جمهوريـــــــة العراق
وزارة التعليم العالي والبحث العلمي
جـــــامعة كـــــــــــربلاء
كلية الإدارة والاقتصـــــاد
قســـــم الإحـــــصاء
الدراسات العليا

التصنيف الحصين بإستعمال التحليل التمييزي اللبي اللامعلمي مع تطبيق عملي

رسالة

مقدمــة الى مجلس كلية الإدارة والاقتصاد في جامعة كربلاء

وهي جزء من متطلبات نيل درجة ماجستير في علوم الإحصاء

للباحث

جعفر علي فرحان

إشـــــــراف

أ. م. د. إيناس عبد الحافظ محمد

1445هـ 2024 م

بسم الله الرحمن الرحيم

﴿ إِنَّ فِـــي خَلْـــقِ السَّـــمَاوَاتِ وَالْأَرْضِ وَاخْـــتِلَافِ اللَّيْـــلِ وَالنَّهَـــارِ لَآيَـــاتٍ لِأُولِـــي الْأَلْبَـــابِ ۝ الَّـــذِينَ يَـــذْكُرُونَ اللَّـــهَ قِيَامًـــا وَقُعُـودًا وَعَلَـى جُنُـوبِهِمْ وَيَتَفَكَّـرُونَ فِـي خَلْـقِ السَّـــمَاوَاتِ وَالْأَرْضِ رَبَّنَـــا مَـــا خَلَقْـــتَ هَـــذَا بَاطِلًا سُبْحَانَكَ فَقِنَا عَذَابَ النَّارِ ۝ ﴾

صدق الله العلي العظيم

(آل عمران)

إقرار المشـرف

أشهد أن إعداد هذه الرسالة الموسومة (**التصنيف الحصين باستعمال التحليل التمييزي اللبي اللامعلمي مع تطبيق عملي**) والتي تقدم بها الطالب " **جعفر علي فرحان منصور**" قد جرت بإشرافي في قسم الاحصاء - كلية الادارة والاقتصاد - جامعة كربلاء، وهي جزء من متطلبات نيل درجة ماجستير علوم في الاحصاء.

أ.م.د ايناس عبد الحافظ محمد

التاريخ / /2024

توصية رئيس قسم الاحصاء

بناءً على توصية الاستاذ المشرف، أرشح الرسالة للمناقشة.

أ.م.د ايناس عبد الحافظ محمد

رئيس قســم الاحصاء

التاريخ: / / 2024

إقرار الخبير اللغوى

أشهد أن الرسالة الموسومة بـ (**التصنيف الحصين بإستعمال التحليل التمييزي اللبي اللامعلمي مع تطبيق عملي)** للطالب **جعفر علي فرحان منصور /** قسم الاحصاء قد جرت مراجعتها من الناحية اللغوية حتى اصبحت خالية من الاخطاء اللغوية والاسلوبية ولأجله وقعت.

الخبير اللغوي

د. محمد سلام مظهر

جامعة كربلاء – كلية الإدارة والاقتصاد

إقرار لجنة المناقشة

نشهد نحن أعضاء لجنة المناقشة بأننا قد اطلعنا على الرسالة الموسومة (**التصنيف الحصين باستعمال التحليل التمييزي اللبي اللامعلمي مع تطبيق عملي**) والمقدمة من قبل الطالب " **جعفر علي فرحان منصور**" وناقشنا الطالب في محتوياتها وفيما له علاقة بها، ووجدنا بأنه جدير بنيل درجة ماجستير علوم في الإحصاء بتقدير(جيد جداً).

أ.د باسم شليبة مسلم
رئيساً

أ.م.د صدى فياض محمد
عضواً

م.د سهاد احمد احمد
عضواً

أ.م.د ايناس عبد الحافظ محمد
عضواً ومشرفاً

إقرار رئيس لجنة الدراسات العليا

بناءً على اقرار المشرف العلمي والخبير اللغوي على رسالة الماجستير للطالب " **جعفر علي فرحان منصور** " الموسومة بـ (**التصنيف الحصين بإستعمال التحليل التمييزي اللبي اللامعلمي مع تطبيق عملي**) ارشح هذه الرسالة للمناقشة.

أ.د علي احمد فارس
رئيس لجنة الدراسات العليا

مصادقة مجلس الكلية

صادق مجلس كلية الادارة والاقتصاد/ جامعة كربلاء على قرار لجنة المناقشة.

أ.م.د هاشم جبار الحسيني

عميد كلية الادارة والاقتصاد- جامعة كربلاء

/ /2024

الإهداء...

الى ...

من خصهم الله تعالى بالكرامة وحباهم بالرسالة محمد وآله الطيبين الطاهرين.

مَن أذكره ولا أنساه الجبل الصامد ادامك الله لي روحاً وقلباً وفخراً (والدي الحبيب).

التي غذتني طعم الحياة ونفس الدنيا .

(أمي الغالية).

الذين وقفوا بجانبي وشدوا من أزري طِوال أيام الدراسة .

(اخوتي واخواتي).

الكثير ممن يتمنون لي كل خير ويدعون لي في ظهر الغيب، وكانوا خير اخوة انجبتهم لي الحياة

(زملائي وزميلاتي).

الشموع التي أضاءت لي الطريق

(أساتذتي الفضلاء).

أهدي ثمرة جهدي المتواضع هذا

جعفر ...

شكر وإمتنان ...

بعد التوجه بالحمد والشكر للملك القدوس الذي وفقني لإنجاز هذه الرسالة .
يسرني ان أتقدم بعظيم شكري وامتناني إلى أستأذتي الفاضلة (أ.م.د إيناس عبد الحافظ محمد) لقبولها الاشراف على هذه الرسالة والتي كان لتوجيهاتها القيّمة وملاحظاتِها السديدة الأثر العميق في تذليل الكثير من الصعوبات، وفّقَها الله وجزاها خيرُ الجزاء.
كما أتقدم بجزيل الشكر والتقدير الى أساتذتي الكرام رئيس لجنة المناقشة وأعضائها لتفضلهم بالموافقة على مناقشة هذه الرسالة وتقويمها.

و وافر شكري وامتناني إلى السيدة رئيس قسم الإحصاء و أساتذتي الأجلاّء في قسم الإحصاء كافة لما قدموه لي من علم وعون كريمين، والى كل زملائي وزميلاتي ولكل مَنْ أعانني بنصحٍ بنّاء شكري وامتناني.

ومن الله التوفيق .....

الباحث...

قائمة المحتويات

قائمة الجداول

قائمة الاشكال

المصطلحات المستعملة في هذا البحث

المصطلح باللغة الانكليزية	المصطلح باللغة العربية
Discrimenant analysis	التحليل التمييزي
Linear Discriminant Function	دالة التمييز الخطي
Non – Linear Discriminant Function	الدالة التمييزية اللاخطية
Parametric Discriminant Analysis	التحليل التمييزي المعلمي
Linear Discriminant Analysis	التحليل التمييزي الخطي
Quadratic Discriminant Analysis	التحليل التمييزي التربيعي
Kernel Estimator	المقدر اللبي
kernel Functions	الدوال اللبية
Bandwidth	عرض الحزمة
Bandwidth Parameter	معلمة عرض الحزمة
Classification data	تصنيف المشاهدات
Misclassification rate	معدل خطأ التصنيف
Kernel Density Estimation	تقدير الكثافة اللبية
Non-Parametric Discriminant Analysis	التحليل التمييزي اللامعلمي

التحليل التمييزي اللبي	**Kernel Discriminant Analysis**
التحليل التمييزي اللبي الحصين	**Robust Kernel Discriminant Analysis**
طريقة التحقق المتقاطع الممهد	**Smoothed Cross –Validation**

المستخلص:

ان غالبية البيانات في عالمنا الواقعي تنحرف عن الافتراضات المثالية التي تتطلبها الأساليب الإحصائية التقليدية والتي يتسبب معها انتهاك افتراض الحالة الطبيعية في البيانات ، او ان هنالك بيانات تم تجميعها تمثل بيانات غير خطية ونتيجة لذلك قد نواجه مشكلة في التصنيف لايمكن للتحليل التمييزي التقليدي مواجهة هذه المشكلة فلابد من البحث عن طريقة حصينة تتعامل مع هذه المشكلة لذلك هدفت هذه الرسالة الى استعمال اسلوب التحليل التمييزي اللبي الحصين (Robust Kenel Discrimenant Analysis RKDA) في حالة انحراف البيانات عن الحالة الطبيعية لها ومقارنته مع التحليل التمييزي اللبي التقليدي والتحليل التمييزي الخطي والتربيعي باستعمال معيار معدل خطأ التصنيف $\widehat{MR}$ لاختيار افضل اسلوب في التصنيف وذلك من خلال جانبين ، في الجانب التجريبي وباستعمال تجارب محاكاة مونت-كارلو تبين بان اسلوب التحليل التمييزي الخطي هو الافضل من باقي اساليب التحليل التمييزي عند دوال الكثافة الهدف التي تتوزع طبيعياً (D, E) وان اسلوب التحليل التمييزي اللبي حقق افضلية عند دول الكثافة الكاوسية (D, E) عند حجم العينة (n=1000, 5000) . وحقق اسلوب التحليل التمييزي اللبي افضلية على باقي الاساليب عند دالة الكثافة (K) بنسبة قليلة. وكذلك حقق اسلوب التحليل التمييزي اللبي الحصين افضلية على باقي الاساليب عند دوال الكثافة المنحرفة عن التوزيع الطبيعي بنسبة افضلية عالية. اما الجانب التطبيقي الذي تم فيه الاعتماد على سجلات وحدة المختبر في مستشفى الحسين التعليمي في محافظة كربلاء المقدسة لغرض الحصول على المتغيرات التي لها علاقة بمرض إبيضاض الدم اللمفاوي (Lymphocytic leukemia) والتي تضمنت (100) مشاهدة من الذكور والإناث وقد قسمت المشاهدات إلى مجموعتين الأولى شملت الأشخاص غيرالمصابين بالمرض بحجم (50) مشاهدة والثانية شملت الأشخاص المصابين بالمرض بحجم (50) مشاهدة وكانت متغيرات التطبيق هي Y : متغير مثل الاصابة ام عدم الاصابة بالمرض ، اما المتغيرات التوضيحية فهي X_1 : جنس المصاب ، X_2 : خلايا الدم البيضاء WBC (White Blood Cells) ، X_3 : خلايا الدم الحمراء RBC (Red Blood Cells) ، X_4 : نسبة هيموجلوبين الدم HGB (Hemoglobin Blood) و X_5 : نسبة الصفائح الدموية PLT (Blood Platelets) وتم التوصل فيه الى ان اسلوب التحليل التمييزي اللبي الحصين اعطى نسبة خطأ التصنيف للمجموعة الأولى $\widehat{MR}_1$ (0.12) وللمجموعة الثانية $\widehat{MR}_2$ (0.56) ، وبذلك تكون نسبة خطأ التصنيف الكـــلي ($\widehat{MR}$) بلغ (0.34) وهي نسبة خطأ قليلة تدل على دقة التصنيف.

الفصل الأول

منهجية الرسالة
والإستعراض المرجعي

1-1 المقدمة (Introduction)

يعرف التحليل التمييزي (Discriminant Analysis DA) بانه عملية استكشافية وتفكيكية تستعمل في العديد من المجالات، بما في ذلك العلوم الاجتماعية، والعلوم الطبية، والإحصاء، وعلوم الحاسوب، والاقتصاد، وغيرها بهدف فصل وتحديد العوامل المختلفة التي تؤثر في الظواهر أو البيانات المدروسة. اما التحليل التمييزي الحصين (Robust Discriminant Analysis) فهو أسلوب إحصائي يستعمل للتصنيف وتقليل الأبعاد عند التعامل مع البيانات التي قد تحتوي على قيم متطرفة أو البيانات التي تنحرف عن التوزيع الطبيعي والذي يعد امتداد للتحليل التمييزي التقليدي (Discriminant Analysis) يهدف إلى تقديم نتائج أكثر حصانة عند مواجهة الانحرافات عن الافتراضات الاساسية كتوزيع البيانات الطبيعي وتجانس مصفوفة التغاير.

وعليه تكمن أهمية التحليل التمييزي الحصين في قدرته على تقديم نتائج أكثر موثوقية ودقة في وجود بيانات تنتهك افتراضات DA التقليدية. من خلال احتساب القيم الشاذة وتخفيف افتراضات التوزيع ، تقدم RDA بديلاً حصيناً ومرنًا لمهام التصنيف وتقليل الأبعاد ، مما يجعله اسلوب جيد في مختلف المجالات حيث قد تكون البيانات عرضة لعدم اليقين والتنوع.

لذلك جاءت هيكلة الرسالة متضمنة خمسة فصول:

الفصل الأول منهجية الرسالة تضمن المقدمة ، مشكلة الرسالة ، هدف الرسالة والإستعراض المرجعي لأهم البحوث وبعض الدراسات السابقة ذات الصلة بموضوع الرسالة.

والفصل الثاني تضمن الجانب النظري الذي تطرق لأهم المفاهيم الاساسية للتحليل التمييزي وانواعه وكذلك تصنيف المشاهدات وخطأ التصنيف وعرض الحزمة وخصائصه وطرائق اختيار عرض الحزمة.

والفصل الثالث شمل الجانب التجريبي، اذ تضمن تجارب محاكاة مونت -كارلو لاختبار افضلية اسالسب التحليل التمييزي المستعملة في هذه الرسالة.

والفصل الرابع الجانب التطبيقي، اذ تم الاعتماد على سجلات وحدة المختبر في مستشفى الحسين التعليمي في محافظة كربلاء المقدسة لغرض الحصول على المتغيرات التي لها علاقة بمرض إبيضاض الدم اللمفاوي (Lymphocytic leukemia) والتي تضمنت (100) مشاهدة من الذكور والإناث.

والفصل الخامس شمل أهم الاستنتاجات والتوصيات التي تمخضت عنها الرسالة وتم التوصل اليها في الجانبين التجريبي والتطبيقي.

2-1 مشكلة الرسالة (Problem of the thesis)

غالبًا ما تنحرف البيانات عن الافتراضات المثالية التي تتطلبها الأساليب الإحصائية التقليدية ففي هذه المواقف يتم فيها انتهاك افتراض الحالة الطبيعية في البيانات (التوزيع الطبيعي للبيانات) ، او ان هنالك مجموعات من البيانات غير خطية تؤدي الى مواجهة مشكلة في التصنيف لايمكن للتحليل التمييزي التقليدي مواجهة هذه المشكلة فلابد من البحث عن طريقة حصينة تتفي عامل مع هذه المشكلة اذ يمكن حل هذه المشكلة باستعمال التحليل التمييزي اللبي الحصين (RKDA) .

3-1 هدف الرسالة (Aim of the thesis)

تهدف الرسالة الى استعمال التحليل التمييزي اللبي الحصين في حالة وجود تلوث في البيانات ومقارنته مع التحليل التمييزي اللبي التقليدي والتحليل التمييزي الخطي والتربيعي باستعمال معيار معدل خطأ التصنيف لاختيار افضل اسلوب في التصنيف.

4-1 الاستعراض المرجعي (Literature Review)

تناولت العديد من الابحاث والدراسات موضوع التحليل التمييزي بصورة في عامة وكذلك التحليل التمييزي اللبي ولكن هنالك ندرة في الدراسات التي تناولت موضوع التحليل التمييزي اللبي الحصين على حد علم الباحث ندرج بعضاً من الدراسات والبحوث والتي تناولت موضوع التحليل التمييزي اللبي الحصين وهي:

- **في عام (2009) اقترح الباحث (Nudurupati)** تحليل تمييزي لا معلمي أقل حساسية من التحليل التمييزي التقليدي للانحرافات عن الافتراضات المعتادة كاعتدالية البيانات باستعمال منهجية متابعة اسقاط المجموعات الداخلة في التصنيف حيث يكون مؤشر الإسقاط هو احتمال ترشيح مجموعتين لتخصيص المشاهدة الجديدة باستعمال مسافات تقليدية بسيطة من المراكز المتوقعة بناءً على مركزية النقطة الجديدة المقاسة باستخدام تحويلين: تحويل متماثل من مجموعتين وتحويل لاستبدال مجموعة النقاط. ومن خلال تجارب المحاكاة تبين أن الطريقة المقترحة توفر معدلات تصنيف خاطئ اقل من الإجراءات التقليدية مثل تحليل التمييز الخطي وتحليل التمييز التربيعي . [19]
- **في عام (2011) اقترح (.You Di & et al) واخرون** طريقة جديدة في تصميم الطرائق اللبية ، وهي العثور على المعلمات اللبية التي تتفي عامل مع المشكلة الخطية بشكل واضح كي تصبح دالة (Kernel) تمثل المصنف البيزي الخطي والتي يمكن ان تطبق بنجاح في العديد من خوارزميات لأسلوب التحليل التمييزي اللبي (KDA) وبينت النتائج فائدة هذا الأسلوب المقترح وان صيغة (Kernel) لتصنيف التحليل التمييزي تعطى اعلى معدلات التعرف على الانماط. [28]

- **في عام (2012) اقترح (.Stefanos & et al) واخرون** اسلوباً قوياً في التمييز اللبي الذي يجمع بين تنظيم ذاتية الطيف (Eigenspectrum) مع مستوى الميزة (ER- KDA) eigenspectrum regularization على اساس استخراج خاصية التعرف و التحقق على الوجه بناءاً على اسلوب تحليل التمييزي اللبي (KDA). فقد تم الجمع بين الطريقة المقترحه (ER-KDA) والبية الحصينة غيرالخطي بشكل مناسب للتعرف على الوجه والتأكيد على الطبقات التي تتطلب الحصانة مقابل القيم المتطرفة. [30]

- **في عام (2014) أقترح الباحثان (Zhang Xiao &Yang Guan)** الشبكة اللاسلكية الموزعة على أساس نظام التعرف على الوجه من خلال استعمال اسلوب التحليل التمييزي اللبي المتعدد (KDA) مع شبكات جهاز الاستشعار اللاسلكي بالاعتماد على معيار التعظيم الحدي Margin Maximization Criterion (MMC) اذ قاما باجراء مخطط متكرر وبشكل منفصل في مجال الحاسبات وفي قواعد بيانات الوجه (CMU , PIE , FERE) من أجل تحسين معلمة تمهيد Kernel لكل وحده وبين الباحثان في الجانب التجريبي (المحاكاة) إن اطار العمل لـ (Kernel) المتعدد هو الأجراء الأمثل لتحقيق الأداء ، واعطى نتائج عالية مقارنة مع دالة (Kernel) المستندة على طريقة (KDDA) . [32]

- **في عام (2017) اقترح (.Li et al) وآخرون** التحليل التمييزي اللبي اللامعلمي المحلي ((local kernel nonparametric discriminant analysis)) (LKNDA) والتي تدمج التحليل التمييزي التقليدي مع الاحصاء اللامعلمي وتم مقارنة الطريقة مع طرائق التحليل التمييزي التقليدي باستعمال ست تجارب محاكاة والتي اثبتت جميعها ان طريقة (LKNDA) لها دقة تصنيف اعلى وتعد حلاً بديلاً للحالات التمييزية لاستخراج الميزات غير الخطية المعقدة أو استخراج الميزات غير المعروفة. وتم تطبيق الطريقة على بيانات سوق الاوراق المالية. [14]

- **في عام (2019) اقترح (Yu et al.) واخرون** انموذج تحليل تمييزي بيزي لامعلمي جديد يقوم باختيار المتغير وتصنيفه ضمن إطار عمل بسيط. يتم تعيين مقدمات شجرة Polya للتوزيعات المجهولة لمجموعة مشروطة لحساب عدم الدقة ، والسماح للمعتقدات السابقة حول التوزيعات ليتم دمجها ببساطة كمعلمات Hyperparameters . تم التوصل الى ان اعتماد الاستدلال البيزي في ظل التحليل التمييزي يؤدي الى تكلفة حسابية اقل. واظهرت الطريقة المقترحة أداءً جيدًا عند مقارنتها بالأساليب التقليدية.[29]
- **في عام (2020) استعملت الباحثة (جاسم)** اسلوبين في تصنيف البيانات و هما اسلوب التحليل التمييزي الخطي (LDA) و اسلوب التحليل التمييزي اللبي (KDA) بهدف ايجاد الدالة التمييزية لكل منها واستعمالها كدالة تصنيف (تمييز) للاسلوبين بين المرضى ، باستعمال بيانات حقيقية لمجموعتين من المرضى المصابين وغير المصابين بمرض اللوكيميا وتوصلت الى ان اسلوب التحليل التمييزي الخطي هو الافضل لكونه يعطى اقل خطأ تصنيف للبيانات . اذ تم اجراء المقارنة بين الاسلوبين وفق معيار احتمال خطأ التصنيف (Misclassification). [2].
- **في عام (2022) اقترح (Obudho et al.)** وآخرون دالة تمييز لبية حصينة لامعلمية تم بواسطتها معادلة مشكلة التصنيف في الحالات التي يتم فيها انتهاك شرط التوزيع الطبيعي للبيانات المستعملة اذ حساب معدلات التصنيف الخاطئ لمختلف مصفوفات النطاق الترددي وقارنوا الطريقة المقترحة مع دوال التصنيف المعلمية مثل التمييز الخطي والتمييز التربيعي باستعمال تجارب المحاكاة. وتم التوصل الى ان الطريقة المقترحة تؤدي أداءً جيدًا من حيث معدلات التصنيف الخاطئ لمصنف تمييز اللب عند تحديد النطاق الترددي الصحيح بالمقارنة مع المصنفات الموجودة الأخرى المستعملة. [20]
- **في نفس العام (2022) استعمل (Gupta et al.) وآخرون** التحليل التمييزي بأبعاد مختلفة مثل الخطي والتربيعي للتصنيف الثنائي لتحليل متلازمة تكيس المبايض ومقارنته باسلوب التعلم الآلي

وطريقة تقليل الأبعاد الخاضعة للإشراف وتم التوصل الى ان باستخدام التحليل التمييزي يحقق دقة اعلى وتباين أقل مع دقة تدريب تصل إلى 97.37٪ ودقة اختبار 95.92٪ باستخدام التحليل التربيعي التمييزي مقارنة بباقي الطرائق . [11] .

استكمالاً لما تقدم من دراسات وبحوث التي تناولت موضوع التحليل التمييزي بصورة في عامة وكذلك التحليل التمييزي اللبي فنلاحظ بانه لاتوجد دراسات عربية تناولت موضوع التحليل التمييزي اللبي الحصين باستعمال دوال كثافة لبية منحرفة عن التوزيع الطبيعي فالشرط الاساسي للتحليل التمييزي اللبي هو ان تكون دالة الكثافة اللبية المستعملة لها توزيع كاوسي ولكن في الكثير من الحالات تنحرف البيانت عن هذا الافتراض فلابد من استعمال اسلوب تحليل تمييزي يتفي عامل مع هذه الحالات فلذلك تم استعمال دوال كثافة لبية كاوسية ودوال كثافة لبية تنحرف عن التوزيع الكاوسي والمقارنة بين الاساليب التمييزية .

الفصل الثاني

الجانب النظري

التمهيد (Preface)

تم في هذا الفصل التطرق الى بعض المفاهيم الأساسية المتعلقة بالتحليل التمييزي وعلاقته بالدوال اللبية (Kernel Functions) وتصنيف المشاهدات وقياس الدقة باستعمال معدل خطأ التصنيف ($\widehat{MR}$) مع الطرائق اللبية (kernel) التي تسمى بطرائق عرض الحزمة (Bandwidth methods) .

1-2 التحليل التمييزي (Discriminant Analysis)

يعد اسلوب التحليل التمييزي من ألاساليب المهمة في تحليل البيانات متعددة المتغيرات ، اذ أنه يعتمد على نوع المشكلة ونوع البيانات سواء كانت هذه البيانات (كمية او نوعية) وله عدة تطبيقات عملية مهمة ، اذ يستعمل في مختلف المجالات من أهمها الزراعية ، الطبية كتصنيف الأمراض ومعرفة شدة الإصابة بها.

يعتمد اسلوب التحليل الاحصائي لمتعدد المتغيرات على الظواهر التي لها ابعاد ومتغيرات متعددة لوصفها وتحليلها. ان التحليل التمييزي يهتم بمسألة التمييز بين مجموعتين او اكثر والتي تشترك فيما بينها بمجموعة من الصفات والخصائص بدرجة مختلفة وذلك باستعمال دالة خاصة تسمى الدالة التمييزية (Discriminant Function). تتبعها عملية التصنيف وهي عملية تلي عملية تكوين الدالة التمييزية، إذ يتم الاعتماد على هذه الدالة في التنبوء وتصنيف المفردة الجديدة لإحدى المجموعات قيد الدراسة بأقل خطأ تصنيف ممكن.[31]

يعتمد انموذج التحليل التمييزي على الوصول إلى دالة التمايز التي تعمل على تعظيم الفروق بين متوسط المجموعات وتقليل التشابه في أخطاء التصنيف في الوقت ذاته، وذلك من خلال إيجاد مجموعات خطية من المتغيرات والتي غالباً ما يطلق على المتغيرات الكمية في التحليل التمييزي متغيرات مستقلة أو منبئة ،ويشار أيضاً لمتغير انتماء المجموعة بالمتغير التابع أو المتغير الحمكي

التصنيفي. ولا تحتاج بيانات التحليل التمييزي لأن تكون معيارية، أي أن يكون لها وسط صفر وتباين يساوي الواحد، وذلك لأن نتيجة تحليل التمييزي لا تتأثر بكثرة بتغير مفردات المتغيرات [10].

ومن الدوال التمييزية التي يمكن استعمالها :

1) دالة التمييز الخطي (Linear Discriminant Function)

تسمى بدالة فيشر (Fisher) نسبة الى الباحث الذي قام باشتقاقها ، وتستعمل عندما تكون العلاقة بين المتغيرات خطية .[12]

لنفترض ان البيانات الأصلية في X مقسمة الى C من الأصناف كأن تكون $X = [X_1, X_1, \dots, X_C]$ اذ ان $X \in R^{n+n_i}$ تحتوي على نقاط البيانات للصنف i بحيث أن $N = \sum_{i=1}^{c} n_i$ فانه يمكن تعريف معيار (Fisher) كالآتي:[12]

$$\max_{\lambda} J(\lambda) = \frac{\lambda' \delta_b \lambda}{\lambda' \delta_t \lambda} \qquad \dots (2\text{-}1)$$

λ : متجه مميز

وان :

$$\delta_b = \frac{1}{N} \sum_{i=1}^{c} n_i \, (\delta_i - \delta_0)(\delta_i - \delta_0)' \qquad \dots (2\text{-}2)$$

δ_b مصفوفة الانتشار بين الاصناف المعرفة في فضاء المتغيرات

$$\delta_t = \frac{1}{N} \sum_{i=1}^{N} (\delta(x_i) - \delta_0)(\delta(x_i) - \delta_0)' \qquad \dots (2\text{-}3)$$

δ_t مصفوفة الانتشار الكلية المعرفة في فضاء المتغيرات

وان :

δ_i يمثل متجه المتوسطات لعينات التدريب في الصنف i

δ_0 يمثل متجه المتوسطات لكل عينات التدريب (لكل الاصناف)

$\delta(x_i)$ الصنف i في فضاء المتغيرات

2) الدالة التربيعية او الدالة اللاخطية (Non – Linear Discriminant Function)

تستعمل عندما تكون العلاقة بين المتغيرات غير خطية (تربيعية ، او ذات درجة أعلى). ان اهم خطوة في التحليل التمييزي هي حساب التمييز اذ يهدف الى تكوين صيغة خطية او غير خطية بين المتغيرات لتصنيف المفردات الى المجموعات تنتمي اليها .

ان دالة التمييز الخطي تستند الى تركيب خطي للمتغيرات ، وفي حالة كون البيانات لا تتوفر فيها شروط التمييز الخطي اي عدم تساوي مصفوفة التباين والتباين المشترك للمجموعات سوف تستعمل دالة التمييز التربيعي .

فاذا كان هنالك v من المجموعات (Groups) المقابلة لدوال الكثافة $f_1, f_2, \dots, f_v$ وان الهدف هو تعيين جميع نقاط x من فضاء العينة لواحدة من تلك المجموعات او دوال الكثافة ، سوف نقارن القمم الموزونة لدوال الكثافة للحصول على قاعدة التمييز البيزية (Bayes Discriminant rule) الآتية: [12]

$$x \text{ is allocated to group } j_0 \text{ if } j_0 = \max_{j \in 1,\dots,v} \pi_j f_j(x) \quad ; \qquad \dots(2\text{-}4)$$

اذ ان:

$j = 1,2 \dots, v$: يمثل عدد المجموعات المدروسة

π_j : تمثل الاحتمال المسبق (Prior Probabilities) لدالة الكثافة $f_j(x)$

$f_j(x)$: تمثل دالة الكثافة الاحتمالية

وبحصر جميع قيم x في فضاء العينة ، فاذا كان الجزء $P = \{P_1, P_1, \ldots\ P_V\}$ من العينة فان x تنتمي للـ P_j اذا كانت x مخصصة للمجموعة j اي أن: [7]

$$x \ \in P_j \ if \ x \ \ is \ allocated \ to \ group \ j \qquad \ldots(2\text{-}5)$$

فان قاعدة التمييز المعرفة في المعادلة (2-1) تتضمن دوال الكثافة المجهولة والاحتمالات المسبقة (الممكنة) .

ولنفرض انه تم جمع بيانات ، فانه يمكن تعديل هذه القاعدة الملخصة الى قاعدة تمييز عملية باستعمال قاعدة التمييز البيزية والتقسيم . [20]

2-2 التحليل التمييزي المعلمي (Parametric Discriminant Analysis)

هنالك نوعان اساسيان من التحليل التمييزي المعلمي هما التحليل التمييزي الخطي والتربيعي ، وهما الأكثر استخدامًا. إن سهولة حسابهم ناتجة عن افتراض الحالة الطبيعية للمجموعات، والذي لا ينطبق بالضرورة على معظم مجموعات البيانات.

3-2 التحليل التمييزي الخطي (Linear Discriminant Analysis)

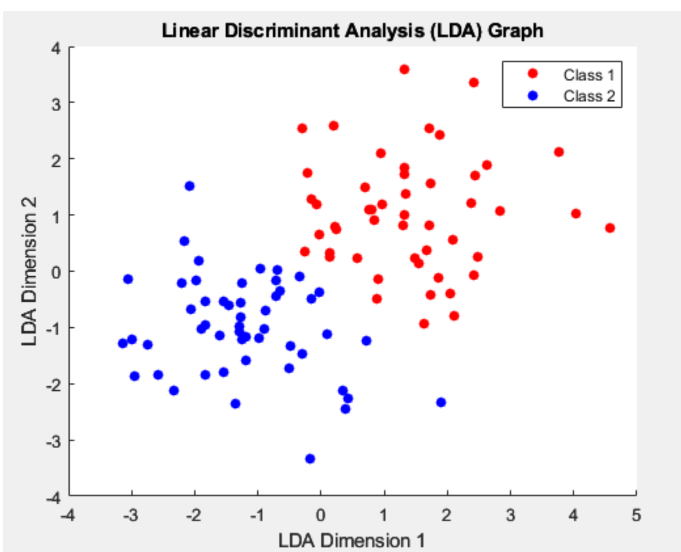
وهو خوارزمية تعلم خاضعة للإشراف تستعمل لتقليل الأبعاد ومهام التصنيف. يتم استخدامه بشكل أساسي في التعرف على الأنماط والتعلم الآلي للعثور على مجموعة خطية من الميزات التي تفصل أو تميز بشكل أفضل بين فئتين أو مجموعات أو أكثر. يفترض أن متغيرات التنبؤ (الميزات) تتبع توزيعًا طبيعيًا متعدد المتغيرات ولديها مصفوفة تغاير مشتركة لكل مجموعة ويحسب التركيبة الخطية لمتغيرات التوقع التي تفصل بين المجموعات بشكل أفضل بحيث التحليل التمييزي الخطي إلى تعظيم التباين بين المجموعة مع تقليل التباين داخل المجموعة. [6][7]

لنفرض انه لدينا f_i دالة توزيع طبيعي بقيم مختلفة من متجه المتوسطات μ_i ومصفوفة تباين-تباين مشترك Σ أي أن $f_i \sim N(\mu_i, \Sigma)$. وعليه يمكن كتابة دالة التمييز الخطية كالاتي: [24]

$$x \text{ is allocated to group } j_0 \text{ if } j_0 = \underset{j \in \{1,\dots,v\}}{\operatorname{argmax}} \; \log(\pi_j) - \frac{1}{2}(x - \mu_j)^T \Sigma^{-1} (x - \mu_j) \quad (2\text{-}6)$$

وإن Argmax هي دالة رياضية تستعمل في التحسين أو نظرية القرار. إنه يرمز إلى "وسيطة الحد الأقصى" ويشير إلى قيمة الإدخال التي تزيد من دالة معينة. بمعنى آخر، عندما يكون لديك دالة تأخذ بعض قيم الإدخال وترجع المخرجات، فإن الحد الأقصى لهذه الدالة هو قيمة الإدخال التي تنتج الحد الأقصى من المخرجات.

من المعادلة (2-3) ، يمكن ملاحظة أن التقسيم الناتج يتم الحصول عليه عن طريق تقاطعات الإهليجات (ellipsoids) ذات المراكز المختلفة وبنفس الاتجاه هذا يعطي حدود التقسيم التي هي مستويات.ويمكن أن يُنظر إلى الإهليجات على أنها أشكال بيضاوية الشكل يتم تحديدها بواسطة مركزها وطول نصف القطر في الاتجاهين الرئيسيين. في تحليل التمييز الخطي ، يتم استخدام الإهليجات ذات نصف قطر مختلف لتمثيل فئات مختلفة. يتم تقسيم البيانات إلى فئات مختلفة عن طريق تحديد نقاط البيانات التي تقع داخل أو خارج الإهليجات [2] [24].



شكل رقم (2-1) انموذج تصنيف التحليل التمييزي الخطي لصنفين من المشاهدات

يبين الشكل (2-1) التحليل التمييزي الخطي لمجموعتين والذي تم الحصول عليه من استعمال $\overline{x}_j$ كتقدير للـ μ_j و s كتقدير للـ Σ وان s_j وهو تباين العينة.

4-2 التحليل التمييزي التربيعي (Quadratic Discriminant Analysis)

هو خوارزمية تعلم خاضعة للإشراف تستعمل لتصنيف وهو مشابه لـ LDA ولكنه يخفف من افتراض مصفوفة التغاير المشتركة لكل مجموعة وفي حالة كون البيانات المستعملة لاتتوزع طبيعياً او غير خطية بحيث يسمح لكل مجموعة أن يكون لها مصفوفة التغاير الخاصة بها ، مما يجعلها أكثر مرونة ولكنها تتطلب المزيد من البيانات لتقدير المعلمات بدقة فأن دالة الكثافة يمكن تقديرها بشكل مباشر من البيانات وفق اسلوب يعرف بالمقدر اللبي (Kernel) . اذ ان التقدير اللبي يعتمد على كثافة المجتمع للحصول على قاعدة اكثر شيوعاً للتخصيص وهي قاعدة التمييز البيزية (BDR) . [26]

يتم الحصول على قاعدة التمييز اللبية (KDR) من قاعدة التمييز البيزية من خلال استبدال دالة الكثافة الاحتمالية بتقدير دالة الكثافة اللبية $\hat{f}_j$ (x ; H_j) ، وكما يأتي :[6]

$$G_i = \pi_i f_i (\underline{x}) \qquad \ldots(2\text{-}7)$$

$$KDR : d_j(x) = \underset{j \in \{1,\ldots,v\}}{\operatorname{argmax}}\ \hat{\pi}_j \hat{f}_j (x ; H_J) \qquad \ldots(2\text{-}8)$$

اذ ان

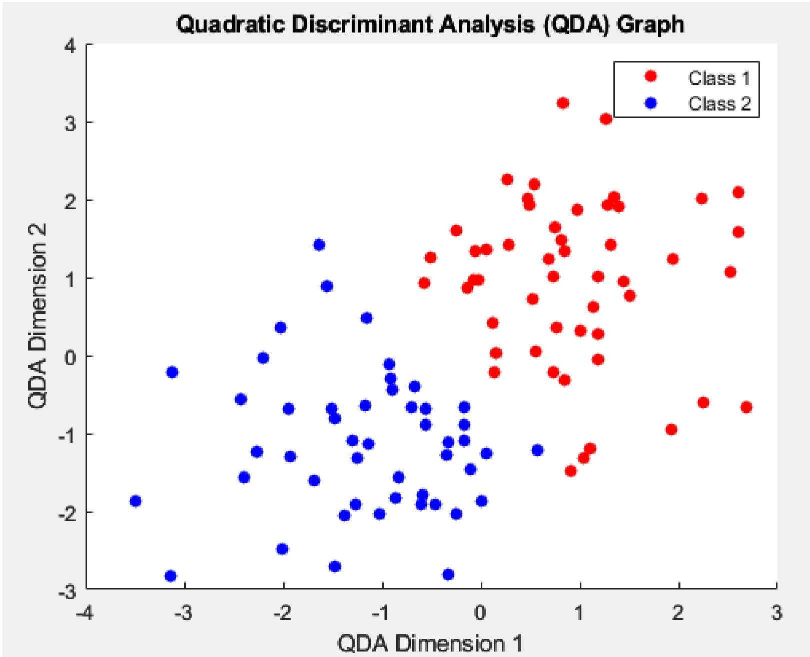
$j = 1,2 \ldots, v$: يمثل عدد المجموعات المدروسة

$\hat{\pi}_j$: تمثل الاحتمالات السابقة المقدرة

$\hat{f}_j (x ; H_j)$: تمثل مقدر دالة الكثافة الاحتمالية .

وتكون دالة التحليل التمييزي التربيعي كالآتي :

$$x \text{ is allocated to group } j_0 \text{ if } j_0 = \underset{j \in \{1,\ldots,v\}}{\operatorname{argmax}}\ \log(\pi_j) - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(x - \mu_j)^T \Sigma_j^{-1}(x - \mu_j) \qquad \ldots(2\text{-}9)$$



شكل رقم (2-2) انموذج تصنيف التحليل التمييزي التربيعي

ان استبدال المعلمات المجهولة بتقديرات العينة هو خطوة ضرورية لاستخدام فعال لقواعد التمييز المعلمية. ومع ذلك ، يجب أن نضع في الاعتبار أن تقديرات العينة غير دقيقة ، ويجب اختيار عينة تمثيلية للبيانات لضمان نتائج دقيقة .[2]

5-2 المقدر اللبي (Kernel Estimator)

المعروف أيضًا باسم تقدير كثافة اللب (Kernel density estimator KDE) ، هو طريقة لامعلمية تستعمل لتقدير دالة كثافة الاحتمال (PDF) لمتغير عشوائي مستمر من مجموعة معينة من نقاط البيانات على عكس الطرائق المعلمية ، مثل ملائمة توزيع محدد للبيانات ، لا يفترض تقدير كثافة اللب أي شكل دالي محدد للتوزيع الأساسي اذ هو مقدر يعتمد على دالة وزن (Weight function) تستعمل في تقدير الدالة اللامعلمية وهو يعطي اوزان لنقاط البيانات المجاورة في اجراء التقدير. والدوال اللبية (Kernel functions) تكون مستمرة ومحدودة ومتماثلة حول نقطة الصفر وقيمتها حقيقية وتكاملها مساو للواحد، وهي دالة الكثافة الاحتمالية المتماثلة ، و يرمزلها برمز K(x) .[14]

يستخدم تقدير اللب على نطاق واسع في تحليل البيانات ، وتصور البيانات ، وفي العديد من التطبيقات الإحصائية ، مثل تقدير الكثافة ، واكتشاف القيم الشاذة في البيانات ، والذي يكون مفيدًا بشكل خاص عندما يكون توزيع البيانات الأساسي غير معروف أو عندما لا يتم نمذجة البيانات بسهولة عن طريق توزيع حدودي بسيط. ومع ذلك ، مثل أي طريقة إحصائية ، من الضروري النظر في الاختيار المناسب للنواة وعرض النطاق الترددي لتجنب التمهيد الناقص أو الإفراط في تجانس التقدير. يمكن أن تساعد تقنيات التحقق المتبادل في تحديد النطاق الترددي الأمثل لمجموعة بيانات معينة.[14]

ان دالة اللب تحقق الشروط الاتية :[1][2]

1) $K(x) = K(-x)$ …(2-10)

2) $\int_{R} K(x)\, dx = 1$ …(2-11)

3) $\int_{-\infty}^{\infty} x\, K(x)\, dx = 0$ …(2-12)

4) $\int_{-\infty}^{\infty} x^2 K(x)\, dx = K > 0$ for some constant k. …(2-13)

$K(x) \geq 0$ for all $x \geq 0$ …(2-14)

الفكرة الأساسية وراء تقدير اللب هي تمثيل كل نقطة بيانات على أنها دالة "نواة" صغيرة تتمحور حول تلك النقطة. يتم بعد ذلك تلخيص دالة اللب الفردية لإنتاج تقدير مستمر للـ PDF. ويلعب اختيار دالة kernel ، والمعروف أيضًا باسم داالة التمهيد ، وعرض النطاق الترددي الخاص بها (أو عرض الحزمة) أدوارًا مهمة في جودة التقدير.ان صيغة دالة الكثافة اللبية يمكن ان تكتب بالصورة الآتية:[2]

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K(.) \left(\frac{x - x_i}{h}\right) \quad \dots (2\text{-}15)$$

اذ ان :

$\hat{f}(x)$ دالة الكثافة اللبية المقدرة عند النقطة x

n عدد نقاط البيانات

x_i نقطة بيانات فردية

$K(.)$ الدالة اللبية

تضمن دوال اللب الشائعة الاستخدام دالة اللب الكاوسية (*Gaussian*) ، ودالة *Epanechnikov* ، واللبية المستطيلة (*Regtanular*) ، وغيرها من الدوال. [1]

h عرض الحزمة (معلمة التمهيد Smoothing parameter)

تتحكم معلمة التمهيد h ، في عرض النواة ، وبالتالي مستوى التمهيد. سينتج عرض الحزمة الأكبر تقديرًا أكثر سلاسة ولكن يحتمل أن يكون أقل حساسية ، في حين أن عرض الحزمة الأصغر سوف يلتقط تفاصيل أدق في البيانات ولكن قد يتأثر أكثر بالضوضاء.[1]

تتضمن عملية تقدير *Kernel* وضع نواة في كل نقطة بيانات ، وقياس النواة من خلال عرض الحزمة h وتلخيص المساهمات من جميع النواة للحصول على ملف *PDF* المقدر.[6]

6-2 اختيار الدوال اللبية (Selection of kernel function)

تعد تقديرات الكثافة اللبية (Kernel Density) من دوال الانحدار الأكثر شيوعاً والتي تعد من تقديرات الكثافة اللامعلمية. والدالة اللبية (Kernel Function) هي دالة الاوزان الموحدة ومهمة جداً في انتشار الكثافة الاحتمالية ، فمن المعروف ان اسلوب هذه التقديرات تعتمد اساساً على اختيار معلمة عرض الحـــــزمة (Bandwidth) والتي تسيطر على تمهيد التقدير وعلى اختيار دالة Kernel.

ان الدوال اللبية الاكثر استخداماً هي دالة Gaussian بمتوسط (0) وتباين (1) . [13]

ان سبب اختيار دالة (Gaussian) يعود الى وجود العديد من الخصائص المرغوب بها لان الكثير من البيانات لها توزيع طبيعي مما يجعل سهولة ملائمتها للتوزيع الطبيعي، مما يجعل اختيار عرض الحزمة اسهل. [13]

7-2 عرض الحزمة (Bandwidth)

تشير معلمة التمهيد او عرض الحزمة او عرض النطاق الترددي او سعة القيد او حجم النافذة او معلمة الانتشار إلى المعلمة الحاسمة التي تتحكم في عرض دوال kernel الموضوعة في كل نقطة بيانات أثناء عملية التقدير. وان اختيار عرض الحزمة له تأثير كبير على جودة وخصائص مخرجات KDA. وينتج عن عرض الحزمة الأصغر نواة أضيق تكون أكثر موقعية حول نقاط البيانات الفردية. يمكن أن يؤدي ذلك إلى مستوى عالٍ من التفاصيل في ملف PDF التقديرية ، مما يؤدي إلى التقاط الاختلافات الصغيرة في البيانات. ومع ذلك ، فإنه قد يجعل التقدير أكثر حساسية للضوضاء ، مما يؤدي إلى تمثيل غير دقيق.[14] ويرمز لها بالرمز (H) اذا كانت تستعمل لمتعدد المتغيرات ، ويرمزلها (h) اذا كانت تستعمل لأحادي المتغير ، وسبب هذه التسمية جاءت من خلال الاتي : [4]

1) كونها اهم في عامل يؤثر على اداء KDA (Kernel Discriminant Analysis).

2) تؤدي الى توجيه الدالة اللبية (Kernel Function) .

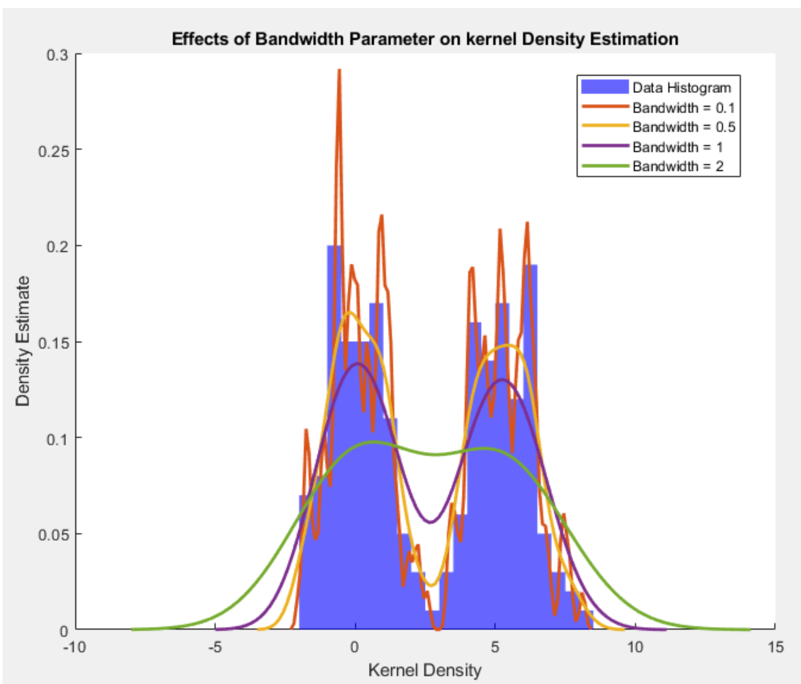
3) تتحكم بانتشار الدالة اللبية (Kernel Function) .

8-2 اختيار معلمة عرض الحزمة (Selection of Bandwidth Parameter)

يعد تقديردالة الكثافة اللبية من التقنيات المهمة جداً في تمهيد البيانات ، اذ كان تطبيقها ناجحاً لمعظم بيانات احادي المتغير في حين كان تنفيذها وتطويرها محدوداً نسبياً بالنسبة للبيانات متعددة المتغيرات . ان اختيار عرض الحزمة في حالة احادي المتغير يتضمن اختيار معلمة مفردة تسيطر على مقدار من التمهيد ، اما في حالة متعدد المتغيرات فان مصفوفة عرض الحزمة تسيطر على كل من

درجة واتجاه التمهيد ولذلك يكون اختيارها اكثر صعوبة . اذاً من المهم جداً اختيار عرض الحزمة الأمثل (Optimal Smoothing) وليس كقيمة صغيرة جداً او كبيرة ، فأن القيم الصغيرة من مصفوفة عرض الحزمة (H) يؤدي الى تقديرات شائكة جداً قليلة التمهيد (under smoothing) في حين القيم الكبيرة تؤدي الى تمهيد عالي جـــداً (over smoothing). [4]

عرض الحزمة (Bandwidth) الخاص باللبية هو معلمة حرة لها تأثير قوي على التقدير الناتج. يتطلب تجانس اللبية اختيار معلمة عرض الحزمة. هذا الاختيار مهم للغاية ، حيث يمكن أن يؤدي التمهيد أو الإفراط في التجانس إلى تقليل الدقة بشكل كبير. والشكل (2-1) يبين تأثيرات تغيير معلمة عرض الحزمة على تقديرات الكثافة اللبية .[4].



شكل رقم (2-3) تأثيرات معلمة عرض الحزمة على تقديرات الكثافة اللبية

ان المعيار الأمثل والأكثر شيوعاً لاختيار عرض الحزمة هو دالة المخاطرة او مايدعى بمتوسط مربعات الخطأ التكاملي MISE والذي يعتبر توقع لدالة المخاطرة كما يأتي: [20]

$$MISE(b) = E[\int (f_b(x) - f(x))^2\, dx] \quad \ldots (2\text{-}16)$$

وفي ظل الافتراضات الضعيفة حول f, K فان صيغته تكون كالآتي:

$$MISE(b) = AMISE(b) + o\left(\frac{1}{nb} + b^4\right) \quad \ldots (2\text{-}17)$$

اذ ان o هو الحرف الصغير للـ O ، فان $AMISE(b)$ كالآتي:

$$AMISE(b) = \frac{R(K)}{nb} + \frac{1}{4} m_2(K)^2 b^4 R(f'') \quad \ldots (2\text{-}18)$$

اذ ان:

$$R(K) = \int g(x)^2 \quad \ldots (2\text{-}19)$$

g دالة في

$$m_2(K) = \int x^2 K(x) dx \quad \ldots (2\text{-}20)$$

f'' المشتقة الثانية للدالة f

فان اقل AMISE هو حل للمعادلة الآتية:

$$\frac{\partial}{\partial b} AMISE = -\frac{R(K)}{nb^2} + m_2(K)^2 b^3 R(f'') = 0$$

$$b_{AMISE} = \frac{R(K)^{\frac{1}{5}}}{m_2(K)^{\frac{2}{5}} R(f'')^{\frac{1}{5}} n^{\frac{1}{5}}} \quad \ldots (2\text{-}21)$$

لا يمكن استخدام صيغتي MISE(b) و AMISE (b) مباشرةً لأنها تتضمن دالة الكثافة المجهولة f أو مشتقتها الثانية f'' ، لذلك تم تطوير مجموعة متنوعة من الأساليب التلقائية القائمة على البيانات لتحديد عرض الحزمة . ومن الطرائق الاكثر شيوعاً التي استعملت لاختيار عرض الحزمة هي طريقة التحقق المتقاطع الممهد (Smoothed Cross validation method) [20]

يختلف عرض الحزمة باختلاف دالة اللب المختارة اذ لا يمكن النظر إلى عرض الحزمة الأمثل لدالة اللب بنفس الطريقة لدالة أخرى. لهذا السبب ، أجرى العديد من الباحثين دراسات تهدف إلى تحديد تقنيات الحصول على نطاقات التردد التي تقلل من وظائف MSE أو AMSE التي يمكن استخدامها مع دوال اللب المختلفة.[20]

في في عام (1993) اعتبر الباحثان (Wand & Jonse) ان معلمات مصفوفة عرض الحزمة لثنائي المتغيرات تشمل مصفوفة معرفة موجـــــــــبة قطرية والتي تم استعمالها في هذه الرسالة وبالشكل الآتي : [2]

$$\mathbf{H} = \begin{bmatrix} h_1^2 & 0 \\ 0 & h_2^2 \end{bmatrix} \qquad \dots (2\text{-}22)$$

9-2 تصنيف المشاهدات (Classification data)

يشير مصطلح تصنيف المشاهدات إلى نوع من البيانات التي تستعمل في مهام التعلم الخاضع للإشراف في تعلم الآلة. في التعلم الخاضع للإشراف ، يتم تزويد الخوارزمية بمجموعة بيانات مصنفة ، حيث تحتوي كل نقطة بيانات على كل من الميزات المدخلة (المتغيرات المستقلة) والعناوين المستهدفة المقابلة (المتغيرات التابعة). الهدف من مهمة التصنيف هو تعلم تعيين من الميزات المدخلة إلى العلامات المستهدفة ، مما يسمح للخوارزمية بتوقع عنوان لبيانات جديدة غير ظاهرة. في تصنيف المشاهدات ، تكون العلامات المستهدفة منفصلة وتمثل فئات أو فئات مختلفة. الهدف هو بناء نموذج يمكنه تصنيف مثيلات جديدة بدقة إلى واحدة من هذه الفئات المحددة مسبقًا[15].

أهم خطوات معالجة بيانات التصنيف:

1. تنظيف البيانات: يجب تنظيف البيانات أولاً لإزالة أي أخطاء أو قيم غير صالحة.
2. تمثيل البيانات: يجب تمثيل البيانات في شكل يمكن أن تفهمه الخوارزمية.
3. اختيار الخوارزمية: يجب اختيار الخوارزمية المناسبة لنوع البيانات والمشكلة التي يتم حلها.
4. تدريب الخوارزمية: يجب تدريب الخوارزمية على البيانات المصنفة.
5. تقييم النموذج: يجب تقييم النموذج الجديد باستخدام مجموعة بيانات اختبار لتحديد أدائها.

في التحليل التمييزي اللبي (KDA) ، الهدف هو إجراء التحليل التمييزي في مساحة ميزات متحولة باستخدام دالة لبية دون حساب متجه الميزات المحولة اذ يكون KDA مفيدًا بشكل خاص عندما لا تكون البيانات قابلة للفصل خطيًا في مساحة الميزات الأصلية . ان هدف التحليل التمييزي

الخطي القياسي (LDA) إلى إيجاد تركيبة خطية من الميزات تفصل أفضل بين الفئات بدلاً من إسقاط البيانات على مساحة أقل أبعادًا باستخدام تركيبات خطية للميزات ، فإنه يستخدم دالة لبية لتحويل البيانات ضمنيًا إلى مساحة أعلى الأبعاد حيث يمكن أن تكون قابلة للفصل خطيًا. لذلك يتم حساب عدد التصنيفات الصحيحة ومعدل أخطاء التصنيف (Misclassifications rate) والنتائج يمكن عرضها في جدول التصنيف او مصفوفة التداخل (Confusion Matrix). والجدول (2-1) يوضح عملية تصنيف مشاهدات لمجموعتين (G_1) و (G_2) :[15]

جدول رقم (2-1) نتائج التصنيف (Classification) لمجموعتين

المجموعة الفعلية Actual Group	عدد المشاهدات Number of observation	المجموعة المتوقعة Predicted Group	
		1	2
1	n_1	A	B
2	n_2	C	D

اذ ان (n_1) يمثل عدد المشاهدات في المجموعة الأولى (G_1) و (A) يمثل التصنيف الصحيح في المجموعة الأولى (G_1) و (B) يمثل خطأ التصنيف في المجموعة الاولى (G_1) ، وان ($n_1 = A + B$)

والشيء نفسه بالنسبة الى (n_2) تمثل عدد المشاهدات في المجموعة الثانية (G_2) و (C) يمثل خطأ التصنيف في المجموعة الثانية (G_2) و (D) يمثل التصنيف الصحيح في المجموعة الثانية (G_2) وان ($n_2 = C + D$) . فأن معدل خطأ التصنيف (Misclassification Rate) يكتب كالآتي: [2]

$$\widehat{MR} = \frac{n_{12}+n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}} \quad \ldots (2\text{-}23)$$

وان معدل التصنيف الصحيح الظاهري (Apparent Correct Classification) يكتب كالآتي :

$$\widehat{ACC} = \frac{n_{11}+n_{22}}{n_1+n_2} \qquad \dots (2\text{-}24)$$

ويمكن حساب معدل خطأ التصنيف الظاهري بالشكل الاتي :

$$MR = 1- \widehat{ACC} \qquad \dots(2\text{-}25)$$

وبإستعمال الاحتمالات السابقة ($\widehat{\pi_i}$) و دوال تقدير الكثافة اللبية (KDE) لمجتمعين ($\hat{f}$ (. ; H_2) و $\hat{f}$ (. ; H_1) فأن قاعدة التصنيف الأمثل (Optimal Classification Rate) هي ان تقلل من معدل خطأ التصنيف وبالشكل الاتي : [6]

تعيين المشاهدة (i) الى المجموعة الأولى اذا كان:

$$\widehat{\pi_1}\ \hat{f}(. ; H_1) > \widehat{\pi_2}\ \hat{f}(. ; H_2) \qquad \dots(2\text{-}26)$$

والحالات الأخرى تعين المشاهدة (i) الى المجموعة الثانية (G_2) .

هناك نقطتين مهمتين يجب ان تاخذ بنظر الاعتبار في حالة التحليل التمييزي باستعمال تقديرات الكثافة اللبية (KDE) :

1) اختيار عرض الحزمة (H) يجب ان تعتمد على مشاهدات معينة (محددة) لتصنيفها بالاضافة الى اعتمادها على كثافة المجتمع .

2) في مشكلة تمييز تعدد الطبقات عوضاً عن استعمال عرض حزمة واحدة (ثابتة) لجميع تقديرات الكثافة للمجتمع يكون من المفيد إستعمال عرض الحزمة متغيرة لتقدير الكثافة المصنفة. [2]

10-2 معدل خطأ التصنيف (Misclassification rate)

هذا المعدل هو النسبة المئوية للنقاط التي يتم تعيينها إلى المجموعة غير الصحيحة بناءً على قاعدة تمييز. ويُعرف باسم معدل الخطأ في التصنيف ، وهو مقياس تقييم شائع يستخدم لتقييم أداء نموذج التصنيف. يمثل نسبة الحالات المصنفة خطأً إلى إجمالي عدد الحالات في مجموعة البيانات.

أشار الباحثان (Hall and Wand) في عام (1988) ان ايجاد افضل عرض الحزمة بشكل مباشر تكون عن طريق معدل خطأ التصنيف كالاتي :[15][16]

$$1 - MR = Pr(\,Y\ is\ classifed\ correcctly)$$

$$= E_Y[1\{Y\ is\ classified\ correctly\}]\ \}$$

$$= E_X\big[E_Y[1\ \{Y\ is\ classified\ correctly\}]\big]\,\big|\,X_1, X_1, \ldots X_v\}$$

$$= 1 - \frac{TP+TN}{TP+FP+TN+FN} \qquad \ldots (2\text{-}27)$$

اذ ان :

E_Y التوقع بالنسبة لـ Y او $\sum_{j=1}^{v} \pi_j f_j$

E_X التوقع بالنسبة لـ $X_1, X_1, \ldots X_v$ او $\pi_1 f_1, \pi_2 f_2, \ldots, \pi_v f_v$

TP (True Positive) من المتوقع أن تكون الملاحظة إيجابية وهي في الواقع إيجابية.

TN (True Negative) من المتوقع أن تكون الملاحظة سلبية وهي في الواقع سلبية.

FP (False Positive) من المتوقع أن تكون الملاحظة إيجابية وهي سلبية في الواقع.

FN (False Positive) من المتوقع أن تكون الملاحظة سلبية وهي في الواقع إيجابية.

ان استعمال نهج اعلاه يعتمد على تقديرات كثافة التوزيعات الفردية والذي له ثلاثة مزايا:

- ان التقديرات الدقيقة لدوال الكثافة الفردية مفيدة في حد ذاتها. يمكن استخدامها لإنشاء مخططات كثافة الاحتمالية التي توفر معلومات مفيدة حول توزيع البيانات.
- يمكن استخدام تقديرات الكثافة الدقيقة في مشكلات التمييز الأخرى الأكثر تعقيدًا والتي تبحث في مقاييس أخرى غير معدل الخطأ.
- إن التحسين المباشر فيما يتعلق بمعدل الخطأ يطرح العديد من العقبات الرياضية الصعبة. من الصعب إيجاد تقديرات دقيقة لدوال الكثافة الفردية التي تصغر معدل الخطأ[15].

وان هذا النهج على الرغم من تمتعه بالعديد من المزايا ، إلا أنه لا يخلو من العيوب. أحد العيوب هو أنه يتطلب تقديرات دقيقة لدوال الكثافة الفردية. قد يكون هذا صعبًا إذا كانت البيانات غير منتظمة أو لاتتوزيع طبيعيا.اضافة الى أن هذا النهج قد يكون بطيئًا للغاية بالنسبة للمجموعات الكبيرة من البيانات[15][16].

2- 11 تقدير الكثافة اللبية (Kernel Density Estimation)

وهي طريقة شائعة لتقدير دالة الكثافة اللامعلمية ولها تطبيق معروف في التحليل التمييزي اللبي في مشكلة التصنيف لـ (J) اصناف.

اذا كانت لدينا عينة تدريب كالآتي:

$$S = \{(x_i, c_i\ ), x_i \in R^d\ , c_i \in (1,2, \dots J), i\ = 1,2, \dots , n\} \qquad \dots (2\text{-}28)$$

لـ n من المشاهدات

فإن تقدير الكثافة اللبية يكون عن طريق الصيغة الآتية :

$$\hat{f}_{jb} = \frac{1}{n_j b^d} \sum_{i:c_i=j}^{n} K\left\{\frac{(x-x_i)}{b}\right\} \qquad \dots (2\text{-}29)$$

اذ ان:

$\hat{f}_{jb}$ تقدير الكثافة اللبية

n_j عدد المشاهدات في الصنف j^{th} بحيث ان $\sum n_j = n$

K دلة الكثافة بالبعد d متماثلة حول الصفر

b معلمة التمهيد

ان تقدير الكثافة اللبي يمكن ان يستعمل لانشاء قاعدة التمييز اللبية (Kernel Discremenan Rule) الآتية:

$$\text{KDR: is allocated to group } j_0 \text{ if } j_0 = argmax\ \hat{\pi}_{j\in\{1,2,\dots.v\}}\ \hat{f}_j(x_i, B_i) \qquad \dots (2\text{-}30)$$

$\hat{f}_j(x_i, B_i)$ تقدير الكثافة اللبية المقابلة للمتجه j^{th}

$\hat{\pi}$ الاحتمال السابق للمتجه j^{th}

اذا كان التوزيع الأولي غير معروف فإنه عادة مايتم تقديره باستعمال عينه التدريب عن طريق الصيغة الآتية:

$$\hat{\pi}_{j\in\{1,2,\dots.v\}} = \frac{n_j}{n} \quad ; \quad j = 1,2, \dots\dots J \qquad \dots (2\text{-}31)$$

ان اختيار عرض الحزمة المناسب امر بالغ الأهمية من ناحية يمكن محاولة العثور على عرض حزمة امثل لتقديرات الكثافة اللبية المثلى، ومن ناحية أخرى يمكن العثور على عرض حزمة امثل يعمل بشكل مباشر على تحسين معدل التصنيف الخاطئ.

12-2 التحليل التمييزي اللامعلمي (Non-Parametric Discriminant Analysis)

التحليل التمييزي اللامعلمي (NDA) هو نهج للتصنيف لا يقدم افتراضات واضحة حول التوزيع الأساسي للبيانات كما في طرائق التحليل التمييزي التقليدية، مثل التحليل التمييزي الخطي (LDA) أو التحليل التمييزي التربيعي (QDA)، غالبًا ما تفترض أن البيانات تتبع توزيعًا طبيعيا متعدد المتغيرات ، في المقابل، فإن الأساليب اللامعلمية تضع افتراضات أقل حول توزيع البيانات ويمكن أن تكون أكثر مرونة في التقاط الأنماط المعقدة لذلك تعد الطرائق اللامعلمية مفيدة بشكل خاص عندما لا يتم استيفاء افتراضات الطرائق المعلمية أو عند التفي عامل مع البيانات التي لا تتبع توزيعًا محددًا.

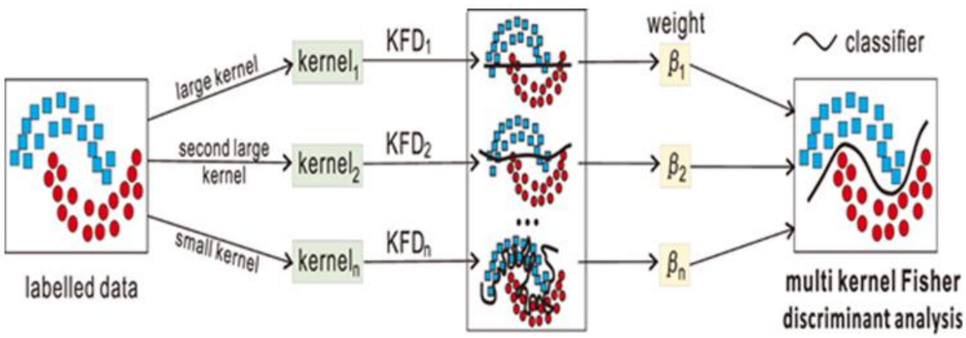
تتضمن بعض الطرق اللامعلمية الشائعة الاستعمال للتحليل التمييزي ما يلي:

12-2- 1 التحليل التمييزي اللبي (Kernel Discriminant Analysis)

ان التحليل التمييز اللبي ((KDA) Kernel discriminant analysis) تقنية قوية لمهام التصنيف ، خاصة عند التفي عامل مع هياكل البيانات المعقدة وغير الخطية. ومع ذلك ، فإنه يتطلب التحديد الدقيق لدالة اللب (kernel function) وضبط المعلمة لتحقيق الأداء الأمثل. بالإضافة إلى ذلك ، يمكن أن تكون KDA مكلفة من الناحية الحسابية بالنسبة لمجموعات البيانات الكبيرة ، لأنها تتضمن حساب مصفوفة النواة والقيمة الذاتية. اذ يعد دالة تقدير الكثافة اللبية اسلوب مهم في تحليل البيانات التمييزية، اذ ان استعمال تقديرات الكثافة اللبية في التحليل التمييزي معروف تماماً من قبل الباحثين للتعرف على الانماط الاحصائية و يتم الاعتماد على تقديرات الكثافة اللبية لبناء طريقة التصنيف اللامعلمية ، و تسمى بالتحليل التمييزي اللبي (KDA).

الفكرة الأساسية وراء (KDA) هي ايجاد فضاء فرعي مميز يزيد من التباين بين الفئة ويقلل من التباين داخل الفئة في مساحة الميزة ذات الأبعاد الأعلى الناتجة عن دالة اللب. فبذلك يمكن لـ (KDA) التفي عامل بفعالية مع البيانات غير القابلة للفصل الخطي وتحسين إمكانية فصل الفئات المختلفة.

والشكل (2-2) يوضح مبدأ (KDA):



الشكل (2-4) توضيح التحليل التمييزي اللبي [2]

ان اداء تقديرات الكثافة اللبية تعتمد بشكل كبير على الأختيار المناسب لحجم التمهيد وهي معلمة مصفوفة عرض الحزمة (H)، اذ ان القيمة المثلى لعرض الحزمة هي التي تجعل متوسط مربع الخطأ (MR Sequare Erorr) اقل مايمكن ، ومن ناحية اخرى الطرائق الاساسية لأختيارمعلمة مصفوفة عرض الحزمة (H) يكون فيها التركيز الرئيسي على تقليل نسبة خطأ التصنيف $(\widehat{MR})$

الى جانب ذلك فأن مشكلة التصنيف ستكون اكثر اهمية ووضوحاً بان تكون معلمة عرض الحزمة مختلفة عند المقارنة مع مختلف دوال الكثافة المصنفة (اي وجود عدة مجموعات ولكل مجموعة توجد تقدير كثافة، وبعد اجراء التصنيف نقوم بمقارنة بين دوال الكثافة تسمى دوال الكثافة المصنفة) واختيارها بشكل امثل لتقدير الكثافة تعتمد بشكل كبير على الكثافة المصنفة $(\hat{f}_j)$ و الاحتمال السابق لها ($\hat{\pi}_j$) ولذلك في مشكلة تعدد الطبقات (استعمال عدة متغيرات) . وعليه فان حسن اختيارمعلمة عرض الحزمة ينبغي ايضاً ان يعتمد على المشاهدة الخاصة لتصنيفها ونتيجة لذلك وعوضاً عن التركيز على عرض حزمة مثلى واحدة لتقدير الكثافة يكون اكثر فائدة ، اذاً تم الاعتماد على نتائج المستويات المختلفة من التمهيد لتقديردالة الكثافة اللبية .

لنفرض انه لدينا مجموعات مؤلفة من N من النقاط لكل منها متجه ميزات بـ n بعد أي أن:

$$X = \{x_1, x_2, \dots x_n\}$$

اذ أن $X_j \in R^d$ وان $Y = \{y_1, y_2, \dots y_n\}$ تمثل الاصناف المقابلة اذ ان $y_i \in \{1,2, \dots c\}$ والتي تمثل انتماء الصفة لكل مجموعة نقاط i^{th} .

اختيار دالة لبية معينة *k(x,x')* والتي تعبر عن مقياس التشابه بين نقاط البيانات *xx'* في فضاء الميزة الاصلي ومنه الدوال الشائعة هي الدالة الغاوسية (*Gaussain*) والدالة متعددة الحدود (*Multinomial*) والدالة السينية (*Sigmoid*) ... الخ .

ومن ثم نحسب المصفوفة اللبية K كالآتي:

$$K(i,j) = k(x_i, x_j)$$

وهي مقياس التشابه بين النقط x_i, x_j

ومن ثم نقوم بجعل مصفوفة اللب K معيارية للتأكد من خاصية *Kernel Trick properity* وكالآتي:

$$H = I\_N - \left(\frac{1}{N}\right) * 1\_N * 1\_N^T \qquad \dots (2\text{-}32)$$

اذ ان:

I_N مصفوفة الوحدة بالبعد N

1_N متجه عمودي بالواحدات بالبعد N

فان مصفوفة اللب المركزية تحسب كالآتي:

$$KC = HKH' \qquad \dots (2\text{-}33)$$

اذ ان KC تمثل القيم الذاتية (*Eigen Values*) على مصفوفة اللب المركزية للحصول على القيم الذاتية (λ) والمتجهات الذاتية المقابلة لها (a) ونقوم بترتيبها تنازلياً.

بعد ذلك نقوم باستخراج الميزات بتحديد المتجهات الذاتية k الأعلى المقابلة لأكبر متجه مميز لتشكيل فضاء فرعي جديد مخفض الأبعاد. وتمثل هذه المتجهات الذاتية $a_1, a_2, \ldots a_k$ ، $\mathrm{a_k} \in \mathrm{R}^N$ الاتجهات في مساحة الميزة ذات الابعاد الاعلى التي تفصل بين الفئات بشكل افضل.

ثم نقوم بعملية التصنيف (*Classification*) لتنصنيف نقاط البيانات الجديدة باسقاطها على مساحة الميزة ذات الابعاد المخفضة باستعمال المتجهات الذاتية (*Eigen Vectors*) المحددة . ثم نطبق خوارزمية تصنيف متجه آلآت الدعم (*Support Vector Machine*) على مساحة الميزة المخفض لعمل تنبؤات والمقصود بالتنبؤات اي ان تنبؤ الاسلوب التمييزي بالمشاهدة الى اي مجموعة تنتمي.

تتمثل الخطوة الرئيسة في الانموذج الرياضي في استخدام دالة اللب للعمل ضمنًا في مساحة ميزة ذات أبعاد أعلى دون حساب متجهات الميزات المحولة وهذا يسمح لـ *KDA* بالتفي عامل مع البيانات غير الخطية القابلة للفصل وتحقيق أداء أفضل مقارنة بـ *LDA* في مثل هذه الحالات. يعد اختيار دالة اللب وعدد المتجهات الذاتية العليا (k) معلمات تشعبية أساسية تؤثر على أداء نموذج *KDA*. غالبًا ما يتم استخدام الضبط الصحيح والتحقق من الصحة لتحديد القيم المثلى لهذه المعلمات.[33]

2-12-2 التحليل التمييزي اللبي الحصين (Robust Kernel Discriminant Analysis)

يعد التحليل التمييزي اللبي الحصين (Robust Kernel Discriminant Analysis RKDA) امتدادًا للتحليل التمييزي اللبي (KDA)) الذي يهدف إلى تحسين أداء وحصانة KDA ، خاصة عند التفي عامل مع البيانات التي بها ضوضاء أو التالفة. تتمثل الفكرة الرئيسية وراء RKDA في تقديم

الحصانة لخطوة تحويل البيانات التي تقوم بها KDA ، مما يسمح لها بالتفي عامل مع القيم الشاذة والعينات الضوضائية بشكل أكثر فعالية.[25][15]

يتضمن معيار KDA حساب مصفوفات الانتشار (داخل الطبقة وبين الطبقة) في مساحة النواة ، متبوعًا بحل مشكلة القيمة الذاتية المعممة للحصول على اتجاهات الإسقاط. ومع ذلك ، عند وجود قيم متطرفة أو عينات ضوضائية في البيانات ، يمكن أن تتأثر مصفوفات التبعثر بهذه النقاط الشاذة ، مما يؤدي إلى اتجاهات إسقاط دون المستوى الأمثل.[9]

في RKDA ، يتم استخدام متغير قوي لتقدير مصفوفة الانتشار لتقليل تأثير القيم الشاذة. أحد الأساليب الشائعة هو استخدام تقدير مصفوفة التغاير الحصين ، مثل محدد التغاير الأدنى (Minimum Covaruance Determeniant) أو مقدر هوبر ، لحساب مصفوفات التبعثر. هذه التقديرات الحصينة أقل حساسية للقيم الشاذة ويمكن أن تنتج اتجاهات إسقاط أكثر موثوقية.

تتشابه الخطوات الفي عامة للتحليل التمييزي الحصين مع KDA القياسي ، مع التعديل فقط في حساب مصفوفة التبعثر وكالآتي:

1. ادخال البيانات (Data Input) :

X: البيانات الأصلية مع العينات الموجودة في الصفوف والمعالم في الأعمدة.

Y: تسميات الفئة المقابلة لكل عينة في X.

2. دالة اللب (Kernel function):

يتم إختيار دالة kernel مناسبة (على سبيل المثال ، Gaussian ، متعدد الحدود) تقيس التشابه بين عينتين في مساحة الميزة الأصلية.

3. مصفوفة اللب (Kernel Matrix) :

يتم حساب مصفوفة اللب ، حيث تمثل (K (i,j، التشابه بين العينات (X (i، و (X (j

4. توسيط البيانات في مساحة اللب:

تحسب مصفوفة التمركز <u>H</u> ، والتي تضمن أن البيانات في مساحة النواة لها متوسط صفري.

5. مصفوفات التشتت الحصينة:

نستخدم تقدير مصفوفة التغاير الحصينة (Huber) لحساب مصفوفة التشتت داخل الفئة Sw والمصفوفة المبعثرة بين الفئة Sb في مساحة النواة باستخدام البيانات المركزية.

6. حل مشكلة القيمة الذاتية العمومية:

نبحث عن القيم الذاتية والمتجهات الذاتية لمسألة القيمة الذاتية العمومية

Sb * alpha = lambda * Sw * alpha.

7. حدد أعلى المتجهات الذاتية:

يتم اختيار أعلى متجهات ذاتية لـ k المطابقة لأكبر قيم ذاتية لـ k لتشكيل مصفوفة الإسقاط.

8. مشروع البيانات:

نقوم بإسقاط البيانات الأصلية X في مساحة الميزة الجديدة باستخدام مصفوفة الإسقاط W_rkda.

9. التصنيف:

تطبيق مصنف (على سبيل المثال ، k-Nearest Neighbours ، دعم آلة المتجه (Support Vector Machine)) على البيانات المتوقعة من أجل التصنيف.

من خلال دمج تقدير مصفوفة التبعثر الحصينة ، يمكن لـ RKDA معالجة البيانات التي بها الضوضاء أو الملوثة بشكل أكثر فعالية من طريقة KDA القياسية . [9][15]

13-2 طريقة التحقق المتقاطع الممهد (Smoothed Cross –Validation)

تم تنفيذ طريقة التحقق المتقاطع التي اقترحها Stone بواسطة (Jaksa, Nejad, 2017) لتقسيم البيانات إلى ثلاث مجموعات: التدريب والاختبار والتحقق من الصحة. تستعمل مجموعة التدريب لضبط الاوزان ، بينما تستعمل مجموعة الاختبار للتحقق من أداء الانموذج في مراحل مختلفة من التدريب ولتحديد وقت إيقاف التدريب لتجنب فرط الملائمة. يتم استخدام مجموعة التحقق من الصحة لتقدير أداء الشبكة المدربة في البيئة المنتشرة. [27]

تهدف طرق التحقق المتقاطع إلى توفير تقديرات حصينة لأداء الانموذج من خلال تقسيم البيانات بطرائق مختلفة لتدريب الانموذج والتحقق من صحته. وهي طريقة جديدة للتحقق من الصحة (Validity) وهي تقنية إعادة أخذ العينات المستعملة لتقييم أداء الانموذج التنبئي والتخفيف من مخاطر الملائمة المفرطة (Overfitting). يتضمن تقسيم مجموعة البيانات إلى مجموعات فرعية متعددة ، وتدريب النموذج على بعض هذه المجموعات الفرعية (مجموعات التدريب) ، ثم تقييم أدائها

على المجموعة الفرعية المتبقية (مجموعة التحقق من الصحة). تتكرر العملية عدة مرات للحصول على متوسط مقياس الأداء.

تم تقديم طريقة التحقق المتقاطع لأحادي المتغير(SCV) من قبل الباحثون Hall , Marron and Park في عام (1992) مع ممهد تجريبي افضل g المستقل عن h. اما بعد ذلك قدم ممهد SCV لمتعدد المتغيرات من قبل الباحثون (Sain , Baggerly and Scott (1994)) . عمل بها في البداية من صيغة معدله بشكل قليل من طريقة LSCV في المعادلة (2-38) المعروفة باسم صيغة – diagonal Leave – in للحصول على عينات البيانات التي ليس لها قيم متكرره. [2]

$$Lscv(H) = n^{-1} (4\pi)^{-d/2} |H|^{-1/2} + n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} (K_{2H} - 2K_{H} + K_0) (X_i - X_j) \quad \ldots(2\text{-}34)$$

اذ ان :

$K_0 \rightarrow$ Dirac delta function

لتشكيل (SCV) قبل تمهيد فروقات البيانات ($X_i - X_j$) بواسطة K_{2G} ، اي استبدال ($X_i - X_j$) بالإلتواء مع K_{2G} ($X_i - X_j$) :

$$SCV(H) = n^{-1} (4\pi)^{-d/2} |H|^{-1/2} + n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} (K_{2H+2G} - 2K_{H+2G} + K_{2G})(X_i - X_j) \quad \ldots(2\text{-}35)$$

اذ ان :

G: تمثل مصفوفة عرض الحزمة التجريبي

الفصل الثالث

الجانب التجريبي

تمهيد (Preface)

تم في هذا الفصل استعمال تجارب محاكاة مونت-كارلو لغرض بيان افضلية اسالسيب التحليل التمييزي المستعلمة في الجاني النظري والمقارنة فيما بينها باستعمال معيار معدل خطأ التصنيف عند دوال كثافة لها توزيع طبيعي ودوال كثافة منحرفة عن التوزيع الطبيعي.

1-3 مفهوم المحاكاة : (Simulation Concept)

بعد التطور الكبير الذي حصل في مجال الحاسبات الإلكترونية اصبح استخدام أسلوب المحاكاة كطريقة لحل ودراسة لكثير من المشكلات الصعبة والمعقدة والتي لاتكون هنالك إمكانية لتحقيقها في المجال العملي ولاسيما في حالة عدم توافر البيانات الكافية عن الظاهرة المدروسة وصعوبة الحصول على تلك البيانات فأنه يتم اللجوء الى أسلوب المحاكاة مما يوفر كثيرا من الجهد والوقت والمال،وباستخدام الحاسبات الإلكترونية يتم توليد البيانات المطلوبة نظريا من دون الحصول عليها عمليا" وذلك من دون الإخلال بدقة النتائج المطلوبة .[28]

وبصورة في عامة (فأن أسلوب المحاكاة يتلخص بكونه أسلوبا يتم من خلاله أيجاد إنموذج جديد مماثل الى الأنموذج الحقيقي من دون محاولة الحصول على الأنموذج الحقيقي) .

ويمكن القول بأن عملية المحاكاة هي أسلوب رقمي لإنجاز تجارب على الحاسبات الإلكترونية والتي تتضمن انواعا" معينة من العمليات المنطقية والرياضية الضرورية لوصف سلوك وهيكلية النظام الحقيقي المعقد خلال مدة زمنية معينة . [29]

وتوجد طرائق مختلفة للمحاكاة هي الطريقة التناظرية (Analoge Procedure) والطريقة المختلطة (Mixed procedure) وطريقة مونت- كارلو(Monte-carlo procedure) ، وتعد

طريقة مونت- كارلو من أهم هذه الطرائق واكثرها شيوعاً وتستعمل لتوليد مشاهدات لمعظم التوزيعات الاحتمالية الكثيرة الاستخدام والتي تمتلك دالة كثافة احتمالية معروفة، ويتلخص هذا الأسلوب بكونه يتم بواسطة أساليب العينات التي تؤخذ من مجتمع نظري يحاكي المجتمع الحقيقي حيث يتم صياغة الارقام العشوائية. وتمتاز عملية المحاكاة بالمرونة اذ تعطي القدرة على التجريب والاختبار من خلال تكرار العملية لمرات عديدة بتفسير المدخلات الخاصة بعمليات التقدير في كل مرة وكذلك تأتي أهمية عملية المحاكاة في العشوائية ،اذ ان سلسلة الارقام العشوائية التي تستعمل في التجربة الاولى تكون مستقلة عن سلسلة الارقام العشوائية في التجربة الثانية وهكذا.[30]

2-3 خطوات تجارب المحاكاة (Steps of experiments Simulation)

تمت عملية بناء انموذج محاكاة لدراسة سلوك طرائق التحليل التمييزي المدروس باعتماد معايير معينة في تقدير النماذج المفترضة وكما الخطوات الآتية:

اولاً: تحديد احجام العينات :

تعد هذه الخطوة من الخطوات المهمة التي يعتمد عليها لتنفيذ باقي خطوات تجارب المحاكاة والتي تتضمن الآتي:

1) اختيار احجام عينات التدريب (Triannning Sample size)

تم تحديد احجام عينات التدريب الافتراضية لغرض اجراء طرائق التحليل التمييزي المستعملة في هذه الرسالة والتي هي:

$$n=100, 500, 1000, 5000$$

2) اختيار احجام عينات الاختبار (Test Sample size)

تم تحديد حجم عينة الاختبار الافتراضية لغرض اجراء طرائق التحليل التمييزي المستعملة في هذه الرسالة والتي هي:

$$k=1000, 2000, 3000, 5000$$

ثانياً : توليد العينات (Samples Generating)

تم في هذه الرسالة استعمال مجوعتين لغرض اختبار طرائق التحليل التمييزي اذ تم توليد بيانات المجموعة الاولى والمجموعة الثانية بإستعمال طريقة (Box-Muller) بالاضافة إلى الدالة المكتبية (Randn) في برنامج ماتلاب وكما يأتي :

حيث تستعمل لتوليد متغير عشوائي ثنائي يتبع التوزيع الطبيعي القياســــــــــي N_2 (0, 1) وتعتمد طريقة (Box – Muller) على الأسلوب الآتي:

1. توليد عددين عشوائيين مستقلين U_1، U_2 بحيث يتبعان التوزيع المنتظم للفترة (0 ،1) حيث يتم توليد متجه معين من هذين العددين بحجم العينة المطلوبة (n) أي إن:

$$U_i = \text{rand}\ (1, n) \quad \dots (3\text{-}1)$$

2. يمكن تحويل هذين العددين إلى التوزيع الطبيعي القياسي وفقاً لما يأتي:

$$X_1 = (-2\ln(U_1))^{1/2} \cos(2\pi U_2) \quad \dots (3\text{-}2)$$

$$X_2 = (-2\ln(U_1))^{1/2} \sin(2\pi U_2) \quad \dots (3\text{-}3)$$

حيث أن X_1، X_2 متغيران عشوائيان طبيعيان مستقلان وبذلك فان الدالة المشتركة لهما هي:

$$f(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}\left(x_1^2 + x_2^2\right)} \quad \dots (3\text{-}4)$$

3. تم توليد عدد من المتغيرات التي لها توزيع طبيعي قياسي متعدد متغيرات في كل مجموعة باستعمال الخوارزمية الآتية في ماتلاب:

1) تحديد حجم العينة (n) في كل متغير كالآتي:

```
numSamples = n;
```

2) تحديد عدد المتغيرات في كل مجموعة وكالآتي:

```
numVariables = K; % Change this to the desired number of variables
```

3) تحديد متجه المتوسطات الصفري بعدد المتغيرات المطلوبة وكالآتي:

```
mu = zeros(1, numVariables);
```

4) تحديد مصفوفة التبيان-التباين المشترك للتوزيع الطبيعي القياسي متعدد المتغيرات وكالآتي:

```
covMatrix = eye(numVariables);
```

5) توليد المتغيرات العشوائية القياسية الطبيعية المتعددة في كل مجموعة كالآتي:

$\underline{X}$ = mvnrnd(mu, covMatrix, numSamples);

ثالثاً: اختيار مصفوفة عرض الحزمة (Bandwidth Selection)

في هذه الرسالة تم اختيار مصفوفة عرض الحزمة القطرية باستعمال طريقة العبور الشرعي

رابعاً : اختيار دالة الكثافة اللبية الهدف (Target Kernel Density Selection)

في هذه الرسالة تم اختيار اربعة دوال كثافة لبية لتحقيق صيغة التمييز اللبي البيزية مع توزيعاتها الاولية الافتراضية بحيث تكون بعض من هذه الدوال الهدف تبتعد عن التوزيع الطبيعي لغرض انتهاك افتراض التوزيع الطبيعي لاختبار طرائق التقدير عند هذه الدوال الهدف وكالآتي:[22]

1) دالة الهدف الكاوسية D : والتي تكون كالآتي:

$$\boldsymbol{D} = \begin{cases} \pi_1 = \frac{1}{2}\ \mathbf{f_1 \sim N_2}\ \left(\begin{bmatrix}\mathbf{1}\\ \mathbf{-1}\end{bmatrix}, \begin{bmatrix}\frac{4}{9} & \frac{14}{45}\\ \frac{14}{45} & \frac{4}{9}\end{bmatrix}\right) \\ \pi_2 = \frac{1}{2}\ \mathbf{f_2 \sim N_2}\ \left(\begin{bmatrix}\mathbf{-1}\\ \mathbf{1}\end{bmatrix}, \begin{bmatrix}\frac{4}{9} & \mathbf{0}\\ \mathbf{0} & \frac{4}{9}\end{bmatrix}\right) \end{cases} \quad \text{.... (3-5)}$$

2) دالة الهدف الكاوسية E: والتي تكون كالآتي:

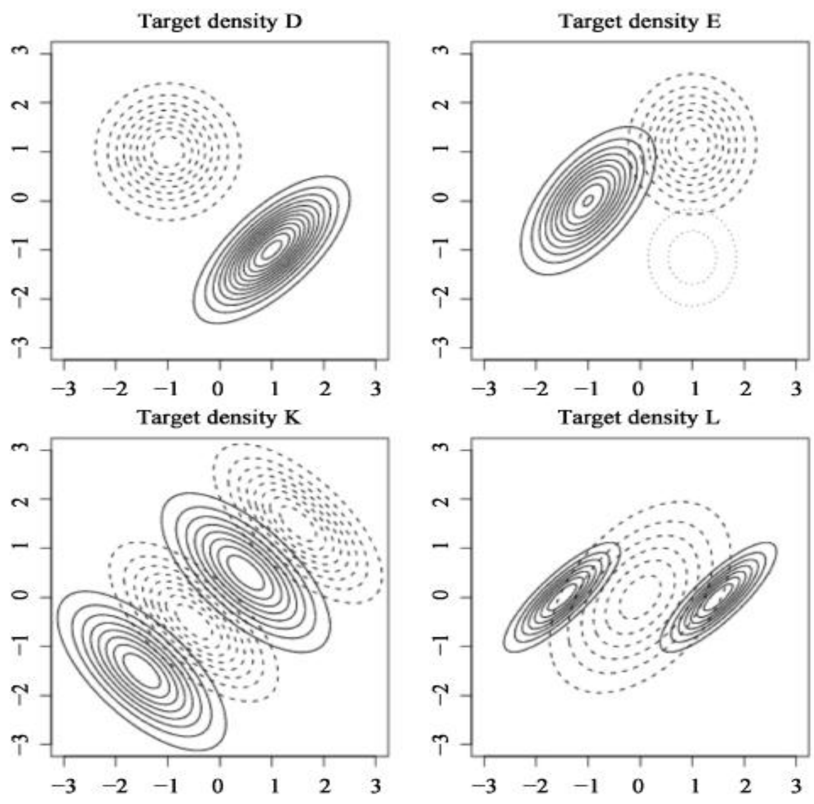
$$\boldsymbol{E} = \begin{cases} \pi_1 = \frac{3}{7}\ \mathbf{f_1 \sim N_2}\ \left(\begin{bmatrix}\mathbf{-1}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\frac{9}{25} & \frac{63}{250}\\ \frac{63}{250} & \frac{49}{100}\end{bmatrix}\right) \\ \pi_2 = \frac{3}{7}\ \mathbf{f_2 \sim N_2}\ \left(\begin{bmatrix}\mathbf{1}\\ \frac{2}{\sqrt{3}}\end{bmatrix}, \begin{bmatrix}\frac{9}{25} & \mathbf{0}\\ \mathbf{0} & \frac{49}{100}\end{bmatrix}\right) \end{cases} \quad \text{.... (3-6)}$$

3)دالة الهدف K والتي تكون كالآتي:

$$K=\begin{cases}\pi_1=\frac{1}{2}\ \mathbf{f_1}\sim\mathbf{N_2}\left(\begin{bmatrix}\frac{-3}{2}\\ \frac{-3}{2}\end{bmatrix},\begin{bmatrix}\frac{4}{5} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{4}{5}\end{bmatrix}+\frac{1}{2}\mathbf{N_2}\begin{bmatrix}\frac{1}{2}\\ \frac{1}{2}\end{bmatrix},\begin{bmatrix}\frac{4}{5} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{4}{5}\end{bmatrix}\right)\\ \pi_1=\frac{1}{2}\ \mathbf{f_2}\sim\mathbf{N_2}\left(\begin{bmatrix}\frac{3}{2}\\ \frac{3}{2}\end{bmatrix},\begin{bmatrix}\frac{4}{5} & \frac{1}{2}\\ \frac{1}{2} & \frac{4}{5}\end{bmatrix}+\frac{1}{2}\mathbf{N_2}\begin{bmatrix}-\frac{1}{2}\\ -\frac{1}{2}\end{bmatrix},\begin{bmatrix}\frac{4}{5} & -\frac{1}{2}\\ -\frac{1}{2} & \frac{4}{5}\end{bmatrix}\right)\end{cases}\qquad \text{.... (3-7)}$$

4)دالة الهدف L والتي تكون كالآتي:

$$L=\begin{cases}\pi_1=\frac{1}{3}\ \mathbf{f_1}\sim\ \frac{1}{2}\mathbf{N_2}\left(\begin{bmatrix}-\frac{3}{2}\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\frac{3}{10} & \frac{1}{4}\\ \frac{1}{4} & \frac{3}{10}\end{bmatrix}\right)+\frac{1}{2}\mathbf{N_2}\left(\begin{bmatrix}\frac{3}{2}\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\frac{3}{10} & \frac{1}{4}\\ \frac{1}{4} & \frac{3}{10}\end{bmatrix}\right)\\ \pi_2=\frac{2}{3}\ \mathbf{f_2}\sim\mathbf{N_2}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\frac{4}{5} & \frac{2}{5}\\ \frac{2}{5} & \mathbf{1}\end{bmatrix}\right)\end{cases}\qquad \text{.... (3-8)}$$



شكل (3-1) الرسم المحيطي (Contour Plot) دوال الكثافة الهدف (D, E, K, and L)

[22]

ويبين الشكل (3-1) انواع دوال الكثافة المستعملة وهي دالة الهدف الكاوسية D و دالة الهدف الكاوسية E التي يكون لها توزيع طبيعي بالصفات المحددة في ارقام المعادلات (3-5) و (3-4). ودالة الهدف K و L هي الدوال المنحرفة عن التوزيع الطبيعي.

وبذلك يكون ملخص النماذج المستعملة كما في الجدول (3-1) الآتي:

جدول (3-1) ملخص النماذج المفترضة في جداول المحاكاة

Model	n	K	Traget density
1	100	1000	D
	500		E
	1000		K
	5000		L
2	100	2000	D
	500		E
	1000		K
	5000		L
3	100	3000	D
	500		E
	1000		K
	5000		L
4	100	5000	D
	500		E
	1000		K
	5000		L

سادساً : المقارنة بين اساليب التحليل التمييزي (Compairing between Approches)

تمت المقارنة بين طرائق التحليل التمييزي باستعمال معدل خطأ التصنيف والانحراف المعياري لخطأ التصنيف عند كل دالة كثافة لبية والذي صيغته كالآتي:

$$\widehat{MR} = 1 - m^{-1} \sum_{j=1}^{m} 1 \{ Y_j \text{ is correctly classified using KDA} \} \quad \ldots (3\text{-}9)$$

ومن ثم تحديد أفضل أسلوب من بين الأساليب المستعملة في التقدير التي تمتلك أقل نسبة الخطأ $(\widehat{MR})$.

3-3: تحليل نتائج المحاكاة (Analysis of Experiments Results)

تمت المقارنة بين اساليب التحليل التمييزي الآتية :

التحليل التمييزي الخطي (LDA)

التحليل التمييزي التربيعي (QDA)

التحليل التمييزي اللبي (KDA)

التحليل التمييزي اللبي الحصين (RKDA)

وكما في الجداول الآتية :

جدول (2-3) المعدل والانحراف المعياري لخطأ التصنيف وفق أساليب التحليل التمييزي للانموذج الأول

Target Density	n	k	method	LDA	QDA	KDA	RKDA
D	100	1000	MR	0.05892	0.08992	0.19092	0.22343
			STD	0.00152	0.00671	0.00211	0.00115
E			MR	0.04456	0.05675	0.15322	0.21134
			STD	0.00133	0.00455	0.00115	0.00132
K			MR	0.13331	0.15643	0.05462	0.00462
			STD	0.00235	0.00364	0.00143	0.00111
L			MR	0.16794	0.14788	0.00835	0.00342
			STD	0.00366	0.00227	0.00127	0.00045

Target Density	n	k	method	LDA	QDA	KDA	RKDA
D	500	1000	MR	0.04792	0.07892	0.17992	0.21243
			STD	0.00448	0.00429	0.00689	0.00685
E			MR	0.04575	0.03356	0.14222	0.20022
			STD	0.00567	0.00645	0.00785	0.00761
K			MR	0.12231	0.14543	0.00533	0.04362
			STD	0.00265	0.00236	0.00152	0.00142
L			MR	0.15694	0.13688	0.00265	0.00258
			STD	0.00734	0.00873	0.00973	0.00151

Target Density	N	k	method	LDA	QDA	KDA	RKDA
D	1000	1000	MR	0.14344	0.11891	0.01131	0.11845
			STD	0.01245	0.01673	0.00343	0.01433

E			MR	0.12895	0.11234	0.05534	0.11324
			STD	0.04553	0.02346	0.00114	0.01146
K			MR	0.02253	0.04456	0.00335	0.00226
			STD	0.00454	0.00466	0.00131	0.00245
L			MR	0.03433	0.11322	0.00167	0.00118
			STD	0.00542	0.02123	0.00253	0.00103
Target Density	**N**	**k**	**method**	**LDA**	**QDA**	**KDA**	**RKDA**
D	5000	1000	MR	0.11244	0.08791	0.01969	0.08745
			STD	0.01855	0.01427	0.00757	0.03667
E			MR	0.06795	0.05134	0.00431	0.08224
			STD	0.01453	0.00754	0.02922	0.01954
K			MR	0.00847	0.01356	0.00662	0.00271
			STD	0.02646	0.02634	0.00961	0.00151
L			MR	0.00333	0.08222	0.02933	0.00024
			STD	0.02558	0.00977	0.02847	0.00191

نلاحظ من جدول (3-2) ما يأتي:

1) عندما **(n=100, k=1000)** كان اسلوب التحليل التمييزي الخطي هو الافضل عند دوال الهدف (D, E) بأقل خطأ تصنيف بلغ **(0.05892)** و **(0.04456)** على التوالي، يليه اسلوب التحليل التميز التربيعي عند نفس دوال الهدف ، وان اسلوب التحليل التمييز اللبي الحصين كان هو الافضل عند الدوال (K, L) باقل خطأ تصنيف بلغ **(0.00462)** و **(0.00342)** على التوالي.

2) عندما (**n=500, k=1000**) كان اسلوب التحليل التمييزي الخطي هو الافضل عند دالة الهدف (D) بأقل خطأ تصنيف بلغ (**0.01131**) على التوالي ، يليه اسلوب التحليل التمييز التربيعي باقل خطأ تصنيف بلغ (**0.03356**) عند دالة الهدف (E) ، وان اسلوب التحليل التمييز اللبي كان هو الافضل عند دوال الهدف (K) باقل خطأ تصنيف بلغ (**0.00533**). يليه اسلوب التحليل التمييزي اللبي الحصين عند دالة الهدف (L) بلغ (**0.00258**).

3) عندما (**n=1000, k=1000**) كان اسلوب التحليل التمييزي اللبي هو الافضل عند دوال الهدف (D, E) بأقل خطأ تصنيف بلغ (**0.04792**) و (**0.05534**) على التوالي ، يليه اسلوب التحليل التمييز التربيعي باقل خطأ تصنيف عند دالة الهدف (E) باقل خطأ تصنيف بلغ (**0.05534**) ، وان اسلوب التحليل التمييز اللبي الحصين كان هو الافضل عند دوال الهدف (K, L) باقل خطأ تصنيف بلغ (**0.00226**) و (**0.00118**) على التوالي.

4)عندما (**n=5000, k=1000**) كان اسلوب التحليل التمييزي اللبي هو الافضل عند دوال الهدف (D, E) بأقل خطأ تصنيف بلغ (**0.01969**) و (**0.00431**) على التوالي ، يليه اسلوب التحليل التمييز التربيعي باقل خطأ تصنيف عند دالة الهدف (E) باقل خطأ تصنيف بلغ (**0.05534**) ، وان اسلوب التحليل التمييز اللبي الحصين كان هو الافضل عند دوال الهدف (K, L) باقل خطأ تصنيف بلغ (**0.00271**) و (**0.00024**) على التوالي.

جدول (3-3) المعدل والانحراف المعياري لخطأ التصنيف وفق أساليب التحليل التمييزي للانموذج الثاني

Target Density	n	K	Method	LDA	QDA	KDA	RKDA
D	100	2000	MR	0.04762	0.05893	0.17941	0.23424
			STD	0.00148	0.00329	0.00789	0.04983
E			MR	0.04552	0.03344	0.11222	0.19033
			STD	0.00467	0.00245	0.00482	0.00662
K			MR	0.12231	0.14543	0.04362	0.00638
			STD	0.01165	0.02336	0.00657	0.00289
L			MR	0.15694	0.13688	0.00265	0.00252
			STD	0.03734	0.02873	0.00273	0.00055
Target Density	**N**	**K**	**Method**	**LDA**	**QDA**	**KDA**	**RKDA**
D	500	2000	MR	0.02662	0.03793	0.15841	0.21324
			STD	0.01952	0.01771	0.01311	0.02883
E			MR	0.02452	0.01244	0.09122	0.16933
			STD	0.01633	0.01855	0.01618	0.01438
K			MR	0.10131	0.12443	0.02262	0.01462
			STD	0.00935	0.00236	0.01443	0.01811
L			MR	0.33594	0.10588	0.01535	0.01482
			STD	0.01634	0.00773	0.01827	0.02045
Target Density	**N**	**K**	**Method**	**LDA**	**QDA**	**KDA**	**RKDA**
D	1000	2000	MR	0.05643	0.02402	0.03538	0.15128

			STD	0.03312	0.03775	0.00255	0.03318
E			MR	0.10733	0.13244	0.02923	0.03242
			STD	0.03439	0.04855	0.03618	0.00231
K			MR	0.06241	0.09466	0.02738	0.02441
			STD	0.02233	0.03816	0.00443	0.00231
L			MR	0.27394	0.04388	0.03537	0.01487
			STD	0.03634	0.02773	0.03827	0.01042
Target Density	**N**	**K**	**Method**	**LDA**	**QDA**	**KDA**	**RKDA**
D	5000	2000	MR	0.39743	0.36502	0.17631	0.49228
			STD	0.37412	0.37875	0.34355	0.37418
E			MR	0.44833	0.47344	0.11024	0.37342
			STD	0.37539	0.38955	0.21711	0.34331
K			MR	0.40341	0.43566	0.36838	0.12541
			STD	0.36333	0.37916	0.34543	0.24331
L			MR	0.61494	0.38488	0.37637	0.05581
			STD	0.37734	0.36873	0.37927	0.25142

نلاحظ من جدول (3-3) ما يأتي:

1)عندما (**n=100, k=2000**) كان اسلوب التحليل التمييزي الخطي هو الافضل عند دوال الهدف (D) بأقل خطأ تصنيف بلغ (**0.04762**)، يليه اسلوب التحليل التميز التربيعي عند عند دالة الهدف (E) بأقل خطأ تصنيف بلغ (**0.03344**)، وان اسلوب التحليل التمييز اللبي الحصين كان هو الافضل عند دوال الهدف (K, L) باقل خطأ تصنيف بلغ (**0.00638**) و (**0.00252**) على التوالي.

2)عندما (**n=500, k=2000**) كان اسلوب التحليل التمييزي الخطي هو الافضل عند دالة الهدف (D) بأقل خطأ تصنيف بلغ (**0.02662**) و (**0.02452**) على التوالي ، وان اسلوب التحليل التمييز اللبي الحصين كان هو الافضل عند دوال الهدف (K, L) باقل خطأ تصنيف بلغ(**0.01462**) و (**0.01482**) على التوالي.

3)عندما (**n=1000, k=2000**) كان اسلوب التحليل التمييزي اللبي هو الافضل عند دالة الهدف (D) بأقل خطأ تصنيف بلغ (**0.03538**)، ، وان اسلوب التحليل التمييز اللبي الحصين كان هو الافضل عند دوال الهدف (E, K, L) باقل خطأ تصنيف بلغ (**0.03242**) و (**0.02441**) و (**0.01487**) على التوالي.

4)عندما (**n=5000, k=2000**) كان اسلوب التحليل التمييزي اللبي هو الافضل عند دوال الهدف (D, E) بأقل خطأ تصنيف بلغ (**0.17631**) و (**0.11024**) على التوالي ، يليه اسلوب التحليل التمييز التربيعي باقل خطأ تصنيف عند دالة الهدف (E) باقل خطأ تصنيف بلغ (**0.05534**) ، وان اسلوب التحليل التمييز اللبي الحصين كان هو الافضل عند دوال الهدف (K, L) باقل خطأ تصنيف بلغ (**0.12541**) و (**0.05581**) على التوالي.

جدول (3-4) المعدل والانحراف المعياري لخطأ التصنيف وفق أساليب التحليل التمييزي للانموذج الثالث

Target Density	N	K	method	LDA	QDA	KDA	RKDA
D	100	3000	MR	0.00662	0.01793	0.13841	0.19324
			STD	0.03952	0.03771	0.03311	0.00883
E			MR	0.00452	0.00756	0.07122	0.14933
			STD	0.03633	0.03855	0.03618	0.03438
K			MR	0.08131	0.10443	0.00262	0.01461
			STD	0.02935	0.01764	0.03443	0.03811
L			MR	0.11594	0.09588	0.03835	0.00148
			STD	0.00466	0.01327	0.00273	0.00051
Target Density	N	K	method	LDA	QDA	KDA	RKDA
D	500	3000	MR	0.03438	0.02307	0.09741	0.15224
			STD	0.00148	0.00329	0.00789	0.03217
E			MR	0.03648	0.03344	0.03022	0.10833
			STD	0.00467	0.00245	0.00482	0.00662
K			MR	0.04031	0.06343	0.03838	0.02639
			STD	0.01165	0.02336	0.00657	0.00289
L			MR	0.07494	0.05488	0.00265	0.00151
			STD	0.03634	0.02773	0.00827	0.00049
Target Density	N	K	method	LDA	QDA	KDA	RKDA
D	1000	3000	MR	0.00652	0.01793	0.05641	0.11124
			STD	0.00952	0.03771	0.03311	0.00883

E			MR	0.00452	0.00756	0.01078	0.06733
			STD	0.03633	0.03855	0.03618	0.03438
K			MR	0.01461	0.02243	0.00262	0.00069
			STD	0.02811	0.01764	0.03443	0.00935
L			MR	0.03949	0.03835	0.02394	0.00018
			STD	0.04051	0.03273	0.00466	0.00327
Target Density	**N**	**K**	**method**	**LDA**	**QDA**	**KDA**	**RKDA**
D	5000	3000	MR	0.03438	0.02307	0.00541	0.07024
			STD	0.03148	0.00329	0.00189	0.03217
E			MR	0.03648	0.03344	0.00022	0.02633
			STD	0.00467	0.00245	0.00112	0.00662
K			MR	0.02639	0.01857	0.03838	0.00031
			STD	0.01289	0.02336	0.00657	0.00165
L			MR	0.00151	0.00265	0.01706	0.00012
			STD	0.00049	0.00827	0.03634	0.00103

نلاحظ من جدول (3-4) ما يأتي:

1)عندما (**n=100, k=3000**) كان اسلوب التحليل التمييزي الخطي هو الافضل عند دوال الهدف (D, E) بأقل خطأ تصنيف بلغ (**0.00662**) و (**0.00452**) على التوالي ، يليه اسلوب التحليل التمييز اللبي عند دالة الهدف (K) باقل خطأ تصنيف بلغ (**0.00262**)، وان اسلوب التحليل التمييز اللبي الحصين كان هو الافضل عند دوال الهدف (L) باقل خطأ تصنيف بلغ (**0.00148**) .

2)عندما (**n=500, k=3000**) كان اسلوب التحليل التمييزي التربيعي هو الافضل عند دالة الهدف (D) بأقل خطأ تصنيف بلغ (**0.02307**) ، اسلوب التحليل التمييزي اللبي هو الافضل عند دالة الهدف

(E) بأقل خطأ تصنيف بلغ (**0.03022**) ، وان اسلوب التحليل التمييز اللبي كالحصينان هو الافضل عند دوال الهدف (K, L) باقل خطأ تصنيف بلغ (**0.02639**) و (**0.00151**) على التوالي.

3)عندما (**n=1000, k=3000**) كان اسلوب التحليل التمييزي الخطي هو الافضل عند دالة الهدف (D, E) بأقل خطأ تصنيف بلغ (**0.000652**) و (**0.00452**) على التوالي ، ، وان اسلوب التحليل التمييز اللبي الحصين كان هو الافضل عند دوال الهدف (K, L) باقل خطأ تصنيف بلغ (**0.00069**) و (**0.00018**) على التوالي.

4)عندما (**n=5000, k=3000**) كان اسلوب التحليل التمييزي اللبي هو الافضل عند دوال الهدف (D, E) بأقل خطأ تصنيف بلغ (**0.00541**) و (**0.00022**) على التوالي ، وان اسلوب التحليل التمييز اللبي الحصين كان هو الافضل عند دوال الهدف (K, L) باقل خطأ تصنيف بلغ (**0.00031**) و (**0.00012**) على التوالي.

جدول (3-5) المعدل والانحراف المعياري لخطأ التصنيف وفق أساليب التحليل التمييزي للانموذج الرابع

Target Density	N	k	method	LDA	QDA	KDA	RKDA
D	100	5000	MR	0.03792	0.06892	0.16992	0.20243
			STD	0.01948	0.01429	0.01889	0.01985
E			MR	0.02356	0.03575	0.13222	0.19034
			STD	0.01967	0.01645	0.01985	0.01968
K			MR	0.11231	0.13543	0.03362	0.01638
			STD	0.01865	0.01736	0.01957	0.01989
L			MR	0.14694	0.12688	0.01265	0.01058
			STD	0.01734	0.01873	0.01973	0.02055
Target Density	**N**	**k**	**method**	**LDA**	**QDA**	**KDA**	**RKDA**
D	500	5000	MR	0.01692	0.04792	0.14892	0.18143
			STD	0.00152	0.00671	0.00211	0.00115
E			MR	0.00256	0.01475	0.11122	0.16934
			STD	0.00133	0.00455	0.00115	0.00132
K			MR	0.09131	0.11443	0.01262	0.00462
			STD	0.00235	0.00364	0.00143	0.00111
L			MR	0.12594	0.10588	0.00835	0.00042
			STD	0.00366	0.00227	0.00127	0.00045
Target Density	**N**	**K**	**method**	**LDA**	**QDA**	**KDA**	**RKDA**
D	1000	5000	MR	0.01338	0.01559	0.00207	0.04924
			STD	0.01048	0.01911	0.01771	0.01117

E			MR	0.01548	0.01244	0.00124	0.00533
			STD	0.01633	0.01855	0.00588	0.01438
K			MR	0.00539	0.00243	0.01738	0.00069
			STD	0.00811	0.00236	0.01443	0.01935
L			MR	0.01949	0.01835	0.00394	0.00018
			STD	0.02051	0.01273	0.01534	0.00394
Target Density	**N**	**K**	**method**	**LDA**	**QDA**	**KDA**	**RKDA**
D	5000	5000	MR	0.00762	0.01893	0.00541	0.02824
			STD	0.01052	0.00329	0.00189	0.00983
E			MR	0.01976	0.00856	0.00552	0.01567
			STD	0.01512	0.00245	0.00467	0.00662
K			MR	0.01561	0.01857	0.02031	0.00411
			STD	0.01289	0.01864	0.00165	0.00657
L			MR	0.01706	0.02082	0.00265	0.00112
			STD	0.00566	0.01706	0.00827	0.00049

نلاحظ من جدول (3-5) ما يأتي:

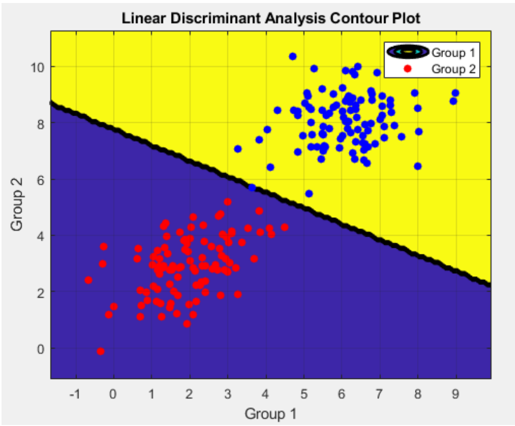
1) عندما **(n=100, k=3000)** كان اسلوب التحليل التمييزي الخطي هو الافضل عند دوال الهدف (D, E) بأقل خطأ تصنيف بلغ **(0.03792)** و **(0.02356)** على التوالي ، يليه اسلوب التحليل التمييز اللبي الحصين عند دالة الهدف (K, L) باقل خطأ تصنيف بلغ **(0.01638)** و **(0.01058)** على التوالي .

2) عندما **(n=500, k=3000)** كان اسلوب التحليل التمييزي الخطي هو الافضل عند دالة الهدف (D, E) بأقل خطأ تصنيف بلغ **(0.01692)** و **(0.00242)** على التوالي ، ، وان اسلوب

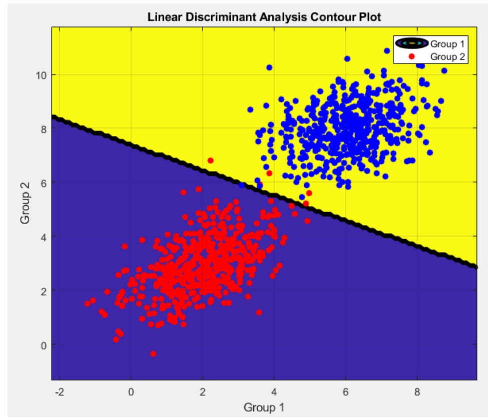
التحليل التمييز اللبي الحصين هو الافضل عند دوال الهدف (K, L) باقل خطأ تصنيف بلغ (**0.00462**) و (**0.00042**) على التوالي.

3) عندما (**n=1000, k=3000**) كان اسلوب التحليل التمييزي الخطي هو الافضل عند دالة الهدف (D, E) بأقل خطأ تصنيف بلغ (**0.00207**) و (**0.00124**) على التوالي ، ، وان اسلوب التحليل التمييز اللبي الحصين كان هو الافضل عند دوال الهدف (K, L) باقل خطأ تصنيف بلغ (**0.00069**) و (**0.00018**) على التوالي.

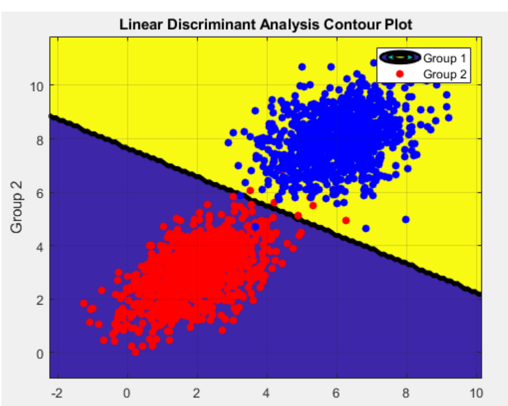
4) عندما (**n=5000, k=3000**) كان اسلوب التحليل التمييزي اللبي هو الافضل عند دوال الهدف (D, E) بأقل خطأ تصنيف بلغ (**0.00541**) و (**0.00552**) على التوالي ، وان اسلوب التحليل التمييز اللبي الحصين كان هو الافضل عند دوال الهدف (K, L) باقل خطأ تصنيف بلغ (**0.00411**) و (**0.00112**) على التوالي.



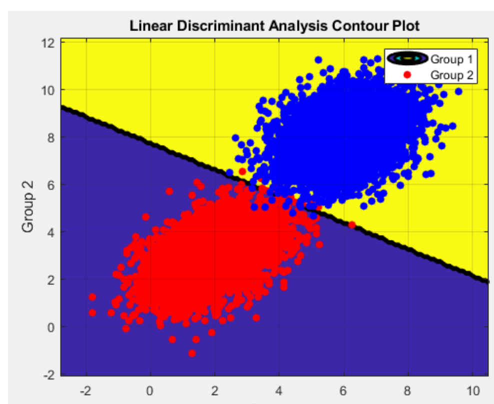
شكل (3-2) التصنيف وفق التحليل التمييزي الخطي عندما n=100, k=1000



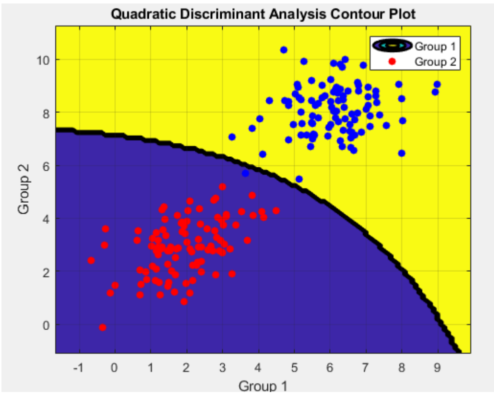
شكل (3-3) التصنيف وفق التحليل التمييزي الخطي عندما n=500, k=1000



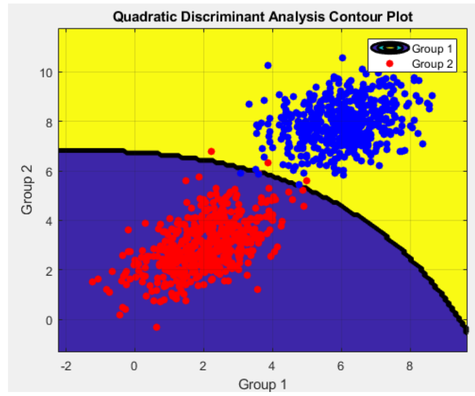
شكل (3-4) التصنيف وفق التحليل التمييزي الخطي عندما n=1000, k=1000



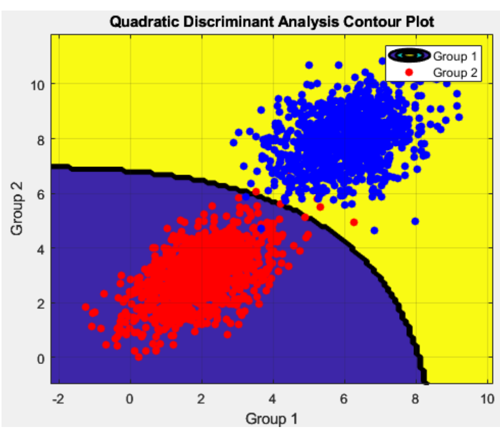
شكل (3-5)التصنيف وفق التحليل التمييزي الخطي عندما n=5000, k=1000



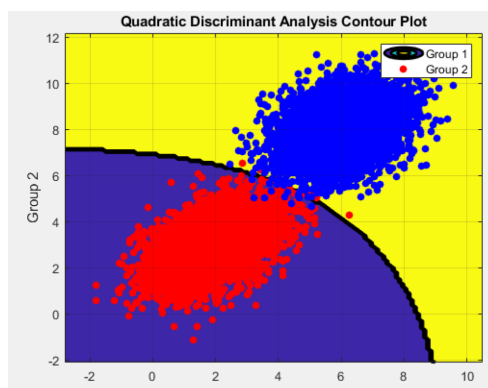
شكل (3-6) التصنيف وفق التحليل التمييزي التربيعي عندما n=100, k=1000



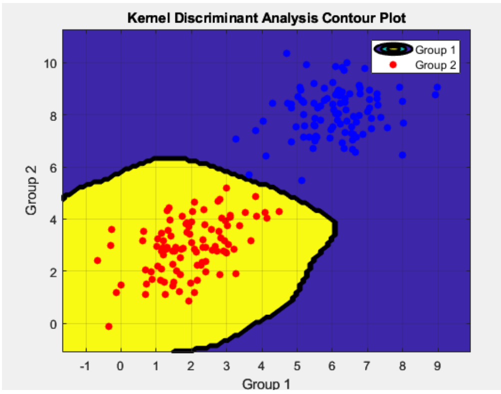
شكل (3-7) التصنيف وفق التحليل التمييزي التربيعي عندما n=500, k=1000



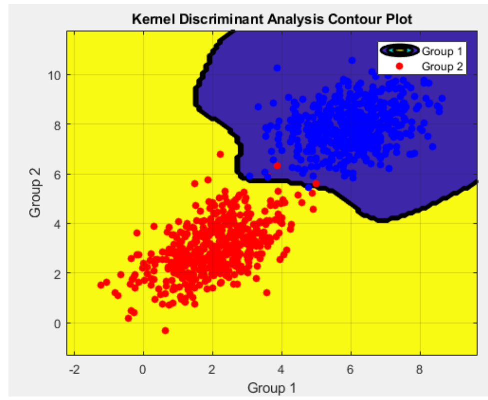
شكل (3-8) التصنيف وفق التحليل التمييزي التربيعي عندما n=1000, k=1000



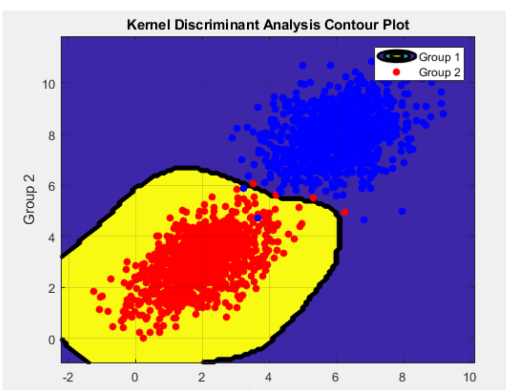
شكل (3-9) التصنيف وفق التحليل التمييزي التربيعي عندما n=5000, k=1000



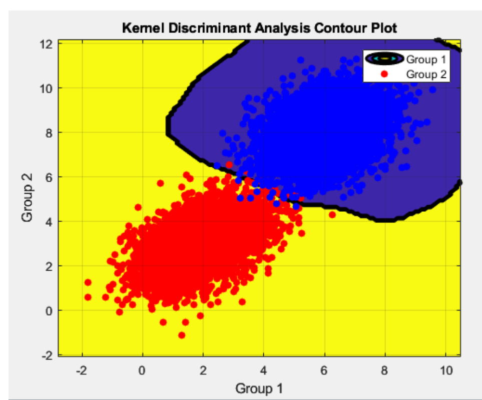
شكل (3-10) التصنيف وفق التحليل التمييزي اللبي عندما n=100, k=1000



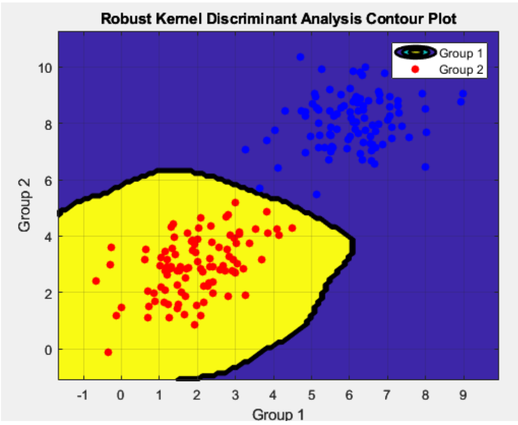
شكل (3-11) التصنيف وفق التحليل التمييزي اللبي عندما n=500, k=1000



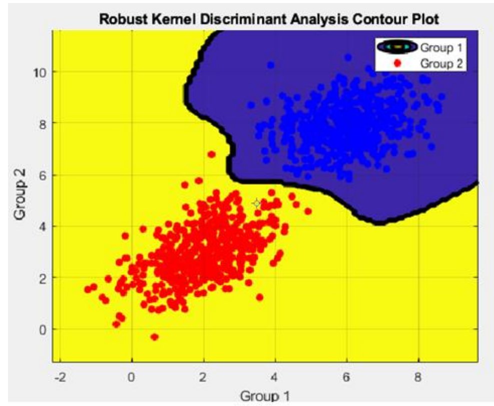
شكل (3-12) التصنيف وفق التحليل التمييزي اللبي عندما n=1000, k=1000



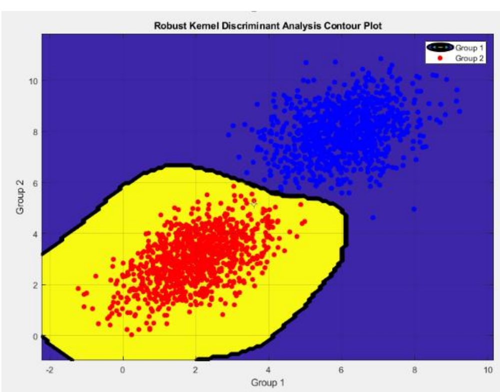
شكل (3-13) التصنيف وفق التحليل التمييزي اللبي عندما n=5000, k=1000



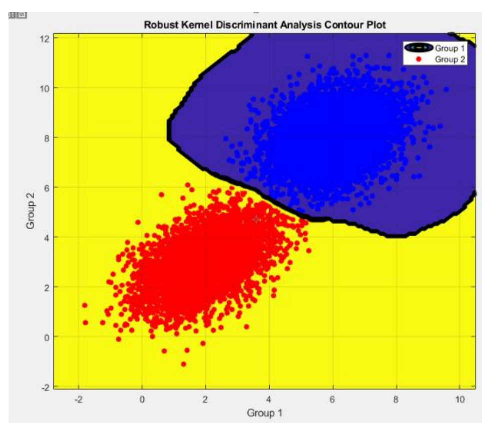
شكل (3-14) التصنيف وفق التحليل التمييزي اللبي الحصين عندما n=100, k=1000



شكل (3-15) التصنيف وفق التحليل التمييزي اللبي الحصين عندما n=500, k=1000



شكل (3-16) التصنيف وفق التحليل التمييزي اللبي الحصين عندما n=1000, k=1000



شكل (3-17) التصنيف وفق التحليل التمييزي اللبي الحصين عندما n=5000, k=1000

تبين الأشكال (3-2) الى (3-17) ان اسلوب التحليل التمييزي اللبي الحصين قد زاد من دقة تصنيف المشاهدات بين المجوعات اذ تم الفصل وفق معدل خطأ تصنيف قليل جداً مقرمة بالاساليب الاخرى كما نلاحظ من المنطقة الزرقاء والمنطقة الصفراء اذ هنالك فصل تام بين نقاط المجموعتين ولاتوجد نقاط تتداخل مع كلا المجموعتين. على كس التحليل التمييزي اللبي اذ نلاحظ ان هنالك بعض نقاط المشاهدات تداخلت بين المنطقة الصفراء والمنطقة الزرقاء اي ان الفصل ليس تام بين المجموعتين.

ولبيان افضلية اساليب التحليل التمييزي ندرج ادناه جدول الأفضلية الذي يبين عدد مرات الافضلية ونسب الافضلية لكل اسلوب وعند كل حجم عينة تدريب وعينة اختبار

جدول (3-6) عدد مرات الافضلية ونسب الافضلية لكل اسلوب وعند كل دالة هدف

Traget density	No.	Ratio	No.	Ratio	No.	Ratio	No.	Ratio
	LDA		QDA		KDA		RKDA	
D	8	13	0	0	7	11	0	0
E	6	9	2	3	6	9	1	2
K	0	0	0	0	3	5	15	23
L	0	0	0	0	0	0	16	25

نلحظ من جدول (3-6) ما يأتي:

1- حقق اسلوب التحليل التمييزي الخطي افضلية على باقي اساليب التحليل التمييزي عند دوال الكثافة التي تتوزع طبيعياً (D, E) وبنسبة (13%) و (9%) لكل دلة كثافة على التوالي، وحقق اسلوب التحليل التمييزي اللبي افضلية عند هذه الدوال عند حجم العينة (n=1000, 5000) بنسبة (11%)

2- حقق اسلوب التحليل التمييزي اللبي افضلية على باقي الاساليب عند دالة الكثافة (K) بنسبة قليلة بلغت (5%) .

3- حقق اسلوب التحليل التمييزي اللبي الحصين افضلية على باقي الاساليب عند دوال الكثافة المنحرفة عن التوزيع الطبيعي بنسبة افضلية (23%) و (25%) لكل دالة كثافة على التوالي.

الفصل الرابع

الجانب التطبيقي

تمهيد (Preface)

تم في هذا الفصل استعمال بيانات حقيقية لمجموعتين من المصابين وغير المصابين بابيضاض الدم اللمفاوي وتطبيق الطريقة التي تبين افضليتها في الجانب التجريبي وهي طريقة التحليل التمييزي اللبي الحصين بعد ان تم اختبار البيانات

1-4 ابيضاض الدم اللمفاوي (Lymphocytic Leukemia)

ابيضاض الدم اللمفاوي هو حالة تحدث عندما يكون هناك زيادة في عدد خلايا الدم اللمفاوية في الجسم. الخلايا اللمفاوية هي جزء من جهاز المناعة وتلعب دورًا مهمًا في مكافحة العدوى والأمراض. ويتميز ابيضاض الدم اللمفاوي بزيادة عدد خلايا اللمفاوية في الدم بشكل غير طبيعي.

هناك عدة أسباب محتملة لابيضاض الدم اللمفاوي، ومنها:

1. الإلتهابات: يمكن أن تؤدي الالتهابات المزمنة إلى زيادة إنتاج الخلايا اللمفاوية.
2. الأمراض النقيضية: بعض الأمراض النقيضية تسبب زيادة في إنتاج الخلايا اللمفاوية، مثل مرض ليمفوما هودجكين واللوكيميا.
3. اضطرابات في نخاع العظم: بعض الاضطرابات في نخاع العظم يمكن أن تسبب زيادة في إنتاج الخلايا اللمفاوية.
4. استجابة مناعية مفرطة: بعض الاضطرابات المناعية يمكن أن تؤدي إلى ازدياد نشاط الخلايا اللمفاوية.
5. أمراض الدم الوراثية: بعض الأمراض الوراثية يمكن أن تسبب زيادة في إنتاج الخلايا اللمفاوية.

لتحديد السبب الدقيق وعلاج ابيضاض الدم اللمفاوي، يجب استشارة الطبيب وإجراء الفحوصات والاختبارات اللازمة. العلاج يعتمد على السبب والحالة الصحية الفي عامة للشخص، ويمكن أن يشمل العلاج الأدوية أو العلاج الإشعاعي أو زراعة نخاع العظم في بعض الحالات.

تشمل أعراض ابيضاض الدم اللمفاوي التعب، والحمى، والاورام اللمفية المتضخمة، وفقدان الوزن غير المبرر. يمكن تشخيص هذه الحالة من خلال فحص دم واستشارة الطبيب. يعتمد العلاج على السبب الرئيسي للحالة وقد يشمل علاج الالتهابات أو العلاج الكيميائي إذا كان سبباً للمشكلة هو وجود ورم لمفي.

هناك أنواع مختلفة من الأبيضاض اللمفاوي، منها:

1. **لمفوما هودجكين :(Hodgkin Lymphoma)** هذا النوع من الأبيضاض اللمفاوي يتميز بوجود خلية ريد ستيرنبرغ (Reed-Sternberg cell) في الأنسجة المصابة. يمكن علاجه بنجاح في معظم الحالات.
2. **لمفوما غير هودجكين :(Non-Hodgkin Lymphoma)** هذا النوع يتضمن العديد من الأشكال المختلفة من الأبيضاض اللمفاوي ويمكن أن يكون أكثر تنوعًا وصعوبة في العلاج.
3. **اللوكيميا :(Leukemia)** اللوكيميا هي نوع من السرطان الذي يؤثر على الدم والنقي العظمي. تحدث عندما تتكاثر خلايا الدم اللمفاوية بشكل غير طبيعي وتمتلئ النقي العظمي بالخلايا السرطانية.

2-4 عينة التطبيق: (Appled Sample)

تم الاعتماد على سجلات وحدة المختبر في مستشفى الحسين التعليمي في محافظة كربلاء المقدسة لغرض الحصول على المتغيرات التي لها علاقة بمرض إبيضاض الدم اللمفاوي (Lymphocytic

leukemia) والتي تضمنت (100) مشاهدة من الذكور والإناث وقد قسمت المشاهدات إلى مجموعتين وكالأتي:

1-المجموعة الأولى : شملت الأشخاص الغيرالمصابين بالمرض بحجم (50) مشاهدة ورمزنا لهم بالرمز(1).

2-المجموعة الثانية: شملت الأشخاص المصابين بالمرض بحجم (50) مشاهدة ورمزنا لهم بالرمز(2)

وكانت متغيرات التطبيق كالآتي:

Y : متغير مثل الاصابة ام عدم الاصابة بالمرض ،اذ ان الرمز (1) إذا كان الشخص غيرمصاب ، و (2) إذا كان الشخص مصاب بالمرض .

ان اغلب الدراسات الطبية التي تناولت مرض ابيضاض الدم اللمفاوي حددت عدة عوامل تؤثرعلى الإصابة بالمرض وهي كالأتي:

X_1 : جنس المصاب اذ يرمز (1) للذكور و (2) للإناث.

X_2 : خلايا الدم البيضاء WBC (White Blood Cells) ونســـــــبتها الطبيعية هي 4.00 (– 11.0).

X_3 : خلايا الدم الحمراء RBC (Red Blood Cells) ونسبتها الطبيعية هي (3.90 – 6.50)

X_4 : نسبة هيموجلوبين الدم HGB (Hemoglobin Blood) ونسبتها الطبيعية هي (11.5 – 17.5)

X_5 : نسبة الصفائح الدموية PLT (Blood Platelets) ونسبتها الطبيعية هي (150.0 - 450)

وان:

$\widehat{MR}$ نسبة خطأ التصنيف الكلي، اذ ان:

$\widehat{MR}$ 1 نسبة خطأ التصنيف للمجموعة الأولى

$\widehat{MR}$ 2 نسبة خطأ التصنيف للمجموعة الثانية.

والجدولين (4-1) و (4-2) يمثلان البيانات الحقيقية من المصابين بالمرض وغير المصابين اذ ضمت المجموعة الأولى (5) متغيرات والمجموعة الثانية (5) متغيرات.

جدول (4-1) البيانات الحقيقية للمجموعة الأولى التي تمثل المرضى غير مصابين

N	Y	X_1	X_2	X_3	X_4	X_5
1	1	1.00	4.77	4.44	11.81	445.00
2	1	1.00	6.55	5.56	16.55	361.00
3	1	2.00	5.56	6.11	15.65	232.00
4	1	1.00	5.22	4.89	14.45	167.00
5	1	2.00	6.44	4.67	11.90	168.00

6	1	1.00	6.78	4.89	13.40	182.00
7	1	1.00	9.44	5.00	15.67	199.00
8	1	2.00	6.89	4.12	14.23	233.00
9	1	1.00	8.44	6.01	11.12	432.00
10	1	2.00	5.22	3.71	10.45	156.00
11	1	2.00	10.33	4.78	13.50	254.00
12	1	1.00	9.78	5.55	16.30	355.00
13	1	2.00	8.56	4.39	14.45	311.00
14	1	1.00	10.33	4.56	15.34	212.00
15	1	1.00	6.45	4.44	13.13	432.00
16	1	1.00	8.45	5.33	13.44	166.00
17	1	1.00	7.67	5.01	13.22	355.00
18	1	1.00	6.66	5.15	12.22	189.00
19	1	1.00	9.54	6.05	12.99	179.35
20	1	1.00	8.55	4.26	12.56	118.00
21	1	1.00	1.11	6.78	3.11	185.00
22	1	1.00	7.44	5.93	15.90	222.00

23	1	2.00	6.44	4.51	12.70	381.00
24	1	1.00	11.99	4.22	15.80	178.50
25	1	1.00	1.67	4.80	12.40	295.00
26	1	2.00	6.54	3.98	12.13	356.50
27	1	1.00	9.78	3.99	12.00	322.00
28	1	2.00	7.66	4.81	12.10	231.00
29	1	1.00	10.51	5.21	13.80	342.00
30	1	1.00	7.56	4.57	14.40	123.00
31	1	2.00	10.21	4.78	12.80	33.00
32	1	2.00	8.11	4.56	12.90	273.00
33	1	1.00	9.13	4.56	11.99	438.00
34	1	2.00	7.32	4.44	1.10	294.00
35	1	2.00	8.55	5.78	15.60	411.00
36	1	2.00	9.44	5.56	16.34	162.00
37	1	2.00	7.34	4.66	14.01	313.00
38	1	1.00	10.12	5.22	13.55	113.00
39	1	2.00	9.11	7.00	12.51	326.00

40	1	2.00	4.44	5.94	12.60	414.00
41	1	1.00	10.11	5.25	15.10	401.00
42	1	2.00	9.89	1.32	7.31	49.00
43	1	2.00	5.15	2.55	14.55	374.00
44	1	2.00	2.99	4.56	13.78	355.00
45	1	2.00	4.99	4.89	12.676	426.00
46	1	2.00	9.12	4.88	12.78	294.00
47	1	2.00	9.11	4.67	14.61	498.00
48	1	1.00	9.56	4.44	16.78	421.00
49	1	2.00	10.55	4.45	15.04	212.00
50	1	2.00	8.11	4.66	14.42	152.00

جدول (2-4) البيانات الحقيقية للمجموعة الأولى التي تمثل المرضى مصابين

N	Y	X_1	X_2	X_3	X_4	X_5
1	2	2.00	1.21	4.23	10.55	109.00
2	2	1.00	7.70	4.76	12.34	133.00
3	2	1.00	0.99	4.51	9.33	167.00

4	2	2.00	7.60	5.17	13.66	182.00
5	2	1.00	5.60	5.27	12.10	296.00
6	2	1.00	2.52	3.32	14.20	79.00
7	2	1.00	10.20	5.83	17.20	201.00
8	2	2.00	2.70	4.01	12.80	108.00
9	2	2.00	3.60	5.26	14.55	32.00
10	2	2.00	2.00	4.04	11.34	204.00
11	2	2.00	3.00	4.01	9.67	343.00
12	2	2.00	0.58	4.73	10.44	118.00
13	2	1.00	1.66	1.44	3.22	64.80
14	2	1.00	8.30	3.17	12.22	83.00
15	2	2.00	4.90	3.87	10.32	212.00
16	2	2.00	3.60	4.60	12.10	261.00
17	2	2.00	3.48	4.13	9.11	124.00
18	2	1.00	2.73	2.14	7.33	42.00
19	2	2.00	2.30	5.25	13.33	127.22
20	2	1.00	2.60	5.05	12.78	384.00

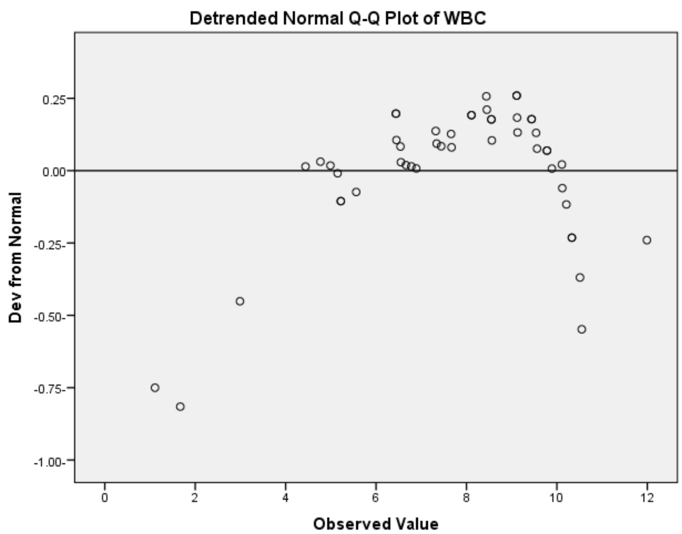
21	2	1.00	3.97	4.64	12.20	412.00
22	2	1.00	1.10	3.93	10.77	168.00
23	2	2.00	1.40	7.46	14.22	124.00
24	2	2.00	1.33	3.09	9.23	211.00
25	2	2.00	2.60	4.42	12.21	376.00
26	2	2.00	3.00	5.91	11.11	407.00
27	2	2.00	1.34	2.73	6.35	376.00
28	2	2.00	3.50	4.77	11.80	108.00
29	2	2.00	3.82	3.51	3.87	293.00
30	2	2.00	1.52	2.59	8.23	74.30
31	2	1.00	15.10	4.20	1.32	210.00
32	2	2.00	35.60	2.84	2.52	176.00
33	2	1.00	11.60	3.23	9.51	41.00
34	2	2.00	13.80	4.82	13.10	42.00
35	2	2.00	2.90	3.58	10.70	212.00
36	2	2.00	2.20	4.78	13.10	22.80
37	2	2.00	2.10	4.38	2.50	37.00

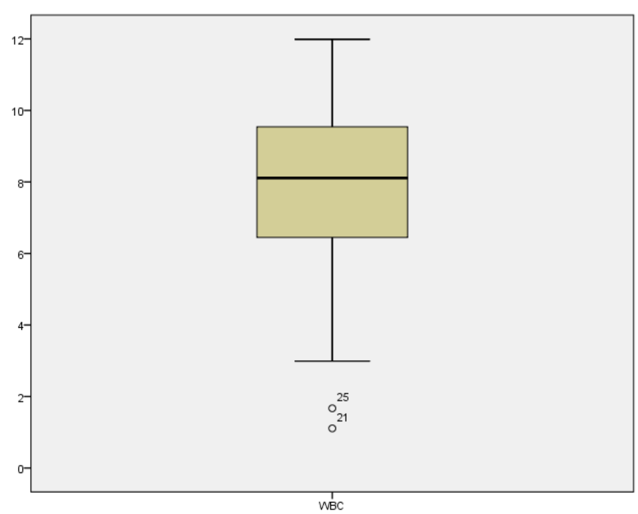
38	2	1.00	2.51	3.61	8.01	75.00
39	2	2.00	2.14	3.59	10.50	255.00
40	2	2.00	1.60	4.51	12.70	312.00
41	2	2.00	2.93	4.51	12.20	188.00
42	2	2.00	2.87	4.09	8.70	135.00
43	2	1.00	1.32	5.83	12.60	165.00
44	2	2.00	2.29	4.73	11.10	121.00
45	2	2.00	1.94	5.63	8.60	85.00
46	2	2.00	2.51	3.53	8.90	113.00
47	2	2.00	1.78	4.52	1.55	15.00
48	2	1.00	1.60	3.66	8.87	184.00
49	2	2.00	2.81	4.60	10.31	149.00
50	2	1.00	2.60	4.20	5.23	242.00

3-4 اختبار البيانات: (Data Test)

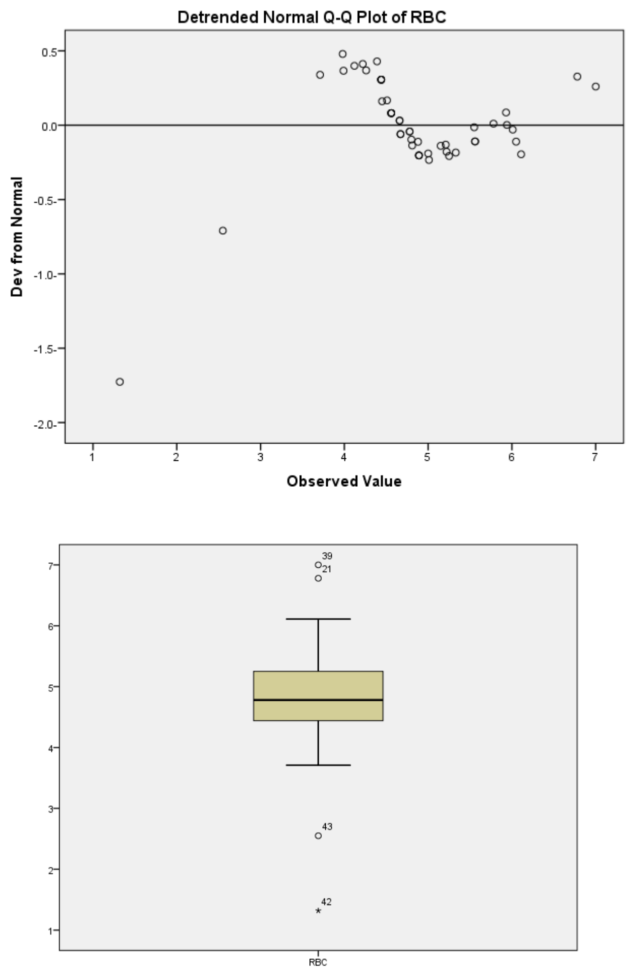
تم اختبار خطية البيانات باستعمال الرسم (Quantile - Quantile Plot (Q-Q-Plot للتاكد من انها غير خطية وكذلك رسم الشكل الانتشاري لاكتشاف وجود قيم شاذة من عدمها باستعمال البرنامج الاحصائي .SPSS Ver23 وكالآتي:

اولاً: بالنسبة لبيات الغير مصابين:

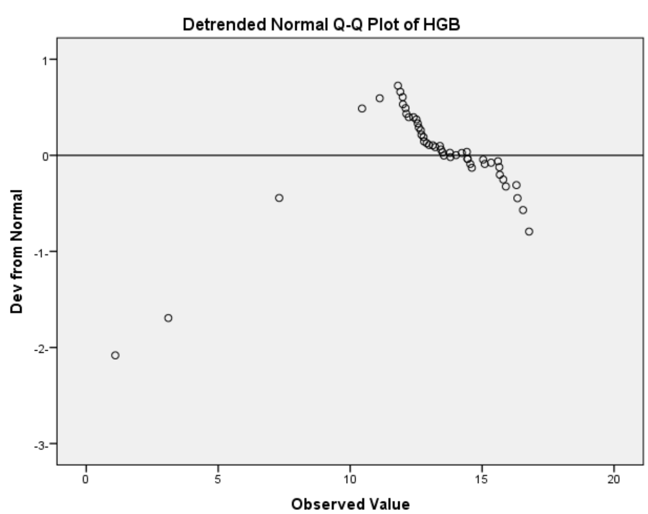


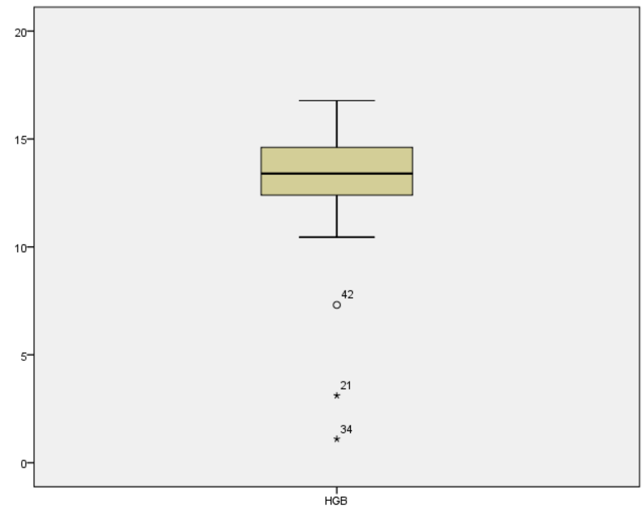


شكل (4-1) انتشار البيانات للمتغير WBC لمجموعة الغير مصابين

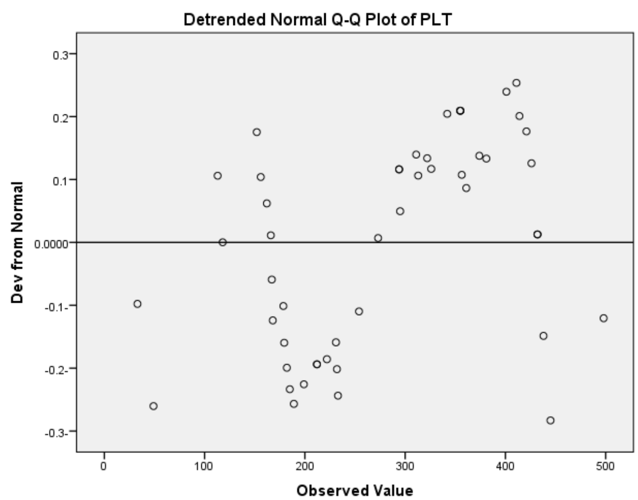


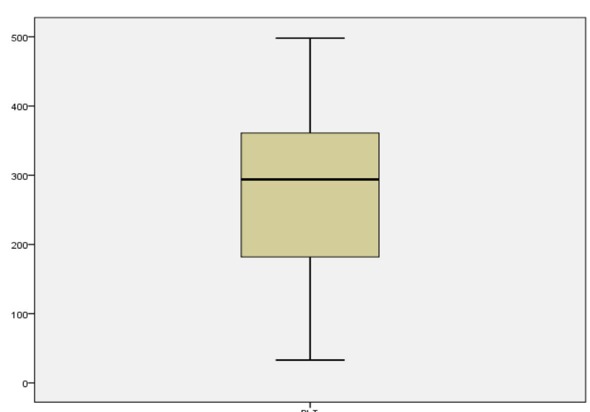
شكل (2-4) انتشار البيانات للمتغير RBC لمجموعة الغير مصابين





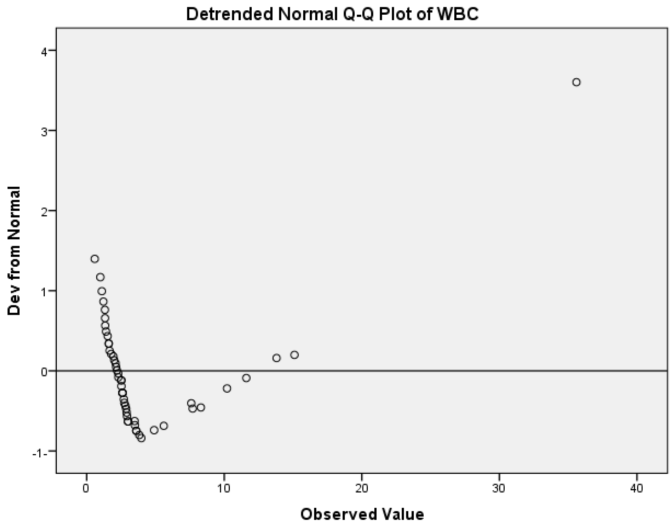
شكل رقم (4-3) انتشار البيانات للمتغير HGB لمجموعة الغير مصابين

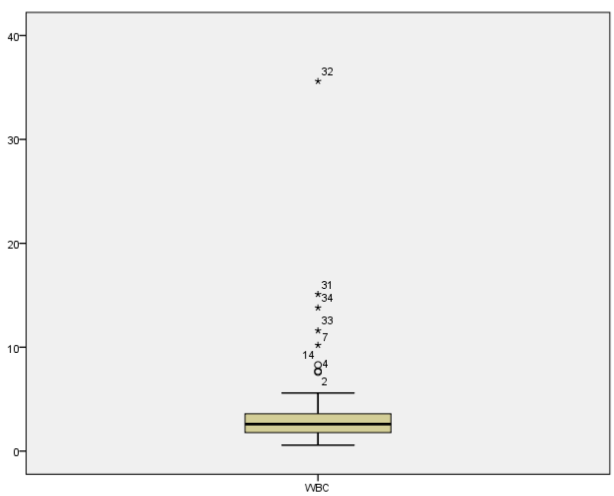




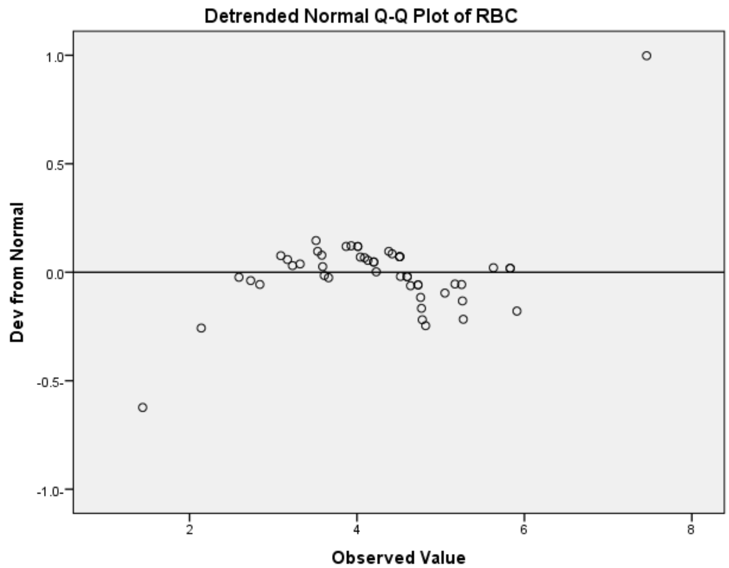
شكل رقم (4-4) انتشار البيانات للمتغير PLT لمجموعة الغير مصابين

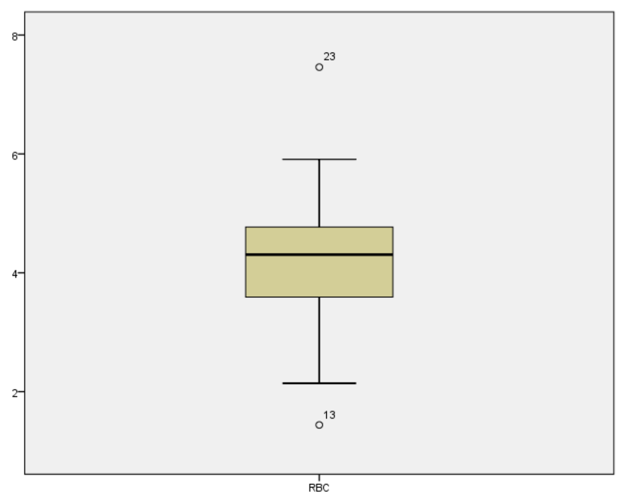
ثانياً: بالنسبة لبيات المصابين:



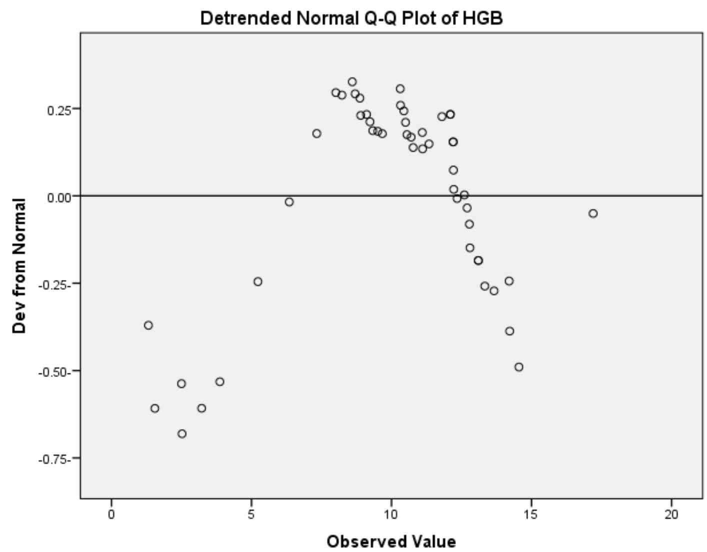


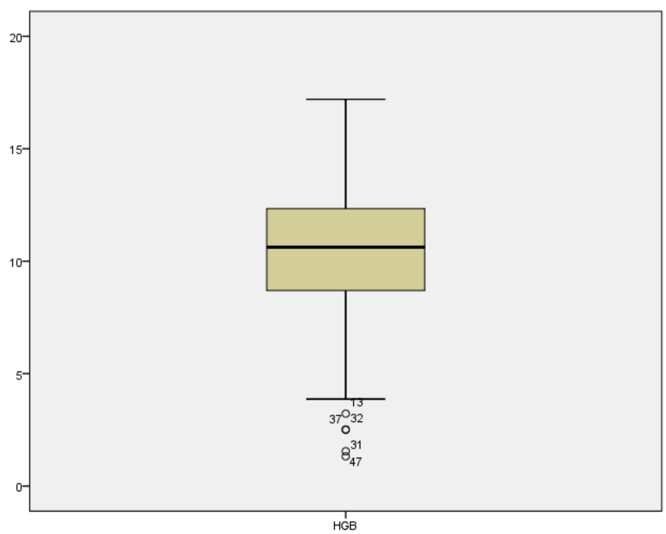
شكل (4-5) انتشار البيانات للمتغير WBC لمجموعة المصابين



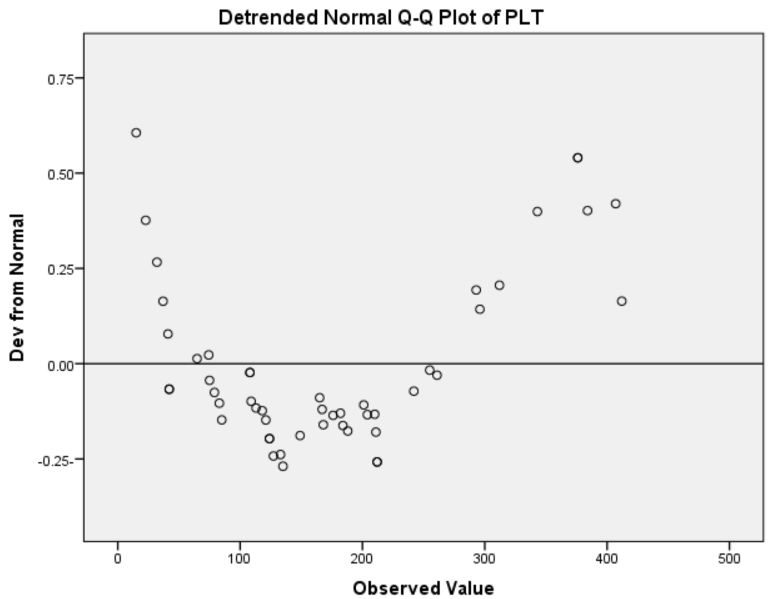


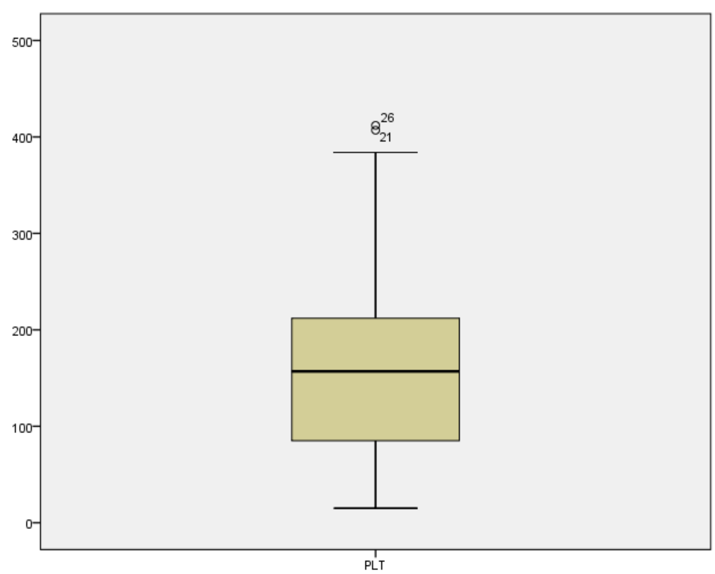
شكل (4-6) انتشار البيانات للمتغير RBC لمجموعة المصابين





شكل رقم (4-7) انتشار البيانات للمتغير HGB لمجموعة المصابين





شكل رقم (8-4) انتشار البيانات للمتغير PLT لمجموعة المصابين

نلاحظ من الاشكال (1-4) الى (8-4) ان مجموعة المصابين وغير المصابين تمثل بيانات لا خطية ولا تتبع التوزيع الطبيعي وكذلك تحتوي على قيم شاذة وخاصة مجموعة المصابين فهي تحتوي على قيم شاذة كثيرة وتبتعد عن التوزيع الطبيعي .

4-4 تحليل البيانات: (Data Analysis)

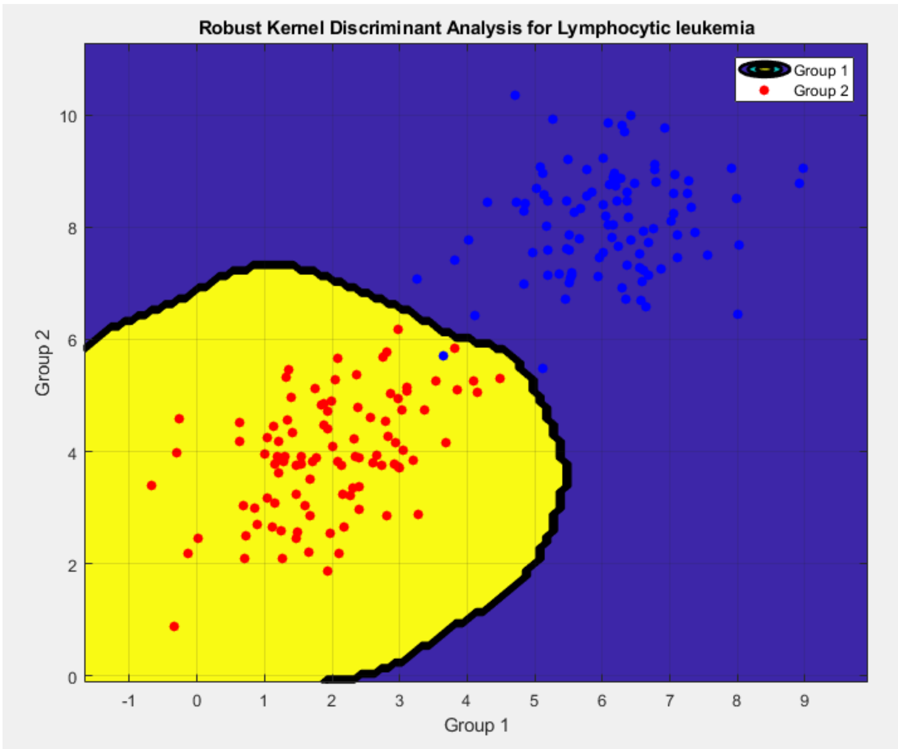
تبين في الجانب التجريبي افضلية اسلوب التحليل التمييزي اللبي الحصين (RKDA) على باقي اساليب التحليل التمييزي عند دوال الكثافة المنحرفة عن التوزيع الطبيعي وبما ان البيانات الحقيقية غير خطية ولا تتبع التوزيع الطبيعي وتحتوي قيم شاذة فيمكن تطبيق هذه الطريقة لغرض ايجاد نسبة خطأ التصنيف وتصنيف المشاهدات بدقة عالية وكما مبين في جدول (3-4) الآتي:

جدول (3-4) نسبة خطأ التصنيف ($\widehat{MR}$) للبيانات الحقيقية باستعمال إسلوب التحليل التمييزي اللبي الحصين

RKDA			$\widehat{MR}$	
Estimate / True	Estimate 1	Estimate 2	$\widehat{MR}$ 1	$\widehat{MR}$ 2
True 1	44	6	0.12	0.56
True 2	33	17	0.34	

يوضح الجدول رقم (3-4) معدل خطأ التصنيف ($\widehat{MR}$) والقيم التقديرية لعدد المشاهدات الصحيحة وغير الصحيحة لكل مجموعة فقد تبين ان اسلوب التحليل التمييزي اللبي الحصين اعطى نسبة خطأ

التصنيف للمجموعة الأولى $\widehat{MR}_1$ (0.12) وللمجموعة الثانية $\widehat{MR}_2$ (0.56) ، وبذلك تكون نسبة خطأ التصنيف الكلي ($\widehat{MR}$) بلغ (0.34).



شكل (4-9) التحليل التمييزي اللبي الحصين للبيانات الحقيقية

نلحظ من الشكل (4-9) ان هنالك تمييز تام بين مشاهدات المجموعتين (المصابين وغير المصابين) اذ ان المنطقة الصفراء المحددة باللون الاسود تمثل مجموعة الغير المصابين بمعزل عن مجموعة المصابين التي تمثلها المنطقة الزرقاء اي ان التحليل التمييزي قام بالتمييز بصورة دقيقة بين مشاهدات المجموعتين .

الفصل الخامس

الاستنتاجات

و

التوصيات

5-1 الإستنتاجات (Conclusions)

بالاعتماد على ما تم التوصل اليه من نتائج في الجانبين التجريبي والتطبيقي تم ادراج الاستنتاجات الآتـية:

1- اسلوب التحليل التمييزي الخطي هو الافضل من باقي اساليب التحليل التمييزي عند دوال الكثافة التي تتوزع طبيعياً (D, E)

2- حقق اسلوب التحليل التمييزي اللبي افضلية عند دول الكثافة (D, E) عند حجم العينة (n=1000, 5000) .

3- حقق اسلوب التحليل التمييزي اللبي افضلية على باقي الاساليب عند دالة الكثافة (K) بنسبة قليلة.

4- حقق اسلوب التحليل التمييزي اللبي الحصين افضلية على باقي الاساليب عند دوال الكثافة المنحرفة عن التوزيع الطبيعي بنسبة افضلية عالية.

5- ان اسلوب التحليل التمييزي اللبي الحصين اعطى نسبة خطأ التصنيف للمجموعة الأولى $\widehat{MR}_1$ (0.12) وللمجموعة الثانية $\widehat{MR}_2$ (0.56) ، وبذلك تكون نسبة خطأ التصنيف الكلي ($\widehat{MR}$) بلغ (0.34) وهي نسبة خطأ قليلة تدل على دقة التصنيف.

2-5 التوصيات (Recommendations)

من خلال ما تم التوصل اليه من استنتاجات ندرج التوصيات الآتية:

1- ضرورة استعمال اسلوب التحليل التمييزي اللبي الحصين في حالة كون البيانات تبتعد عن التوزيع الطبيعي او توجد قيم شاذة ضمنها.

2- استعمال التحليل التمييزي البيزي في حالة كون البيانات تحتوي على قيم شاذة

3- استعمال دوال كثافة غير الكاوسية مثل دالة ايبانكتشوف لتطبيق اسلوب التحليل التمييزي اللبي الحصين.

4- استعمال عرض الحزمة القطرية غير طريقة العبور الشرعي كان تكون طريقة متجه آلات الدعم (Support Vector Machine)

المصادر

- **المراجع**

القرآن الكريم

اولاً : المصادر العربية:

1. حميد ، آلاء عماد ، داود ، باسل خضر ذنون ، باسل يونس ، ، (2009)، " دراسة إحصائية للتبخر الحر في منطقة الموصل بطريقة المقدر اللبي" قاعدة بيانات الملخصات العلميه، الموصل، مجلة هندسة الرافدين. Vol.17 No.5
2. جاسم ، سكينة شامل ، (2020) .، دارسة مقارنة بين اسلوب التحميل التمييزي الخطي و اسلوب التحليل التمييزي اللبي" ، المجلة العراقية للعلوم الادارية . المجلد (14) ، العدد (55) .
3. بسيوني ، عبد الرحيم عوض عبد الخالق، (2021) ، " استخدام التحليل التمييزي في التصنيف والتنبؤ "، مجلة التجارة والتمويل ، المجلد 41، العدد 3 سبتمبر 2021 الصفحة 325-297
4. محمد ، لقاء علي ,عبود ، أمير علي، (2020) ، (مقارنة مقدرات عرض الحزمة(معلمة التمهيد)باستخدام الدوال اللبية في تحليل المركبات الرئيسية)، مجلة كلية التراث الجامعة العدد العشرون ، 412

ثانياً : المصادر الأجنبية:

5. Adolfo Hern?ndez and Santiago Velilla , (2016), " Dimension Reduction in Nonparametric Kernel Discriminant Analysis ", Journal of Computational and Graphical Statistics, Volume 14, Number 4, Pages 847-866 DOI: 10.1198/106186005X79610

6. Benyamin Ghojogh, Mark Crowley, (2019), " Linear and Quadratic Discriminant Analysis: Tutorial ", arXiv:1906.02590v1 [stat.ML] 1 Jun 2019.

7. Cai, L., Liu, Y., & Liu, H. (2019). Discriminant Analysis on Riemannian Manifold for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 57(7), 4562-4576. (Discusses Discriminant Analysis applied to hyperspectral image classification

8. Chen, S., Zhang, D., & Zhang, H. (2019). Discriminant analysis-based neighborhood repulsed metric learning for face recognition. Neurocomputing, 333, 472-482. (An application of Discriminant Analysis in face recognition)

9. David Anthony Mercer,(2013), Nonparametric Discriminant Analysis in Forensic Ancestry Estimation: An Assessment of Utilized and Alternative Statistical Methods", TRACE: Tennessee Research and Creative Exchange.

10. Duda, R. O., Hart, P. E., & Stork, D. G. (2012). Pattern classification. John Wiley & Sons. (Chapter 4 provides an introduction to Linear Discriminant Analysis)

11. Gupt , Abhishek M.; Soni , Himanshu H.; Joshi , Raunak M.; Laban, Ronald Melwin, (2022), " DISCRIMINANT ANALYSIS IN

CONTRASTING DIMENSIONS FOR POLYCYSTIC OVARY SYNDROME PROGNOSTICATION", A PREPRINT - JANUARY

12. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media. (Chapter 4 covers
13. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer
14. Li , Quanbao, Wei, Fajie, ; Zhou, Shenghan , (2017), " Local kernel nonparametric discriminant analysis for adaptive extraction of complex structures ", Open Phys. 2017; 15:270–279. DOI 10.1515/phys-2017-00303
15. Li-Pang Chen, (2022), " Classification and prediction for multi-cancer data with ultrahigh-dimensional gene expressions ", PLOS ONE.
16. Macdonald G. Obudho, George O. Orwa, Romanus O. Otieno, Festus A. Were, (2021), " Classification of Stateless People through a Robust Nonparametric Kernel Discriminant Function ", Open Journal of Statistics, 2022, 12, 563-580 https://www.scirp.org/journal/ojs ISSN Online: 2161-7198 ISSN Print: 2161-718X

17. McLachlan, G. J. (2004). Discriminant analysis and statistical pattern recognition. John Wiley & Sons. (A comprehensive book on Discriminant Analysis)

18. North, M. J., Macal, C. M., & Vos, J. R. (2019). Agent-Based Modeling and Simulation. The MIT Press.

19. Nudurupati , Sai Vamshidhar , (2009), " Robust Nonparametric Discriminant Analysis Procedures ", A Dissertation Submitted to the Graduate Faculty of Auburn University in Partial Ful_llment of the Requirements for the Degree of Doctor of Philosophy Auburn, Alabama.

20. Obudho , Macdonald G., Orwa , George O., Otieno, Romanus O., Were, Festus A. , (2022), "Robust Classification through a Nonparametric Kernel Discriminant Analysis", Open Journal of Statistics, 2022, 12, 443-455 https://www.scirp.org/journal/ojs ISSN Online: 2161-7198 ISSN Print: 2161-718X.

21. Obudho, M. , Orwa, G. , Otieno, R. and Were, F. (2022) Robust Classification through a Nonparametric Kernel Discriminant Analysis. Open Journal of Statistics, 12, 443-455. doi: 10.4236/ojs.2022.124028.

22. Pidd, M. (2018). Computer Simulation in Management Science. Wiley.

23. Robinson, S., & Gerasimov, V. (2019). Continuous System Simulation. Springer International Publishing.

24. Sapkal, A. U., & Kale, K. K. (2012). Study and comparison of linear discriminant analysis and principal component analysis for feature extraction. International Journal of Engineering Research and Applications, 2(4), 2484-2489.

25. Seung-Jean Kim Alessandro Magnani Stephen P. Boyd, (2023), (Robust Fisher Discriminant Analysis), Advances in Neural Information Processing Systems 18, p659-666.

26. Srivastava, M. S., & Kubokawa, T. (2008). Topics in multivariate approximation and covariance matrix analysis. World Scientific Publishing Co. Pte. Ltd. (Chapter 9 discusses Quadratic Discriminant Analysis)

27. Wand, M. P., & Marron, J. S. (1991). "Ridge Estimators for the Smoothing of Differenced Data." Journal of Time Series Analysis, 12(2), 97-112.

28. You, D. , Onur, C.H. and Aleix M.M. (2011) , " Kernel optimization in discriminant analysis ",Published in Xplore Digital Library , Volume: 33, Issue :3 .

29. Yu, Weichang,_, Azizi, Lamiae, Ormerod , John T., (2019), " Variational Nonparametric Discriminant Analysis", Preprint submitted to Computational Statistics and Data Analysis

30. Zafeiriou,S. , Tzimiropoulos,G. , Petrou,M. and Stathaki,T. (2012), " Regularized kernel discriminant analysis with a Robust kernel for Fase Recognition and Verification", Published in IEEE Transactions on neural networks and learning systems, Vol. 23. Issue . 3.PP. 526-534.

31. Zhang, D., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition, 40(7), 2038-2048. (Shows an application of Discriminant Analysis in multi-label learning)

32. Zhang,X. L. and Yang ,G. (2014) , " Distributed Face Recognition Using Multiple Kernel Discriminant Analysis in Wireless Sensor Networks", Published in International Journal of Distributed Sensor Networks,Vol . 2014, Article ID 242105, 7 pages.

33. Zhifeng Li, Xiaoou Tang, (2009), " Nonparametric Discriminant Analysis for Face Recognition ", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 31, NO. 4, APRIL 2009

Abstract

The majority of data in our real world deviates from the ideal assumptions required by traditional statistical methods, which causes a violation of the assumption of normality in the data, or there is data collected that represents non-linear data, and as a result we may face a problem in classification. Traditional discriminant analysis cannot confront this problem, so it must From searching for a robust method that deals with this problem, therefore, this thesis aimed to use the Robust Kenel Discriminant Analysis (RKDA) method in case the data deviate from its normal state and compare it with traditional Robust Kenel Discriminant Analysis and quadratic discriminant analysis using the classification error rate criterion. (MR)ˆ To choose the best classification method, through two aspects: the experimental aspect, and using Monte-Carlo simulation experiments. It was found that the linear discriminant analysis method is better than the rest of the discriminant analysis methods when the target density functions are normally distributed (D, E), and that the method Core discriminant analysis achieved an advantage in Gaussian density states (D, E) at sample size (n=1000, 5000). The core discriminant analysis method achieved an advantage over the rest of the methods when the density function (K) was achieved by a small percentage. The hippocampal core discriminant analysis method also achieved an advantage over other methods when density functions deviate from the normal distribution with a high percentage of preference.

In applied side, we depend on the reports of the laboratory unit at Al-Hussein Teaching Hospital in the Holy Governorate of Karbala for the purpose of obtaining variables related to lymphocytic leukemia, which included 100 observations from males and females. The observations were divided into two groups, the first It included people who did not have the disease with a size of (50) views, and the second included people with the disease with a size of (50) views. The application variables were Y: a variable such as having or not having the disease. The explanatory variables are X1: the sex of the infected person, X2: white blood cells (WBC). Blood Cells), X3: RBC (Red Blood Cells), X4: HGB (Hemoglobin Blood) percentage, and The classification for the first group is $\widehat{MR}1$ (0.12) and for the second group is $\widehat{MR}2$ (0.56). Thus, the overall classification error rate ($\widehat{MR}$) was (0.34), which is a small error rate that indicates the accuracy of the classification.

Republic of Iraq
Ministry of Higher Education
And Scientific Research
University of Karbala
Faculty of Management
And Economics
Department of Statistics
Graduate Studies

Robust classification Using Nonparametric Kernel Discriminant Analysis with an Application

A thesis

Submitted to the council of the college of Administration &Economics\ University of Karbala as partial fulfillment of the requirements for the Master degree in Statistics Sciences

By

Ja'afar Ali Farhan

Supervision

Asst. Prof. Dr. Enas Abdul Hafedh Mohammed

A.H. 1445 **A.D. 2024**