مقارنة بين أسلوبي C-H والامكان الأعظم لتصنيف البيانات وذلك بالتطبيق علي مرضي الفشل الكلوي

أ.د محمد توفيق البلقيني
رئيس قسم الاحصاء والتأمين
كلية التجارة- جامعة المنصورة

أ.د البيومي عوض طاقية
أستاذ الإحصاء التطبيقي
كلية التجارة- جامعة المنصورة

أ.م 0د محمد مجدي زيدان
أستاذ مساعد طب الأطفال
كلية طب الأطفال- جامعة المنصورة

د. هناء طه الجوهري
مدرس الإحصاء التطبيقي
كلية التجارة- جامعة المنصورة

الباحثة/ رانيا السيد محمد القواصي

ملخص

استهدفت تلك الدراسة المقارنة بين طريقتين من طرق التحليل التمييزي الأول هو أسلوب التوليفة الخطية (C-H classifier) Chung and Hanوهو عبارة عن دمج دالتين تمايز خطية واحدة للمشاهدات المكتملة والاخري للمشاهدات غير المكتملة, وأسلوب تصنيف الإمكان الأعظم MLE classifier وهو مشتق من دالة التمييز الخطي التقليدية ولكن اعتمد علي تقدير المعالم باستخدام دالة الامكان الأعظم . وكلا الأسلوبين يتطلبا نمط خاص للبيانات وهو أن تحتوي البيانات على مشاهدات مفقودة علي وتيرة واحدة ويجب ان يكون للمجتمعين نفس النمط. كما تهدف الدراسة إلي التنبؤ بمتجه مشاهدات جديدة ذو بعد $1 \times p$ إلي إحدي المجتمعين قيد الدراسة, و قامت الدراسه لتقييم كفاءة الأسلوبين باستخدام معيارمقدر البوتستراب المعلمي لفرق معدل الخطأ المتوقع a) parametric bootstrap estimator of Expected Error Rate (Difference لمعرفة أيهما أفضل في التصنيف, ولقد تمت الدراسة التطبيقية علي مجموعة من بيانات مرضي الفشل الكلوي المتاحة بوحدة أمراض الكلي والغسيل الكلوي بمستشفي الأطفال الجامعي (جامعة المنصورة) وقد توصلت الدراسه إلى أن أسلوب C-H هو الأفضل .

كلمات افتتاحية : التحليل التمييزي, أسلوب تصنيف C-H, أسلوب تصنيف MLE , فقد البيانات علي وتيرة واحدة, البوتستراب المعلمي, معدل الخطأ المتوقع

Abstract:

This study aims to comparing two discriminant analysis methods, namely the linear combination classifier of Chung and Hun (C-H classifier), which is a linear combination of two discriminant functions, one based on the complete observations and the other based on the incomplete observations. Also, the Maximum Likelihood Estimation substitution classifier is the general rule of discrimination based on parameters estimators via MLE estimator. It will be assumed that there are two populations are multivariate normal with equal covariance matrix; one of them is with the same monotone pattern. We consider the problem of classifying a $p \times 1$ observation into one of two population. We examine the two classifiers to know which is better in the classification by using a parametric bootstrap estimator of the Expected Error Rate Differences, The applied study was done on a set of data of patients with kidney failure available in the Kidney Diseases and Kidney Dialysis Unit at the University Children's Hospital (Mansoura University) the result shows that the C-H classifier is more efficiency to the MLE classifier when the proportion of observations with missing data is substantial.

Key words: Discriminant analysis, C-H classifier, MLE classifier, Monotone missing data, Parametric bootstrap, Expected error rate

مقدمة:

يعتبر التحليل التمييزي من الأساليب الهامة في التحليل متعدد المتغيرات حيث يتم بموجبه استخدام مجموعة من المتغيرات للفصل (التمييز) بين مجموعتين أو أكثر عن طريق دوال تمييزية وقد تكون خطية أو تربيعية وهي عبارة عن توليفة خطية للمتغيرات المستقلة، وتعمل هذه الدالة علي زيادة متوسط مربعات الفروق بين المجموعات ومن ثم تقلل من أخطاء التصنيف.

وتعتبر عملية التصنيف، هي العملية اللاحقة لعملية تكوين الدالة التمييزية، إذا يتم الاعتماد علي هذه الحالة في التنبؤ أو تحديد المتغيرات التي تساهم بشكل مؤثر في التمييز بين مجموعتين فأكثر وفي تصنيف مفردة جديدة لإحدي المجموعات قيد الدراسة بأقل خطأ تصنيف ممكن، ويمكن استخدام تحليل التمايز في حالة المجتمعات ذات التباينات المتجانسة وغير المتجانسة.

فقد البيانات من أهم المشكلات التي تواجه الباحثيين، وتحدث ظاهرة فقد البيانات عندما تفقد بعض القياسات لبعض المفردات لاي سبب من الأسباب. قد يحدث الفقد في شكل انقطاع والتي تنسحب فيه بعض المفردات من الدراسة قبل انتهائها ويسمي أيضا فقد متكرر علي وتيرة واحدة Monotone أو يحدث في شكل نمط متقطع Intermittent missing data وذلك عندما تتبع القيمة المفقودة قيمة أو مجموعة من القيم المشاهدة ويسمي هذا النوع من الفقد بالفقد غيرالمتكرر علي وتيرة واحدة non-.monotone [2].

وعندما تحتوي عينة من البيانات علي متجهات مشاهدات غير مكتملة ، فإن هناك العديد من الطرق لمعالجة القيم المفقودة في التحليل التمييزي منها أن يتم تجاهل متجهات المشاهدات

غير المكتملة في بناء قاعدة التصنيف ولكن هذه الطريقة غير فعالة. وهناك طرق أخري وهي دمج متجهات المشاهدات الغير مكتملة في بناء قاعدة التصنيف وتقدير معدل الخطأ (Chan and Dunn, 1972; 1974; Bohannon and Smith, 1975; Twedt and Gill, 1992; Anderson, 1957).

في عام 2000 قدم Chung and Hun قاعدة تصنيف جديدة وسهلة في الاستخدام بدلا من تقدير المعالم وهي عبارة عن توليفة خطية لدالتي تمايز واحدة للمشاهدات المكتملة والاخري للمشاهدات غير المكتملة ولكن هذا الاسلوب يشترط وجود نمط خاص للبيانات وهو أن يكون الفقد علي وتيرة واحدة monotone [5].

وتكمن المشكلة محل الدراسة في تصنيف بيانات تحتوي علي مجموعة من المشاهدات المفقودة بالإضافة إلي أن تكلفة خطأ التصنيف في المجالات التطبيقية وخاصة المجال الطبي مرتفعة جدا حيث تتثمل في إتباع أسلوب علاجي ليس من المفروض اتباعه (أي عندما نتنبأ بأن مشاهدة ما تنتمي لمجتمع معين وهي في الحقيقة تنتمي لمجتمع أخر).

هذا ويتم التصنيف باستخدام عدة طرق وأساليب من أهمها تحليل التمايز, وعلي الرغم من الانتشار الواسع في استخدام هذا الأسلوب, إلا أننا نجد في الواقع العملي تعدد طرق استخدامه واختلاف نتائج دقة تصنيفها.

و يهدف هذا البحث إلي قياس دقة التصنيف من خلال طرق التحليل التمييزي بوجود متغيرات تتبع التوزيع الطبيعي وبعض المتغيرات بها مشاهدات مفقودة علي وتيرة واحدة, وذلك من خلال االسعي نحو تحقيق الأهداف التالية:

1. مقارنة بين مصنف C-H ومصنف الإمكان الأعظم وتوضيح النموذج الذي قد يكون هو البديل الأفضل لتحقيق أعلي معدل لدقة التصنيف.
2. حساب نسبة دقة التنصيف من خلال استخدام معيار مقدر البوتستراب المعلمي لفرق معدل الخطأ المتوقع a) parametic bootstrap estimator of expected (error rate difference
3. تطبيق أساليب التصنيف علي بيانات حقيقية طبية.

1- النماذج المستخدمة:

هناك نموذجين للتصنيف:

أ- أسلوب تصنيف C-H
ب- أسلوب تصنيف الإمكان الأعظم

1-2 أسلوب تصنيف C-H (C-H classifier)[3][5][7].

قدم Chung and Han (2000) قاعدة للتصنيف تعرف بمصنف C-H (C-H Classifier) ،هوعبارة عن توليفة خطية من دوال التمايز الخطي, ويفترض هذا الأسلوب أن هناك نمط خاص من البيانات في هذا البحث، وهو أنه يحتوي علي مشاهدات مفقودة علي وتيرة واحدة، بمعني أنه يوجد انقطاع (drop-out) بعض المشاهدات لبعض المتغيرات المستقلة. هذا الأسلوب عالج مشكلة الفقد للتحليل التمييزي بدلا من تقدير المعالم , وذلك تم إنشاء دالتي تمايز من البيانات المكتملة والبيانات غير المكتملة علي التوالي, ثم بعد ذلك تم عمل توليفة خطية للدالتين لتوضيح قاعدة التصنيف. وتكون مصفوفة المشاهدات ذو أبعاد $P \times n_i$ كالتالي :

$$\begin{bmatrix} Y_{i1} & Y_{i2} \\ Z_i & . \end{bmatrix},$$

ونفترض أيضا تجزئة متجه المشاهدة X ذو بعد $P \times 1$ كتالي

$$\mathrm{X} = \begin{bmatrix} Y \\ Z \end{bmatrix}$$

حيث أن Y متجه ذو بعد $(k \times 1)$ و Z متجه ذو بعد $(p - k) \times 1$ $(1 \leq k < p), k)$.

نفترض أن العينات العشوائية التي حجمها m_i لايوجد بها قيم مفقودة

$$\mathrm{X_{ij}} = \begin{bmatrix} Y_{ij} \\ Z_{ij} \end{bmatrix}, \quad i = 1,2; \quad j = 1,2,\ldots,m_i \qquad (2)$$

حيث أن X_{ij} تتبع توزيع طبيعي كالتالي

$$N_p(\mu_i, \Sigma) = N_p\left(\begin{bmatrix} \mu_{iy} \\ \mu_{iz} \end{bmatrix}, \begin{bmatrix} \Sigma_{yy} & \Sigma_{zy} \\ \Sigma_{yz} & \Sigma_{zz} \end{bmatrix}\right)$$

فعندما تكون عينة البيانات للمجتمعين$(i = 1,2)$ تحتوي علي متجهات من المشاهدات الغير مكتملة كما موضح في الشكل التالي(2006) Batsidis and Zografos ،

$$
\begin{array}{cccccccc}
Y_{11}^{(i)} & Y_{12}^{(i)} & Y_{13}^{(i)} & \cdots & Y_{1m_i}^{(i)} & Y_{1(m_i+} & \cdots & Y_{1n}^{(i} \\
Y_{21}^{(i)} & Y_{22}^{(i)} & Y_{23}^{(i)} & \cdots & Y_{2m_i}^{(i)} & Y_{1(m_i+}^{(i)} & \cdots & Y_{1n_i}^{(i)} \\
\vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\
Y_{k1}^{(i)} & Y_{k2}^{(i)} & Y_{k3}^{(i)} & \cdots & Y_{km_i}^{(i)} & Y_{K(m_i}^{(i)} & \cdots & Y_{Kn_i}^{(i)} \\
Z_{11}^{(i)} & Z_{12}^{(i)} & Z_{13}^{(i)} & \cdots & Z_{1m_i}^{(i)} & & & \\
Z_{21}^{(i)} & Z_{22}^{(i)} & Z_{23}^{(i)} & \cdots & Z_{2m_i}^{(i)} & & & \\
\vdots & \vdots & \vdots & & \vdots & & & \\
Z_{(P-k}^{(i)} & Z_{(P-k}^{(i)} & Z_{(P-k}^{(i)} & \cdots & Z_{(P-k)1}^{(i)} & & &
\end{array}
$$

الشكل (1)

ويلاحظ من الشكل أن العينات العشوائية التي حجمها ($n_i - m_i$) تحتوي فقط علي عدد k من المفردات, حيث $Y_{ij(k\times1)}, i =$ $.1, 2$ ؛ $j = m_i + 1, \dots, n_i$

ونشير إلي أن $(x_{ij}, i = 1, 2; j = 1, 2, \dots, m_i)$هو متجه المشاهدات المكتملة وسبق التوضيح له في المعادلة (2) وأن $(Y_{ij}, i = 1, 2; j = 1, 2, \dots, n_i)$ يشيرإلي متجه المشاهدات الغير مكتملة

فعندما تكون العينة تحتوي علي متجهات من المشاهدات الغيرمكتملة كما موضح فى الشكل (1)

هناك العديد من طرق معالجة البيانات في التحليل التمييزي :

1) استخدام فقط الأفراد (المفردات) لجميع المتغيرات الموجودة وتعرف هذه الحالة بأسلوب تحليل الحالة الكاملة أوالحذف بطريقة القائمة (*listwise deletion*) أو حذف الحالة (*case-wise deletion*).
2) استخدام فقط المتغيرات لجميع الأفراد الذين كل مشاهداتهم بها قيم وتسمي هذه الحالة بطريقة شطب المتغيرات (*variable-wise deletion method*).

مع الأخذ في الاعتبار النقطتيين الأولي والثانية، أسلوب التصنيف تم إنشاؤه من خلال دمج الطريقتين[5].

أولا: دالة التمايز الخطي التي تعتمد علي المشاهدات المكتملة $\mathrm{X_{ij}}, i = 1,2, j = 1,2, \dots, m_i$

اعتمدت هذه الدالة علي أسلوب معالجة تحليل الحالة الكاملة وتكون البيانات علي هذا الشكل

$\boldsymbol{Y_{11}^{(i)}}$	$\boldsymbol{Y_{12}^{(i)}}$	$\boldsymbol{Y_{13}^{(i)}}$	$\cdots$	$\boldsymbol{Y_{1m_i}^{(i)}}$
$\boldsymbol{Y_{21}^{(i)}}$	$\boldsymbol{Y_{22}^{(i)}}$	$\boldsymbol{Y_{23}^{(i)}}$	$\cdots$	$\boldsymbol{Y_{2m_i}^{(i)}}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\boldsymbol{Y_{k1}^{(i)}}$	$\boldsymbol{Y_{k2}^{(i)}}$	$\boldsymbol{Y_{k3}^{(i)}}$	$\cdots$	$\boldsymbol{Y_{km_i}^{(i)}}$
$\boldsymbol{Z_{11}^{(i)}}$	$\boldsymbol{Z_{12}^{(i)}}$	$\boldsymbol{Z_{13}^{(i)}}$	$\cdots$	$\boldsymbol{Z_{1m_i}^{(i)}}$
$\boldsymbol{Z_{21}^{(i)}}$	$\boldsymbol{Z_{22}^{(i)}}$	$\boldsymbol{Z_{23}^{(i)}}$	$\cdots$	$\boldsymbol{Z_{2m_i}^{(i)}}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\boldsymbol{Z_{(P-k)1}^{(i)}}$	$\boldsymbol{Z_{(P-k)2}^{(i)}}$	$\boldsymbol{Z_{(P-k)3}^{(i)}}$	$\cdots$	$\boldsymbol{Z_{(P-k)m_i}^{(i)}}$

تكون الدالة كالتالي :

$$W_x = (\bar{X}^{(1)} - \bar{X}^{(2)})' S_{xx}^{-1} \left[X - \frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})\right],$$

Where:

$$\bar{X}^{(i)} = \frac{1}{m_i}\sum_{i=1}^{2}\sum_{j=1}^{m_i} X_{ij} = \begin{bmatrix}\bar{Y}_{i1}\\ \bar{Z}_i\end{bmatrix}$$

$$\bar{Y}_{i1} = \frac{1}{m_i}\sum_{i=1}^{2}\sum_{j=1}^{m_i} Y_{ij}\ , \bar{Z}_i = \frac{1}{m_i}\sum_{i=1}^{2}\sum_{j=1}^{m_i} z_{ij} = \begin{bmatrix}\bar{Y}_{i1}\\ \bar{Z}_i\end{bmatrix}, i = 1,2 \qquad (4)$$

$$S_{xx} = \sum_{i=1}^{2}\sum_{j=1}^{m_i}(X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)'/v_x, \quad v_x = m_1 + m_2 - 2 \qquad (5)$$

حيث

$\bar{X}_{\mathrm{i}}$هو متوسط عينة االمشاهدات المكتملة

S_{xx}هو مصفوفة تغاير المشاهدات المكتملة

ثانيا: دالة التمايز الخطي التي تعتمد علي المشاهدات الغير مكتملة$Y_{\mathrm{ij(k\times 1)}}, \mathrm{i} = 1,2\ ; j = m_i + 1, \dots, n_i$

اعتمدت هذه الدالة علي أسلوب معالجة تحليل حذف المتغيرات *(variable-wise deletion method)* وتكون البيانات علي هذا الشكل :

$$\begin{array}{cccccccc}
Y_{11}^{(i)} & Y_{12}^{(i)} & Y_{13}^{(i)} & \cdots & Y_{1m_i}^{(i)} & Y_{1(m_i+1)}^{(i)} & \cdots & Y_{1n_i}^{(i)} \\
Y_{21}^{(i)} & Y_{22}^{(i)} & Y_{23}^{(i)} & \cdots & Y_{2m_i}^{(i)} & Y_{1(m_i+1)}^{(i)} & \cdots & Y_{1n_i}^{(i)} \\
\vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \\
Y_{k1}^{(i)} & Y_{k2}^{(i)} & Y_{k3}^{(i)} & \cdots & Y_{km_i}^{(i)} & Y_{K(m_i+1)}^{(i)} & \cdots & Y_{Kn_i}^{(i)}
\end{array}$$

تكون الدالة كالتالي

$$W_y = (\bar{Y}_1 - \bar{Y}_2)' S_{yy}^{-1} \left[Y - \frac{1}{2}(\bar{Y}_1 + \bar{Y}_2)\right],$$

Where:

$$\bar{Y}_i = \frac{1}{n_i}[m_i \bar{Y}_{i1} + (n_i - m_i)\bar{Y}_{i2}], \qquad (5)$$

$$\bar{Y}_{i1} = \frac{1}{m_i}\sum_{j=1}^{m_i} Y_{i1j}$$

$$\bar{Y}_{i2} = \frac{1}{n_i - m_i}\sum_{j=m_i+1}^{n_i} Y_{i2j} \ , i = 1,2$$

$$S_{yy} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)(Y_{ij} - \bar{Y}_i)'/v_y, \ \ v_y = n_1 + n_2 - 2 \qquad (7)$$

ثالثا: قاعدة التصنيف

هي دمج دالتي التمايز الخطيW_x, W_y للحصول علي قاعدة التصنيف والتي هي عبارة عن توليفة خطية لدالتي W_x و W_y وتسمي بإحصائية التوليفة الخطية W_c.

$$W_c = cW_x + (1 - c)W_y, \qquad 0 \le c \le 1 \qquad (8)$$

هذه القاعدة تتميز بسهولة الاستخدام في تصنيف المشاهدة x إلي المجتمع الأولπ_1 إذا كانت

$$W_c \ge 0$$

وغير ذلك تصنف إلي المجتمع الثانيπ_2.

وهذه القاعدة يطلق عليها طريقة تصنيف التوليفة الخطية *the linear combination classification procedure)* وهي تعود إلي*Chung and Han2000* وتعرف أيضا بمصنف *C-H*(*C-H Classifier).*

القاعدة السابقة تعتمد *علي قيمة* C، والتي تكون علي الصيغة التالية

$$c = \frac{\left(\frac{1}{m_1}+\frac{1}{m_2}\right)^{-1} D_x^2}{\left(\frac{1}{m_1}+\frac{1}{2}\right)^{-1} D_x^2 + \left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{-1} D_y^2} \qquad (9)$$

حيث أن

$$D_x^2 = (\bar{X}_1 - \bar{X}_2)' S_x^{-1} (\bar{X}_1 - \bar{X}_2)$$

$$D_y^2 = (\bar{Y}_1 - \bar{Y}_2)' S_y^{-1} (\bar{Y}_1 - \bar{Y}_2)$$

حيث يشير كلا من (D_x^2 وD_y^2) إلي مربع مسافة مهالونوبيس *Distance* Mahalanobis .

اقترح (*Chung and Han 2000*) قيمة C بناءا علي معرفة أن معدل الخطأ الشرطي ومعدل الخطأ المتوقع يعتمد علي مسافة مهالونوبيس (Distance Mahalanobis) للمشاهدات المكتملة وغير المكتملة وحجم العينة المناظر لهم ($(m_i, n_i,\ i = 1,2)$, لذلك تم استخدام c لتكون متعلقة بحجم العينات ومسافة مهالونوبيس للبيانات المكتملة وغير المكتملة.

احتمالية خطأ التصنيف [8][10].

أحد الأساليب الهامة للحكم على كفاءة طرق التصنيف هو حساب معدل الخطأ الخاص بها أي حساب احتمال خطأ التصنيف. ويعتبر خطأ التصنيف عامل هام لاثبات كفاءة الدالة التمييزية، أي أن الدالة التمييزية التي تعطي أقل خطأ تصنيف هي الدالة الأكثر كفاءة وهي الأفضل للتنبؤ الأمثل لمشاهدة جديدة في مجموعة من المجموعات المحددة. ومن أجل الحصول علي احتمال خطأ تصنيف الأسلوب المستخدم يفترض التالي

$$W_x = a'X + b$$

حيث أن

$$a'_{(1\times p)} = \left(\bar{X}^{(1)} - \bar{X}^{(2)}\right)' S_{xx}^{-1}$$

$$b = \frac{-1}{2}\left(\bar{X}^{(1)} - \bar{X}^{(2)}\right)' S_{xx}^{-1}\left(\bar{X}^{(1)} + \bar{X}^{(2)}\right)$$

وايضا بافتراض أن

$$W_y = \mathrm{d}'\mathrm{Y} + \mathrm{e}$$

حيث أن

$$d' = \left(\bar{Y}^{(1)} - \bar{Y}^{(2)}\right)' S_{yy}^{-1}$$

$$\mathrm{e} = -\frac{1}{2}\left(\bar{\mathrm{Y}}^{(1)} - \bar{\mathrm{Y}}^{(2)}\right)' \mathrm{S}_{\mathrm{yy}}^{-1}\left(\bar{\mathrm{Y}}^{(1)} + \bar{\mathrm{Y}}^{(2)}\right)$$

$$\because W_c = cW_x + (1-c)W_y$$

Where; $\mathrm{X} = \begin{bmatrix} Y \\ Z \end{bmatrix}$

$$\therefore W_c = c\left(a_1'Y + a_2'Z + b\right) + (1-c)(d'Y + e)$$

$$= A'\mathrm{Y} + B'\mathrm{Z} + \mathrm{F} = H'X + F$$

حيث أن

$$A = ca_1 + (1-c)d, B = ca_2, F = cb + (1-c)e, H = \begin{bmatrix} A_{(k\times 1)} \\ B_{(p-k)\times 1} \end{bmatrix} \quad (10)$$

الاحتمال الشرطي لخطأ تصنيف x مشاهدة من المجتمع الأول π_1 إلي المجتمع الثاني π_2 أو العكس باستخدام W_c

$$CER_{12}(W_C) = \Pr(w_c < 0 \mid x \in \pi_1)$$

$$= \Pr\left(T < \frac{-H'\mu^{(1)} - F}{\sqrt{H'\Sigma H}}\right)$$

$$= \Phi\left[\frac{-\boldsymbol{H'\mu^{(1)}} - \boldsymbol{F}}{\sqrt{\boldsymbol{H'\Sigma H}}}\right] \qquad (11)$$

حيث أن Φ دالة كثافة التوزيع الطبيعي متعدد المتغيرات

وبالمثل في الحالة العكسية من المجتمع الثاني للمجتمع الأول

$$CER_{21}(W_C) = \Pr\left(\mathrm{w}_c \geq 0 \mid x \in \pi_2\right)$$

$$= \Pr\left(T \geq \frac{-H'\mu^{(2)} - F}{\sqrt{H'\Sigma H}}\right)$$

$$= 1 - \Pr\left(T < \left[\frac{-H'\mu^{(2)} - F}{\sqrt{H'\Sigma H}}\right]\right.$$

$$= \Phi\left[\frac{H'\mu^{(2)} + F}{\sqrt{H'\Sigma H}}\right] \qquad (12)$$

وبالتالي يصبح معدل الخطأ الشرطي

(The conditional error rate) لإحصائية التوليفة الخطية ,مع تساوي الاحتمالات القبلية (priori probabilities):

$$CER(W_c) = \frac{1}{2}[CER_{12}(W_c) + CER_{21}(W_c)] \qquad (13)$$

وبفرض أن $\tilde{\theta}$ هي عبارة عن معالم دالتي التمايز للمشاهدات المكتملة وغير المكتملة

$$\tilde{\theta} = \left[\bar{Y}_1 : \bar{Y}_2 : S_y : S_x : \bar{X}_1 : \bar{X}_2\right]$$

معدل الخطأ المتوقع (Expected Error Rate) لخطأ تصنيف متجه x من المشاهدات من المجتمع الأول إلي المجتمع الثاني هو كالتالي

$$EER(W_C)_{12} = E_{\tilde{\theta}}\left[\Phi\left[\frac{H'\mu^{(1)} - F}{\sqrt{H'\Sigma H}}\right]\right]$$

وبالمثل EER لخطأ تصنيف متجه x من المشاهدات من المجتمع الثاني إلي المجتمع الاول

هو

$$EER(W_C)_{21} = E_{\tilde{\theta}}\left[\Phi\left[\frac{H'\mu^{(2)} + F}{\sqrt{H'\Sigma H}}\right]\right]$$

ومرة أخري مع افتراض تساوي الاحتمالات القبلية فإن معدل الخطأ المتوقع المعادلة (6) هو

$$EER(W_c) = \frac{1}{2}[EER_{12}(W_c) + EER_{21}(W_c)]$$

2-2 أسلوب تصنيف دالة الامكان الأعظم (Maximum Likelihood Estimation [4][9] Classifier)

هذا الأسلوب هو طريقة مطورة لتقدير المعالم في التوزيع الطبيعي متعدد المتغيرات عندما تحتوي البيانات علي مشاهدات غير مكتملة تتبع نمط فقد وتيرة واحدة Hacking and Smith (1968) ويمكن تلخيصه كالتالي:

1. تقسيم البيانات إلي مجموعة من المتغيرات بالإضافة إلي أن بعض المتغيرات بها مشاهدات مفقودة
2. الحصول علي التقديرات الأولية للمعالم لكل مجموعة البيانات التي تكون كل متجهات المتغيرات مكتملة
3. الحصول علي باقي التقديرات الأولية للمعالم من مجموعة البيانات التي متجهات متغيراتها غير مكتملة

يفترض أن المجتمع الأول والثاني يتبعا توزيع طبيعي متعدد المتغيرات $(\mu_i, \Sigma,)$ عند $i = 1,2$ حيث

$$\mu_i = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \end{bmatrix} \qquad (14)$$

$$\Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{zy} \\ \Sigma_{yz} & \Sigma_{zz} \end{bmatrix} \qquad (15)$$

ونفترض أيضا

$$A_{yy,n_i,i} = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)'$$

$$A_{yy,m_i,i} = \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)'$$

$$A_{zy,m_i,i} = \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)(z_{ij} - \bar{z}_i)'$$

$$A_{zz,m_i,i} = \sum_{j=1}^{m_i} (z_{ij} - \bar{z}_i)(z_{ij} - \bar{z}_i)'$$

حيث أن $z_{ij} \in Z_i$ و $y_{ij} \in [Y_{i1}:Y_{i2}]$ و Y_{i1}, Y_{i2}, Z_i موضحة في المعادلة (1).

دالة الامكان الأعظم $(MLEs)$ للمعادلتين (14)و(15) هما علي التوالي

$$\hat{\mu}_i = \begin{bmatrix} \hat{\mu}_{i1} \\ \hat{\mu}_{i2} \end{bmatrix} \qquad \hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_{yy} & \hat{\Sigma}_{zy} \\ \hat{\Sigma}_{yz} & \hat{\Sigma}_{zz} \end{bmatrix} \quad (16)$$

حيث أن

$$\hat{\Sigma}_{yy} = \frac{\sum_{i=1}^{2} A_{11,n_i,i}}{\sum_{i=1}^{2} n_i}$$

$$\hat{\Sigma}_{yz} = \frac{1}{\sum_{i=1}^{2} n_i}\left[\sum_{i=1}^{2} A_{yy,n_i,i}\right]\left[\sum_{i=1}^{2} A_{yy,m_i,i}\right]^{-1}\left[\sum_{i=1}^{2} A_{yz,m_i,i}\right] \qquad (17)$$

$$\hat{\Sigma}_{zz} = \frac{1}{(\sum_{i=1}^{2} m_i)}\sum_{i=1}^{2} A_{zz.1,m_i,i} + \frac{1}{(\sum_{i=1}^{2} n_i)}\left[\sum_{i=1}^{2} A_{zy,m_i,i}\right]\left[\sum_{i=1}^{2} A_{yy,m_i,i}\right]^{-1} \times \left[\sum_{i=1}^{2} A_{yy,n_i,i}\right]\left[\sum_{i=1}^{2} A_{yy,m_i,i}\right]^{-1}\left[\sum_{i=1}^{2} A_{yz,m_i,i}\right] \qquad (18)$$

مع $\hat{\mu}_{i1} = \bar{y}_i$ حيث أن $\bar{y}_i$ سبق تعريفها في (6)

$$\hat{\mu}_{i2} = \bar{z}_i - \left[\hat{\Sigma}_{yz}\hat{\Sigma}_{zz}^{-1}\right](\bar{y}_{i1} - \bar{y}_{i2}),$$

$$\bar{z}_i = \frac{1}{m_i}\sum_{j=1}^{n_i} z_{ij}$$

$$\sum_{i=1}^{2} A_{zz,m_i,i} = \sum_{i=1}^{2} A_{zz,m_i,i} - \left[\sum_{i=1}^{2} A_{zy,m_i,i}\right]\left[\sum_{i=1}^{2} A_{yy,m_i,i}\right]^{-1}\left[\sum_{i=1}^{2} A_{yz,m_i,i}\right]$$

حيث أن $\bar{y}_{i1}$, $\bar{y}_{i2}$, $\hat{\Sigma}_{yz}$ و $\hat{\Sigma}_{zz}$ تم توضيحهم في (6) ,(16) و(17)علي التوالي عند $i = 1,2$

إحصاءه تصنيف دالة الامكان الأعظم هي

$$W_{MLE} = (\hat{\mu}_2 - \hat{\mu}_1)'\hat{\Sigma}^{-1}\left[x - \frac{1}{2}(\hat{\mu}_2 + \hat{\mu}_1)\right] \qquad (19)$$

حيث $\hat{\mu}_2, \hat{\mu}_1$ و$\hat{\Sigma}$ تم توضيحهم (16), و$x \in \mathbb{R}_{p\times 1}$ هو متجه مشاهدة الغير مسجلة (unlabeled observation vector), ويتم تصنيف مشاهدات هذا المتجه إلي المجتمع الأول إذا كانت

$$W_{MLE} \leq 0 \qquad (20)$$

وغير ذلك تنتمي إلي المجتمع الثاني.

احتمالية خطأ التصنيف [10]

يتم تقييم أسلوب التصنيف من خلال أخطاء التصنيف الناتجة عن الاسلوب المتبع بمعني إذا كانت مشاهدة من المجتمع الأول وتم تصنيفها علي أنها تتبع المجتمع الثاني والعكس.حيث تم التركيز علي حالة انه يوجد مجتمعين, معدل الخطأ الشرطي اعتمد علي $\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}$

احتمال خطأ التصنيف للمشاهدة x من المجتمع الأول إلي المجتمع الثاني باستخدام قاعدة التصنيف W_{MLE} هو

$$CER_{12}(\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}) = P\left[W_{MLE} > 0 | \hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}; x \in \pi_1\right] = 1 - \Phi(w_1)$$

وبالمثل احتمال خطأ التصنيف للمشاهدة x من المجتمع الثاني إلي المجتمع الأول هو

$$CER_{21}(\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}) = P\left[W_{MLE} \leq 0 | \hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}; x \in \pi_2\right] = \Phi(w_2)$$

حيث أن

$$w_i = \left[\hat{\delta}'\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\delta}\right]^{-1/2}\left[\hat{\delta}'\hat{\Sigma}^{-1}\left(\frac{1}{2}(\hat{\mu}_1+\hat{\mu}_2)\right)-\mu_i\right] \quad i=1,2 \qquad (21)$$

Where

$$\hat{\delta} = \hat{\mu}_1 - \hat{\mu}_2$$

إذا وبالاخذ في الاعتبار تساوي الاحتمالات القبلية فإن معدل الخطأ الشرطي يكون كالتالي:

$$\mathrm{CER}(\hat{\mu}_1,\hat{\mu}_2,\hat{\Sigma}) = \frac{1}{2}[1-\Phi(w_1)+\Phi(w_2)] \qquad (22)$$

وبالتالي معدل الخطأ المتوقع هو

$$EER(\hat{\mu}_1,\hat{\mu}_2,\hat{\Sigma}) = \frac{1}{2}[1-E_{\tilde{\theta}}(\Phi(w_1))+E_{\tilde{\theta}}(\Phi(w_2))]$$

3-2 معدل الخطأ المتوقع البوتسترابي لاسلوبي التصنيف C-H ,MLE

Bootstrap Expected Error Rate for the C-H and MLE Classifiers[10][7][1]

يعتبر أسلوب البوتستراب أحد الأساليب الاحصائية التي قدمها Efron(1979) واستخدمها في تقدير التباينات والأخطاء المعيارية وفترات الثقة والقيمة الاحتمالية في بادئ الأمر ثم قام بتطويرها كلا من Efron and Tibshirani(1993) كطريقة من طرق إعادة المعاينة, حيث يعتمد أسلوب البوتستراب علي توليد البيانات عن طريق السحب بإرجاع من البيانات الأصلية أي اعادة استخدام العينة محل الدراسة بكفاءة , حيث يمكن الحصول علي عينة البوتستراب$X^* = (x_1^*, x_2^*, x_3^*, \dots, x_k^*)$ من خلال

سحب عينة حجمها n بالارجاع من البيانات الأصلية لعدد k من المرات.

تم استخدام أسلوب البوتستراب لتقدير معدل الخطأ المتوقع لأساليب التصنيف المستخدمة C-H and MLE. حيث أن $\hat{\mu}_1, \hat{\mu}_2$و$\hat{\Sigma}$ هم مقدرات الامكان الأعظم للمعالم μ_1, μ_2, Σ علي التوالي وسبق التعرف عليهم في (2-2) وأيضا $\hat{\mu}_1^*, \hat{\mu}_2^*$و$\hat{\Sigma}^*$ هم مقدرات البوتستراب للمعالم μ_1, μ_2, Σ , تم حسابهم من خلال عينة البوتستراب المعلمي من بيانات العينة محل الدراسة

$$\begin{bmatrix} Y_{i1}^* & Y_{i2}^* \\ Z_i^* & . \end{bmatrix} \qquad (23)$$

حيث تم توليد العينة من التوزيع الطبيعي متعدد المتغيرات$N_P(\hat{\mu}_i, \hat{\Sigma})$ حيث $i = 1,2$.

إذا معدل الخطأ الشرطي البوتستراربي لمصنف C-H هو

$$CER_{ig}^*(W_c^*) = \Phi\left[\frac{(-1)^{2-i}H^{*'}\hat{\mu}_i + (-1)^{2-i}F^*}{\sqrt{H^{*'}\hat{\Sigma}H^*}}\right]$$

حيث $i, g = 1,2,\ i \neq g$ وW_c^*, H^*وF^*هما نفس تعريف W_c, HوF في كلا من (8)(10)علي التوالي إلا أنه تم استخدام بوتستراب بيانات طبيعية متعددة

المتغيرات في (23).وبالتالي مع افتراض تساوي الاحتمالات القبلية فإن CER_{Boot} لأسلوب تصنيف C-H هو كالتالي

$$CER^*(W_c^*) = \frac{1}{2}[CER_{12}^*(W_C^*) + CER_{21}^*(W_C^*)]$$

وبالمثل CER_{Boot} لأسلوب تصنيف الامكان الأعظم MLE هو

$$CER_{ig}^*(\hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\Sigma}^*) = \mathrm{P}\left[(-1)^{2-g} W_{MLE}^* > 0 | \hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\Sigma}^*; x \in \prod_i\right]$$

حيث أن W_{MLE}^* هي نفس تعريف W_{MLE} في المعادلة (19)

إذا احتمال خطأ التصنيف للمشاهدة x من المجتمع الأول إلي المجتمع الثاني هي

$$CER_{12}^*\left(\hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\Sigma}^*\right) = 1 - \Phi(w_1^*)$$

إذا احتمال خطأ التصنيف للمشاهدة x من المجتمع الثاني إلي المجتمع الأول هي

$$CER_{21}^*\left(\hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\Sigma}^*\right) = \Phi(w_2^*)$$

حيث أن

$$w_i^* = \left[\hat{\delta}^{*'}\hat{\Sigma}^{*-1}\hat{\Sigma}\hat{\Sigma}^{*-1}\hat{\delta}^*\right]^{-1/2}\left[\hat{\delta}^{*'}\hat{\Sigma}^{*-1}\left(\frac{1}{2}(\hat{\mu}_1^* + \hat{\mu}_2^*)\right) - \hat{\mu}_i\right]$$

$$i = 1,2$$

$$\hat{\delta}^* = \hat{\mu}_1^* - \hat{\mu}_2^*$$

وبالأخذ في الاعتبار تساوي الاحتمالات القبلية فإن

$$CER^*\left(\hat{\mu}_1^*, \hat{\mu}_2^*, \hat{\Sigma}^*\right) = \frac{1}{2}[1 - \Phi(w_1^*) + \Phi(w_2^*)]$$

وبذلك توصل Young. and Ounpraseuth إلي مقدر البوتستراب المعلمي لمعدل الخطأ المتوقع (the estimated parametric bootstrap EERD) لكلا من أسلوبي التصنيف C-H و MLE

$$\widehat{EERD}_{Boot} = \frac{1}{K}\sum_{j=1}^{K}\left(\widehat{CER}_{jBoot(C-H)} - \widehat{CER}_{jBoot(MLE)}\right) \qquad (24)$$

حيث K العدد الاجمالي لمحاكاة العينة محل الدراسة و j تشير إلي j^{th} حيث $j \in \{1,2,\dots,k\}$

2- تطبيق النموذج علي البيانات الفعلية

اعتمد البحث علي البيانات المتوفرة بوحدة أمراض الكلي والغسيل الكلوي بمستشفي الاطفال الجامعي (جامعة المنصورة) وتم اختيار عينة عشوائية حجمها 85 مريض وفقا لمعادلة ستيفن ثامبسون من مجتمع حجمه 110 مريض مصابون بمرض الفشل الكلوي المزمن ويأخذون حقن الاريثوبويتين لتحسين نسبة الهيموجلوبين وذلك في الفترة من يناير الي فبراير2018 نظرا

لأن المريض يسحب منه عينة دم كل شهر أو شهرين لمعرفة نتيجة حقن الاريثوبويتين وتم استخدام برنامج لغة البرمجة R لتطبيق الأساليب المستخدمة لقياس دقة التصنيف وأيضا الوصول إلي الأسلوب الذي يحقق أقل خطأ تصنيف وأيضا التنبؤ بنتيجة جرعة الاريثروبيتين.

مجموعة البيانات لمجتمعين، المجتمع الأول هو عدم تحسن نسبة الهيموجلوبين والمجتمع الثاني تحسن نسبة الهيموجلوبين، ولكل مجتمع ثمانية متغيرات مستقلة تم الاتفاق عليها مع الأطباء أصحاب التخصص وهي

- العمر بالسنوات.
- النوع وهو متغير ثنائي يأخذ القيمة (1) تعني أنثي,القيمة (2) تعني ذكر.
- الغسيل الكلوي Dialysis وهو نوع من أنواع العلاج المتوافرة التي يستخدمها الأطباء كحل بديل عندما تفشل الكلية في أداء وظائفها. هو متغير ثنائي يأخذ القيمة (1) إذا كان يستخدم الغسيل الكلوي, القيمة(0) إذا كان لا يستخدم الغسيل الكلوي.
- الحديد Iron وهو تحليل يشير إلي كمية الحديد في جسم الإنسان وإذا ما كان الشخص يعاني من نقص أو فائض في عنصر الحديد. هو متغير ثنائي يأخذ القيمة (1) إذا كان الشخص يأخذ

دواء حديد, القيمة (0) إذا كان لا يأخذ دواء حديد.

- االاريثربويتين Erythropoietin وهو الهرمون المسئول عن زيادة كرات الدم الحمراءعند مرضي الفشل الكلوي لعدم إفراز الكلية الهرمون بكمية كافية ونقصه يؤدي إلي الانيميا أو فقر الدم, ويعطي عن شكل حقن تحت الجلد أو عن طريق الوريد لمنع حدوث فقر الدم. هو متغير ثلاثي القيمة يأخذ القيمة (1) إذا كان يأخذ نوع أرانسب Aranasep والقيمة(2) إذا كان يأخذ نوع إيبركس Eprex والقيمة (3) إذا كان يأخذ نوع ريكرمونRecormon.
- جرعة الاريثروبيوتن ولكل مريض جرعته الخاصه به.
- الفيريتين Ferritin هو بروتين يوجد داخل الخلايا ويتحكم في تخزين وإطلاق الحديد. يعكس الفيريتين حالة الحديد في الجسم فكلما زادت نسبته في المصل دل ذلك علي زيادة الحديد في الجسم والعكس. لكل مريض قيمة الفيريتين الخاص به بناءا علي تحليل الفيريتين.
- تشبع الترانسفرين Transferrin Saturation(TSAT) هو قيمة مختبرية طبية تقاس بالنسبة المئوية وهي نسبة بين حديد مصل الدم Iron والسعة الرابطة للحديد الكلي Total Iron Binding Capacity(TIBC) وهذه القيمة

تعطي فكرة للأطباء حول كمية الحديد المرتبط في مصل الدم بالترانسفرين

- أما المتغير التابع وهو نسبة الهيموجلوبين في الدم Haemoglobin(HB) وهو متغير ثنائي يأخذ القيم (1) إذا تحسنت نسبة الهيموجلوبين, والقيم (0) إذا لم تتحسن نسبة الهيموجلوبين .

جدول (4-1)

المتغير	الاسم بالعربي	الاسم بالانجليزي	القيم التي تم ادخالها
X_1	العمر	Age	لكل مريض عمره الخاص
X_2	النوع	Gender	١ = أنثى, ٢ =ذكر
X_3	الغسيل الكلوي	Dialysis	١ = يأخذ, ٠ = لايأخذ غسيل
X_4	الحديد	Iron	١ = يأخذ, ٠ = لا يأخذ حديد
X_5	الاريثوبيوتين	Erythropoietin	١ =أر انسب, ٢ = إبركس, ٣=ريكرمون
X_6	جرعة الاريثوبيوتين		لكل مريض جرعته الخاصة به
X_7	الفيريتين	Ferritin	لكل مريض قيمته الخاصة به
X_8	تشبع الترانسفرين	Transferrin Saturation(TSAT)	لكل مريض نسبته الخاصة به
Y	الهيموجلوبين	Hemoglobin (HB)	١ = تحسن نسبة الهيموجلوبين, ٠ = عدم تحسن نسبة الهيموجلوبين

4-1 نتائج تطبيق أسلوب C-H

أولا نتائج دالة التمايز للبيانات المكتملة

$$W_x = \left(\bar{X}^{(1)} - \bar{X}^{(2)}\right)' S_{xx}^{-1}\left[X - \frac{1}{2}\left(\bar{X}^{(1)} + \bar{X}^{(2)}\right)\right], \quad (25)$$

$$X_j^{(i)} = \begin{bmatrix} Y_j^{(i)} \\ Z_j^{(i)} \end{bmatrix}, \quad i = 1,2; \quad j = 1,2,\ldots,m_i$$

جدول (1)

الوسط الحسابي لمتغيرات الدراسة

Mean		المجتمع الأول	المجتمع الثاني
	Y	0	1
Y	x_1	12.25	11.25
	x_2	1.5	1.5263
	x_3	0.96875	0.92105
	x_4	0.1875	0.39474
	x_5	2.21875	2.13158
	x_6	227.1875	330.26316
Z	x_7	174.875	194.3684
	x_8	33.109375	27.84211

من خلال الجدول نلاحظ أن متوسطات كل من المتغير الاول(العمر) والثاني(النوع) والثالث(الغسيل الكلوي) والخامس(الاريثوبيوتين) تقريبا متساوين في كلا المجتمعين، اما بالنسبة لمتوسطي المتغيرين السادس(جرعة الاريثوبيوتين) والسابع(الفيريتين) في

المجتمع الثاني أكبر من المجتمع الاول، وهذا يعني أنه كلما زادت قيم تلك المتغيرات كلما أدي ذلك إلي تحسين نسبة الهيموجلوبين وأخيرا متوسط المتغير الثامن(تشبع الترانسفرين) للمجتمع الاول أكبر من المجتمع الثاني

جدول (2) مصفوفة التغاير لدالة التمايز الأولي

S_x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	13.77022	-0.29412	0.404412	0.047794	0.308824	-49.6691	-220	-7.95404
x_2	-0.29412	0.256966	-0.02825	0.001548	0.056889	-20.0298	-22.1378	-0.33959
x_3	0.404412	-0.02825	0.054881	0.020172	0.023728	4.676132	-1.44147	0.79948
x_4	0.047794	0.001548	0.020172	0.020502	0.010497	-4.66283	-33.9379	-2.8057
x_5	0.308824	0.056889	0.023728	0.010497	0.408983	-16.5681	16.82401	0.735645
x_6	-49.6691	-20.0298	4.676132	-4.66283	-16.5681	243.47.71	12204.05	203.918
x_7	-220	-22.1378	-1.44147	-33.9379	16.82401	12204.05	62482.95	1157.546
x_8	-7.95404	-0.33959	0.79948	-2.8057	0.735645	203.918	1157.546	155.748

يمثل الجدول (2) تقدير مصفوفة التباين والتباين المشترك والموضحة في المعادلة (5) فالأعداد الواقعة علي القطر الرئيسي هي تقديرات التباين للمتغيرات المستقلة من x_1 إلي x_8 ,أما الأعداد أعلي القطر الرئيسي تشير إلي تقديرات التباين المشترك بين المتغيرات المستقة .

وبناءا علي النتائج السابقة للحصول علي نتيجة متجه متوسطات المجتمع الأول والثاني ومصفوفة التغايرلدالة التمايز للمشاهدات المكتملة في المعادلة (25) تكون قيمة $W_X = 0.39872$

4-2 نتائج دالة التمايز الثانية

$$W_y = \left(\bar{Y}^{(1)} - \bar{Y}^{(2)}\right)' S_{yy}^{-1}\left[Y - \frac{1}{2}\left(\bar{Y}^{(1)} + \bar{Y}^{(2)}\right)\right], \qquad (26)$$

جدول (3)

الوسط الحسابي لمتغيرات الدالة الثانية

Mean		المجتمع الأول	المجتمع الثاني
	Y	0	1
Y	x_1	12.63158	10.89362
	x_2	1.52632	1.531915
	x_3	0.97368	0.872340
	x_4	0.18421	0.36170
	x_5	2.23684	2.14894
	x_6	239.86842	336.38298

نلاحظ أن متوسط المتغيرين الثاني(النوع) والثالث(الغسيل الكلوي) والخامس(الاريثوبيوتين) متساويين في كلا المجتمعين أي أن ليس لهم تأثير في التمييز ، وأن متوسط المتغير السادس(جرعة الارثوبيوتين) والرابع(الحديد) في المجتمع التاني أكبر من المجتمع الاول، بمعني أنه كلما زادت قيم هذين المتغيرين كلما أدي ذلك الي تحسين نسبة الهيموجلوبين, ومتوسط المتغير الاول(العمر) في المجتمع الاول أكبر من المجتمع الثاني .

جدول(4)

مصفوفة التغاير لدالة التمايز الثانية

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	14.26277	-0.31292	0.566184	0.131174	0.211572	6.41654
x_2	-0.31292	0.25513	-0.0275	0.01534	0.030599	-13.6981
x_3	0.566184	-0.0275	0.074792	0.028367	0.01362	11.54417
x_4	0.131174	0.01534	0.028367	0.199537	0.009761	2.559162
x_5	0.211572	0.030599	0.01362	0.009761	0.443685	-7.27014
x_6	6.41654	-13.6981	11.54417	2.559162	-7.27014	28336.26

يمثل الجدول (4) تقدير مصفوفة التباين والتباين المشترك والموضحة في المعادلة (7) فالأعداد الواقعة علي القطر الرئيسي هي تقديرات التباين للمتغيرات المستقلة من x_1 إلي x_6, أما الأعداد أعلي القطر الرئيسي تشير إلي تقديرات التباين المشترك

بين المتغيرات المستقة .وبناءا علي النتائج السابقة من خلال الحصول علي نتيجة متجه متوسطات المجتمع الأول والثاني ومصفوفة التغاير لدالة التمايز للمشاهدات الغير مكتملة في المعادلة (26) تكون قيمة $W_Y = 0.23$

تم الحصول علي قيمة C= 0.5234 من خلال المعادلة (9)

وبما أن ($W_X = (W_Y = 0.23)$ (C= 0.5234 0.39872)

إذا

$W_c = 0.3183$ $W_c = cW_x + (1-c)W_y,\ 0 \le c \le 1$-

3-4 نتائج أسلوب MLE

جدول رقم 5الوسط الحسابي للمتغيرات

Mean		المجتمع لأول	المجتمع لثاني
	Y	0	1
Y	x_1	12.63158	10.89362
	x_2	1.526316	1.531915
	x_3	0.973684	0.87234
	x_4	0.184211	0.361702
	x_5	2.236842	2.148936
	x_6	239.8684	336.383
Z	x_7	174.8756	194.3681
	x_8	33.4671	27.6812

نلاحظ أن متوسط المتغير الثاني(النوع) والخامس(الارثوبيوتين) متساويين في كلا المجتمعين، أي أن ليس لهم تأثير في التمييز وأن متوسط المتغير السادس(جرعة الارثوبيوتين) والرابع(الحديد) في المجتمع الثاني أكبر من المجتمع الاول وهذا يعني أنه كلما زادت قيم تلك المتغيرات كلما أدي ذلك إلي تحسين نسبة الهيموجلوبين، ومتوسط المتغيريين الاول(العمر) والثامن(تشبع الترانسفرين) في المجتمع الاول أكبر من المجتمع الثاني

جدول (6)

مصفوفة التغاير المشترك

$\hat{\Sigma}$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	13.92718	-0.30555	0.552862	0.128088	0.206594	6.265562	-212.206	-10.6421
x_2	-0.30555	0.249127	-0.02685	0.14979	0.29879	-13.3758	22.7962	-0.54455
x_3	0.552862	-0.02685	0.073032	0.027699	0.0133	11.27254	-2.81733	-1.14391
x_4	0.128088	0.14979	0.027699	0.194842	0.009332	2.498946	-31.7534	-2.78061
x_5	0.206594	0.29879	0.0133	0.009332	0.433246	-7.09996	26.60733	1.147036
x_6	6.265562	-13.3758	11.27254	2.498946	-7.09996	27669.52	11865.31	59.43154
x_7	-212.206	-22.7962	-2.81733	-31.7534	26.60733	11865.31	5.96E+10	2.49E+09
x_8	-10.6421	-0.54455	-1.14391	-2.78061	1.147036	59.43154	7.43E+10	31273100

يمثل الجدول (6) تقدير مصفوفة التغاير المشترك فالأعداد الواقعة علي القطر الرئيسي هي تقديرات التباين للمتغيرات المستقلة من x_1 إلي x_8, أما الأعداد أعلي القطر الرئيسي تشير إلي تقديرات التباين المشترك بين المتغيرات المستقة .

4-4 المقارنة بين أسلوب C-H وMLE

نحن الان اننا بصدد نموذجين نود المفاضلة بينهم ومقارنة كل من الأسلوبين لمعرفة أيهما أكثر كفاءة لتصنيف بيانات تحتوي علي فقد علي وتيرة واحدة وتم الاعتماد علي EERD الموضح في المعادلة (22) لتقييم كلا الأسلوبين لمعرفة أيهم أفضل في التصنيف ونجد أنه إذا كان قيمة EERD بقيمة سالبة فإن أسلوب C-Hأفضل من MLE والعكس صحيح. ولحساب $\widehat{EERD_{Boot}}$ تم توليد 10000 مفردة من العينة.

معلمات البوتستراب لمعلمات التوزيع الطبيعي متعدد المتغيرات حيث أن معلمات التوزيع الطبيعي متعدد المتغيرات(المتوسط - التباين) إلي MLEs هي كالتالي:

جدول(7)

Mean		المجتمع الأول	المجتمع الثاني
	Y	0	1
Y	x_1	12.46554	10.8459
	x_2	1.58452	1.58239
	x_3	0.98945	0.89583
	x_4	0.17954	0.34564
	x_5	2.30942	2.0421
	x_6	240.042	337.1299
Z	x_7	173.9453	195.3453
	x_8	33.9745	28.96465

يوضح جدول (7) تقدير معلمة المتوسط لجميع متغيرات الدراسة لكلا من المجتمع الأول والثاني

ومن خلال الجدول نلاحظ أن متوسطات كل من المتغير الثاني(النوع) والثالث(الغسيل الكلوي) والخامس(الاريثوبيوتين) تقريبا متساوين في كلا المجتمعين، اما بالنسبة لمتوسطي المتغيرين السادس(جرعة الاريثوبيوتين) والسابع(الفيريتين) في المجتمع الثاني أكبر من المجتمع الاول، وهذا يعني أنه كلما زادت قيم تلك المتغيرات كلما أدي ذلك إلي تحسين نسبة الهيموجلوبين وأخيرا متوسط المتغير الثامن(تشبع الترانسفرين) للمجتمع الاول أكبر من المجتمع الثاني.

جدول(8)

$\hat{\Sigma}$	X1	X2	X3	X4	X5	X6	X7	X8
X1	13.0465	-0.30125	0.5126	0.124920	0.295432	6.925142	-210.945	-11.523
X2	-0.3150	0.24126	-0.01965	0.019432	0.02093	-12.91374	-21.7510	-0.5569
X3	0.5404	-0.02826	0.02154	0.026794	0.01987	11.488218	-2.08939	-1.14652
X4	0.12192	0.013985	0.02158	0.18412	0.010091	2.41254	-31.6429	-2.78529
X5	0.21890	0.02493	0.01284	0.049529	0.43456	-7.41996	25.42286	1.109456
X6	6.249602	-13.1752	10.21569	2.5018	-7.59412	27670.42	11870.96	58.93218
X7	-211.952	-22.1832	-2.21619	-31.296	25.5923	11800.96	0.9E+10	0.03E+08
X8	-10.0451	-0.53285	-1.21920	-2.5681	1.90544	58.12389	0.03E+08	3125919

يوضح جدول (8) تقدير معلمة التباين المشترك لجميع المتغيرات محل الدراسة , فالأعداد الواقعة علي القطر الرئيسي هي تقديرات التباين للمتغيرات المستقلة من x_1 إلي x_8 ,أما الأعداد أعلي القطر الرئيسي تشير إلي تقديرات التباين المشترك بين المتغيرات المستقة .

وبعد ذلك تم الحصول علي قيمة مقدر البوتستراب لمعدل الخطأ المتوقع في المعادلة (24) - $\widehat{EERD_{Boot}}$= 0.0985931 بخطأ معياري $\widehat{s.e.EERD_{Boot}}$ = 0.000957

نلاحظ أنه إشارة $\widehat{EERD_{Boot}}$ سالبة وهذا يعني أن أسلوب C-H أفضل من MLE في تصنيف البيانات التي تحتوي علي فقد في شكل انقطاع (فقد علي وتيرة واحدة).

النتائج :

يعتبر تحليل التمايز أحد أهم أشكال النماذج الإحصائية التي تستخدم في تصنيف البيانات ويتعلق هذا البحث بدراسة كل من أسلوب C-Hو MLE وتحديد أفضلهم استخداما في تصنيف البيانات بحيث يحقق أفضلهم أعلي معدل دقة تصنيف وتم تطبيق كل منهما علي بيانات مرضي الفشل الكلوي بمستشفي الأطفال الجامعي بالمنصورة ومن ثم تم استخلاص النتائج التالية

ويمكن تلخيص النتائج الدراسية فيما يلي:

1. تم تطبيق التحليل التمييزي بأسلوبي C-H و MLE علي بيانات مرضي الفشل الكلوي والمقارنة بينهم من حيث أيهم أفضل في دقة التصنيف وتم استخدام أسلوب مقدر البوتستراب لمعدل الخطأ

2. أثبت التطبيق العلمي أن أسلوب C-H أفضل في التصنيف من أسلوب MLE لأنه يتميز بخطأ
3. تصنيف أقل وذلك استنادا علي قيمة $\widehat{EERD_{Boot}}$ = -0.0985931
4. أهم المتغيرات التي تساعد علي تحسين نسبة الهيموجلوبين في الدم هي جرعة الاريثوبيوتين, Iron و Ferritin

التوصيات

اعتمادا علي النتائج التي تم التوصل إليها يمكن إيجاز أهم توصيات الدراسة فيما يلي:

1. استخدام أسلوب C-H Classifier المعرف باسم أسلوب التوليفة الخطية أو الأسلوب المدمج في تصنيف البيانات التي مشاهداتها فيها فقد علي وتيرة واحدة.
2. توصي الباحثة بزيادة مساحة المقارنة بين أساليب البوتستراب في حساب معدل الخطأ المتوقع للتصنيف

المراجع

أولا: المراجع العربية

[1] أحمد. أحمد شمس الدين (2018) تقدير معلمات نموذج الانحدار الخطي البسيط باستخدام طريقة البوتستراب في حالة عدم ثبات التباين بنمطي الدالة التربيعية والجذرية. مجلة كلية التجارة للبحوث العلمية, جامعة الاسكندرية.

[2] صبري. حنين ناجي (2015) المداخل البديلة للتعامل مع مشاكل القيم المفقودة في البيانات الطولية. رسالة دكتوراة في الإحصاء التطبيقي كلية التجارة, جامعة المنصورة.

[3] مصطفي. مها وائل البكري (2014) مقارنة النماذج الخطية والمختلطة لتحليل التمايز في تصنيف البيانات وذللك بالتطبيق علي مرضي حصوات الكلي. رسالة ماجستيرفي الإحصاء التطبيقي, كلية التجارة, جامعة المنصورة.

ثانيا: المراجع الإنجليزية

[4] Anderson, T.W. and Olkin, I. (1985) Maximum-Likelihood Estimation of the Parameters of a Multivariate Normal Distribution. Linear Algebra and Its Applications, 70, 147-171.

[5] Batsidis, A. and Zografos, K. (2006) Discrimination of Observations into one of two Elliptic Populations

Based on Monotone Traning Samples, Metrika, 64: 221–241

[6] Chung, H.-C. and Han, C.-P. (2000) Discriminant Analysis When a Block of Observations Is Missing. *Annals of the* Institute of Statistical Mathematics, 52, 544-556

[7] Chung, H-C. and Han, C-P. (2009) Bootstrap confidence intervals for classification error rate in circular models when a block of observations is missing Journal of the Korean data and information science society 24(4), 757-764

[8] Chung, H-C. and Han, C-P. (2013) Conditional bootstrap confidence intervals for classification error rate when a block of observations is missing. Journal of the Korean data and information science society 24(1), 189-200.

[9] Hocking, R.R. and Smith, W.B. (2000) Estimation of Parameters in the Multivariate Normal Distribution with Missing Observations. Journal of the American Statistical Association, No. 63, 159-173.

[10]Young, P.D., Young, D.M. and Ounpraseuth, S.T.(2016) A Comparison of Two Linear Discriminant Analysis Methods That Use Block Monotone Missing Training Data. Open Journal of Statistic, 6, 172-185