Comparison of linear function classification and nearest neighbor function
مقارنة تصنيف الدالة الخطية ودالة الجار الاقرب

أ.م. د. شروق عبد الرضا السباح مريم مهدي عناد الخزعلي

جامعة كربلاء / كلية الادارة و الاقتصاد
(بحث مستل من رسالة ماجستير)

المستخلص:-

تم تطبيق دالة التمييز لتشخيص ثلاث أنواع من الأورام السرطانية وهي سرطان الثدي وسرطان العظم وسرطان الرئة لكثرة شيوعها في مجتمعنا حالياً، ولغرض دراسة هذا الموضوع تم تسجيل قيم المشاهدات لخمس متغيرات وهي (الجنس، العمر، مهنة المريض، حالة خروج المريض، فترة بقاء المريض في المستشفى) من عينة عشوائية بسيطة بحجم (270 مريض) درست في نموذجين وهي النموذج المعلمي (دالة التمييز الخطية) والنموذج اللامعلمي (دالة الجار الاقرب)، واستخرجت النتائج باستعمال البرنامج الاحصائي **Stata** فضلاً عن الحزمة الاحصائية الجاهزة **SPSS** ، وتم استعمال معيار خطأ التصنيف كمعيار للمقارنة بين النماذج، بهدف الوصول إلى أفضل نموذج باقل خطأ تصنيفي ممكن، وتم الوصول الى ان النموذج اللامعلمي (الجار الاقرب) ذو تفوق على النموذج المعلمي (دالة التمييز الخطية) من حيث قلة نسبة التصنيف الخاطئ.

Abstract:-

The function of discrimination has been applied to diagnose three types of tumors, namely breast tumors, bone cancer and lung cancer, for their prevalence in our society. For the purpose of this study, the values of observations were recorded for five variables (sex, age, patient's profession, the sample random of size (270 patients) studied in two models: the parametric model (linear discrimination function) and the nonparametric model (the nearest neighbor function). The results were extracted using the statistical program **Stata** as well as the statistical package **SPSS**. With a view to reaching Li best model the lowest rank possible error, was reached that the nonparametric model (nearest neighbor) is superior to the parametric model (linear discrimination function) in terms of the lack of proportion of the wrong classification.

المقدمة :

يستعمل تحليل التمايز ***Discriminate Analysis*** عادة لتصنيف مفردة واحدة أو أكثر في مجموعات اعتماداً على متغيرات تحمل صفات تمييزية و تختلف من حيث المقاييس والصفات وتكون المفردة أو المشاهدة الواحدة تنتمي فقط إلى مجموعة واحدة .

قدمت دالة التمييز الخطية (***Linear Discriminate***) من قبل (***Fisher***) عام (**1936**)، الذي أقترح بأن التصنيف يجب أن يستند إلى تركيبة خطية للمتغيرات التمييزية من خلال تعظيم فروق المجموعات من جهة وتقليل التباينات داخل المجموعات من جهة أخرى، وهذه الطريقة مناسبة عندما يكون خطأ التصنيف فيها اقل ما يمكن.

اما طريقة الجار الأقرب ***KNN-Nearest Neighbor*** فهي من الطرق المبنية على الذاكرة، بمعنى انها لا تتطلب توفيق نموذجي للبيانات على خلاف الطرق الاحصائية الاخرى. وتستند على فكرة بديهية تتلخص في ان المشاهدات القريبة يجب ان تقع

في نفس الفئة. فهي اسلوب تصنيف يقرر في اي فئة سنضع المشاهدة الجديدة باختبار عدد وليكن (k) في معظم الحالات المشابهة ويلجأ الباحث لهذه الطريقة عند عمل مقارنة البيانات باستعمال اساليب اختزال البيانات.

فرضية البحث ***Search Hypothesis***

H_0: عدم وجود مشاكل في النموذج الاحصائي

H_1: وجود مشاكل في النموذج الاحصائي

هدف البحث ***Research Aim***

يهدف البحث إلى:

تصنيف البيانات باستعمال دالة التمييز الخطية والدالة اللامعلمية (دالة الجار الاقرب) بهدف الوصول إلى أفضل تصنيف لبعض أنواع الأورام السرطانية على أساس معيار احتمال اقل خطأ تصنيفي ممكن.

الجانب النظري

مفهوم تحليل التمايز ***Concept of Discriminant Analysis***

إن تحليل التمايز (*Discriminate Analysis*) هو أسلوب إحصائي لتحليل البيانات متعددة المتغيرات. ودالة التمييز عبارة عن توليفة خطية للمتغيرات التوضيحية (التفسيرية) تصنف مفردات العينة إلى مجموعتين أو أكثر، فهي التي تقوم بعملية التمييز وتليها عملية التصنيف (*Classification process*) التي نعتمد عليها في تصنيف المفردات الجديدة إلى إحدى المجموعات تحت الدراسة بأقل خطأ تصنيف ممكن **[6]**.

اولاً: دالة التمييز الخطية ***Linear Discriminate Function***

تعد هذه الطريقة من الطرق المعلمية ويستعمل هذا النوع من الدوال عندما يكون المجتمع تحت الدراسة ذو توزيع طبيعي متعدد المتغيرات وان تكون التباينات (مصفوفة التباين والتباين مشترك $\sum$) متساوية ، وبمتوسطات مختلفة $\underline{\mu}_1, \underline{\mu}_2$.

1- دالة التمييز الخطية في حالة مجموعتين[2]

Linear Discriminate Function (LDF) –Tow Groups

إن دالة التمييز هي نموذج يمكن صياغته اعتمادا على مؤشرات العينة التي تم اختيار مفرداتها ووضعت في مجموعتين مختلفتين ، وبواسطة هذه الدالة نستطيع أن نختبر المفردة ونحدد عائديتها إلى أي مجموعة.

نفرض أن مجال العينة R يكون مقسم إلى قسمين R_1 ينتمي إلى المجموعة الأولى و R_2 ينتمي إلى المجموعة الثانية والنقطة الفاصلة بين المجموعتين R_1, R_2 يمكن أن تنتمي إلى المجموعة الأولى أو الثانية وفي مثل هذه الحالة سيكون **[2]**:

$$f_1(\underline{X}) = f_2(\underline{X}) \qquad (1)$$

وبافتراض إن مصفوفة التباين والتباين المشترك متساوية $\sum$ وبمتوسطات مختلفة $\underline{\mu}_1, \underline{\mu}_2$.

$$(\underline{X} - \mu_1)' \Sigma^{-1}(\underline{X} - \mu_1) = (\underline{X} - \mu_2)' \Sigma^{-1}(\underline{X} - \mu_2) \qquad (2)$$

$$\underline{X}' \Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2) = 0 \qquad (3)$$

وهي دالة التمييز عند النقطة الفاصلة بين المجموعتين عندما تكون معالم المجتمع معروفة وفي حالة المعالم غير معروفة فتقدر بالاعتماد على قيم العينتين n_2 , n_1 وبأوساط حسابية $\overline{X}_1, \overline{X}_2$.

وبتقدير مصفوفات التباين والتباين المشترك تصبح الدالة كما يأتي **[5]**:

$$\underline{X}'S^{-1}\left(\overline{\underline{X}}_1 - \overline{\underline{X}}_2\right) - \frac{1}{2}(\overline{\underline{X}}_1 - \overline{\underline{X}}_2)' S^{-1}\left(\overline{\underline{X}}_1 + \overline{\underline{X}}_2\right) \qquad (4)$$

$$S = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} \qquad (5)$$

إذ أن:

$\overline{X}_1, \overline{X}_2$: تمثل تقدير الإمكان الأعظم لـ $\underline{\mu}_1, \underline{\mu}_2$.

S_1^2, S_2^2 : تمثل تقدير مصفوفة التباين والتباين المشترك للعينتين الأولى والثانية على التوالي.

S : يمثل تقدير الإمكان الأعظم لمصفوفة التباين والتباين المشترك $\sum$.

2- دالة التمييز الخطية في حالة أكثر من مجموعتين[6],[1]

Linear Discriminate Function (LDF)-More Than Two Groups

نفترض أن لدينا K من المجموعات، وكل مجموعة تحتوي على n من المشاهدات، وكل مشاهدة تتضمن P من المتغيرات.

إذ إن n_{i} هو حجم العينة المسحوبة من المجموعة i .

$$n = \sum_{i=1}^{k} n_{\mathrm{i}} \qquad (6)$$

نفترض أن T تمثل مجموع المربعات الكلية .

$$T = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(X_{ij} - \overline{\mathrm{X}}\right)\left(X_{ij} - \overline{\mathrm{X}}\right)' \qquad (7)$$

وكذلك نفرض إن W_{i} تمثل (مجموع المربعات للمجموعة i).

$$W_{\mathrm{i}} = \sum_{i=1}^{n_i}\left(X_{ij} - \overline{\mathrm{X}}\right)\left(X_{ij} - \overline{\mathrm{X}}\right)' \qquad (8)$$

وان مصفوفة التباين والتباين المشترك داخل المجموعات تساوي W:

$$W = W_1 + W_2 + \ldots\ldots + W_K \qquad (9)$$

وان مصفوفات التباين والتباين المشترك بين المجموعات هي:

$$B = T - W \qquad (10)$$

$$T = \begin{pmatrix} S_{11T} & S_{12T} & S_{1PT} \\ S_{21T} & S_{22T} & S_{2PT} \\ S_{P1T} & S_{P2T} & S_{PPT} \end{pmatrix}$$

$$B = \begin{pmatrix} S_{11B} & S_{12B} & S_{1PB} \\ S_{21B} & S_{22B} & S_{2PB} \\ S_{P1B} & S_{P2B} & S_{PPB} \end{pmatrix}, \qquad W = \begin{pmatrix} S_{11W} & S_{12W} & S_{1PW} \\ S_{21W} & S_{22W} & S_{2PW} \\ S_{P1W} & S_{P2W} & S_{PPW} \end{pmatrix}$$

وبهدف الحصول على مجموعة من التراكيب الخطية والتي هي:

$$Y = [Y_1, Y_2, \dots \dots Y_r] \qquad (11)$$

التي تعظم مقياس التمييز عن طريق تعظيم λ بالنسبة لكل b.

$$\lambda = \frac{between\ groups}{within\ groups}$$

$$\lambda = \frac{\underline{b}' B\ \underline{b}}{\underline{b}'\ W\ \underline{b}} \qquad (12)$$

ولجعل λ اعظم ما يمكن نأخذ الاشتقاق الجزئي بالنسبة لـ b .

$$\frac{\partial \lambda}{\partial \underline{b}} = \frac{2[\underline{b}'W\ \underline{b}(B\ \underline{b}) - \underline{b}'B\ \underline{b}(W\ \underline{b})]}{(\underline{b}'W\ \underline{b})^2}$$

$$\frac{\partial \lambda}{\partial \underline{b}} = 0$$

$$(\underline{b}'W\ \underline{b})B\ \underline{b} - (\underline{b}'B\ \underline{b})W\ \underline{b} = 0 \qquad (13)$$

بقسمة طرفي معادلة (12) على $\underline{b}'W\ \underline{b}$ وبالتعويض عن λ بما يساويها نحصل على :

$$B\ \underline{b} - \lambda W\ \underline{b} = 0$$

$$(B - \lambda W)\underline{b} = 0$$

$$(W^{-1}\ B - \lambda I)\underline{b} = 0 \qquad (14)$$

بعد إيجاد قيم λ ، اكبر قيمة إلى λ هي اكبر جذر مميز ***(Eigen Value)*** لمصفوفة $W^{-1}\ B$ والذي يقابل اكبر متجه مميز b_1 ***(Eigen Vector)*** .

$$b_1 = (b_{11}, b_{12}, \dots \dots, b_{1P}) \qquad (15)$$

b_1 تمثل مقياس التمييز للدالة الأولى Y_1 والتي تساوي :

$$Y_1 = b_{11}\mathrm{X}_1 + b_{12}\mathrm{X}_2 + \ldots\ldots + b_{1\mathrm{P}}\mathrm{X}_P \qquad (16)$$

وثاني اكبر جذر مميز لمصفوفة $\mathrm{W}^{-1}\,\mathrm{B}$ هو λ_2 والذي يقابل ثاني اكبر متجه مميز b_2 الذي يمثل مقياس التمييز للدالة الثانية والتي تساوي :

$$Y_2 = b_{21}\mathrm{X}_1 + b_{22}\mathrm{X}_2 + \ldots\ldots + b_{2\mathrm{P}}\mathrm{X}_P \qquad (17)$$

Y_1 من الضروري إن تكون غير مرتبطة مع Y_2 .

Y_3 تمتلك ثالث اكبر متجه مميز.

$$Y_3 = b_{31}\mathrm{X}_1 + b_{32}\mathrm{X}_2 + \ldots\ldots + b_{3\mathrm{P}}\mathrm{X}_P \qquad (18)$$

Y_3 غير مرتبطة مع Y_1 و Y_2 .

وهكذا نستمر إلى Y_r والتي تكون توضيحية عن $Y_1, Y_2, \ldots\ldots Y_{\mathrm{r}-1}$ ويطلق على الدوال $(Y_1, Y_2, \ldots\ldots Y_{\mathrm{r}})$ الدوال الخطية المميزة والتي يمكن نعبر عنها بشكل مصفوفة وكما يلي :

$$\underline{Y} = \underline{b}\mathrm{x} \qquad (19)$$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_r \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1P} \\ b_{21} & b_{22} & & b_{21} \\ & \vdots & \ddots & \vdots \\ b_{r1} & b_{r1} & \cdots & b_{rP} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_P \end{pmatrix}$$

تحدد r بعدد الدوال المميزة معتمداً على رتبة المصفوفة المركبة $\mathrm{W}^{-1}\,\mathrm{B}$.

اذ ان رتبة المصفوفة $\mathrm{P} = W_{\mathrm{PXP}}$
ورتبة W^{-1} = رتبة W
ورتبة مصفوفة B يكون اقل من $(P, K-1)$ ودائما يكون $(K-1)$ اقل من (P) ، وبهذا تكون رتبة المصفوفة $\mathrm{W}^{-1}\,\mathrm{B}$ تساوي :

$$rank\ (\mathrm{W}^{-1}\,\mathrm{B}) = \min(P, K-1) \qquad (20)$$

أي يكون عدد الدوال المميزة لـ K من المجموعات و P من المتغيرات هو:

$$discrimnat\ function = \min(P, K-1)$$

والتصنيف يكون عن طريق تعويض قيم المتغيرات الخاصة بأي مفردة يراد تصنيفها في جميع الدوال التمييزية ، ويتم تصنيف المفردة إلى الدالة المقابلة لأكبر مقدار .

ثانياً: دالة الجار الاقرب *Nearest Neighbor Function (KNN)*

تعد طريقة الجار الاقرب من الطرق اللامعلمية و إن الفكرة الأساسية لهذه الطريقة تكمن في تصنيف الحالات غير المصنفة أو (غير المرئية) إلى الحالات الأقرب لها ضمن حجم معين، وعليه نحتاج إلى تحديد قيمة k لتعيين المشاهدة التي تتلقى اكبر احتمال من بين ki الأقرب من الجيران**[3]**.

1- خوارزمية الجار الأقرب للتصنيف[7]:

لتقدير دالة الكثافة الاحتمالية للقيمة x نحتاج إلى تحديد منطقة صغيرة حولها حجمها v ، وعليه فان احتمالية وقوع x داخل المنطقة v هو :

$$\theta = \int_{v(x)} P(x)\, dx \qquad (21)$$

إذ أن التكامل يكون على الحجم v، ولحجم صغير فان:

$$\theta \sim P(x)v \qquad (22)$$

إن الاحتمالية يمكن أن تقترب من نسبة العينات التي تقع داخل المنطقة المحددة v حول x ، فإذا كانت k تمثل عدد العينات (الحالات) المأخوذة من n من العينات داخل المنطقة v ، فان:

$$\theta \sim \frac{k}{n} \qquad (23)$$

ومن الصيغتين (22) و (23) نحصل على تقدير دالة الكثافة لـ x كما يلي :

$$\hat{P}(x) = \frac{k}{nv} \qquad (24)$$

إذ إن :

v : حجم المنطقة حول المشاهدة x .
k : تمثل عدد الحالات داخل المنطقة v.
n : تمثل العدد الكلي للمشاهدات .

2- قاعدة قرار بيز[7]: *Bayesian Decision Rule*

إن قاعدة القرار المعتمدة على نظرية بيز في التعرف على الأنماط مستندة على أن الاحتمالات اللاحقة *Posterior Probability* تعتمد على الاحتمالات السابقة *Prior Probability* . أي أن التعرف على نمط (المشاهدة x) يعتمد على أعظم قيمة للاحتمالات اللاحقة $\hat{P}(w|x)$، وعلية يمكن تأشير النمط x إلى الفئة أو الصنف w_m إذا تحقق الشرط التالي :

$$\hat{P}(w_m|x) > \hat{P}(w_i|x) \quad , \forall\, i \in \mathrm{c} \qquad (25)$$

إذ إن c تمثل عدد الأصناف أو الفئات ، وباستعمال نظرية بيز :

$$\hat{P}(w_j|x) = \frac{P(x|w_j).P(w_j)}{P(x)} \qquad (26)$$

نحصل على قاعدة القرار الخاصة بالتعرف على الأنماط :

$$P(x|w_m).P(w_m) > P(x|w_j).P(w_j) \qquad (27)$$

إذ إن $P(x)$ حذف من طرفي المتباينة.

مما تقدم يمكن استخدام دالة الكثافة الاحتمالية في الصيغة (24) لتحديد قاعدة للقرار للتصنيف باستعمال دالة الجار الأقرب، فلو فرضنا أن أول أقرب مجموعة من الأنماط عددها k هناك w_m تقع في الصنف أو الفئة w_m بحيث أن :

$$\sum_{m=1}^{c} k_m = k \qquad (28)$$

ولو فرضنا أن العدد الكلي للعينات في الفئة w_m هو n_m بحيث أن :

$$\sum_{m=1}^{c} n_m = n \qquad (29)$$

وعليه يمكن تقدير دالة الكثافة الشرطية للفئة *Class-Conditional Density* كما يلي:

$$\hat{P}(x|w_m) = \frac{k_m}{n_m \,.\, v} \qquad (30)$$

أما الاحتمال الأولي *Prior Probability* والذي يمثل احتمال الفئة، فهو:

$$\hat{P}(w_m) = \frac{n_m}{n} \qquad (31)$$

وبتعويض كل من المعادلتين (30) و (31) في قاعدة قرار بيز (27) نحصل على الآتي :

$$\frac{k_m}{n_m \,.\, v}\,\frac{n_m}{n} > \frac{k_i}{n_i \,.\, v}\,\frac{n_i}{n} \quad ,\forall\, i \in \mathrm{c} \qquad (32)$$

وعلية سيتم تحديد النمط x إلى الفئة أو الصنف w_m إذا تحقق الشرط التالي :

$$k_m > k_i \quad ,\forall\, i \in \mathrm{c} \qquad (33)$$

هذا يعني انه يتم التعرف على نمط مشاهدة معينة بالاعتماد على عدد الجيران الأقرب ضمن فئة معينة، أي أن دالة التمييز *Discriminate Function* لخوارزمية الجار الأقرب تمثل نسبة عدد المشاهدات الخاصة بفئة معينة إلى عدد المشاهدات الكلية داخل المنطقة v ، أي إن :

$$h(x) = \frac{k_i}{k} \quad ,\forall\, i \in \mathrm{c} \qquad (34)$$

3- مقاييس التشابه والمسافة [3]: *Similarity and Distance Measures*

تستعمل مقاييس التشابه والمسافة لتحديد الجار الأقرب للحالة المدروسة غير المرئية ، فضلاً عن استعمالها في قياس التقارب والتماثل بين العناصر التي تمتلك أعلى قيمة تشابه داخل العنقود الواحد ، ففي حالة تكوين العناقيد فان أزواج القيم تكون متشابه (تمتلك أقل مسافة) ضمن العنقود الواحد ومختلفة (لها مسافات كبيرة) مع قيم في عناقيد أخرى .

إن استعمال فكرة مقاييس التشابه لا يكون سهلاً في عملية العنقدة ، لذا يستعمل في أغلب الأحيان بدلاً من مقاييس التشابه ، مقياس عدم التشابه *Dissimilarity* ، أو المسافة *Distance* ، ويرمز للمسافة بين x_i و x_j بالرمز $D(x_i, x_j)$.

وهنالك عدة طرائق لقياس المسافة بين عنصرين أو مشاهدتين في n من الخواص أو الصفات، أهمها وأكثرها شيوعًا المسافة الاقليدية *Euclidean Distance* والتي تعود الى العالم الرياضي الاسكندينافي اليوناني أقليديس ، وكما في الصيغة التالية :

$$D(x_i, x_j) = \sqrt{\sum_{t=1}^{n} (x_{it} - x_{jt})^2} \qquad (35)$$

أما شروط المسافة فهي :

1. المسافة بين النقطة واخرى يساوي صفر، أي أن: $D(x_i, x_j) = 0$

2. المسافة بين النقطتين x_i و x_j هي نفسها المسافة بين النقطتين x_j و x_i اي ان :

$$D(x_i, x_j) = D(x_j, x_i)$$

3. المسافة الأقصر بين أي نقطتين تمثل بخط مستقيم، أي أن:

$$D(x_i, x_j) \leq D(x_i, x_z) + (x_z, x_j)$$

ثالثاً: عملية التصنيف *Classification*

إن عملية التصنيف (***Classification***) هي العملية اللاحقة بعد تكوين دالة التمييز حيث يتم الاعتماد على هذه الدالة بالتنبؤ وتصنيف المفردة الجديدة لإحدى المجموعات قيد الدراسة بأقل خطأ تصنيف ممكن ، ويشترط تساوي التباينات للمجموعات قيد البحث ، وهناك تمييز خطي في حالة مجموعتين ، وتمييز خطي في حالة أكثر من مجموعتين ، أما التمييز غير الخطي فيستخدم في حالة عدم تساوي التباينات **[1]**.

احتمال خطأ التصنيف

إن خطأ التصنيف هو عامل مهم لإثبات كفاءة الدالة التمييزية، أي أن الدالة التمييزية التي تعطي أقل خطأ تصنيف هي الدالة الأكثر كفاءة وتكون الأفضل من بين دوال التمييز.

وهناك نوعين من احتمال خطأ التصنيف[4]:

أولاً: احتمال خطأ التصنيف P_{ij} وهو احتمال تصنيف المشاهدة الى المجموعة **j** وهي تعود إلى المجموعة ***i*** .

$$P_{ij} = \Phi\left(-\frac{\delta}{2}\right) \qquad (36)$$

ثانياً: احتمال خطأ التصنيف P_{ji} وهو احتمال تصنيف المشاهدة إلى المجموعة ***i*** وهي تعود إلى المجموعة ***j*** .

$$P_{ji} = \Phi\left(-\frac{\delta}{2}\right) \qquad (37)$$

إذ أن:

Φ : تمثل دالة التوزيع الطبيعي القياسي.
δ : تمثل جذر مقياس مسافة مهانلوبس D^2.

إذ إن :

$$\delta^2 = D^2 = (\overline{X}_1 - \overline{X}_2)' S^{-1} (\overline{X}_1 - \overline{X}_2)$$

وسيكون احتمال خطأ التصنيف كالتالي:

$$P_{ij} = P_{ji} = \Phi\left(-\frac{D}{2}\right) \qquad (38)$$

إذ إن D هو جذر مقياس مهانلوبس.

الجانب التطبيقي

تم جمع البيانات من خلال اخذ عينة عشوائية بسيطة حجمها **270** من طبلات المرضى الراقدين في مستشفى (مدينة مرجان الطبية) في محافظة بابل لإمراض السرطان للتميز بين اشخاص مصابين بأورام الثدي والعمود الفقري (سرطان العظم) والجهاز التنفسي (سرطان الرئة) وذلك لكثرة المصابين بهذه الانواع خلال عام 2016، ولغرض تحليل البيانات تم استعمال البرنامج الاحصائي Stata وكذلك الحزمة الاحصائية SPSS، وتم دراسة المتغيرات التالية لكل مجموعة:(الجنس، العمر، مهنة المريض، تشخيص المرض، حالة خروج المريض، فترة بقاء المريض في المستشفى).

تعريف المتغيرات *Definition of variables*

اعتمدنا في تكوين دالة التمييز على عدد من المتغيرات تم جمعها عن كل مشاهدة (مريض) من مشاهدات العينة ، وحيث إن جزء من المتغيرات هي متغيرات متقطعة وأخرى متصلة .

أ- المتغير المتعمد (Y) والذي يمثل نوع أو تشخيص المرض:

سرطان الثدي =1 سرطان العظم =2 سرطان الرئة =3

ب- المتغيرات التوضيحية تمثل ما يلي:

1- $\mathbf{X_1}$ = الجنس: شملت العينة كلا الجنسين وكانت نسبتهم إلى مجموع العينة متفاوتة ويمثل متغير ثنائي (ذكر=1 ، أنثى=2)

2- $\mathbf{X_2}$ = العمر: تتراوح أعمار المرضى المشمولين ضمن العينة بين (4 – 95) سنة.

3- $\mathbf{X_3}$ = مهنة المريض: أعطينا رموز لكل مهنة وذلك لتسهيل تحليل البيانات

(طفل=1، ربة بيت أو كاسب=2 ، طالب =3 ، عاجز=4 ، متقاعد=5 ، موظف =6)

4- $\mathbf{X_4}$ =حالة خروج المريض :(وفاة=1 ، إحالة=2 ، خرج على مسؤوليته=3 ، تحسين=4)

5- $\mathbf{X_5}$ = فترة بقاء المريض في المستشفى.

اختبار قدرة الدالة التمييزية على التمييز

1- مقياس ويلكس *Wilks' Lambda*

عندما يراد التمييز بين ثلاث مجاميع فأكثر وتكوين دوال تمييزية مقبولة إحصائياً بمستوى معنوية فانه لابد من اختبار معنوية الفروق بين متوسطات المجاميع قيد الدراسة وذلك بالاعتماد على الفرضية التالية:

$\mathbf{H_0}: \mu_1 = \mu_2 = \mu_3$
$\mathbf{H_1}: \mu_1 \neq \mu_2 \neq \mu_3$

هناك عدد من المقاييس لاختبار الفرضية أعلاه منها مقياس **(*Wilks' Lambda*)** وتكون صيغته كما في المعادلة (39) وهي تتوزع تقريبا $\boldsymbol{\chi^2}$ بدرجة حرية $p(k-1)$ ومستوى معنوية (α) وتم اختبار معنوية الفروق بين المتوسطات المجاميع الثلاث لأورام السرطان بموجب هذه الصيغة وكانت نتائج الاختبار كما مبين في **الجدول (1)**:

$$\Lambda = \frac{|W|}{|\mathrm{T}|} = \frac{|W|}{|\mathrm{W}+\mathrm{B}|} \qquad (39)$$

الجدول (1) اختبار معنوية الدالة التمييزية

Test of Function(s)	Wilks' Lambda(Λ)	Chi-square($\boldsymbol{\chi^2}$)	df	Sig.
1 through 2	**.179**	**458.572**	**4**	**<.0001**
2	**.729**	**84.243**	**1**	**<.0001**

أظهرت نتائج الاختبار من خلال **الجدول (1)** وجود فروق معنوية بين المتوسطات الثلاث إذ إن قيمة P-value لـ($\boldsymbol{\chi^2}$) اقل من (0.05) أي نرفض فرضية العدم ونرجح الفرضية البديلة القائلة عدم تساوي المتوسطات الثلاثة وذلك دليل على إن هناك فروق معنوية بين متوسطات المجاميع نستنتج من ذلك إن الدوال التمييزية لها القدرة على التمييز أي يمكن الاعتماد عليها لتصنيف أي مفردة إلى إحدى المجموعات الثلاثة.

2- اختبار F

بالنسبة لإحصائية اختبار F فنستخدم الصيغة التالية:

$$F = \frac{1-\Lambda^{1/5}}{\Lambda^{1/5}} * \frac{ms-2\lambda}{\mathrm{p(k-1)}} \qquad (40)$$

$$m = N - \frac{1}{2}(\mathrm{P}+\mathrm{K}) = 270 - \frac{1}{2}(5+3) = 266$$

$$s = \left[\frac{P^2(1-K)^2-4}{(1-K)^2+P^2-5}\right]^{1/2} = \left[\frac{5^2(1-3)^2-4}{(1-3)^2+5^2-5}\right]^{1/2} = 2$$

$$\lambda = \frac{\mathrm{p(k-1)}-2}{4} = \frac{5(3-1)-2}{4} = 2$$

$$F = \frac{1 - 0.179^{1/5}}{0.179^{1/5}} * \frac{2(266) - 2(2)}{5(3-1)} = 21.68$$

علماً ان Λ يمثل مقياس **Wilks' Lambda** والذي يساوي 0.179 من **الجدول(1)**

بدرجات حرية

$$df_1 = \mathrm{p}(\mathrm{k}-1) = 5(3-1) = 10$$
$$df_2 = ms - 2\lambda = 536$$

إذ إن قيمة F الجدولية بدرجات حرية (10,536) ومستوى معنوية 0.05 تساوي:

$$F_{(0.05,10,536)} = 1.843$$

ظهرت قيمة F المحسوبة اكبر من قيمة F الجدولية وهذا يدل على إن الدالة التمييزية الخطية لها القابلية على التمييز بالاعتماد على قيم المتغيرات التوضيحية .

3- القيم الذاتية

الجدول (2) يبين القيم الذاتية والارتباط القانوني

Eigen values				
Function	**Eigen value**	**% of Variance**	**Cumulative %**	**Canonical Correlation**
الدالة الاولى	**3.074[a]**	**89.2**	**89.2**	**.869**
الدالة الثانية	**.372[a]**	**10.8**	**100.0**	**.521**

الجدول (2) يبين قيمة ***Eigen values*** للدالة التمييزية كانت (3.074) مما يشير الى ان للدالة التمييزية مقدرة عالية على التمييز حيث ان قيمة ***Eigen values*** اكبر من الواحد الصحيح . وما يؤكد ان 100% من التباين كان مفسراً . وتحسب ***Eigen values*** بقسمة مجموع مربعات التباينات بين المجموعات (SSB) على جذر مجموع مربعات التباينات داخل المجموعات (SSW) . اما في ما يتعلق بالارتباط القانوني ***Canonical Correlation*** فقد بلغ (0.869) ويدل ذلك على جودة توفيق الدالة التمييزية.

اما مصفوفة الارتباطات للمتغيرات المدروسة مبينة في **الجدول (3)**:

الجدول (3) يبين مصفوفة الارتباطات

Pooled Within-Groups Matrices[a]					
	(X_1) الجنس	**(X_2) العمر**	**(X_3) مهنة المريض**	**(X4) حالة خروج المريض**	**(X_5) فترة بقاء المريض**
(X_1) الجنس	**1.000**	**-.148-**	**-.280-**	**.011**	**-.039-**
(X_2) العمر	**-.148-**	**1.000**	**.104**	**-.139-**	**-.025-**
(X_3) مهنة المريض	**-.280-**	**.104**	**1.000**	**-.083-**	**-.026-**
(X_4) حالة خروج المريض	**.011**	**-.139-**	**-.083-**	**1.000**	**.015**
(X_5) فترة بقاء المريض	**-.039-**	**-.025-**	**-.026-**	**.015**	**1.000**

من **الجدول** اعلاه نلاحظ معاملات الارتباط الثنائي بين المتغيرات التوضيحية وقد كان معامل الارتباط بين متغير الجنس(X_1) ومتغير مهنة المريض(X_3) اعلى ارتباط اذ بلغت قيمته (0.280)، ويليه معامل الارتباط بين متغير الجنس(X_1) ومتغير العمر(X_2) اذ بلغت قيمته (0.148)، ويليه معامل الارتباط بين متغير العمر(X_2) ومتغير حالة خروج المريض(X_4) اذ بلغت قيمته (0.139) وهكذا...، وتعني الإشارة السالبة وجود علاقة عكسية بين المتغيرين.

اما المتغيرات الداخلة في التحليل فهي موضحة بالجدول التالي:

الجدول (4) يبين المتغيرات الداخلة في التحليل

Variables in the Analysis				
Step		**Tolerance**	**F to Remove**	**Wilks' Lambda**
1	**العمر (X_2)**	**1.000**	**398.876**	
2	**العمر(X_2)**	**.978**	**407.790**	**.728**
	الجنس(X_1)	**.978**	**53.386**	**.251**

يبين **الجدول (4)** المتغيرات التوضيحية التي بقيت في التحليل وذلك لكل خطوة من خطوات تحليل التمايز التدريجي، حيث ان اول متغير دخل التحليل في الخطوة الاولى هو متغير العمر **(X_2)** حيث كانت قيمة F عاليا جدا اذ بلغت (398.876)، وفي الخطوة الثانية تم ادخال متغيري **(X_1)** الجنس و **(X_2)** العمر وذلك لكون قيم F المحسوبة لهما والبالغة (407.79، 53.386) اكبر من الحدود الدنيا اللازمة لإدخال المتغير في التحليل. اضافة الى ذلك فان القيمة المرتفعة نسبيا لمؤشر(Tolerance) يبين بان هذه المتغيرات لا تعاني من مشكلة الارتباط الخطي بينهما .

و المتغيرات الغير داخلة في التحليل فهي كما يلي:

الجدول (5) يبين المتغيرات الغير داخلة في التحليل

Variables Not in the Analysis					
Step		**Tolerance**	**Min. Tolerance**	**F to Enter**	**Wilks' Lambda**
0	**الجنس (X_1)**	**1.000**	**1.000**	**49.986**	**.728**
	العمر (X_2)	**1.000**	**1.000**	**398.876**	**.251**
	مهنة المريض (X_3)	**1.000**	**1.000**	**18.740**	**.877**
	حالة خروج المريض (X_4)	**1.000**	**1.000**	**9.177**	**.936**
	فترة بقاء المريض (X_5)	**1.000**	**1.000**	**4.025**	**.971**
1	**الجنس (X_1)**	**.978**	**.978**	**53.386**	**.179**
	مهنة المريض (X_3)	**.989**	**.989**	**2.462**	**.246**
	حالة خروج المريض (X_4)	**.981**	**.981**	**.551**	**.250**
	فترة بقاء المريض (X_5)	**.999**	**.999**	**1.357**	**.248**
2	**مهنة المريض (X_3)**	**.918**	**.908**	**2.610**	**.175**
	حالة خروج المريض (X_4)	**.980**	**.959**	**.482**	**.178**
	فترة بقاء المريض (X_5)	**.997**	**.976**	**.588**	**.178**

يشير **الجدول (5)** إلى المتغيرات المحذوفة من التحليل والى الخطوات الثلاث التي اتبعت لتحديد المتغيرات المستبعدة من التحليل إذ بدأت الخطوة ما قبل الأولى باستخراج قيمة F to Remove للمتغيرات الخمسة وانتهت الخطوة الأخيرة باستخراج قيمة F to Remove للمتغيرات الثلاث المفترض إخراجها من التحليل كون قيمة F لهذه المتغيرات قليلة.

ولاختبار معنوية دالة التمييز بالتفصيل تحتسب قيمة *Wilks' Lambda* واختبار F وكما يلي:

الجدول (6) يبين اختبار معنوية دالة التمييز التفصيلي واختبار F

Wilks' Lambda(Λ)									
Step	Number of Variables	Lambda	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	1	.251	1	2	267	398.88	2	267.00	<.0001
2	2	.179	2	2	267	181.41	4	532.00	<.0001

في **الجدول (6)** كانت Wilks' Lambda (Λ) في الخطوة الأولى للمتغير الأول في التحليل 0.251 بينما في الخطوة الثانية بلغت للمتغيرين الأول والثاني الداخلين في التحليل 0.179 أي نلاحظ إن قيمتها انخفضت فهي تقل كلما أضفنا متغير مؤثرا في التحليل وهذا يدل على وجود فروق بين المجموعتين فقد كانت قيمة F في كلا الخطوتين أكبر من قيمتها الجدولية ومما يؤكد ذلك إن مستوى الدلالة الإحصائية (P-Value) في كل خطوة منها كانت اقل من (0.0001).

حساب احتمال التصنيف الصحيح

The probability of correct classification

إن عملية التصنيف قد تؤدي الى الوقوع فيما يعرف بخطأ التصنيف (*misclassification*) وهو احتمال تصنيف مفردة معينة الى المجموعة الأولى بينما هي في الحقيقة تعود للمجموعة الثانية أو الثالثة وبالعكس والجداول التالية تمثل التصنيف بحسب دالة التمييز الخطية ودالة تمييز الجار الاقرب.

اولاً: تصنيف دالة التمييز الخطية

الجدول (7) يبين تصنيف الدالة الخطية

التصنيف الصحيح (Y)	التصنيف			المجموع
	المجموعة الاولى	المجموعة الثانية	المجموعة الثالثة	
المجموعة الاولى	85	0	7	92
	92.39	0.00	7.61	%100
المجموعة الثانية	1	52	1	54
	1.85	96.30	1.85	%100
المجموعة الثالثة	41	2	81	124
	33.06	1.61	65.32	%100
المجموع	127	54	89	270
	47.04	20.07	32.96	

يبين **الجدول (7)** إلى مدى دقة النتائج النهائية للتصنيف إذ يتبين إن (85) حالة من المجموعة الأولى وبنسبة 92.39 % قد تم تصنيفها بشكل صحيح وان (7) حالات وبنسبة 7.61% تم تصنيفها بشكل خاطئ إذ صنفت بأنها ضمن المجموعة الأولى وهي تعود إلى المجموعة الثالثة. وفي المجموعة الثانية يتبين إن (52) حالة وبنسبة 96.30% قد تم تصنيفها بشكل صحيح وبناءا عليه هناك حالة واحدة وبنسبة (1.85) تم تصنيفها بشكل خاطئ إذ صنفت بأنها ضمن المجموعة الأولى وهي تعود إلى المجموعة الثانية و كذلك هناك حالة واحدة وبنسبة (1.85) صنفت بشكل خاطئ إذ تم تصنيفها ضمن المجموعة الثانية وهي تعود إلى المجموعة الثالثة. وفي المجموعة الثالثة يتبين إن (81) حالة وبنسبة 65.32% تم تصنيفها بشكل صحيح وبناءا عليه فان هناك (41) حالة ونسبة 33.06% تم تصنيفها بشكل خاطئ إذ صنفت ضمن المجموعة الثالثة وهي تعود إلى المجموعة الأولى وكذلك هناك حالتان وبنسبة 1.61% صنفت بشكل خاطئ إذ صنفت بأنها ضمن المجموعة الثالثة وهي تعود المجموعة الثانية ، وكنتيجة عامة فقد دلت النتائج بأن الحالات المصنفة تصنيفا صحيحا كانت (218) حالة أي ما نسبته 80.74% من حالات العينة البالغة (270) .

ثانياً: تصنيف دالة الجار الأقرب

الجدول (8) يبين تصنيف دالة الجار الأقرب

التصنيف الصحيح (Y)	التصنيف			المجموع
	المجموعة الاولى	المجموعة الثانية	المجموعة الثالثة	
المجموعة الاولى	92	0	0	92
	%100	0.00	0.00	%100
المجموعة الثانية	0	54	0	54
	0.00	%100	0.00	%100
المجموعة الثالثة	10	0	114	124
	8.06	0.00	91.94	%100
المجموع	102	54	114	270
	37.78	20.00	42.22	

يبين **الجدول (8)** مدى دقة النتائج النهائية للتصنيف إذ يتبين إن (92) حالة من المجموعة الأولى وبنسبة 100 % قد تم تصنيفها بشكل صحيح ، وفي المجموعة الثانية يتبين إن (54) حالة وبنسبة 100% قد تم تصنيفها بشكل صحيح ، وفي المجموعة الثالثة يتبين إن (114) حالة وبنسبة 92.94% تم تصنيفها بشكل صحيح وبناءا عليه فان هناك (10) حالات ونسبة 8.06% تم تصنيفها بشكل خاطئ إذ صنفت ضمن المجموعة الثالثة وهي تعود إلى المجموعة الأولى، وكنتيجة عامة فقد دلت النتائج بأن الحالات المصنفة تصنيفا صحيحا كانت (260) حالة أي ما نسبته 96.30% من حالات العينة البالغة (270) وهذا يدل على جودة نتائج التصنيف.

الاستنتاجات:

1. عند اختبار المعنوية بين متوسطات المجموعات الثلاثة لمرضى السرطان من خلال احتساب قيمة (χ^2) إن الدوال التمييزية لها القدرة على التمييز أي يمكن الاعتماد عليها لتصنيف أي مفردة إلى إحدى المجموعات الثلاثة.

2. لوحظ من خلال اختبار معنوية المتغيرات التوضيحية ان جميع المتغيرات لها تأثير واهمية في تكوين وبناء الدالة التميزية.

3. عند اختبار القيم الذاتية تبين ان قيمة ***Eigen values*** للدالة التمييزية كانت **(3.074)** مما يشير الى ان للدالة التمييزية مقدرة عالية على التمييز حيث ان قيمة ***Eigen values*** اكبر من الواحد الصحيح . اما في ما يتعلق بالارتباط القانوني ***Canonical Correlation*** فقد بلغ **(0.869)** ويدل ذلك على جودة توفيق الدالة التمييزية .

4. عند اختبار معاملات الارتباط الثنائي بين المتغيرات التوضيحية تبين ان معامل الارتباط بين متغير الجنس**(X_1)** ومتغير مهنة المريض**(X_3)** اعلى ارتباط اذ بلغت قيمته (0.280-)، ويليه معامل الارتباط بين متغير الجنس**(X_1)** ومتغير العمر**(X_2)** اذ بلغت قيمته (0.148-)، ويليه معامل الارتباط بين متغير العمر**(X_2)** ومتغير حالة خروج المريض**(X_4)** اذ بلغت قيمته (0.139-) ، وتعني الإشارة السالبة وجود علاقة عكسية بين المتغيرين .

5. عند اختبار نسبة التصنيف الصحيح تبين ان دالة الجار الاقرب افضل النماذج اذ كانت اعلى نسبة فقد بلغت نسبتها (96.30) تليها دالة التمييز الخطية فقد بلغت نسبتها (80.74) ثم دالة التمييز اللوجستية فقد بلغت نسبتها (79.62) واخيراً دالة التمييز التربيعية فقد بلغت نسبتها (79.26).

6. عند المقارنة بين افضل نموذج للتصنيف تبين ان النموذج اللامعلمي افضل النماذج في تمثيل بيانات الدراسة كون البيانات كانت تعاني من مشاكل في التوزيع الطبيعي وتجانس التباينات.

التوصيات

1. نوصي بتطبيق التصنيف وفق نموذجي الدالة الخطية ودالة الجار الاقرب في مجالات أخرى غير الطبية.

2. نوصي بأستخدام الدوال التمييزية الأخرى (دالة كيرنل، دالة الرتبة، الدالة اللوجستية التربيعية، نايف بيز للتصنيف).

3. توسيع عدد المتغيرات في الدراسة لتشمل (التدخين، الوراثة، تناول المشروبات الكحولية، التعرض المكثف للأشعة الضارة، الإصابة بعدوى فيروسية أو بكتيرية،....) من خلال استمارة شاملة كل المعلومات.

4. نؤكد عند تحليل اي مشكلة يجب التأكد من سلامة البيانات ومن توافقها مع الشروط الاساسية (التوزيع الطبيعي، وجود قيم شاذة، التجانس، الاستقلالية، ....) للأسلوب الاحصائي المستخدم للوصول الى نموذج احتمالي باقل خطأ ممكن.

5. من الضروري تطوير مراكز الإحصاء في المستشفيات فضلاً عن معلومات جديدة للمرضى المصابين.

6. ضرورة تشخيص المرض وإجراء العمليات الجراحية اللازمة في وقت مبكر.

Reference

1. الجبوري، شلال و حمزة، صلاح، "تحليل متعدد المتغيرات"، دار الكتب لجامعة بغداد، (2000).

2. Anderson, T.W., "An introduction to Multivariate Statistical Analysis" by John Wiley & Sons, Inc, (1918).

3. Bramer, M., "Principles of Data Mining", Springer-Verlag London Limited, 21-34, (2007).

4. Hardle, W. and Simar, L., "Applied Multivariate Statistical Analysis", Berlin and Louvain-la-Neuve, Germany, p332, (2003).

5. Marshall, R. J. and Chisholm, E. M., "Hypothesis testing in the polychotomous logistic model with an application to detecting gastrointestinal cancer", Statist. Med., 4, 337-344, (1985).

6. Rencher, A. C., "Methods of Multivariate Analysis", John Wiley & sons, New York, USA, (1995).

7. Webb, Andrew R., "Statistical Pattern Recognition", Second Edition, John Wiley & Sons, Ltd.,93-97, (2002).