاومة الأنسولين

أو كليهما معاً، يصنّف كنمط أول أو نمط ثاني، ومرضى النمط الأول يحتاجون الأنسولين بصورة يومية ودائمة، أمّا النمط الثاني بحاجة لتناول الأدوية الفموية المخفضة لنسبة السكر في الدم.

عينة البحث هي عينة بسيطة تتضمن نوعين من المتغيرات الثنائية والمتصلة لمجموعتين من أنواع مرض السكري (النمط الأول، النمط الثاني) والتي تمثل المتغير التابع Y (1 تدل على النمط الأول، 2 تدل على النمط الثاني)، وتمّ دراسة المتغيرات التالية لكل مجموعة:

X_1 : العمر X_2 : الجنس X_3 : الوزن X_4 : الطول

X_5 : كتلة الجسم (BMI)

الجدول (1) متغيرات التطبيق الأول

المجموعات	أسماء المجموعات	عدد الحالات	الاحتمال
1	النمط الأول	36	0.29
2	النمط الثاني	87	0.71
		123	1

التصنيف باستخدام تحليل التمايز الخطي لفيشر:

باستخدام تحليل التمايز الخطي على المتغيرات وباستخدام برنامج R وبعد التأكد من شروط تحليل التمايز، حصلنا على النتائج التالية:

* يظهر الجدول التالي نتائج دالة التمايز ومقياس Wilks' Lambda الذي تتراوح قيمته بين 0 و 1، فإذا كانت قيمته قريبة أو مساوية للواحدة فإنّ ذلك يشير إلى أنّ متوسطات المجموعات متساوية وبذلك لا يوجد تمايز بين المجموعات المدروسة ويعني هذا أنّ دالة التمايز المحسوبة فاشلة، إذا كانت قيمته قريبة أو مساوية للصفر هذا يدل على قوة التمايز ويمكن اعتماد الدالة في التصنيف.

الجدول (2) ملخص نتائج دالة التمايز– المصدر من إعداد الباحث

الدالة	Wilks' Lambda	Chi-square	Df	Sig
1	0.4	108.444	5	0.000

يتضح من الجدول السابق معنوية النموذج (0.000) حيث بلغت قيمة إحصاء Wilks' Lambda (0.4) وهو قريب من الصفر وبالتالي يمكن الاعتماد على

هذا النموذج للتمايز بين المجموعتين.

* الجدول التالي يظهر معاملات التصنيف لتابع فيشر.

الجدول (3) معاملات التصنيف لتابع فيشر–المصدر من إعداد الباحث باستخدام برنامج R

الدالة	معاملات التصنيف
1	
X_1 : العمر	0.07 *
X_2 : الجنس	-0.11
X_3 : الوزن	0.03 *
X_4 : الطول	-0.01 *
X_5 : كتلة الجسم	-0.03 *

* يظهر الجدول التالي كفاءة التصنيف الصحيح ونسبته المئوية.

الجدول (4) كفاءة التصنيف الصحيح ونسبته المئوية لدالة التمايز–المصدر من إعداد الباحث باستخدام برنامج R

Y	التصنيف الصحيح		الإجمالي
	1	2	
العدد الأصلي 1 2	31 5	5 82	36 87
النسب المئوية 1 2	%86.11 %5.75	%13.89 %94.25	%100 %100
كفاءة التصنيف الصحيح 91.87%			
نسبة التصنيف الخاطئ 8.13%			

يتضح من الجدول السابق أنّ:

- 31 مشاهدة من المجموعة الأولى تمّ تصنيفها تصنيفاً صحيحاً أي بنسبة 86.11%.
- 5 مشاهدات من المجموعة الأولى تمّ تصنيفها في المجموعة الثانية أي تصنيف خاطئ بنسبة 13.89%.
- 82 مشاهدة من المجموعة الثانية تمّ تصنيفها تصنيفاً صحيحاً أي بنسبة 94.25%.
- 5 مشاهدات من المجموعة الثانية تمّ تصنيفها في المجموعة الأولى أي تصنيف

خاطئ بنسبة 5.75%.

- نسبة التصنيف الصحيحة هي 91.87% وهي نسبة جيدة وبالتالي فإنّ تصنيف فيشر هو تصنيف جيد.

التصنيف باستخدام الطريقة المقترحة:

بعدما تمّ إيجاد دالة التمايز الخطية لفيشر وكانت بالشكل التالي:

$$Y = 0.07\,X_1 - 0.11\,X_2 + 0.03\,X_3 - 0.01\,X_4 - 0.03\,X_5$$

نقوم بتعويض عينة البحث في هذه الدالة لإيجاد المتغير الأول Z_1.

$z_{11} = 0.07\,(x_{11}) - 0.11\,(x_{12}) + 0.03\,(x_{13}) - 0.01\,(x_{14}) - 0.03\,(x_{15})$
$z_{21} = 0.07\,(x_{21}) - 0.11\,(x_{22}) + 0.03\,(x_{23}) - 0.01\,(x_{24}) - 0.03\,(x_{25})$
$\vdots$
$z_{N1} = 0.07\,(x_{N1}) - 0.11\,(x_{N2}) + 0.03\,(x_{N3}) - 0.01\,(x_{N4}) - 0.03\,(x_{N5})$

نوجد انحرافات قيم العينة للمجموعة الأولى والثانية عن متوسطاتها بالشكل التالي:

$m_{11} = x_{11} - \bar{x}_{11}$	$m_{12} = x_{12} - \bar{x}_{12}$	$m_{13} = x_{13} - \bar{x}_{13}$	$m_{14} = x_{14} - \bar{x}_{14}$	$m_{15} = x_{15} - \bar{x}_{15}$
$m_{12} = x_{21} - \bar{x}_{11}$	$m_{22} = x_{22} - \bar{x}_{12}$	$m_{32} = x_{23} - \bar{x}_{13}$	$m_{42} = x_{24} - \bar{x}_{14}$	$m_{52} = x_{25} - \bar{x}_{15}$
$\vdots$				
$m_{n_1 1} = x_{n_1 1} - \bar{x}_{11}$	$m_{n_1 2} = x_{n_1 2} - \bar{x}_{12}$	$m_{n_1 3} = x_{n_1 3} - \bar{x}_{13}$	$m_{n_1 4} = x_{n_1 4} - \bar{x}_{14}$	$m_{n_1 5} = x_{n_1 5} - \bar{x}_{15}$
$m_{(n_1+1)\,1} = x_{(n_1+1)\,1} - \bar{x}_{21}$	$m_{(n_1+1)\,2} = x_{(n_1+1)\,2} - \bar{x}_{22}$	$m_{(n_1+1)\,3} = x_{(n_1+1)\,3} - \bar{x}_{23}$	$m_{(n_1+1)\,4} = x_{(n_1+1)\,4} - \bar{x}_{24}$	$m_{(n_1+1)\,5} = x_{(n_1+1)\,5} - \bar{x}_{25}$
$\vdots$				
$m_{N1} = x_{N1} - \bar{x}_{21}$	$m_{N2} = x_{N2} - \bar{x}_{22}$	$m_{N3} = x_{N3} - \bar{x}_{23}$	$m_{N4} = x_{N4} - \bar{x}_{24}$	$m_{N5} = x_{N5} - \bar{x}_{25}$

نقوم بتعويض القيم الناتجة عن الانحرافات في دالة التمايز وإيجاد المتغير الثاني Z_2.

$z_{12} = 0.07\,(m_{11}) - 0.11\,(m_{12}) + 0.03\,(m_{13}) - 0.01\,(m_{14}) - 0.03\,(m_{15})$
$z_{22} = 0.07\,(m_{21}) - 0.11\,(m_{22}) + 0.03\,(m_{23}) - 0.01\,(m_{24}) - 0.03\,(m_{25})$
$\vdots$
$z_{N2} = 0.07\,(m_{N1}) - 0.11\,(m_{N2}) + 0.03\,(m_{N3}) - 0.01\,(m_{N4}) - 0.03\,(m_{N5})$

نقوم باعتبار Z_2 ، Z_1 متغيرات جديدة ونطبّق خوارزمية آلية متجه الدعم عليهما مع الأخذ بعين الاعتبار أنّ Y هو المتغير التابع.

التصنيف باستخدام آلية متجه الدعم:

الجدول التالي يبيّن عدد بيانات التدريب وعدد بيانات الاختبار لبيانات النمط الأول والثاني:

الجدول (5) عدد عينات التدريب والاختبار للنمط الأول والثاني للسكري–المصدر من إعداد الباحث باستخدام برنامج R

البيانات Data الصنف Classes	مجموعة التدريب Training Set	مجموعة الاختبار Testing Set	المجموع Total
النمط الأول	13	23	36
النمط الثاني	16	71	87
المجموع Total	29	94	123
النسبة المئوية لمجموعة البيانات % of Dataset	%24	%76	%100

بعدما تمّ تقسيم البيانات لمجموعة تدريب ومجموعة اختبار، يتم تدريب الشبكة بالاعتماد على مجموعة التدريب ومن ثم اختبارها على مجموعة الاختبار حيث يجب أن يتطابق خرجي الهدف والفعلي للشبكة، فالجدول التالي يبيّن أداء الشبكة بعد تدريبها واختبارها:

الجدول (6) متجهي الخرج والهدف بعد تطبيق آلية متجه الدعم–المصدر من إعداد الباحث

مجموعة الاختبار Testing Set	النمط الأول	النمط الثاني	المجموع Total
العدد الأصلي للخرج الهدف	23 0	0 71	23 71
العدد الأصلي للخرج الفعلي	23 0	0 71	23 71
النسبة المئوية لتطابق الخرجين	%100 %0	%0 %100	%100 %100

الجدول السابق يبيّن أداء الشبكة إذ صنّفت 100% من مجموعة الاختبار بشكل صحيح حيث تطابق متجهي الخرج والهدف.

الجدول التالي يظهر كفاءة التصنيف الصحيح ونسبته المئوية على المتغيرين Z_1 ، Z_2 باستخدام آلية متجه الدعم.

الجدول (7) كفاءة التصنيف الصحيح ونسبته المئوية–المصدر من إعداد الباحث باستخدام برنامج R

Y	التصنيف الصحيح		الإجمالي
	1	2	
العدد الأصلي 1 2	36 0	0 87	36 87
النسب المئوية 1 2	%100 %0	%0 %100	%100 %100
كفاءة التصنيف 100%			
نسبة الخطأ في التصنيف 0%			

يتضح من الجدول السابق أنّ:

- جميع مشاهدات المجموعة الأولى صنّفت تصنيفاً صحيحاً أي بنسبة 100%.
- جميع مشاهدات المجموعة الثانية صنّفت تصنيفاً صحيحاً أي بنسبة 100%.
- نسبة التصنيف الصحيحة كانت 100% باستخدام الطريقة المقترحة.

➢ **التطبيق الثاني:**

هي عينة عشوائية بسيطة تم توليدها عشوائياً باستخدام برنامج الـ Excel وتتضمن نوعين من المتغيرات ثنائية ومتصلة لمجموعتين.

تتكون عينة البحث من 55 مشاهدة، 30 مشاهدة للمجموعة الأولى و 25 مشاهدة للمجموعة الثانية، و 5 متغيرات عشوائية X_1, X_2, X_3, X_4, X_5 حيث فرضنا أنّ المتغيرات هي متغيرات متصلة ويمثل Y المجموعات الأولى والثانية.

الجدول (8) متغيرات التطبيق الثاني

المجموعات	عدد المشاهدات	الاحتمال
1	30	0.55
2	25	0.45
	55	1

التصنيف باستخدام تحليل التمايز الخطي لفيشر:

باستخدام تحليل التمايز الخطي لفيشر على المتغيرات الأصلية وباستخدام برنامج R حصلنا على النتائج التالية:

يظهر الجدول التالي نتائج دالة التمايز ومقياس Wilks' Lambda.

الجدول (9) ملخص نتائج دالة التمايز–المصدر من إعداد الباحث

Sig	Df	Chi-square	Wilks' Lambda	الدالة
0.000	5	43.925	0.42	1

يتضح من الجدول السابق معنوية النموذج (0.000) حيث بلغت قيمة إحصاء Wilks' Lambda (0.42) وهو قريب من الصفر وبالتالي يمكن الاعتماد على هذا النموذج للتمايز بين المجموعتين.

الجدول التالي يظهر معاملات التصنيف لتابع فيشر.

الجدول (10) معاملات التصنيف لتابع فيشر–المصدر من إعداد الباحث باستخدام برنامج R

معاملات التصنيف	الدالة
	1
-0.039 *	X_1
-0.015 *	X_2
-0.015 *	X_3
-0.037 *	X_4
-0.039 *	X_5

يظهر الجدول التالي كفاءة التصنيف الصحيح ونسبته المئوية.

الجدول (11) كفاءة التصنيف الصحيح ونسبته المئوية لدالة التمايز–المصدر من إعداد الباحث باستخدام برنامج R

الإجمالي	التصنيف الصحيح		Y
	2	1	
30 25	2 21	28 4	العدد الأصلي 1 2
%100 %100	%6.67 %84	%93.33 %16	النسب المئوية 1 2
كفاءة التصنيف الصحيح 89.09%			
نسبة التصنيف الخاطئ 10.91%			

يتضح من الجدول السابق أنّ:

- 28 مشاهدة من المجموعة الأولى تمّ تصنيفها تصنيفاً صحيحاً أي بنسبة 93.33%.
- 2 مشاهدتين من المجموعة الأولى تمّ تصنيفها في المجموعة الثانية أي تصنيف خاطئ بنسبة 6.67%.

- 21 مشاهدة من المجموعة الثانية تمّ تصنيفها تصنيفاً صحيحاً أي بنسبة 84%.
- 4 مشاهدات من المجموعة الثانية تمّ تصنيفها في المجموعة الأولى أي تصنيف خاطئ بنسبة 6%.
- نسبة التصنيف الصحيحة هي 89.09% وهي نسبة جيدة وبالتالي فإنّ تصنيف فيشر هو تصنيف جيد.

التصنيف باستخدام الطريقة المقترحة:

بعدما تمّ إيجاد دالة التمايز الخطي لفيشر وكانت بالشكل التالي:

$$Y = -0.039\,X_1 - 0.015\,X_2 - 0.015\,X_3 - 0.037\,X_4 - 0.039\,X_5$$

- نقوم بتعويض عينة البحث في هذه الدالة لإيجاد المتغير الأول Z_1 كما في التطبيق السابق.
- نوجد انحرافات قيم العينة للمجموعة الأولى والثانية عن متوسطاتها، ونقوم بتعويض القيم الناتجة عن الانحرافات في دالة التمايز وإيجاد المتغير الثاني Z_2.
- نقوم باعتبار Z_2 ، Z_1 متغيرات جديدة ونطبّق خوارزمية آلية متجه الدعم عليها مع الأخذ بعين الاعتبار أنّ Y هو المتغير التابع.

التصنيف باستخدام آلية متجه الدعم:

الجدول التالي يبيّن عدد بيانات التدريب وعدد بيانات الاختبار لبيانات المجموعتين الأولى والثانية:

الجدول (12) عدد عينات التدريب والاختبار للنمط الأول والثاني للسكري–المصدر من إعداد الباحث باستخدام برنامج R

البيانات Data الصنف Classes	مجموعة التدريب Training Set	مجموعة الاختبار Testing Set	المجموع Total
المجموعة الأولى	9	21	30
المجموعة الثانية	8	17	25
المجموع Total	17	38	55
النسبة المئوية لمجموعة البيانات % of Dataset	%31	%69	%100

الجدول التالي يبيّن أداء الشبكة بعد تدريبها واختبارها:

الجدول (13) متجهي الخرج والهدف بعد تطبيق آلية متجه الدعم–المصدر من إعداد الباحث

مجموعة الاختبار Testing Set	المجموعة الأولى	المجموعة الثانية	المجموع Total
العدد الأصلي للخرج الهدف	21 0	0 17	21 17
العدد الأصلي للخرج الفعلي	21 0	0 17	21 17
النسبة المئوية لتطابق الخرجين	%100 %0	%0 %100	%100 %100

ويبيّن الجدول السابق أداء الشبكة إذ صنّفت 100% من مجموعة الاختبار بشكل صحيح، حيث تطابق متجهي الخرج والهدف.

الجدول التالي يظهر كفاءة التصنيف الصحيح ونسبته المئوية على المتغيرين Z_1 ، Z_2 باستخدام آلية متجه الدعم.

الجدول (14) كفاءة التصنيف الصحيح ونسبته المئوية–المصدر من إعداد الباحث باستخدام برنامج R

Y	التصنيف الصحيح		الإجمالي
	1	2	
العدد الأصلي 1 2	30 0	0 25	30 25
النسب المئوية 1 2	%100 %0	%0 %100	%100 %100
كفاءة التصنيف 100%			
نسبة الخطأ في التصنيف 0%			

يتضح من الجدول السابق أنّ:

- جميع مشاهدات المجموعة الأولى صنّفت تصنيفاً صحيحاً أي بنسبة 100%.
- جميع مشاهدات المجموعة الثانية صنّفت تصنيفاً صحيحاً أي بنسبة 100%.
- نسبة التصنيف الصحيحة كانت 100% باستخدام الطريقة المقترحة.

النتائج:

1- تمكّنا من دمج طريقة تحليل التمايز الخطي لفيشر وخوارزمية آلية متجه الدعم.

2- بناء قاعدة تصنيف لعينات من مرضى السكري (النمط الأول والنمط الثاني)

باستخدام تحليل التمايز الخطي لفيشر.

3- بناء قاعدة تصنيف أخرى لعينات من مرضى السكري (النمط الأول والنمط الثاني) باستخدام خوارزمية آلية متجه الدعم وذلك بالاعتماد على دالة التمايز الخطية لفيشر كقاعدة لبنائها.

4- كانت نسبة التصنيف الصحيحة باستخدام الطريقة المقترحة 100% وهي أفضل مقارنة مع نسبة التصنيف الصحيحة باستخدام طريقة تحليل التمايز الخطي التي أعطت أخطاء في التصنيف.

التوصيات:

على ضوء ما تقدّم يمكننا اقتراح التوصيات التالية:

1- نوصي باعتماد الطريقة المقترحة في التطبيقات لبناء قواعد تصنيف واستخدامها لأغراض التنبؤ.

2- دراسة حالة وقوع المشاهدات على حد القرار في خوارزمية آلية متجه الدعم.

3- دراسة حالة مساواة قيمة دالة التمايز مع نقطة الفصل في طريقة تحليل التمايز الخطي لفيشر.

المراجع

[1] جبارة ازهار كاظم، -**مقارنه بين التحليل التمييزي الخطي واحتمال الاستجابة في تصنيف البيانات**. مجلة الإدارة والاقتصاد، المجلد الثالث، العدد الثالث عشر، 21.

[2] ILYA Narsky& Frank C. Porter., 2014- **Statistical Analysis Techniques in Particle Physics**. Wiley-VCH, United State of America, 461.

[3] صالح عائدة هادي، 2008-**استخدام التحليل المميز لتشخيص بعض أمراض العيون**. مجلة الإدارة والاقتصاد، العدد السابع والستون، 24.

[4] خانطوماني آية، 2017-**مقارنة التصنيف المتعدد باستخدام طريقة المركبات الأساسية (PCA) وطريقة تحليل التمايز الخطي لفيشر (Fisher-LDA)**. جامعة حلب، سوريا، 149.

[5] SADANORI K., -**Introduction to Multivariate Analysis**. Chuo University, Tokyo, Japan, 299.

[6] أبوزلام أحمد، 2013–**دراسة بعض طرائق الشبكات العصبية وآلة متجه الدعم والطرائق الإحصائية في مسائل التصنيف لدعم القرار الطبي**. جامعة حلب، سوريا، 98.

[7] YIZENG L., QING-SONG X., HONG-DONG L., DONG-SHENG C., 2011- **Support Victor Machines and Their Application in Chemistry and Biotechnology.** United State of America, 194.

[8] GARETH J., DANIELA W., TREVOR H., ROBERT T., -**An Introduction to Statistical Learning with Applications in R.** New York Heidelberg Dordrecht London, 426.