دراسة مقارنة بين اسلوب التحليل التمييزي الخطي و اسلوب التحليل التمييزي اللبي

(دراسة تطبيقية)

م.م سكينه شامل جاسم

جامعة واسط / كلية الاداره والاقتصاد / قسم الاحصاء

ملخص

في هذا البحث تم استعمال اسلوبين في تصنيف (تمييز) البيانات و هما اسلوب التحليل التمييزي الخطي (LDA) و اسلوب التحليل التمييزي اللبي (KDA) و هما من اكثر الاساليب الاحصائية استعمالاً في تحليل البيانات اذ تشترط هذه الاساليب توفر عدد من الفرضيات و من اهمها هي ان تكون المتغيرات التوضيحية تتوزع توزيعاً طبيعياً متعدد المتغيرات. و يهدف هذا البحث الى ايجاد الدالة التمييزية لكل منها و استعمالها كدالة تصنيف (تمييز) للاسلوبين بين المرضى ، اذ تم استعمال بيانات حقيقية لمجموعتين من المرضى المصابين و غير المصابين (بمرض اللوكيميا) و تم الحصول على النتائج باستعمال البرنامج الاحصائي (SPSS) بالنسبة لاسلوب التحليل التمييزي الخطي و برنامج (R – Package) بالنسبة لاسلوب التحليل التمييزي اللبي و تم التوصل 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).

المصطلحات : التحليل التمييزي الخطي، التحليل التمييزي اللبي، عرض الحزمة، خطأ التصنيف

Abstract

In this research, two methods were used to classify data and they are Linear Discriminant Analysis and method Kernel Discriminant Analysis, and they are the two most statistical methods used in data analysisas requiring these methods provide a number of hypothesesThe most important of these is that the explanatory variables are distributed in a multivariate normal distribution. This research aims to find the discriminatory function of each,And used as a function of classification (discrimination) of the two methods among patients,Where the use of real data for the two groups of patients infected and non-infected and The results were obtained using the Statistical Program (SPSS) for the linear discriminant analysis methodand (R – Package) for the kernel discriminant analysis method. andIt was found that the method of linear discriminant analysis is better because it gives less error classification of data. as it has been a comparison between the two modes in accordance with the standard classification error probability of (Misclassification).

Keywords: Linear discriminant analysis, Kernel discriminant analysis, Bandwidth , Classification error.

١- المقدمة

ان تصنيف مشاهدة (مفردة) جديدة لأحدى المجموعات قيد الدراسة يمكن ان تتم بالاعتماد على اسلوب التحليل التمييزي (Discriminant Analysis) و هو من الاساليب الاحصائية المهمة لمتعدد المتغيرات او يتم عن طريق استعمال مجموعة من المتغيرات بين مجموعتين او اكثر عن طريق استعمال دالة التمييز الخطي و دالة التمييز اللبي اذ يعد التحليل التمييزي من الاساليب المهمة جداً لدى الاحصائيين لاسيما في المجال الطبي الذي يتميز بوجود متغيرات كثيرة ويصعب على الباحث تحليلها و استخلاص النتائج منها بدقة، اذ يتم تحديد انتماء كل مشاهدة (مفردة) الى احدى المجموعات وفق اسلوب التحليل التمييزي عن طريق الحصول على نقطة الفصل (Cut – Point) التي تفصل بين المجموعات قيد الدراسة بناءاً على الاوزان او النسب التي يتم الحصول عليها عن طريق ايجاد دالة معينة من المتغيرات المستقلة التي يتم التنبؤ بها وفق الاسلوبين المستخدمين في هذا البحث و هما: اسلوب التحليل التمييزي الخطي و اسلوب التحليل التمييزي اللبي ، ويهدف هذا البحث الى التوصل الى افضل دالة تمييز بالاعتماد على معيار خطأ التصنيف (Misclassification) |، اي يتم الاعتماد على دالة التميز التي تصنف مفردة جديدة الى احدى المجموعات بأقل خطأ تصنيف ممكن.

٢. الجانب النظري_:

2-1اسلوب التحليل التمييزي العادي[6] [5] [3]

يعد التحليل التمييزي من احدى الإجراءات المهمة في التحليل الاحصائي لمتعدد المتغيرات والتي يهتم بكيفية التمييز بين مجموعتين او اكثر سواء أكانوا افراداً ام اشياء، بالاستناد على اسس معينة من المقاييس والتي تكون متشابهه في بعض الصفات ، اذ تنتمي كل مفردة (مشاهدة) الى المجموعة الصحيحة عن طريق استعمال الدالة التمييزية و تعرف بانها عبارة عن تركيب خطي للمتغيرات و يعتمد عليها لغرض التنبؤ و التصنيف للمشاهدة الجديدة لإحدى المجاميع والتي تزودنا بتقدير متكامل لكفاءة قواعد التصنيف و يجب ان تعطى اقل احتمال خطأ التصنيف لكي يتم الحصول على افضل النتائج و هناك عدة دوال تمييزية منها :

١- دالة التمييز الخطي Linear Discriminant Analysis

٢- دالة التمييز التربيعي Quadratic Discriminant Analysis

بالنسبة الى استعمال دالة التمييزي الخطي يشترط تتساوى التباينات للمجموعات القيد الدراسة . اما بالنسبة لدالة التمييزي التربيعي فتستعمل عندما تكون الشروط اللازمة غير متوفرة في الدالة التمييزي الخطي وفي هذه الحالة لا يتحقق شرط التساوي لمصفوفة التباين والتباين المشترك للمجموعات.

٢-٢ دالة التمييز الخطي [10] [5] [4] [2] Linear Discriminant function

قام باشتقاق دالة التمييزي الخطي (LDF) العالم (Fisher) وتستعمل عندما تكون العلاقة بين المتغيرات علاقة خطية ،فلو فرضنا دالة الكثافة الطبيعية (fj) مع متجه متوسطات مختلفة (μj) مع مصفوفة التباين مشتركة لتحليل التمييزي الخطي ،فأن :

$$fj \sim N(\mu j, \sum)$$

اذ تكون لدينا مجموعتين مجموعة الاولى (n_1) و لها (X) من المتغيرات بمتوسط ($\bar{x}_1$) ،ومصفوفة تباين المشتركة (S) ، والمجموعة الثانية (n_2) و(X) من المتغيرات بمتوسط ($\bar{x}_2$) ،ومصفوفة تباين المشترك (S) . ويمكن صياغة دالة التمييزي الخطي بالاعتماد على مؤشرات العينة المختاره .

التركيب الخطي بشكل عام يكون حسب الصيغة الاتية:

$$L = a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_kx_k \dots \quad (1)$$

اذ ان :-

L= يمثل التركيب الخطي بشكل عام

$a_1, a_2, \dots, a_k$ يمثل عدد معلمات الدالة التمييزية

بحيث تعطى اعلى تمييز اي يكون الاختلاف بين قيم المجموعتين (Between Group Variation) اكبر من اختلاف بين القيم داخل كل مجموعة (Within Group Variation) اي:-

$$G = \frac{Between\ groups}{Within\ groups} \dots (2)$$

G : يمثل قيمة الاختلاف بين قيم المجموعتين و القيم داخل المجموعتين.

ولتحديد هذا الاختلاف بين المجموعات يجب الحصول على الاوساط الحسابية لهذه المجاميع بالاعتماد على دالة (١) فيكون لدينا :

$$\bar{L} = a_1\bar{x}_1 + a_2\bar{x}_2 + \ldots + a_k\bar{x}_k \ldots (3)$$

$\bar{L}$: يمثل التركيب الخطي الناتج من الاوساط الحسابية للمجاميع.

اذ يمكن كتابة صيغة مصفوفة اختلافات القيم داخل المجموعات بالشكل الاتي :-

$$Y = \underline{\acute{a}}S\,\underline{a} \ldots (4)$$

S : تمثل مصفوفة التغاير داخل المجموعات

اما بالنسبة لمصفوفة الاختلافات بين قيم المجموعات فيكون لدينا :

$$W = \underline{\acute{a}}\lambda\,\underline{a} \ldots (5)$$

λ : تمثل مصفوفة التغاير بين المجموعات

وتصبح :- $$R = \frac{W}{Y} = \frac{\underline{\acute{a}}\lambda\,\underline{a}}{\underline{\acute{a}}S\,\underline{a}} \quad \ldots (6)$$

R : تمثل نتيجة تقسيم مصفوفة الاختلافات داخل المجموعات و مصفوفة الاختلافات بين المجموعات

اذ يجب تعظيم نسبة R لتقدير المعلمات دالة التمييز باشتقاقها و مساواتها للصفر فنحصل علي :-

$$\hat{a} = S^{-1}(\bar{X}_{ki} - \bar{X}_{pi}) \ldots (7)$$

اذ ان :-

S^{-1} : يمثل معكوس مصفوفة التباين والتباين المشتر للمتغيرات والمجموعات لكافة.

$\bar{X}_{ki}$: يمثل مصفوفة متوسط المتغير K للمجموعة i .

$\bar{X}_{pi}$: يمثل مصفوفة متوسط المتغير P للمجموعة i

٢-٣ نقطة الفصل Cut off – Point[1] [4]

و هي نقطة التي تفصل بين المجموعتين بهدف تصنيف مفردة معينة الى المجموعة الارجح لها و يعبر عنها بالصيغة الاتية :-

$$C.P = 1/2\ (\bar{X}_1 - \bar{X}_2)'S^{-1}(\bar{X}_1 - \bar{X}_2) \ldots (8)$$

فإذا كانت القيمة التمييزية لدالة المشاهدة اكبر من قيمة نقطة الفصل يتم تصنيفها الى المجموعة الاولى ، اما اذا كانت القيمة التمييزية لدالة المشاهدة اصغر (اقل) من نقطة الفصل فانها تتبع الى المجموعة الثانية.

4-2اسلوب التحليل التمييزي اللبي Style Kernel Discriminant Analysis[7] [9] [10] [11] [13] [15]

الممهد اللبي (Kernel Smoothing) هو احد اساليب الاحصائية و الاكثر استعمالاً في تحليل البيانات متعددة المتغيرات بفضل المرونه والسلوك الجيد لها، واللبية تعرف بانها دالة الاوزان الموحدة و من شروط استعمال اللبية يجب ان تكون حقيقية و متماثلة ومحددة و مستمرة و تكاملها يساوي واحد. ومن استعمالاتها المهمة هو في تقدير الكثافة اللبية (Kernel Density Estimation) في التحليل التمييزي ، ان هذه التقديرات تعتمد اساساً على معلمة عرض الحزمة (Bandwidth) و تسمى ايضاً بنافذة العرض او حجم النافذة او معلمة التمهيد وترمز لها برمز (H) و هو مهم جداً الى توجيه اللبية و التحكم بانتشارها و من شروطها ان تكون غير عشوائية ، غير سالبة الحدود و متماثلة .وعادة يستعمل لتقدير الطرائق (Cross – Validation) لتقليل مشاكل نسبة خطأ التصنيف باعتماد على اختيار مصفوفة عرض الحزمة او نقطة عرض الحزمة الامثل التي يقلل من نسبة خطأ التصنيف ($\widehat{MR}$) (Misclassification rate) . وتسمى بمصفوفة عرض عند استعمال عدة متغيرات في البحث اي متعدد المتغيرات ،اما بالنسبة لنقطة عرض الحزمة فيستعمل عندما يكون لدينا متغير احادي قيد البحث.ومن المهم جداً ان لا يتوقف الاختيار المناسب لمعلمات عرض الحزمة على كثافة المجتمع الاساسي و حجم العينة فقط بل ايضاً تعتمد علي الاحتمالات السابقة (π_i) .

ولتقدير اللبية توجد عدة طرائق لمصفوفات عرض الحزمة اذ تم استعمال مصفوفة عرض الحزمة طريقة العبور الشرعي الممهد (Smoothed Cross – Validation)القطرية في هذا البحث .

اي :

$$H = \text{dig}\ (h_1^2, h_2^2, \ldots, h_d^2) \ldots (9)$$

و استعمال الدالة اللبية الكاوسية و صيغتها كالاتي :

$$K(x) = (2\pi)^{\frac{-d}{2}} \exp(\frac{-1}{2} x^T x) \dots (10)$$

و الصيغة العامة لمقدر الكثافة لمتعدد المتغيرات مع مصفوفة عرض الحزمة القطرية هي:

$$\hat{f}(x;H) = n^{-1} \sum_{i=1}^{n} K_x (x - xi) \dots (11)$$

وتقدير الكثافة f(x) هي

$$f(x) = f(x_1 , x_2, \dots, x_k) \dots (١٢)$$

اذ ان :- $X = (x_1 , x_2, \dots , x_d)^T$

هي عينه عشوائية ذات (d) من الابعاد (المتغيرات) .

d : هو عدد المتغيرات المستعملة في البحث و هو d = 5.

n : حجم العينة .

٥-٢ طريقة التحليل التمييزي اللبي مع مصفوفة عرض الحزمة [14] [12] [11] [10] [9]**Scv**

تعرف اختيار المصفوفة (معلمة) عرض الحزمة لمتعدد المتغيرات في هذه الطريقة معدل من طريقة (Least – square – cross– validation) اذ تحصل على عدم تكرارية قيم عينات البيانات ، و صيغته كالاتي:-

$$Scv(H) = n^{-1}(4\pi)^{\frac{-d}{2}}|H|^{-d/2} + n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} (K_{2H+2G} - 2K_{H+2G} + K_{2G})(Xi - Xj) \dots (13)$$

اذ ان :

G : مصفوفة عرض الحزمة التجريبية.

وتكون خوارزمية هذه الطريقة كالاتي :-

<u>الخطوة الاولى</u>:احتساب تقدير الكثافة اللبية :-

$$Xj = \{ Xj_1 , Xj_2, \dots , Xj_{nj}\}$$

$$\hat{f}(x ;H_j) = n_j^{-1} \sum_{i=1}^{nj} K_{Hj} (x - H)$$

<u>الخطوة الثانية</u>:الحصول على احتمالات السابقة :

$$\hat{\pi}_j = nj/n \qquad j = 1,2$$

$\hat{\pi}_1 = \frac{n_1}{n}$, $\hat{\pi}_2 = \frac{n_2}{n}$

الخطوة الثالثة : نخصص نقاط البيانات $(x_1, x_2, \ldots, x_n)$ بالاعتماد على قاعدة التمييز اللبي (KDR)

KDR : x is allocated to group j_o if $j_o = \text{argmax} \hat{\pi}_j \hat{f}_j(x; H_j)$

$j \in \{1, \ldots, v\}$

الخطوة الرابعة : يكون تقدير نسبة خطأ التصنيف كالاتي :-

$\widehat{MR} = 1 - n^{-1} \sum_{j=1}^{n} 1\{ X_k:$ التصنيف الصحيح باستعمال KDR $\}$

٣- معيار المقارنة :[11] [10] [2] [1]

لغرض اجراء المقارنة بين الاسلوبين تم الاعتماد على معيار احتمال خطأ التصنيف (Misclassification)وكما موضح في ادناه :

١.٣: احتمال خطأ التصنيف (Misclassification) لاسلوب التحليل التمييزي الخطي :

يمكن التعرف على خطأ التصنيف بأنه احتمال تصنيف مفردة الاولى الى المجموعة الاولى بينما في الواقع تعود الى المجموعة الثانية ، وهو عامل مهم في اثبات كفاءة الدالة التمييزية التي تعطي اقل خطأ التصنيف ، ويمكن كتابة صياغته كالاتي :

$$P_1 = P_2 = f\left(\frac{-\sqrt{D^2}}{2}\right) \ldots (1٤)$$

اذ ان :

P_1 :يمثل احتمال تصنيف الخطأ للمشاهدة التي تعود الى المجموعة الاولى و صنفت في المجموعة الثانية.

P_2 : : يمثل احتمال تصنيف الخطأ للمشاهدة التي تعود الى المجموعة الثانية و صنفت في المجموعة الاولى.

D^2: و هو مربع المسافة بين مجموعتين وفق الصيغة الاتية:

$$D^2 = \frac{(n_1 + n_2)(n_1 + n_2 - 2)}{n_1 n_2 (n_1 + n_2 - x - 1)} . F \ldots (15)$$

اذ ان :

n_1: يمثل حجم العينة للمجموعة الاولى .

n_2: يمثل حجم العينة للمجموعة الثانية .

x : عدد المتغيرات المستقلة (المستعملة في البحث) .

F: قيمة Fالمحسوبة في تحليل التباين .

٢.٣ : احتمال خطأ التصنيف (Misclassification) لاسلوب التحليل التمييزي اللبي :

يمكن لنا ان نجد افضل مصفوفة (معلمة) عرض الحزمة لتقدير اللبية هي بطريقة ($\widehat{\mathrm{MR}}$) لكي تعطي افضل النتائج وهي اقل نسبة خطأ التصنيف و تكون صيغتها كالاتي :

$$\widehat{MR} = 1- n^{-1} \sum_{j=1}^{n} 1\{ X_k: \quad KDR\text{التصنيف الصحيح باستعمال} \} \dots (17)$$

٤- الجانب التطبيقي

٤-١ جمع البيانات

في هذا الجانب سيتم تطبيق اسلوبين: التحليل التمييزي الخطي و التحليل التمييزي اللبي على بيانات مرض اللوكيميا، وستكون نتائج التحليل والمقارنة بين الاسلوبين قيد الدراسة بالاعتماد على المتغيرات (مرض اللوكيميا) التي تؤثر في المرض و كما موضح ادناه:-

Y: هو المتغير المعتمد يبين وجود اصابة بمرض او عدم الاصابة في المجموعتين وترمز للمجموعة المصابة برمز (١) بحجم (٢٥) مفردة (مشاهدة) . والمجموعة الثانية غير المصابة برمز (٢) و بحجم (٢٥) مفردة ، اما المتغيرات المستقلة فهي :

Sex: يمثل الجنس اذ يرمز للذكور رمز (١) و للاناث برمز (٢).

Age: يمثل العمر.

Mch : يمثل متوسط الكمية الهيموجلوبين لكريات الدم.

Mcv: يمثل فحص حجم كريات الدم .

Esr: يمثل سرعة ترسب كريات الدم .

و سيتم استعمال البرنامجين الاحصائيين (SPSS – Version 20) مستعملة في الاسلوب التحليل التمييز الخطي ، واستعمال حزمة R(R – Version 3.03) في الاسلوب التحليل التمييز اللبي .اذ تم جمع البيانات استناداً على سجلات المرضى المصابين وغير المصابين في المختبر عن مرض اللوكيميا في مستشفى الزهراء في محافظة واسط وكما مبين في الجدول رقم (١).

جدول رقم (١)

يمثل بيانات الحقيقية للمجموعة الاولى المرضى المصابين و التي تحتوي على ٥ متغيرات

n	group	Sex	Age	Mch	Mcv	Esr
1	1	1	40	23.1	83.3	34
2	1	1	62	25	76.4	40
3	1	1	67	24.2	85.4	30
4	1	1	36	26.2	82	80
5	1	2	30	29.8	78.8	28
6	1	2	49	32.5	76.4	35
7	1	2	29	31.4	77.1	65
8	1	1	33	29.8	99.3	37
9	1	2	31	28.7	92.1	34
10	1	1	67	25.2	82.5	26
11	1	1	55	26	88.7	26
12	1	2	55	24.3	86.8	45
13	1	2	51	30.9	98.6	58
14	1	2	45	28.4	62.1	44
15	1	1	36	33.6	89.2	27
16	1	1	63	32.1	86.7	33
17	1	1	48	36.8	87.7	41
18	1	2	30	27.8	81.2	45
19	1	1	56	23.1	91.3	29
20	1	2	34	25.2	84.5	65
21	1	2	35	29.4	89.6	95
22	1	2	57	27.5	87.3	37
23	1	1	30	27.4	78.6	81
24	1	2	30	26.2	90.4	80

25	**1**	**1**	**54**	**25.9**	**90.1**	**27**

جدول رقم (٢)

يمثل بيانات الحقيقية للمجموعة الثانية المرضى غير المصابين و التي تحتوي على ٥ متغيرات

n	group	Sex	Age	Mch	Mcv	Esr
1	2	2	28	22.9	87.8	1
2	2	1	33	31.1	84	12
3	2	1	63	25.2	91.2	14
4	2	2	26	20.6	83.5	20
5	2	2	24	19.5	77.9	19
6	2	2	41	32.4	86.4	25
7	2	1	47	20.7	79.3	22
8	2	1	28	31.2	92.1	1
9	2	2	30	30.4	89.1	5
10	2	2	52	32.9	89.5	17
11	2	1	26	31.8	93.5	10
12	2	2	36	29.8	91.6	17
13	2	1	42	26.7	77	8
14	2	2	28	24.3	70.6	25
15	2	1	27	27.2	90.2	15
16	2	1	37	31.1	91.3	14
17	2	1	20	25.8	82.6	17
18	2	1	22	28.6	90.4	10
19	2	2	48	29.8	83.2	1
20	2	2	59	23.9	82.2	1
21	2	2	57	23.2	85.6	5
22	2	1	24	22.6	88.8	20
23	2	1	43	28.6	89.1	18
24	2	1	21	30.4	84.6	1
25	2	1	36	32.7	73.6	10

٢.٤اختبار تحقق شرط التوزيع الطبيعي

يتم اختبار البيانات لمعرفة فيما اذا كانت المتغيرات التوضيحية للمجموعتين تتوزع توزيعاً طبيعياً ام لا ، باستعمال اختبار (كولموكروف ـ سميرنوف) و بالاعتماد على البرنامج الاحصائي SPSS. و هواختبار إحصائي يُقارن توزيع المجتمع الإحصائي عن طريق عينتين مستقلتين مأخوذتين من هذا المجتمع. ويمكن استعماله لمقارنة أي توزيع

نظري(theoreticldistribution) مع التوزيع المشاهد(observed distribution)يسمى في بعض الأحيان حسن المطابقة (goodness of fit).و كما موضح بالجدول ادناه :

جدول رقم (٣)

يبين اختبار المتغيرات التوضيحية لتوزيع الطبيعي

ت	المتغيرvariable	Komogrov-Smirnov	Sig .*	النتيجة
١	Age	١.٠٤٩	٠.٢٢١	التوزيع طبيعي
٢	Mch	٠.٦٣٩	٠.٨١٠	التوزيع طبيعي
٣	Mcv	٠.٧٠٤	٠.٧٠٥	التوزيع طبيعي
٤	Esr	٠.٨٦٧	٠.٤٤٠	التوزيع طبيعي

نلاحظ من جدول رقم(٣) ان جميع المتغيرات كانت مستوى المعنوية المحسوبة لهما اكبر من (٠.٠٢٥) و ذلك لان الاختبار ذو طرفين،لذا فان البيانات تتوزع تبعاً للتوزيع الطبيعي .

٤-٣ تحليل النتائج باسلوب التحليل التمييز الخطي

تم استعمال اسلوب التحليل التمييزي الخطي و تكوين الدالة الخطية على ٥ متغيرات و ايجاد نقطة الفصل و التي كانت قيمتها (٠.٤١٥١٨) لتصنيف البيانات ضمن المجموعة الصحيحة اليها اوالى المجموعة الثانية ، اي حدوث خطأالتصنيف و كما موضح في جدول ادناه :

جدول رقم (٤)

يمثل تصنيف المشاهدات (المفردات) حسب الدالة التمييزي الخطي

الحالة	التصنيف			
	اصبح المريضعائد للمجموعة(١)	اصبح المريض عائد للمجموعة(٢)	نسبة التصنيف الصحيح	احتمال خطأ التصنيف
المريض يعود للمجموعة(١)	٢٣	٢	٠.٩٢	٠.٠٨
المريض يعود للمجموعة(٢)	١	٢٤	٠.٩٦	٠.٠٤
نسبة التصنيف الكلي	٢٤	٢٦	٠.٩٤	٠.٠٦

يتضح من الجدول (٤) ان المرضى العائدين في المجموعة الاولى والمصنفين بشكل صحيح هي (٢٣) مشاهدة (مفردة) من ضمن (٢٥) مشاهدة و نسبة التصنيف الصحيح لها (٠.٩٢) و هناك مشاهدتين صنفوا بشكل خاطئ و كانت نسبة خطأ التصنيف لها (٠.٠٨) اي يعودان الى المجموعة الثانية والمشاهدتين غير مصابين.اما بالنسبة للمجموعة الثانية فكانت (٢٤) مشاهدة صنفت بشكل صحيح من ضمن (٢٥) مشاهدة و كانت نسبة التصنيف الصحيح لها (٠.٩٦) ، ومشاهدة واحدة صنفت بشكل خاطئ و نسبة خطأ التصنيف لها كانت (٠.٠٤) و هو يعود الى المجموعة الاولى .

و بشكل عام يكون نسبة تصنيف الصحيح لهذه الطريقة هي (٠.٩٤) و احتمال خطأ التصنيف لها(٠.٠٦).

٥-٣ تحليل النتائج باسلوب التحليل التمييز اللبي

تم استعمال اسلوب التحليل التمييزي اللبي العبور الشرعي الممهد (SCV) على (٥) متغيرات و ايجاد معدل خطأ التصنيف للبيانات باستعمال معيار ($\widehat{MR}$) لمجموعتين و كانت نقطة التصنيف لكل مجموعة مقدارها : : Group (1) 0.1759129 , Group(2) : 0.1580436

وكما مبين في جدول ادناه :

جدول رقم (٥)

يمثل تصنيف المشاهدات (المفردات) حسب الدالة التمييزيةاللبية

الحالة	التصنيف			
	اصبح المريض عائد للمجموعة(١)	**اصبح المريض عائد للمجموعة(٢)**	**نسبة التصنيف الصحيح**	**احتمال خطأ التصنيف**
المريض يعود للمجموعة(١)	٢٣	٢	٠.٩٢	٠.٠٨
المريض يعود للمجموعة(٢)	٢	٢٣	٠.٩٢	٠.٠٨
نسبة التصنيف الكلي	٢٥	٢٥	٠.٩٢	٠.٠٨

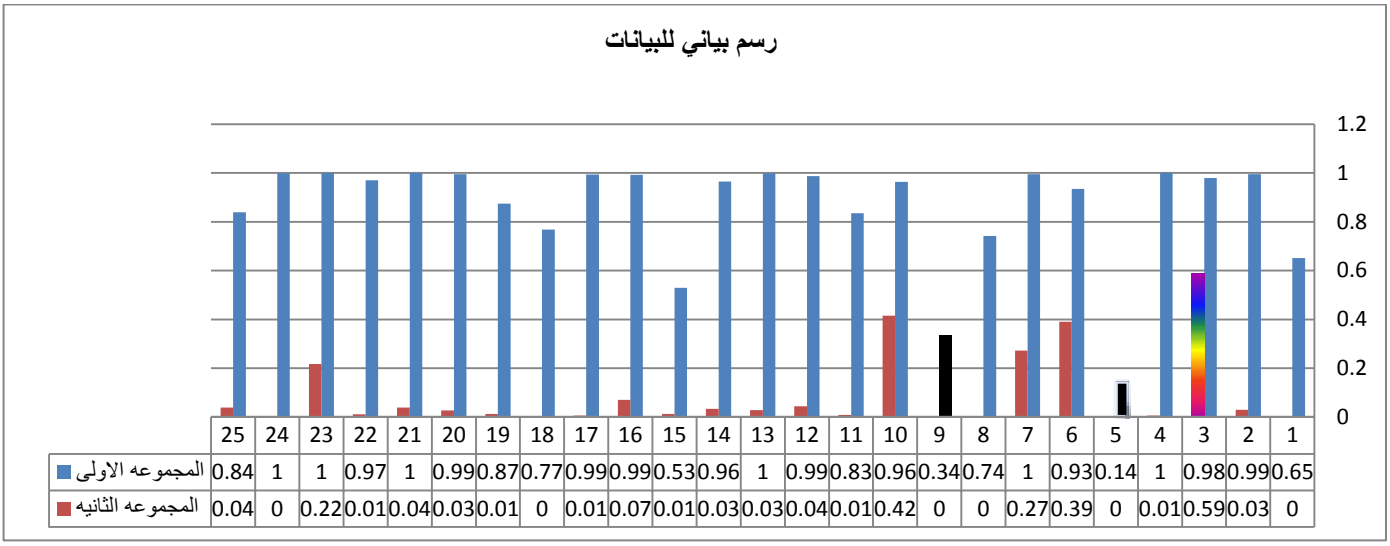
يوضح جدول (٥) ان المرضى العائدين في المجموعة الاولى والمصنفين بشكل صحيح هي (٢٣) مشاهدة (مفردة) من ضمن (٢٥) مشاهدة و نسبة التصنيف الصحيح لها (٠.٩٢) و هناك مشاهدتين صنفوا بشكل خاطئ و كانت نسبة خطأ التصنيف لها (٠.٠٨) اي يعودان الى المجموعة الثانية و المشاهدتين غير مصابين.اما بالنسبة للمجموعة الثانية فكانت ايضاً (٢٣) مشاهدة صنفت بشكل صحيح من ضمن (٢٥) مشاهدة و كانت نسبة التصنيف الصحيح لها (٠.٩٢) ، ومشاهدتين صنفوا بشكل خاطئ و نسبة خطأ التصنيف لها كانت (٠.٠٨) و هما يعودان الى المجموعة الاولى .

و بشكل عام يكون نسبة تصنيف الصحيح لهذه الطريقة هي (٠.٩٢) و احتمال خطأ التصنيف لها(٠.٠٨).

و في رسوم ادناه يوضح المشاهدات التي صنفوا بشكل صحيح و بشكل خاطئ في المجموعتين بطريقتي التحليل التمييز الخطي و التحليل التمييزي اللبي.

شكل رقم (١)

يمثل مشاهدات التي صنفوا بشكل صحيح والخاطئ في المجموعتين بطريقة التحليل التمييزي الخطي

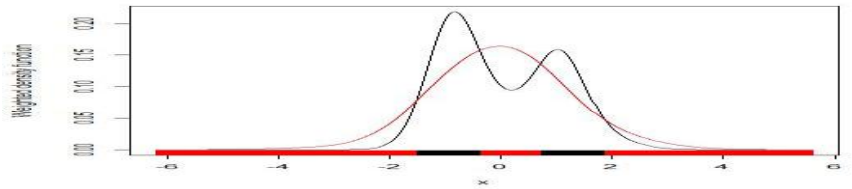


	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
المجموعه الاولى	0.84	1	1	0.97	1	0.99	0.87	0.77	0.99	0.99	0.53	0.96	1	0.99	0.83	0.96	0.34	0.74	1	0.93	0.14	1	0.98	0.99	0.65
المجموعه الثانيه	0.04	0	0.22	0.01	0.04	0.03	0.01	0	0.01	0.07	0.01	0.03	0.03	0.04	0.01	0.42	0	0	0.27	0.39	0	0.01	0.59	0.03	0

يوضح شكل (١) ان مشاهدة رقم (٥) و(٩) في المجموعة الاولى صنفوا بشكل خاطئ اي يجب ان يصنفوا في المجموعة الثانية ، و في المجموعة الثانية مشاهدة رقم (٣) صنفت بشكل خاطئ و يجب ان تصنف في المجموعة الاولى.

شكل رقم(٢)

يمثل تداخل بين مجموعتين بطريقة التحليل التمييزي اللبي



يوضح شكل (٢)مشاهدات التي صنفوا بشكل صحيح والخاطئ في المجموعتين

٥- الاستنتاجات والتوصيات

٥-١ الاستنتاجات :Conclusion

١- تبين عن طريق تصنيف البيانات باستعمال اسلوب التحليل التمييزي الخطي و اسلوب التحليل التمييزي اللبي تم التوصل الى ان طريقة التصنيف حسب دالة التمييز الخطي تعطى نتائج اقل لاحتمال خطأ التصنيف اذ بلغت نسبة التصنيف الصحيح الكلي لها (٩٤%) في حين نسبة التصنيف الخاطئ كانت (٦%) ، بينما كانت نسبة التصنيف الصحيح لدالة التمييزي اللبي الكلي لها (٩٢%) في حين نسبة التصنيف الخاطئ كانت (٨%) مما يدل على قوة الدالة التمييزي الخطي.

٢- يمكن الاعتماد على دالة التمييز الخطي بتصنيف (تمييز) المرضى المصابين وغير المصابين لكونها صنفت المرضى افضل من دالة التمييزي اللبي.

٥-٢ التوصيات :Recommendations

١- نوصي باستعمال دوال التمييز الاخرى (التربيعية، اللوجستي الخطي واللوجستي التربيعي) في تصنيف البيانات .

٢- نوصي باستعمال معالم عرض الحزمة اخرى مثل (Plug- in, Lscv) مع اسلوب التحليل التمييزي اللبي .

٣- نوصي بتشخيص المرضى المصابين واعطاءه العلاج اللازم في وقت مبكر.

المصادر

اولا: المصادر العربية

١. ابراهيم، علي عبد الحافظ، (٢٠٠٨)، " استخدام مقياس متعدد الابعاد وانموذج الدالة التمييزية لتحليل التركيب الكيميائية لنماذج الحليب المجفف"، مجلو جامعة النهرين، كلية العلوم، المجلد ١١(١)، ص: ٤٦-٥٧ .

٢. الوالي ، رنا محمود حسن، (٢٠١٣)،" تقدير الحاجة غير الملباة لتنظيم اداء اسرة في فلسطين"، رسالة ماجستير، كلية الاقتصاد والعلوم الادارية، قسم الاحصاء التطبيقي ، جامعة الازهر- غزة.

٣. جواد، عباس ناجي، (٢٠١٠)، " المفاضلة بين طرق تقدير الدوال الاقتصادية ذات المتغيرات التابعة النوعية"، مجلة تكريت للعلوم الادارية والاقتصادية ، جامعة تكريت، كلية الادارة و الاقتصاد، المجلد-٦/ العدد – ١٨.

٤. حامد، رائد عبد القادر، الفخري، نعمة عبدالله، عزيز، ذكاء يوسف، (٢٠١١)، " تعدين بيانات مشتركي خدمة الانترنيت باستخدام المنطق المضبب والدالة التمييزية"، المجلة العراقية للعلوم الاحصائية، كلية الادارة والاقتصاد، قسم نظم المعلومات الادارية، جامعة الموصل، العدد(١٩)، ص (١٩٧-٢١٨) .

٥. عبد الكريم، انوار ضياء، (٢٠٠٦)، " استخدام الطرائق التمييزية الاحصائية لتشخيص بعض امراض القلب "، مجلة جامعة كركوك، الدراسات العلمية، المجلد (١)، العدد (٢).

٦. مصطفى، خولة موسى، مصطفى، مي صادق، (٢٠١٠)، " التحليل التمييز واستعمالاته طريقة التصنيف"، المجلة العراقية للعلوم الاحصائية، كلية التربية، قسم الرياضيات، جامعة الموصل، العدد (١٨).

ثانياً: المصادر الاجنبية

7.Anil, K. G. and Bandyopadhy, S . (2006) , " Adaptive smoothing in kernel discriminant analysis" , Australian National University, Vol. 18, No.2, PP. 181- 197.

8.Anil,K.G. and Chaudhuri,P. (2004) , "Optimal smoothing in kernel discriminant analysis" , Indian Statistical Institute , Calcutta / statisticasinicaPP. 457-483.

9.Guidoum,A.C. (2013), "Kernel Estimator and Bandwidth selection for Density and its Derivatives",Faculty of Mathematics, University of Science and Technology HouariBoumediene. ,TheKedd Package /version 1.0.0.

10.Duong,T. (2004), "Bandwidth selectors for multivariate kernel density Estimation",This thesis is presented for the degree of Doctor of Philosophy at the University of Western.

11.Duong,T. (2007), " Ks : Kernel density estimation and kernsl discriminant analysis for multivariate data in R", Published by the American statistical Association in Journal of Statistical Software .Vol. 21, Issue. 7.

12.Duong,T. , Hazelton,M . (2004), "Bandwidth selectors for multiveiatekernel density Estimation" , University of Wetern Australia .

13.Ledl, T.(2004) , "Kernel density estimation : Theory and Application in discriminant analysis", university by of Vienna ,Published in Austrian Journal of statistics , vol. 33, NO. 3,PP.267 – 279.

14.Lu,J.Plataniotis, K.N., Venetsanopoulos,A.N.andWang,J. (2005) , "An efficient kernel discriminant analysis method",University of Toronto, Ontario, Canada ,Published by Elsevier B.V, vol. 38, Issue. 10 ,PP. 1788 – 1790.

15.Zhang,X., Maxwell L. K., Rob .J. H . (2004), "Bandwidth selection for multivariate kernel density estimation using CMC", MonashUniversity,Australia.