جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة كربلاء
كلية الإدارة والاقتصاد
قسم الإحصاء

دراسة تطبيقية لتحليل التمايز في بعض النماذج الإحصائية

رسالة مقدمة إلى

مجلس كلية الإدارة والاقتصاد في جامعة كربلاء
وهي جزء من متطلبات نيل درجة الماجستير في علوم الإحصاء

من قبل

مريم مهدي عناد الخزعلي

بأشراف

أ.م.د شروق عبد الرضا سعيد السباح

1439هـ

2018 م

بسم الله الرحمن الرحيم

﴿ ويسألونك عن الروح قل الروح من أمر ربي وما

أوتيتم من العلم إلا قليلا ﴾

صدق الله العلي العظيم

(الاسراء/ ٨٥)

الاهـــداء

إلى الذي أُرسل لهداية البشرية . . ونور وجههِ قمر تدور حوله الأقمار

"رسولنا محمد (ﷺ)"

إلى من أَذهب الله عنهم الرجس وطهرهم تطهيرا . .

"آل البيت (عليهم السلام)"

إلى اللّوحة الرائعة التي رُسمت بدماء طاهرة ولونتها تضحيات كبيرة . .

"أبطال الحشد الشعبي"

إلى النورين اللذين أنارا دربي بدعائهما أطال الله اعمارهما وأعانهما على طاعتهِ . .

"أبي وأمي"

إلى الذين بسمتهم تطل علي باستمرار كأنهم ورود وأشجار . .

"أخوتي وأخواتي"

إلى من لم تبخل علي بالمشورة والمساعدة . .

"مشرفتي الفاضلة"

إلى الذين أرى علمهم وسط الليل نهار . .

"أساتذتي الأعزاء"

مريم

﴿ شكر وتقدير ﴾

الحمد لله ربّ العالمين والصلاة والسلام على سيدنا محمد وعلى آله الطيبين الطاهرين، أنوار الهدى وأعلام الدين.

يسرني في البدء أن اتقدم بالشكر والامتنان الى الدكتورة الفاضلة **(شروق عبد الرضا سعيد)** التي قامت بالإشراف على هذا البحث بحسن توجيهاتها وارشاداتها القيمة وملاحظاتها الرصينة.

وأتقدم بالشكر الجزيل إلى عائلة قسم الاحصاء أساتذتي الكرام لما قدموه لي من علم ومعرفة طوال مدّة دراستي الأولية والعليا كما أشكر زملائي في الدراسة الاعزاء.

ويسرني ان اتقدم بالشكر والتقدير للأستاذين الفاضلين "**الخبير اللغوي**" و "**الخبير العلمي**" لتفضلهم بقبول تدقيق الرسالة جزاهما الله خير الجزاء.

كما اتقدم بالشكر الجزيل إلى السادة رئيس وأعضاء لجنة المناقشة لتفضلهم بقبول مناقشة هذه الرسالة متمنياً أن تنال استحسانهم ورضاهم.

واتقدم بالشكر إلى منتسبوا **(مستشفى مرجان/محافظة بابل)** عينة البحث.

واختم شكري وتقديري للذين اسهموا في توفير الدعم المادي والمعنوي لي طوال مدّة الدراسة والدي ووالدتي وإلى اخوتي واخواتي الذين تحملوا أعباء الدراسة وصعوبتها.

الباحثة

المستخلص

ان التحليل التمييزي أسلوب إحصائي يعتمد على عينة من المشاهدات المسحوبة من مجتمعات معلومة لبناء قاعدة يمكنها المساعدة مستقبلاً في تعيين المجتمع الذي تنتمي إليه المشاهدات الجديدة اعتمادا على متغيرات ذات صفات تمييزية، وطبقت كحالة خاصة لتشخيص ثلاثة أنواع من الأورام السرطانية وهي اورام الثدي والعمود الفقري (سرطان العظم) والجهاز التنفسي (سرطان الرئة) لكثرة شيوعها في مجتمعنا حالياً.

ولغرض دراسة هذا الموضوع تم تسجيل قيم المشاهدات لخمسة متغيرات وهي (الجنس، العمر، مهنة المريض، حالة خروج المريض، مدّة بقاء المريض في المستشفى) من (عينة عشوائية بسيطة) بحجم **270** مريض درست في أربع نماذج احصائية وهي (دالة التمييز الخطية) وغير الخطية (دالة التمييز التربيعية ودالة التمييز اللوجستية) والانموذج اللامعلمي (دالة الجار الاقرب)، واستخرجت النتائج وفق البرنامج الاحصائي **Stata** فضلاً عن الحزمة الاحصائية الجاهزة **SPSS** ، وذلك لتصنيف المشاهدات الى المجموعة التي تنتمي اليها، وتم استعمال معيار خطأ التصنيف كمعيار للمقارنة بين النماذج الاربعة، بهدف الوصول إلى أفضل انموذج باقل خطأ تصنيفي ممكن.

وتم الوصول الى ان انموذج الجار الأقرب اكثر تفوق من بقية الدوال من حيث قلة نسبة التصنيف الخاطئ بالمقارنة مع بقية النماذج.

قائمة المحتويات

قائمة الجداول

قائمة الاشكال

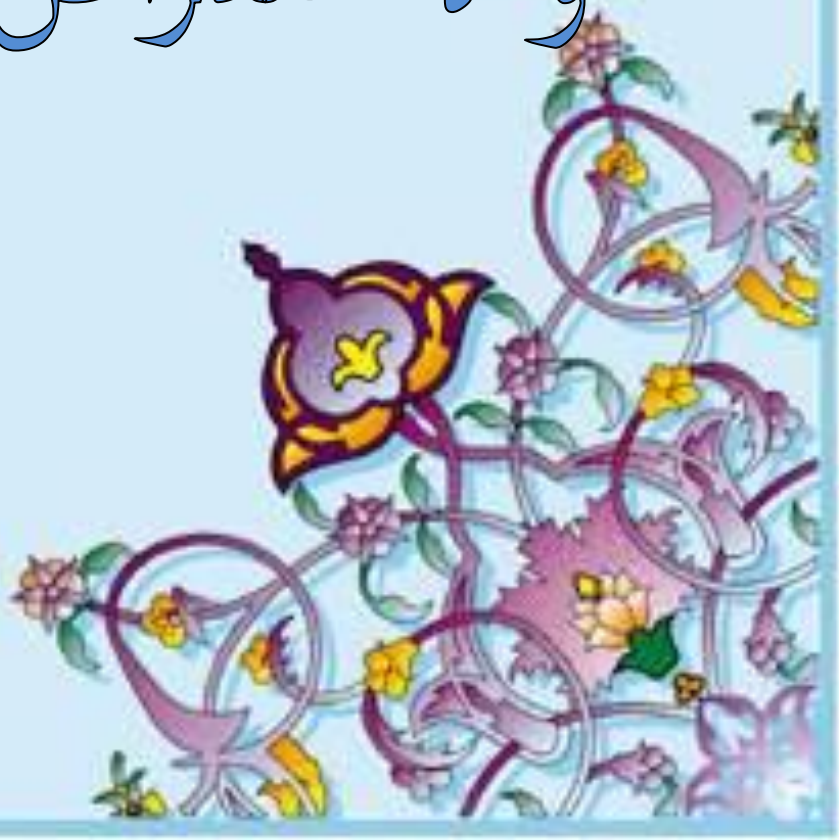
قائمة الرموز

الرمز	مفهومه
Y^*	القيمة التمييزية المعيارية.
x_t	المتغيرات التمييزية المعيارية.
a_t	المعاملات التمييزية المعيارية.
Y	القيمة التمييزية غير المعيارية.
Z_t	المتغيرات التمييزية غير المعيارية.
a_t^*	المعاملات التمييزية غير المعيارية.
t	عدد المتغيرات التمييزية المعيارية المكونة للمعادلة التمييزية.
c	ثابت
$\overline{x}_1, \overline{x}_2$	تقدير الإمكان الأعظم لـ $\underline{\mu}_1, \underline{\mu}_2$.
W, A	مصفوفتا التباين والتباين المشترك داخل المجموعات على التوالي.
λ	الجذور المميزة للمصفوفة $W^{-1}A$.
$\underline{X}$	المتجهات المميزة للمصفوفة $W^{-1}A$.
I	مصفوفة الوحدة.
Λ	مقياس ويلكس لامدا
W	مصفوفة التباين والتباين المشترك داخل المجموعات .
B	مصفوفة التباين والتباين المشترك بين المجموعات .
T	مصفوفة التباين والتباين المشترك الكلي للمجموعات .
S	مصفوفة التباين والتباين المشترك التقديرية
S_i	تباين العينة وهو تقدير غير متحيز لـ Σ_i إذ $(i = 1, 2, \dots . k)$
k	عدد المجاميع .
p	عدد المتغيرات التوضيحية
n	حجم العينة الكلي
$\overline{Y}_1, \overline{Y}_2$	تمثل متوسطي المجموعتين الأولى والثانية.
n_i	حجم العينة للمجموعة i.
P_i	هو احتمال الاستجابة عندما $(Y = 1)$
$1 - P_i$	احتمال عدم الاستجابة عندما $(Y = 0)$
β	قيم المعلمات المراد تقديرها.
L_i	يمثل تحويل $(logt)$ للاحتمال (P_i).
v	حجم المنطقة حول المشاهدة x .
k	تمثل عدد الحالات داخل المنطقة v
Φ	تمثل دالة التوزيع الطبيعي القياسي .
δ	تمثل جذر مقياس مسافة مهانلوبس D^2 .
P	يمثل احتمال خطأ التصنيف
P_1	تمثل نسبة التصنيف الصحيح للمجموعة 1.
P_2	تمثل نسبة التصنيف الصحيح للمجموعة 2.

الفصل الأول

المقدمة

والاستعراض المرجعي



الفصل الأول

المقدمة والاستعراض المرجعي

1.1 : مقدمة *Introduction*

يستعمل تحليل التمايز **(*Discriminate Analysis*)** 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 .

يستفاد من تحليل التمايز في التعرف على المتغيرات المساهمة في التصنيف وكذلك في التنبؤ الذي يعطي تقدير شامل لكفاءة قواعد التصنيف .

استخدمت 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.

افتراضات دالة التمييز الخطية:-

1. إن يكون متجه المتغيرات التوضيحية ذا توزيع طبيعي متعدد المتغيرات.
2. مصفوفة التباين والتباين المشترك متساوية ، ومتجهات متوسطات مختلفة في كل مجموعة من المجاميع.
3. يتم اختيار العينة عشوائيا.
4. عدم وجود ازدواج خطي بين المتغيرات التوضيحية.

ولكن أحيانا نواجه اختلال في بعض الفرضيات وبالتالي تفقد دالة التمييز الخطية بعض خواصها المثلى. فعند عدم تحقق ثبات مصفوفة التباين والتباين المشترك يكون من الضروري استعمال دالة التمييز التربيعية **(*Quadratic Discriminate*)** ، فضلاً عن ذلك ان دالة التمييز الخطية تفقد خواصها المثلى عندما تكون جميع او بعض المتغيرات الاضافية ذات طبيعة ثنائية او مصنفة فتكون دالة التمييز اللوجستية **(*Logistic Discrimination*)** البديل الامثل للتمييز الطبيعي ، وعند عدم توافر شرط التوزيع الطبيعي وحجم العينة صغير نلجأ إلى استعمال دالة تصنيف لامعلمية (الجار الاقرب) .

أما عملية التصنيف **(*Classification*)** هي العملية اللاحقة بعد تكوين الدالة المميزة حيث يتم الاعتماد على هذه الدالة بالتنبؤ وتصنيف المفردة الجديدة لإحدى المجموعات قيد الدراسة بأقل خطأ تصنيف ممكن.

2.1 : مشكلة البحث *Research Importanc*

ايجاد افضل انموذج وبأقل خطأ ممكن لبيانات يكون فيها متغير الاستجابة من النوع الفئوي باستعمال بعض دوال التمييز لغرض التصنيف بأقل خطأ ممكن.

3.1 :منهجية البحث *Research Methodology*

تم العمل على منهج العمل الإستقرائي وفيه نبدأ بملاحظة المشكلة تم وضع الفروض لها ومن ثم اختبارها، اي من خلال سحب عينة عشوائية بسيطة وتعميم نتائجها على المجتمع .

4.1 : هدف البحث *Research Aim*

يتضمن هذا البحث المقارنة بين بعض اساليب التصنيف (دالة التمييز الخطية ، دالة التمييز التربيعية ، دالة التمييز اللوجستية والدالة الجار الاقرب) لمعرفة مدى قابليتهم لتصنيف بيانات بعض الاورام السرطانية بأقل احتمال خطأ تصنيف ممكن.

ولغرض تحقيق اهداف البحث فقد تم تقسيمه إلى اربعة فصول وكالاتي :

الفصل الأول : تضمن المقدمة ومشكلة البحث ،منهجية البحث، هدف البحث والاستعراض المرجعي التي كانت من عام **1968** ولغاية عام **2017**.

الفصل الثاني : تضمن الجانب النظري وفيه تم عرض مفصل عن الدوال التمييزية الخطية والدالة التمييزية التربيعية واللوجستية ودالة التصنيف اللامعلمية (دالة الجار الاقرب) اضف إلى ذلك بعض الاختبارات المهمة التي تستخدم في تحليل التمييز.

الفصل الثالث : تضمن الجانب التطبيقي وقد اشتمل على مبحثين:
المبحث الاول: تناول الجانب الطبي وقد تضمن تعريفا مختصراً عن مرض السرطان وأنواعه والاعراض والعلامات التي تظهر على الشخص عند الاصابة بمرض السرطان وأخيراً تناول هذا المبحث بعض العلاجات التي تعطى للشخص المصاب بالمرض.
المبحث الثاني: قد تناول اسلوب اختيار العينة ووصف البيانات وتعريف متغيرات البحث وعرض المشكلة احصائيا والتحليل الاحصائي باستعمال برنامج **Stata** وكذلك استخدام الحزمة الإحصائية الجاهزة **SPSS** ، فضلاً عن نتائج التطبيق لعملية التمييز والتصنيف.

الفصل الرابع: تضمن أهم الاستنتاجات والتوصيات التي تم الوصول إليها.

5.1 : الاستعراض المرجعي *Reference Review*

لقد تناول العديد من الباحثين موضوع تحليل التمايز وفيما يأتي عرض لبعض ما كتب عن هذا الموضوع من أبحاث.

في عام **(1968)** قام (***Altman***)[33] باستعمال التحليل التمييزي متعدد المتغيرات لتحليل النسب المالية المستخرجة من القوائم المالية للشركات وقام بإجراء دراسة على 33 شركة غير مفلسة و33 شركة مفلسة وكلا الصنفين متساويين في نوع الصناعة وقام بتحليل 22 نسبة مالية استخرجت من القوائم المالية للشركات وكان الانموذج قادرا على التنبؤ بفشل الشركات قبل حصوله بسنتين بدقة بلغت 83% .

وفي العامين **(1972 و 1975)** قام العالم (***Anderson***)[35] بمناقشة إمكانية التمييز باستخدام انموذج Logistic وذلك عندما تكون الدالة غير خطية وعالج الحالة بطريقة معالجة المتغيرات الغير خطية نفسها في حالة الانحدار الخطي.

وفي العام **(1976)** قام الباحثون (***Broffitt, Hogg and Randles***)[40] باستعمال الرتب بدل استعمال القيم الأصلية لغرض معالجة مشكلة الابتعاد عن التوزيع الطبيعي في التمييز الجزئي لمجتمعين. وتم توضيحها بالتطبيق وقارنت هذه الطريقة مع طريقة منطقة التسامح (*Tolerance*) ، باستعمال طريقة مونت كارلو . وتوصلوا إلى إن طريقة الرتب هي الأفضل في السيطرة على احتمالات خطأ التصنيف أو المحافظة نسبياً على احتمالات صغيرة لخطأ التصنيف.

وفي عام **(1977)** كان للتحليل التمييزي دور بارز في البحوث والدراسات في العراق، حيث تناول **(شومان)**[21] تحليل التمايز واستخدمه في إيجاد صيغة ملائمة لتصنيف الطلبة في المرحلة الإعدادية بفروعها المختلفة (العلمي، الأدبي، التجاري، الصناعي، الزراعي) .

وفي العام **(1978)** قام الباحثون (***Brpffitt, Randls, Ramberg, and Hogg***)[41] بمقارنة أسلوب فيشر لإيجاد الدالة المميزة الخطية والطرق الجديدة والتي تعتمد على تقديرات المتوسطات ومصفوفات التباين والتباين المشترك لدوال التمييز الخطية والتربيعية ثم مقارنة هذه الطرق اعتمادا على نتائج مونت كارلو والتي تعطي اصغر احتمال لخطأ التصنيف.

وفي العام **(1980)** درس الباحث (***NG***)[60] طريقة الرتب للتمييز بين مجتمعين أو أكثر حيث استعمل دالة التمييز الخطية والتربيعية على بيانات لا تتبع التوزيع الطبيعي . وقد نوقشت مزايا ومساوئ الطريقة ، وتم تطبيق مونت كارلو لإظهار أداة طرق التمييز ومناقشة أربع طرق للرتب وهي ، طريقة p للحد الأدنى، طريقة الدمج، طريقة الرتب الدنيا و طريقة المسافة الدنيا .

وفي عام **(1981)** قام الباحث (***Titteringtin***)[69] بمناقشة تطبيق أساليب التمييز في بحث تجريبي لمجموعة كبيرة من البيانات التي تتضمن مرضى تعرضوا لإصابات شديدة بالرأس ، بهدف التنبؤ بحالة المريض المستقبلية ، وكان من بين الأساليب التي استخدمت في مقارنة أسلوب التمييز الطبيعي الخطي والتربيعي واللوجستي ، وقد توصلوا إلى نتائج مشابهة جدا إلى الطرق

المختلفة من حيث الأداء التشخيصي . وبينوا إن اختبار أي من هذه الطرق يستند إلى مدى ملائمة المرونة والحضور وان جميع هذه الخصائص تعود لانموذج انحدار اللوجستي .

وفي عام **(1984)** وضح **(*Begg & Gray*)**[37] إن من الممكن إجراء تحليل انحدار اللوجستي الثنائي بدل انموذج الانحدار المتعدد وبينوا مدى ملائمة هذا الأسلوب . فقد قاموا باشتقاق التوزيع ألتقاربي للتقديرات الناتجة عن الطريقة الانفرادية وقارنوا حساب الكفاءة التقاربية الانفرادية مع التحليل المتعدد المجاميع . واستنتجوا إن الكفاءة التقاربية النسبية للتقديرات الانفرادية كانت مرتفعة وكذلك الكفاءة التقاربية النسبية لتقديرات الاحتمالات اللاحقة إلا إنها تقل بعض الشيء في حالة اختبار الفرضيات المركبة للمعالم الناتجة من مجاميع مختلفة ، كذلك تقل الكفاءة النسبية عندما يكون التكرار النسبي لمجموعة الأساس صغيراً؛ لذلك أوصوا باستعمال المجموعة ذات احتمال ظهور الاكبر كمجموعة أساس في حالة التقدير الانفرادي.

وفي عام **(1987)** قام كل من **(*Bull & Donner*)**[42] بمقارنة بين انحدار اللوجستي والتمييز الطبيعي إلى أكثر من مجموعتي استجابة . وفي كلا الأسلوبين تم اشتقاق توزيع العينة الكبيرة لتقدير الإمكان الأعظم للمعالم التي تمثل لوغاريتم نسب الإمكان باستعمال طرائق حساب المصفوفات . بعد ذلك تم تعريف كفاءة التقدير النسبية بأنها معكوس نسبة التباين . وقد بينوا أن مقياس Efron للكفاءة هو حالة خاصة من المقياس الذي تم اشتقاقه من قبلهم. وتم تقييم النتائج على حالتين تكون فيها الكفاءة النسبية معرفة من خلال معالم تكرارات مجاميع الاستجابة ولوغاريتم نسب الإمكان الأعظم حيث عرضت النتائج على شكل خطوط ذات بعدين . وعند إيجاد تقدير الكفاءة النسبية تم التركيز على علاقتها بقيمة لوغاريتم نسبة الإمكان الأعظم لثلاث مجاميع استجابة في حالة متغيرين توضيحيين والاهتمام بصورة ثانوية بتأثير تكرارات مجموعة الاستجابة على الكفاءة النسبية .

وفي عام **(1988)** قام الباحثان **(الزوبعي، المشهداني)**[10] باستعمال تحليل التمييز لتصنيف الطلبة الناجحين في الصف الرابع الثانوي إلى الفرعين العلمي والأدبي واستعملا طريقة *Stepwise* لتحديد المتغيرات المعنوية.

وفي عام **(1991)** قامت الباحثة **(حميد)**[17] بدراسة النماذج التمييزية واختيار الانموذج الأنسب بين عدد من النماذج هي الخطية والتربيعية واللوجستية من خلال تشخيص بعض الأورام السرطانية وتم الوصول إلى إن الانموذج اللوجستي أفضل انموذج.

وفي العام **(1994)** قام الباحث **(*Al-Iathary*)**[31] باستعمال خمسة أساليب في تحليل التمييز وهي (دالة التمييز الخطية DLF، دالة التمييز التربيعية،QDF دالة الرتبة R والترتيب Q ودالة الرتبة والترتيب RQ) لأشخاص مصابين باضطرابات نفسية وتوصل إلى إن دالة التمييز التربيعية هي أكثر دقة في التنبؤ بالمجاميع المختلفة.

وفي عام **(1995)** قام الدكتور **(رشيد)**[18] بصياغة أفضل انموذج لتشخيص الحالة الصحية للطفل الخديج من خلال المقارنة بين عدد من دوال التمييز ، وتوصل إلى إن دالة انحدار اللوجستي هي الأكثر دقة حيث كانت نسبة التصنيف الصحيح هي الأعلى وكذلك سهولة تطبيقها.

وفي عام **(1996)** قامت الباحثة **(البلداوي)**[6] بمقارنة انموذجي التمييز الخطي واللوجستي وتوسيع مجال مقارنة الأساليب الحصينة من هذه الدوال من الحالة الثنائية إلى حالة متعددة المجاميع لبيانات لا تتوزع طبيعيا وتم تطبيقها على المرضى المصابين بإمراض الكبد المزمن وتم الوصول إلى إن تقديرات الانموذج اللوجستي يمكن الاستفادة منها فقط في عملية التصنيف وليس تقديرات الإمكان الأعظم وكذلك استعمال تصحيح التمييز يكون أكثر أهمية في حالة النماذج المتعددة المجاميع من الحالة الثنائية لكثرة عدد المعلمات المقدرة حيث يؤدي إلى تخفيض نسبة المجاميع في العينة.

وفي عام **(1997)** قامت الباحثة **(البكري)**[4] بمقارنة بعض طرائق التقدير للدالة المميزة ومن ضمن هذه الطرق الحصينة ولمجموعتين إذ طبقت للتمييز ما بين الطلبة من حيث درجة التفوق، أي التنبؤ بمستوى الطالب هل هو متفوق أم غير متفوق واستنتجت إن استعمال الدالة المميزة التربيعية هي الأفضل في حالة عدم تساوي التباينات وفي كل الطرق سواء المعلمية واللامعلمية.

وفي عام **(1997)** قام الباحث **(احمد)**[3] بدراسة إحصائية تمييزية حول تسوس الأسنان لدى الأطفال وتوصل الباحث من خلال الدالة التمييزية إلى أمكانية التمييز بين حالات التسوس من عدمه واتضح إن المتغيرات في الدراسة قادرة فعلا على التمييز بين مجموعة التسوس.

وفي عام **(1998)** قام الباحث **(السليماني)**[11] باستعمال الدالة المميزة لتشخيص حالات التهاب الأمعاء عند الأطفال الرضع واستعمل الدالة المميزة في حالة مجموعتين وفي حالة أكثر من مجموعتين إذ استعمل الباحث 20 متغير وتوصل إلى تشخيص المتغيرات ذات الأهمية المعنوية التي تساعد كثيرا على جمع البيانات في المستقبل إذ أنها ستقتصر على المعلومات المطلوبة للتحليل .

وفي عام **(1999)** قام الباحث **(حسين)**[15] باستعمال الدالة المميزة التربيعية في تمييز الأرقام العربية .حيث تضمنت تصميم منظومة الكترونية لتحليل أنماط الأرقام العربية، تطلب ذلك مناقشة تمييز الأنماط وعلاقته بالتحليل الإحصائي. مع عرض لتحويل فورير واشتقاق لدالة المميزة التربيعية المستخدمة في عملية التمييز .واختيار المتغيرات الداخلة في الدالة كذلك المعايير المستخدمة لقياس أداء الدالة المميزة باستخدام تقدير خطأ التصنيف الذي استخدم لأول مرة باعتماد طريقة التعويض. ولإثبات ذلك تم اعتماد عينة من 150 شخص كتبوا الأرقام العربية من (9-0) يدوياً وبذلك تم الحصول على (1250) نمط . أعطى الأسلوب الإحصائي للتمييز نتائج جيدة في تمييز الأرقام العربية من حيث خطأ التصنيف إذ أعطى 15 متغير وهو اقل عدد بين باقي الأساليب ووصل تقدير احتمال خطا التصنيف إلى الصفر . من خلال برامج كتبت بلغة C++.

وفي عام **(2000)** قدم (***Zhang***)[73] مسالة تمييز الأنماط (القوالب) الإحصائية باستعمال تحليل التمييز وقدم طريقتين رياضيتين من تحليل التمييز هما تحليل التمييز الخطي LDA وتحليل التمييز التربيعي QDA واثبت فعالية استخدام تمييز القوالب باستخدام طرائق التمييز في مجال الجينات البشرية.

وفي عام **(2001)** قام **(رشيد وآخرون)**[19] ببناء أفضل انموذج احتمالي لتشخيص بعض أمراض قرحة المعدة ولثلاثة مجاميع باستخدام دالة التمييز اللوجستية وعدة دوال تمييزية معينة تعرض لأول مرة إضافة لدوال غير معينة وتوصلوا إلى أن انموذج اللوجستي أعطى تفوقاً واضح في النسب المئوية للتصنيف الصحيح مما يعطيه الأهمية في اعتماده كأسلوب مساعد من ذوي الاختصاص في عملية تشخيص بعض أمراض قرحة المعدة .

وفي عام **(2003)** قدمت الباحثة **(صالح)**[22] بحث يتضمن استخدام التحليل المميز (التصنيفي) لتحديد أهم العوامل المؤثرة في تسرب ورسوب الطلبة في جميع المراحل الدراسية، وتوصلت إلى إن الدالة المميزة الخطية تعطي اقل احتمال خطأ تصنيف مقارنة بالدالة التربيعية .

وفي عام **(2004)** درس **(الحنيطي وآخرون)**[9] استخدم فيها أسلوب التحليل التمييزي لتمييز الأسر الفقيرة من غير الفقيرة في المناطق النائية التابعة لإقليم جنوب الأردن وذلك بأخذ بعض المتغيرات الخاصة بموضوع الدراسة والتي من خلالها تم التمييز بين الأسر.

وفي عام **(2005)** قدمت الباحثة **(حمودات)**[16] بحثاً حول استعمال الدالة التمييزية وطرق تحديد متغيراتها حيث يهدف البحث إلى تشخيص المتغيرات ذات الأهمية التصنيفية في الدالة التمييزية و استخدمت طريقة *Roy-Bose* واختبار t . وتم استعمال طرائق مقترحة لاختيار المتغيرات في هذا التحليل هي طريقة الانحدار الخطي وتشمل (أسلوب الاختيار الأمامي- أسلوب الحذف العكسي- طريقة الاختيار المتدرج)، وأخيرا طريقةً التحليل العاملي وطريقة المكونات الرئيسية.

وفي عام **(2006)** قدمت الباحثة **(عبد الكريم)**[25] بحث بعنوان **"استخدام الطرائق التمييزية الإحصائية لتشخيص بعض أمراض القلب"** تناول البحث التحليل المميز والذي يعد من الطرائق الإحصائية المهمة في تصنيف مفردة واحدة أو أكثر إلى أحد المجتمعات بالاعتماد على متغيرات ذات صفات تمييزية ، والتطرق إلى بعض الأنواع من الدوال التمييزية المتقدمة في هذا المجال وتطبيقها على نوعين من أمراض القلب لبناء انموذج احتمالي للتمييز بينها بالاعتماد على بعض الإعراض والصفات والمتغيرات المصابة للمرض ، وقد كانت انموذج اللوجستك ذو تفوق على بقية النماذج المستخدمة في البحث من حيث قلة نسبة التصنيف الخاطئ بالمقارنة مع بقية النماذج.

وفي عام **(2007)** قام الباحثان **(الجاعوني وغانم)**[7] باستعمال التحليل الإحصائي متعدد المتغيرات (تحليل التمايز) في (تصنيف وتوزيع الأسر داخل الهيكل الاقتصادي الاجتماعي في المجتمع) وتوصلا إلى مدى كفاءة دوال التمييز لتصنيف وتوزيع الأسر من خلال تقدير معدلات الخطأ واختبار معنوية الفروق بين المجتمعات للتأكد من معنوية تصنيف وتوزيع الأسر.

وفي عام **(2008)** قام كل من **(*Abbas and Azhar*)**[28] بدراسة اثنين من التقنيات الاحصائية (التحليل التمييزي والانحدار اللوجستي)، اذ استخدموا تراكيز من الزرنيخ ومحتويات المعادن الثقيلة من مصبات الانهار في ولاية بيانغ _ماليزيا وتوصلوا إلى ان نتائج انموذج الانحدار اللوجستي مقاربة إلى نتائج التحليل التمييزي.

وفي عام **(2009)** قدمت **(عبد الرزاق)**[24] بحثا بعنوان **(استخدام الدالة التمييز الخطية لتصنيف مرضى الثلاسيميا في مستشفى كركوك العام)** تم تطبيق دالة التمييز الخطية على 134 مريض بالثلاسيميا في مستشفى كركوك العام لغرض بناء منظومة معادلات لغرض التصنيف وتم الوصول إلى أن منظومة المعادلات قد كانت مقبولة بنسبة تصنيف صحيح تبلغ 76.1 %.

وفي عام **(2010)** قام الباحث **(الكاتب)**[12] بدراسة الدالة المميزة الخطية في حالة أكثر من مجموعتين حيث قام باختبار قابلية الدالة على التمييز بين الصور الرقمية ومن ثم إيجاد الدالة التي سيتم استخدامها في التمييز بين الصور الرقمية وفي المرحلة الأخيرة التعرف على أخطاء التصنيف الناتجة من تحليل التمييز.

وفي عام **(2011)** قامت الباحثة **(إحسان)**[2] بتقديم بحث حول استعمال الدالة المميزة للتنبؤ بنتيجة الطالب حيث تشمل الدراسة مرحليتين مرحلة التمييز ومرحلة التصنيف بهدف بناء أو صياغة قاعدة تشتق من الصفات التي تحملها مفردات مصنفة إلى مجموعتين أو أكثر لعينة معينة. وتوصلت الدراسة إلى إن المتغيرات التوضيحية جميعها ذات تأثير معنوي ، وكانت نسبة كفاءة التصنيف الصحيح للراسبين إلى مجاميع الراسبين 81% ونسبة كفاءة التصنيف الصحيح للناجحين 82% وهذه النسب تدعم أسلوب التحليل المميز في التنبؤ.

في عام **(2011)** تناول الباحثون **(*Mirakabad et al*)**[58] مجموعة من البيانات لدارسة الحالة الفسيولوجية للبشر مثل ضربات القلب وضغط الدم ودرجات حرارة الجسم للتنبؤ بها باستخدام طريقة الجار الأقرب (*K-NN*) ، وقد بينت الدراسة أن نتائج طريقة الجار الأقرب متوازية وقريبة من القيم الفعلية.

وفي عام **(2012)** قارن الباحث **(*Krieng*)**[54] بين الانحدار اللوجستي والتحليل التمييزي في تصنيف مجموعة من المصابين بأمراض سرطان الثدي وتوصل إلى ان نتائج الانحدار اللوجستي يعطي تصنيف اكفأ من التحليل التمييزي.

وفي عام **(2013)** قام الباحثون **(*Alkhatib et al*)**[32] في التنبؤ بأسعار الأسهم لعينة من ست شركات كبرى مدرجة في البورصة الأردنية لمساعدة المستثمرين وصناع القرار والإدارة والمستخدمين في اتخاذ القرارات الصحيحة للاستثمار باستعمال طريقة الجار الأقرب(K-NN) فكانت النتائج قوية و منطقية مع نسبة خطأ صغيرة، اما نتائج التنبؤ فكانت تتميز بالثقة ومتوازية تقريبا لأسعار الأسهم الفعلية.

وفي عام **(2014)** قامت الباحثة **(جبارة)**[14] بدراسة مقارنة بين التحليل التمييزي الخطي واحتمال الاستجابة في تصنيف البيانات وكان هدف الدراسة هو استعمال دالة التمييز الخطي كدالة تصنيف خطية وتصنيف البيانات بصيغتين ، الصيغة الأولى تتمثل بدالة التمييز الخطي، والصيغة الثانية تمثل دالة احتمال الاستجابة التي تم اشتقاقها لدالة التمييز الخطي . وتمت المقارنة بين هذه الصيغتين وفق معيار احتمال خطأ التصنيف.

وفي عام **(2015)** قام الباحث **(سليمان)**[20] بمقارنة بين التحليل التمييزي والانموذج اللوجستي الثنائي ونماذج الشبكات العصبية في تصنيف المشاهدات بالتطبيق على دراسة أهم العوامل الاقتصادية والاجتماعية المؤثرة على كفاية دخل الأسرة ، وقد توصل إلى إن طريقة

الشبكات العصبية الاصطناعية قد أعطت نسبة تصنيف أفضل من الانموذج اللوجستي وطريقة دالة التمييز.

وفي عام **(2016)** قام الباحثون **(عباس وآخرون)**[23] بنشر بحث بعنوان **"التصنيف وفقا لبعض القدرات البدنية الخاصة والأداء المهاري المركب والقياسات الجسمية للاعبي كرة السلة"** هدفت الدراسة إلى تصنيف لاعبي كرة السلة وفقا للقدرات البدنية والأداء المهاري والقياسات الجسمية وحسب مراكز اللعب المختلفة وقد استعمل الباحثون المنهج الوصفي كونه المنهج الملائم لحل مشكلة البحث ، وقد توصل الباحثون إلى عدة استنتاجات أهمها ، يضمن التصنيف الاختبار الصحيح لنوع المركز الذي يلاءم مواصفات اللاعب البدنية و المهارية والجسمية والذي يجعله متميز بأدائه.

وفي عام **(2017)** قام الباحثون **(البكري وآخرون)**[5] بنشر بحث بعنوان **"مقارنة بين انموذج الانحدار اللوجستي وانموذج التحليل المميز الخطي بأستعمال المركبات الرئيسية لبيانات البطالة لمحافظة بغداد"** اذ كان هدف الدراسة بيان القدرة التنبؤية الافضل بين انموذج الانحدار الوجستي والدالة المميزة الخطية بأستعمال البيانات الاصلية والمركبات الرئيسية لتقليص الابعاد بين المتغيرات لبيانات المسح الاجتماعي والاقتصادي للاسرة لمحافضة بغداد واتضح من خلال المقارنة ان انموذج الانحدار اللوجستي افضل من انموذج الدالة المميزة الخطية باستعمال البيانات الاصلية، اما بأستعمال المركبات الرئيسية كانت النتائج متساوية للأنموذج الانحدار اللوجستي والدالة المميزة الخطية، وان اداء انموذج الانحدار اللوجستي للبيانات الاصلية تقريبا متساوية من استعمال المركبات الرئيسية بينما اداء انموذج الدالة المميزة الخطية بأستعمال المركبات الرئيسية كان افضل من البيانات الاصلية.

الفصل الثاني

الجانب النظري

الفصل الثاني

الجانب النظري

1.2 : مقدمة *Introduction*

من المعلوم إن إحدى المشاكل الإحصائية المهمة هي تمييز المفردات إلى مجموعتين أو أكثر إذ تتركز هذه المشكلة في طريقة إيجاد الدوال المميزة وفق المتغيرات التي تم الحصول عليها بهدف تصنيف المفردات الجديدة المجهولة الانتماء إلى المجموعة الصحيحة. وعليه فالتمييز *Discriminant* هو التفرقة بين مجتمعات متداخلة أو متشابهة بنفس الخصائص أو الصفات.

يعد التمييز بين المشاهدات من الأساليب الشائعة الاستعمال وذلك لكثرة الظواهر التطبيقية التي يمكن أن يتم تحليلها من خلال أسلوب التمييز بين المشاهدات وتوجد عدة طرائق التي يمكن أن تستعمل للتمييز بين المشاهدات منها المعلمية مثل دالة التمييز الخطية ودالة التمييز التربيعية والدالة المختلطة ودالة التمييز اللوجستية والطرق اللامعلمية مثل دالة الجار الاقرب ودالة كيرنل ودالة الرتبة ، وغيرها من طرائق تحليل التمايز

وفي هذا البحث تم التركيز على الدوال التمييزية المعلمية (دالة التمييز الخطية، دالة التمييز التربيعية، دالة التمييز اللوجستية) و الدالة التمييزية اللامعلمية (دالة الجار الاقرب).

2.2 : 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.

ويعرف التحليل التمييزي هو أحد التقنيات الإحصائية المناسبة لتقدير علاقة بين المتغير التابع عندما يكون هذا المتغير متغيراً تصنيفياً (أسمي أو غير كمي) مثلاً أذا كان يتألف من مجموعتين (ذكرا و أنثى) أو أكثر من مجموعتين (عالياً ، متوسطاً أو منخفضاً) مع مجموعة من المتغيرات التوضيحية التنبؤية (المتغيرات الكمية) تستعمل في التنبؤ بعضوية المجموعة، فهو قادر على التعامل مع المجموعتين أو عدة مجموعات (ثلاثة أو أكثر). فيطلق على الاسلوب بالتحليل التمايز الثنائي عند تناول تصنيفين (مجموعتين) و اسلوب تحليل التمايز المتعدد عند تناول ثلاثة او اكثر من مجموعات [46].

ويهدف التحليل إلى الوصول إلى دالة التمييز (*Discriminant Function*) التي تعمل على تعظيم الفروق بين متوسط المجموعات وتقليل التشابه في أخطاء التصنيف في الوقت ذاته، وذلك من خلال إيجاد تجمعات خطية لمجموعة من المتغيرات[52].

3.2 : مراحل تحليل التمايز[51],[52] *Goals Discriminant Analysis*

1- التمييز *Discrimination* : وهو استعمال للمعلومات أو القياسات الموجودة في مجموعة المشاهدات المصنفة لبناء المصنف أو (قاعدة التصنيف) التي من شأنها أن تمييز بين الطبقات المحددة مسبقا بأكبر قدر ممكن وبالاعتماد على الميزات التفاضلية لتلك للمشاهدات.

2- التصنيف *Classification* : وهو عملية فرز أو التنبؤ بفئة المشاهدة الجديدة إلى واحدة فقط من الطبقات أو الفئات المحددة سابقا ، إذا كان مُعْطىً مجموعة من القياسات على مشاهدة جديدة غير مصنفة مسبقاً باستعمال المُصَنِّف أو قانون التصنيف.

4.2 :انواع تحليل التمايز[50] *Sorts Discriminant Analysis*

1- تحليل التمايز التنبؤي *Predictive Discriminant Analysis*(PDA)

ان تحليل التمايز التنبؤي يفرز تصنيفاً لكل من الوحداِت العدِديّة في مجموعة معينة على أساس مقدار ارتباطها بتركيبات خطية يتم حسابها بعدد المجموعات و ذلك بتحليل قيم الوحدات العددية اعتمادا على متغيرات خصائص الظاهرة تحت الدراِسة. و هذا التحليل هو محور دراستنا لأن الغرض ايجاد وسيلة للتنبؤ بتبعية الوحداِت العدديّة أو للتحقق من سلامة تصنيفها.

2- تحليل التمايز الوصفي *Descriptive Discriminant Analysis*(DDA)

يكون التركيز في تحليل التمايز الوصفي على مسألة تأثيرات متغيرات التجميع على متغيرات الخصائص، أي على مسألة فصل المجموعات حسب متغيرات الخصائص و بالتالي تفسير التركيبات الخطية التي ترتبط بالاختلافات بين المجموعات ومن ثم تفسير البُنية المرتبطة بتأثيرات تحليل التباين المتعدد.

الفرق بين نوعي تحليل التمايز يمكن أن نقول بأن تحليل التمايز التنبؤي قريب لتحليل العلاقة الاعتمادية المُتعددة و أن تحليل التمايز الوصفي مثيل لتحليل الارتباط المتعدد.

5.2 : مجالات تطبيق تحليل التمايز

إن أهمية استخدام تحليل التمايز ستكون أكثر وضوحاً إذا تعرفنا على بعض الاستخدامات التطبيقية لهذه الطريقة الإحصائية، والأمثلة التطبيقية على موضوع التصنيف كثيرة فمثلاً:

1- في الدراسات الطبية الحيوية: التطبيق الرئيس لتحليل التمايز في الطب هو تخمين الحالة الحرجة للمريض والإنذار بنتيجة المرض. على سبيل المثال، أثناء الاستعداد للأحداث تم تقسيم المرضى إلى مجموعات بناءاً على خطورة المرض (بسيط، متوسط، خطير). ثم تم دراسة نتائج

التحاليل المعلمية والسريرية بغرض اكتشاف المتغيرات المختلفة إحصائيا في المجموعات قيد الدراسة. وبإستعمال تلك المتغيرات تم بناء دوال التمييز والتي تساعد في تصنيف الأمراض بشكل فعال للمرضى المستقبلين إلى بسيط ومتوسط وخطير.

2- في التعرف على الوجوه : يتم تمثيل كل وجه عند التعرف على الوجوه الحاسوبي بعدد كبير من البكسلات. يستعمل هنا تحليل التمايز الخطي بشكل رئيسي لتقليل عدد الصفات ليصبح التحكم بها أسهل قبل عملية التصنيف. كل بعد جديد هو تركيبة خطية من البكسلات ، التي تمثل قالب . التركيبات الخطية التي تم الحصول عليها باستعمال التمييز الخطي لفيشر تسمى وجوه فيشر ، بينما تلك التي تم الحصول عليها باستعمال تحليل المكون الرئيس تسمى الوجوه الذاتية .

3- في مجال التعليم : إن توزيع مجموعة من الطلبة على عدد من الاختصاصات أو الحُرف التي تتوافق مع إمكانياتهم الذهنية والبدنية على ضوء معايير معينة وذلك؛ لتجنب الهدر بأعداد الطلبة اللذين قد لا تتوافق إمكانياتهم الذهنية والبدنية مع الحرف أو الاختصاصات المنسبين إليه، أو معرفة كفاءة الطلاب وذلك على أساس الدرجات التي حصلوا عليها في مجموعة من الامتحانات ولمعرفة مدى تأثير الاختلافات في البيئة والجنس (ذكر، أُنثى) والتدريس (خصوصي، عام) على كفاءة الطلاب .

4- في مجال الزراعة : تم استعمال الدالة التمييزية الخطية للتمييز بين أصناف متعددة من النباتات وتنسب كل منها إلى النوع والسلالة ذات الصفات الواحدة، وكذلك الحال في علم الأحياء والزراعة كتصنيف حشرة معينة إلى أحد الأصناف المعروفة بناءً على مواصفات (مقاييس) تلك الحشرة.

5- في مجال الجغرافية و علم الأرض : يمكن استعمال تحليل التمايز لفصل مناطق التغير. على سبيل المثال ، عند إتاحة البيانات المختلفة لمناطق متباينة ،فإن تحليل التمايز يمكنه اكتشاف الأنماط داخل البيانات وتصنيفها بفاعلية . وكذلك في التنبؤ في الكوارث فمثلاً ، مطلوب التمييز بين زلزال وانفجار نووي تحت أرضي سجلا استناداً على أساس إشارات في محطة زلزالية اذ ان قاعدة التخصيص تتكون من الإشارات المسجلة على الإحداث الزلزالية الماضية من التصنيف.

6- في التنبؤ بالإفلاس: كان تحليل التمايز الخطي أول طريقة إحصائية تم تطبيقها في التنبؤ بالإفلاس بالاعتماد على النسب المحاسبية والمتغيرات المالية الأخرى لتقديم تفسير منهجي لدخول الشركات في الإفلاس مقابل النجاة منه . بالرغم من الصعوبات المتمثلة في عدم التوافق المعروف بين النسب المحاسبية وافتراض التوزيع الطبيعي لتحليل التمايز الخطي .

كذلك يمكن استخدام التصنيف في عمليات القروض على المصارف وذلك عند تصنيف طالب قرض معين إلى أنه مضمون أو غير مضمون بالاستناد إلى مواصفات معينة.

فضلاً عن أمثلة كثيرة أخرى لاستخدامات التصنيف في مجالات العلوم البيولوجية وعلم النفس وعلوم الفلك وغيرها، ويتضح من الدراسات التطبيقية للتحليل التمييزي أن تطبيقاتها لا تنحصر في مجال علمي دون أخر وإنما تعددت مجالات استخدامها مما يضيف على هذه الطريقة أهمية كبيرة.

6.2 : شروط تحليل التمايز[1] *Conditions of Discriminat analysis*

1- المتغيرات الكمية موزعة توزيعاً طبيعياً لكل مجتمع، ويحدد هذه المجتمعات مستويات المتغير التصنيفي.

3- يتم اختيار العينة عشوائياً

4- وجود علاقة خطية بين المتغيرات المنبئة، ويمكن التحقيق من ذلك برسم شكل الانتشار لكل زوج من هذه المتغيرات.

5- عدم وجود ازدواج خطي بين المتغيرات التوضيحية، أي إن المتغيرات مستقلة عن بعضها البعض.

6- عدم وجود قيم شاذة.

فمن المعلوم عندما تكون مصفوفات التباين والتباين المشترك غير متساوية تستعمل دالة التصنيف التربيعية وعند عدم تحقق شرط التوزيع الطبيعي وتساوي التباينات ووجود عدد من المتغيرات المستمرة والفئوية تكون دالة التصنيف اللوجستي هي الملائمة ، وعند عدم توفر شرط التوزيع الطبيعي وحجم العينة صغير نلجأ إلى استعمال دالة تصنيف لامعلمية (الجار الاقرب).

7.2 : خطوات اجراء تحليل التمايز *Steps to Conduct a Discrimination*

من اجل اجراء تحليل التمايز، يجب المرور بالمراحل الآتية [64],[61] :-

1- اختيار المتغيرات المكونة للمعادلة التمييزية

يتم اختيار المتغيرات التوضيحية التي يتكون منها الانموذج وذلك باختيار المتغيرات التي يكون لها اعلى قيمة (F) وادنى قيمة (*Wilks Lambda*) . وتمثل قيمة (F) مساهمة المتغيرات التوضيحية في التمييز بين المجاميع، بعد الاخذ بالاعتبار التغيرات التي تحدثها بقية المتغيرات التمييزية . ويحسب مقياس (*Wilks Lambda*) درجة التباعد بين المجموعات.

2- المعاملات التمييزية المعيارية *Standardization Discriminant coefficients*

تتمثل المعاملات التمييزية المعيارية بقيم (a) الظاهرة في المعادلة الآتية :-

$$Y^* = a_1^* x_1 + a_2^* x_2 + \dots\dots\dots + a_t^* x_t \qquad (2-1)$$

وتُستعمل المعادلة التمييزية المعيارية في تحديد اهمية المتغيرات في تكوين دالة التمييز ، حيث ان المتغيرات التي تكون القيمة المطلقة لمعاملها كبيرة . تساهم بشكل كبير في تكوين المعادلة التمييزية ، وتعني اشارة المعامل التمييزي المعياري ان مساهمة المتغير في التمييز هي مساهمة موجبة او سالبة .

ويتم ايضاً باستعمال المعادلة التمييزية المعيارية تحديد الحد الفاصل بين المعاملات التمييزية بين المجاميع ، حيث يمثل الحد الفاصل الوسط الحسابي للمعادلات التمييزية المعيارية للمجاميع .

3- المعاملات التمييزية غير المعيارية *Canonical Discriminant coefficients*

تُستخدم المعاملات التمييزية غير المعيارية في تكوين الدالة التمييزية بدلا من المعاملات التمييزية المعيارية . ذلك؛ لأن المتغيرات التمييزية للمجاميع تظهر بالقيم الحقيقية والنسب وليست بالقيم المعيارية . ومما تجدر الاشارة اليه ان المعاملات التمييزية غير المعيارية لا تعطي الاهمية النسبية للمتغيرات التمييزية؛ لأنها تُشتق من البيانات الخام أي القيم الحقيقية للمتغيرات التمييزية.

وتتمثل المعاملات التمييزية غير المعيارية بقيمة (a) الظاهرة في المعادلة الآتية :-

$$Y^* = a_1^* z_1 + a_2^* z_2 + \dots\dots\dots + a_t^* z_t + c \qquad (2-2)$$

8.2 : اختبار قدرة الدالة التمييزية على التمييز

عندما نريد التمييز بين مجتمعين، فإنه يمكن أن نختبر الفرضية التي تنص على تساوي متوسطات المجموعتين تكون:

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

وعند التمييز بين أكثر من مجموعتين فان الفرضية تكون:

$\mathbf{H_0}: \mu_1 = \mu_2 = \cdots\dots\dots\dots\dots = \mu_K$
$\mathbf{H_1}:$ على الاقل اثنان من المتوسطات مختلفة

لاختبار قدرة الدالة التمييزية للتمييز بين المجاميع يكون ذلك بالاعتماد على الاختبارات الاحصائية الآتية :

1.8.2 : مقياس ويلكس[64] *Wilks-criteria*

لاختبار وجود علاقة خطية بين المتغيرات فأننا نختبر الفرضية الآتية :

$\mathbf{H_0}:$ لا توجد علاقة خطية بين المتغيرات
$\mathbf{H_1}:$ توجد علاقة خطية بين المتغيرات

إحصاءة الاختبار تكون :

$$\Lambda = \frac{|W|}{|\mathrm{T}|} = \frac{|W|}{|\mathrm{W}+\mathrm{B}|} \qquad (2-3)$$

وقيمة (Λ) محصورة بين الصفر والواحد، فإذا كانت قيمته مساوية أو قريبة من الواحد فذلك يدل على إن متوسطات المجموعات متساوية أي لا يوجد تمييز بين المجموعات، أما إذا اقتربت قيمتها إلى الصفر دل ذلك على قوة التمييز.

والصيغة أعلاه تتوزع تقريبا χ^2 بدرجة حرية $p(k-1)$ ، فذا كانت P-value اقل من 0.05 نرفض $\mathbf{H_0}$ اي توجد علاقة خطية بين المتغيرات أي إن الدالة التمييزية لها القدرة على التمييز .

2.8.2 : اختبار[64] χ^2 *chi-square*

يوضح هذا الاختبار معنوية الانموذج أي توجد علاقة معنوية بين المتغير المعتمد والمتغيرات التوضيحية وحسب الفرضية الآتية:

$\mathbf{H_0}$: الانموذج غير معنوي

$\mathbf{H_1}$: الانموذج معنوي

اما صيغته الرياضية تكون كالآتي :

$$\chi^2 = -\mathrm{N}\ \mathrm{Log}t|\Lambda| \qquad (2-4)$$

ويكون توزيعه قريب إلى توزيع χ^2 بدرجة حرية $p(k-1)$ ، فاذا كانت P-*value* اقل من 0.05 نرفض $\mathbf{H_0}$ أي الانموذج معنوي.

وقد طورت هذه الصيغة من قبل ***Barttlete*** إلى الشكل التالي :

$$\chi^2 = -[\mathrm{N} - 1 - \frac{1}{2}(\mathrm{P}+\mathrm{K})]\ \mathrm{Log}\ |\Lambda| \qquad (2-5)$$

بدرجة حرية $p(k-1)$.

3.8.2 : اختبار F [64],[47]

يستخدم اختبار F لاختبار المعنوية الاحصائية لقدرة الدالة التمييزية للفصل بين المجاميع اذ ان إحصاءة الاختبار هي كالاتي :

$$F = \frac{1-\Lambda^{1/5}}{\Lambda^{1/5}} * \frac{ms-2\lambda}{\mathrm{p(k-1)}} \qquad (2-6)$$

بدرجات حرية

$$df_1 = \mathrm{p(k-1)}$$

$df_2 = ms - 2\lambda$

إذ إن:

$$m = N - \frac{1}{2}(\mathrm{P} + \mathrm{K}) \quad , \quad \lambda = \frac{\mathrm{p}(\mathrm{k}-1) - 2}{4}$$

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

4.8.2 : القيم الذاتية[64] *Eigen value*

تستخدم القيم الذاتية (الجذور المميزة) وذلك لمعرفة مدى قدرة الدالة التمييزية بين المجاميع حيث ان القيمة المرتفعة للجذور المميزة تكون مؤشرا على قدرة الدالة على التمييز بين المجاميع .

5.8.2 : معامل الارتباط القانوني[64] *Canonical correlation*

يقيس معامل الارتباط القانوني جودة توفيق دالة التمييز ، حيث ان القيمة المرتفعة لمعامل الارتباط القانوني يكون مؤشرا على جودة توفيق عالية لدالة التمييز ويكون مساوي معامل التحديد.

ويحسب معامل الارتباط بقسمة مجموع مربعات التباينات بين المجموعات على الجذر التربيعي لمجموع مربعات التباينات الكلي .

9.2 : اختبار تساوي مصفوفة التباين والتباين المشترك[36]

يستعمل هذا الاختبار لمعرفة النوع الملائم من النماذج لتمثيل دالة التمييز بين المجموعات .

وتكون الفرضية:

$\mathbf{H_0}: \Sigma_1 = \Sigma_2 = \ldots = \Sigma_K$

$\mathbf{H_1}$: على الاقل اثنان منهم غير متسويان

واختبار ***Barttlete*** هو احد الاختبارات التي تطبق للتأكد من شرط تجانس التباين حيث يختبر الفرضية أعلاه وتكون صيغة إحصاءة الاختبار كالتالي :

$$M = Ln|S| \sum_{i=1}^{K} n_i - \sum_{i=1}^{K} Ln|S_i| \qquad (2-7)$$

$$S = \frac{\sum_{i=1}^{K} n_i S_i}{\sum_{i=1}^{K} n_i} \qquad (2-8)$$

واثبت ***Box*** انه عند ضرب M في ثابت C^{-1} الذي يساوي:

$$C^{-1} = 1 - \frac{2P^2 + 3(P-1)}{6(\mathrm{P}+1)(\mathrm{K}-1)} \left[\sum_{i=1}^{K} \frac{1}{n_i} - \frac{1}{\sum n_i} \right] \qquad (2-9)$$

وعندها نتوصل إلى مقياس يتوزع χ^2 بدرجات حرية $\frac{1}{2}(\mathrm{K}-1)(P-1)$ عندما تكون n_i كبيرة .

لذلك فإن:

$$Box's\ M = MC^{-1} \sim \chi^2_{\frac{1}{2}(\mathrm{K}-1)(P-1)} \qquad (2-10)$$

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

تعد هذه الطريقة من الطرق المعلمية، وهي من ابسط حالات التمييز والتي تتطلب توفر الشروط الآتية [29]:

1. يفترض أن تتوزع المتغيرات التوضيحية توزيعاً طبيعياً متعدد المتغيرات
2. تساوي التباينات لكل المجاميع (مصفوفات التباين والتباين المشترك).
3. تصنيف المشاهدات الموجودة إلى مجاميع (n_1 , n_2 , ...) بشكل دقيق.
4. عدم وجود مشكلة الازدواج الخطي بين المتغيرات التوضيحية.

1.10.2 : دالة التمييز الخطية في حالة مجموعتين

Linear Discriminate Function (LDF) –Tow Groups

إن دالة التمييز هي انموذج 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[34] :

$$f_1(\underline{X}) = f_2(\underline{X}) \qquad (2-11)$$

إذ إن: f_1, f_2 دوال احتمالية تتوزع توزيعا طبيعيا متعدد المتغيرات وللمجموعتين الأولى والثانية على التوالي :

$$ln\, f_1 = ln\, f_2 \qquad (2-12)$$

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

$$(\underline{X} - \mu_1)' \Sigma^{-1}(\underline{X} - \mu_1) = (\underline{X} - \mu_2)'\Sigma^{-1}(\underline{X} - \mu_2) \qquad (2-13)$$

$$(\underline{X}' - \mu'_1)\Sigma^{-1}(\underline{X} - \mu_1) - (\underline{X}' - \mu'_2)\Sigma^{-1}(\underline{X} - \mu_2) = 0$$

$$\underline{X}'\Sigma^{-1}(\underline{X} - \mu_1) - \mu'_1\Sigma^{-1}(\underline{X} - \mu_1) - \underline{X}'\Sigma^{-1}(\underline{X} - \mu_2) - \mu'_2\Sigma^{-1}(\underline{X} - \mu_2) = 0$$

$$\underline{X}'\Sigma^{-1}\underline{X} - \underline{X}'\Sigma^{-1}\mu_1 - \mu'_1\Sigma^{-1}\underline{X} + \mu'_1\Sigma^{-1}\mu_1 - \underline{X}'\Sigma^{-1}\underline{X} + \underline{X}'\Sigma^{-1}\mu_2 - \mu'_2\Sigma^{-1}\underline{X} + \mu'_2\Sigma^{-1}\mu_2 = 0$$

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

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

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

$$\underline{X}'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) = 0 \qquad (2-14)$$

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 [59]:

$$\underline{X}'S^{-1}\left(\overline{\underline{X}}_1 - \overline{\underline{X}}_2\right) - \frac{1}{2}(\overline{\underline{X}}_1 - \overline{\underline{X}}_2)'S^{-1}\left(\overline{\underline{X}}_1 + \overline{\underline{X}}_2\right) \qquad (2-15)$$

$$S = \frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2} \qquad (2-16)$$

والمعادلة (15 – 2) تمثل دالة التمييز الخطية المقدرة عند الحد الفاصل بين مجموعتين ، ويمكن الاعتماد على هذه المعادلة في تصنيف أي مشاهدة جديدة ، فالمشاهدة إما تعود إلى المجموعة الأولى إذا كانت هذه المعادلة اكبر من الصفر، أو تعود إلى المجموعة الثانية إذا كانت اقل من الصفر عدا ذلك فإنها تصنف عشوائيا إلى أي من المجوعتين.

نفرض إن Y تمثل تركيبة خطية للمتغيرات التوضيحية $(X_1, X_2, \ldots \ldots X_{\mathrm{P}})$ تدعى بدالة التمييز الخطي وتكتب بالصيغة الآتية :

$$Y = b_1X_1 + b_2X_2 + \ldots \ldots + b_{\mathrm{P}}X_{\mathrm{P}} \qquad (2-17)$$

وبصيغة المصفوفات تكتب بالصيغة الآتية :

$$Y = \underline{b}' \underline{X} \qquad (2-18)$$

حيث إن $(\mathrm{b}_1, \mathrm{b}_2, \ldots \ldots \mathrm{b}_{\mathrm{P}})$ معاملات يمكن تقديرها بجعل الدالة تعطي أفضل تمييز بين المجاميع وذلك عندما تجعل مربع الفرق بين متوسطي المجموعتين إلى التباين (مجموع المربعات داخل أو ضمن المجموعات) اكبر ما يمكن[64].

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

والكمية المراد تعظيمها هي:

$$Q = \frac{[\bar{Y}_1-\bar{Y}_2]^2}{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2} = \frac{[\underline{b}'\underline{\mu}_1-\underline{b}'\underline{\mu}_2]^2}{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2} \qquad (2-19)$$

إذ إن :

$$\overline{Y}_i = b_1\overline{X}_{\mathrm{i}1} + b_2\overline{X}_{\mathrm{i}2} + \ldots \ldots + b_2\overline{X}_{\mathrm{i}P}$$

وعندما تكون المعاملات مجهولة نستعمل تقديرات العينة $\overline{X}_{\mathrm{i}}$ بدلا من μ_i و S بدلا من Σ

وتكون:

$$\overline{X}_{\mathrm{i}} = \left(\overline{X}_{\mathrm{i}1}, \overline{X}_{\mathrm{i}2}, \ldots \ldots, \overline{X}_{\mathrm{i}P}\right) \qquad (2-20)$$

$$S = \frac{1}{n_1 + n_2 - 2}(X_1'X_1 - X_2'X_2) \qquad (2-21)$$

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

$$Q = \frac{[\,\underline{b}'(\overline{\underline{X}}_1 - \overline{\underline{X}}_2)]^2}{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2} \qquad (2-22)$$

$$d = \overline{y}_1 - \overline{y}_2 \qquad (2-23)$$

$$= \underline{b}'\left(\overline{\underline{X}}_1 - \overline{\underline{X}}_2\right)$$

وبفرض إن $\alpha = \overline{X}_1 - \overline{X}_2$ ينتج :

$$D = \overline{Y}_1 - \overline{Y}_2 = \underline{b}'\underline{\alpha}$$

$$d^2 = (\underline{b}'\underline{\alpha})^2 \qquad (2-24)$$

ومجموع مربعات الفرق داخل أو ضمن المجاميع هو :

$$(Y_{ij} - \overline{Y}_i) = b_1(X_{1ij} - \overline{X}_{1i}) + \cdots \ldots + b_P(X_{Pij} - \overline{X}_{Pi}) \qquad (2-25)$$

وبتربيع طرفي المعادلة $(2-25)$ و اخذ المجموع نحصل على:

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 = \underline{b}'S\,\underline{b} \qquad (2-26)$$

$$q = \frac{(\underline{b}'\underline{\alpha})^2}{\underline{b}'S\,\underline{b}} \qquad (2-27)$$

$$\frac{\partial q}{\partial \underline{b}} = \frac{(\underline{b}'S\,\underline{b})2(\underline{b}'\underline{\alpha})\alpha - (\underline{b}'\underline{\alpha})^2(2S\underline{b}')}{(\underline{b}'S\,\underline{b})^2} = 0$$

$$(\underline{b}'S\,\underline{b})\underline{b}'\underline{\alpha}\underline{\alpha} - (\underline{b}'\underline{\alpha})^2(S\,\underline{b}) = 0 \qquad (2-28)$$

وبقسمة طرفي المعادلة $(30-2)$ على $(\underline{b}'\underline{\alpha})^2$ نحصل على :

$$\frac{\underline{b}'S\,\underline{b}}{\underline{b}'\,\underline{\alpha}}\underline{\alpha} - S\,\underline{b} = 0 \qquad (2-29)$$

أو إن:

$$\frac{\underline{b}'S\,\underline{b}}{\underline{b}'\underline{\alpha}}\,\underline{\alpha} = S\,\underline{b}$$

$$\underline{b}' = S^{-1}\,\frac{\underline{b}'S\,\underline{b}}{\underline{b}'\underline{\alpha}}\,\underline{\alpha} \qquad (2-30)$$

وعندما يكون المقدار $\frac{\underline{b}'S\,\underline{b}}{\underline{b}'\underline{\alpha}}$ ثابت ومساوياً إلى الواحد فتصبح المعادلة كالآتي :

$$\underline{\mathrm{b}}' = S^{-1}\,\underline{\alpha}$$

$$= S^{-1}\,(\overline{X}_1 - \overline{X}_2)' \qquad (2-31)$$

وهي تركيبة خطية ناتجة عن تعظيم نسبة التباين بين المجاميع إلى التباين داخل أو ضمن المجاميع .

وبعد استخراج المعاملات $\underline{\mathrm{b}}'$ ، يتم تصنيف المفردة إلى إحدى المجموعتين اعتمادا على نقطة وسط المجموعتين (L) التي تجعل التصنيف الخاطئ اقل ما يمكن [65].

إذ إن:

$$L = \frac{\overline{Y}_1 + \overline{Y}_2}{2} \qquad (2-32)$$

إذا كانت $\hat{Y} > L$ تصنف المشاهدة إلى المجموعة الأولى.

إذا كانت $\hat{Y} < L$ تصنف المشاهدة إلى المجموعة الثانية.

إذا كانت $\hat{Y} = \mathrm{L}$ تصنف المشاهدة عشوائياً إلى المجموعة الأولى أو الثانية.

و إن:

$$\hat{Y} = (\overline{X}_1 - \overline{X}_2)' \, S^{-1} \, X = \hat{\alpha} \, X$$

$$Y_1 = (\overline{X}_1 - \overline{X}_2)' \, S^{-1} \, X_1$$

$$\overline{Y}_1 = \hat{\alpha}' \, \overline{X}_1 \quad , \quad \overline{Y}_2 = \hat{\alpha}' \, \overline{X}_2$$

إذ إن:

$$\hat{\alpha}' = (\overline{X}_1 - \overline{X}_2) \, S^{-1}$$

2.10.2 : دالة التمييز الخطية في حالة أكثر من مجموعتين[64],[8]

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 .

$$n = \sum_{i=1}^{k} n_{\mathrm{i}} \qquad (2-33)$$

نفترض أن T تمثل مجموع المربعات الكلية .

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(X_{ij} - \overline{\mathrm{X}}\right) \left(X_{ij} - \overline{\mathrm{X}}\right)' \qquad (2-34)$$

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

$$W_{\mathrm{i}} = \sum_{i=1}^{n_i} \left(X_{ij} - \overline{\mathrm{X}}\right) \left(X_{ij} - \overline{\mathrm{X}}\right)' \qquad (2-35)$$

وان مصفوفة التباين والتباين المشترك داخل المجموعات تساوي W:

$$W = W_1 + W_2 + \ldots\ldots + W_K \qquad (2-36)$$

وإن مصفوفات التباين والتباين المشترك بين المجموعات هي:

$$B = T - W \qquad (2-37)$$

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

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

وبهدف الحصول على مجموعة من التراكيب الخطية التي هي:

$$Y = [Y_1, Y_2, \dots \dots Y_r] \qquad (2-38)$$

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

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

$$\lambda = \frac{\underline{b}' B\ \underline{b}}{\underline{b}'\ W\ \underline{b}} \qquad (2-39)$$

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

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

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

$$(\underline{b}' W\ \underline{b}) B\ \underline{b} - (\underline{b}' B\ \underline{b}) W\ \underline{b} = 0 \qquad (2-40)$$

بقسمة طرفي معادلة $(2-39)$ على $\underline{b}' W\ \underline{b}$ وبالتعويض عن λ بما يساويها نحصل على :

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

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

$$(\mathrm{W}^{-1}\,\mathrm{B}-\lambda \mathrm{I})\underline{\mathrm{b}}=0 \qquad (2-41)$$

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

$$b_1=(b_{11},b_{12},\ldots\ldots,b_{1\mathrm{P}}) \qquad (2-42)$$

b_1 تمثل مقياس التمييز للدالة الأولى Y_1 والتي تساوي :

$$Y_1=b_{11}\mathrm{X}_1+b_{12}\mathrm{X}_2+\ldots\ldots+b_{1\mathrm{P}}\mathrm{X}_P \qquad (2-43)$$

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 التي تساوي :

$$Y_2=b_{21}\mathrm{X}_1+b_{22}\mathrm{X}_2+\ldots\ldots+b_{2\mathrm{P}}\mathrm{X}_P \qquad (2-44)$$

Y_1 من الضروري إن تكون غير مرتبطة مع Y_2 .

Y_3 تمتلك ثالث اكبر متجه مميز.

$$Y_3=b_{31}\mathrm{X}_1+b_{32}\mathrm{X}_2+\ldots\ldots+b_{3\mathrm{P}}\mathrm{X}_P \qquad (2-45)$$

Y_3 غير مرتبطة مع Y_1 و Y_2 .

وهكذا نستمر إلى Y_{r} التي 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 التي يمكن نعبر عنها بشكل مصفوفة وكما يأتي:

$$\underline{Y}=\underline{b\mathrm{x}} \qquad (2-46)$$

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

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 = رتبة W^{-1}

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 :

$$rank\ (\mathrm{W}^{-1}\ \mathrm{B}) = \min(P, K-1) \qquad (2-47)$$

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 .

11.2 : دالة التمييز التربيعية *Quadratic Discriminant Function*

تعد هذه الطريقة من الطرق المعلمية إذ يستعمل في حالة عدم تساوي مصفوفة التباين والتباين المشترك للمجموعات ويكون المجتمع قيد الدراسة ذا توزيع طبيعي متعدد المتغيرات وبمتجهات متوسطات مختلفة.

وعليه فان مقياس التمايز V سيكون كالآتي[55]:

$$V = \frac{f_1(x_i)}{f_2(x_j)} <> 1 \qquad (2-48)$$

إذا كانت $V > 1$ المشاهدة $\boldsymbol{x}$ تعود إلى المجموعة الأولى و إذا كانت $V < 1$ فان المشاهدة x تعود إلى المجموعة الثانية وعشوائيا عدا ذلك.

وبأخذ اللوغاريتم للطرفين نحصل على الآتي:

$$G = \ln V = \ln f_1(x_i) - \ln f_2(x_j) = 0 \qquad (2-49)$$

وفي حالة وجود P من المتغيرات $(X_1, X_2, \ldots\ldots, X_P)$ لكل مجموعة من المجموعتين فان مقياس التمييز يكون :

$$G = \ln f_1(x_1, \ldots\ldots, x_P) - \ln f_2(x_1, \ldots\ldots, x_P) \qquad (2-50)$$

وفي حالة إن المعالم الدالة الاحتمالية غير معلومة فسوف تكون دالة التمييز التربيعية التقديرية g والتي قدرت معالمها بطريقة الإمكان الأعظم كالآتي[63] :

$$g = \hat{G} = \frac{1}{2} log \frac{|S_1|}{|S_2|} - \frac{1}{2}\left(\overline{\underline{X}}_1' S_1^{-1} \overline{\underline{X}}_1 - \overline{\underline{X}}_2' S_2^{-1} \overline{\underline{X}}_2\right) + \underline{X}'\left(S_1^{-1}\overline{\underline{X}}_1 - S_2^{-1}\overline{\underline{X}}_2\right)$$
$$- \frac{1}{2}\underline{X}'(S_1^{-1} - S_2^{-1})\underline{X} \qquad (2-51)$$

وسيتم تصنيف المشاهدات الجديدة بأنها تعود إلى المجموعة الأولى إذا كانت $g > 0$ وإلى المجموعة الثانية إذا كانت $g < 0$ وعشوائيا عدا ذلك.

2.11.1 : اشتقاق دالة التمييز التربيعية [72], [68]

تكون الفرضية في حالة وجود مجموعتين من مجتمعين فرضية خاصة تمثل انتماء كل مشاهدة من المشاهدات إلى إحدى المجموعتين وهي:

$\boldsymbol{H_0}: \underline{X} \in f_2(\mathrm{x})$

$\boldsymbol{H_1}: \underline{X} \in f_1(\mathrm{x})$

وباستعمال أسلوب بيز لتقليل الخطورة التي تمثل نسبة الإمكان الأعظم والاحتمالات الأولية P_1 و P_2 وكلف التصنيف الخاطئ c_{12} و c_{21} تكون :

$$\frac{f_1(\mathrm{x})}{f_2(\mathrm{x})} < \frac{c_{12}P_2}{c_{21}P_1} \qquad (2-52)$$

وعند امتلاك كل من f_1 و f_2 توزيعاً طبيعياً و اخذ اللوغاريتم للطرفين ينتج :

$$ln\frac{|\Sigma_1|}{|\Sigma_2|} + \left(\underline{x}_{11} - \underline{\mu}_1\right)' \Sigma_1^{-1}\left(\underline{x}_{1i} - \underline{\mu}_1\right) - \left(\underline{x}_{2i} - \underline{\mu}_2\right)' \Sigma_2^{-1}\left(\underline{x}_{2i} - \underline{\mu}_2\right)$$
$$< -2ln\frac{c_{12}P_2}{c_{21}P_1} \qquad (2-53)$$

وبعد التبسيط تصبح دالة التمييز كما يأتي :

$$h(x) = \underline{x'}A\underline{x}b'x + c \ <> T \qquad (2-54)$$

إذ إن:

$$A = \ \Sigma_1^{-1} - \Sigma_2^{-1}$$

$$b = 2\left(\Sigma_2^{-1}\underline{\mu_2} - \Sigma_1^{-1}\underline{\mu_1}\right)$$

$$c = \left(\underline{\mu}_1' \Sigma^{-1} \underline{\mu}_1 - \underline{\mu}_2' \Sigma_1^{-1} \underline{\mu}_2 + ln \frac{|\Sigma_1|}{|\Sigma_2|} \right)$$

$$T = -2ln \frac{c_{12} P_2}{c_{21} P_1}$$

إذ تصنف المشاهدة x على إنها تنتمي للمجتمع f_1 إذا كانت $h(\underline{x}) < T$ وتنتمي إلى المجتمع f_2 إذا كانت $h(\underline{x}) > T$ وتعرف بدالة التمييز التربيعية .

أما عند تساوي مصفوفة التباين والتباين المشترك Σ_i تصبح الدالة بالشكل الآتي:

$$\mathrm{h}^*(x) = b'^* x + c^* \ <> T \qquad (2-55)$$

إذ أن:

$$b^* = 2\Sigma \left(\underline{\mu}_2 - \underline{\mu}_1 \right)$$

$$c^* = \underline{\mu}_1' \Sigma^{-1} \underline{\mu}_1 - \underline{\mu}_2 \Sigma^{-1} \underline{\mu}_2$$

وهي دالة التمييز الخطية.

الرمز <> يعني تصنيف المشاهدة x على أنها تنتمي إلى المجتمع المميز f_1 إذا كانت قيمتها اقل من T وتنتمي إلى المجتمع f_2 إذا كانت قيمتها اكبر من T.

وعليه تكون كلف التصنيف الخاطئ تساوي دائما واحد ، وان خطورة بيز تتحول إلى احتمال خطا التصنيف لذلك فان الحل الذي يقلل من خطورة بيز يقلل أيضاً من احتمال خطا التصنيف ولهذا فإن :

$$P_1 f_1(\underline{x}) < P_2 f_2(\underline{x}) \qquad (2-56)$$

وعند التعميم واخذ مجموعات متعددة تكون الفرضية:

$$\boldsymbol{H_0}: \underline{x} \in w_i \qquad i = 1, \ldots\ldots, g$$

وفي حالة الكلف تساوي واحد يكون الحل باختبار المجموعة التي تحقق ما يأتي كمجموعة تنتمي إليها المفردات $\underline{x}$:

$$P_k f_k > P_i f_i \qquad (2-57)$$

$$for\ i \neq k \ \ , i = 1, \ldots\ldots, g$$

وبعبارة أخرى فان لدينا g من الدوال التمييزية $h_i(\underline{x})$ والحل هو اختبار الدالة التي تعطي أكبر قيمة ، وكما يلي :

$$h_{\mathrm{k}}(\underline{\mathrm{x}}) = \max h_i(\underline{\mathrm{x}}) \qquad (2-58)$$

إذ أن:

$$h_i(\underline{\mathrm{x}}) = P_i f_i(\underline{\mathrm{x}})$$

وتوجد نظرية بهذا الخصوص تنص على إن دالة متزايدة رتيبة $h_i(\underline{\mathrm{x}})$ يمكن أن تحل بدلا من $h_i(\underline{\mathrm{x}})$ وفي حالة توزيع البيانات طبيعيا يكون [68]:

$$h_i(\underline{x}) = P_i N\left(\underline{x_i}\underline{\mu_1}, \Sigma_i\right) \qquad (2-59)$$

إذ إن :

$N(\underline{x_i}\underline{\mu_1}, \Sigma_i)$: هي دالة التوزيع الطبيعي متعدد المتغيرات

وفي حالة تحقق التوزيع الطبيعي فان الصيغة $h_i(\underline{x})$ تكون :

$$h_i(\underline{x}) = -\left(\underline{x} - \underline{\mu_i}\right)' \Sigma_1^{-1}\left(\underline{x} - \underline{\mu_i}\right) - ln|\Sigma_i| + 2lnP_i \qquad (2-60)$$

$$i = 1, \ldots\ldots, g$$

وهي دالة التمييز التربيعية المتعددة (*Multiple Quadratic Discriminant Function*)

وعند تساوي مصفوفة التباين والتباين المشترك تصبح الدالة بالشكل الآتي:

$$h_{\mathrm{i}}^{*}(\underline{x}) = 2\underline{\mu_i'}\Sigma^{-1}\underline{x} - \underline{\mu_i}\Sigma^{-1}\underline{\mu_i} + 2lnP_i \ , i = 1, \ldots\ldots, g \qquad (2-61)$$

وهي دالة التمييز الخطية .

2.11.2 : تقدير معالم الانموذج [44],[62]

يتم تقدير Σ_i و $\hat{\Sigma}_i$ بطريقة الإمكان الأعظم لوجود خاصية الثبات إذ إن :

$$\hat{\mu}_i = \overline{\underline{x}}_i = \frac{1}{n_i}\sum_{j=i}^{ni} \underline{x_{ij}} \qquad (2-62)$$

$$\hat{\Sigma}_i = S_i^2 = \frac{1}{n_i}\sum_{j=i}^{ni} (\underline{x_{ij}} - \overline{x}_j)(\underline{x_{ij}} - \overline{x}_j)' \qquad (2-63)$$

ويمكن استبدال تقدير Σ_i المتحيز بتقدير Σ_1^{-1} غير المتحيز المشتق من توزيع **(*Wishert*)** وهو :

$$\hat{\Sigma}_1^{-1} = \left(\frac{ni - P - 3}{ni}\right) S_i^{-1} \qquad (2-64)$$

إذ إن S_i هي تقدير الإمكان الأعظم لـ Σ_1^{-1} *وكذلك بدلا من* Σ_i يمكن استعمال التقدير غير متحيز لمسافة مهانلوبس .

$$\widehat{D}^2{}_{ij} = \frac{n_i + n_j - p - 3}{n_i + n_j - p - 2}\left(\overline{\underline{x}}_j - \overline{\underline{x}}_i\right)' S^{-1}\left(\overline{\underline{x}}_j - \overline{\underline{x}}_i\right) - \frac{p\left(n_i - n_j\right)}{n_i n_j} \quad (2-65)$$

وإذا كانت $\widehat{D}^2{}_{ij}$ سالبة فتؤخذ صفراً .

اما بالنسبة لتقدير الاحتمالات الأولية فتقديرها يكون:

$$\hat{P}_i = P_i = \frac{n_i}{n} \quad (2-66)$$

12.2: دالة التمييز اللوجستي *Logistic discriminant Function*

يعد التمييز اللوجستي من بين الكثير من الأساليب المقترحة في التمييز الإحصائي للطرق المعلمية المستندة على التوزيع الطبيعي متعدد المتغيرات ، والطرق اللامعلمية حرة التوزيع كطريقة الجار الأقرب (*Nearest Neighbor*) لهذا السبب غالبا ما تدعى هذه الطريقة بالمعلمية الجزئية أو شبه معلمية . وفقاً لذلك فإن الانموذج اللوجستي احد الأدوات الأكثر جاذبية واستعمالا في حل مسائل التمايز والانحدار اللذان يسيران نحو إحدى أو كلا الاتجاهين الآتيين :

1. تلخيص ووصف الفروقات بين المجتمعات.
2. تحديد موقع مفردات جديدة إلى مجاميعها أي التنبؤ بقيم متغير الاستجابة لمفردات جديدة.

ان هذه الطريقة من الطرائق المهمة وذلك لأنها تعالج مشاكل التمييز في حالة إن تكون المتغيرات (المقاييس) متقطعة أو مستمرة أو مختلطة فضلا عن سهولتها في الاستخدام إذ أن عدد حدودها يساوي عدد حدود الدالة الخطية. وهناك نوعان من الانموذج اللوجستي وهي (1) الانموذج اللوجستي ثنائي الاستجابة (2) الانموذج اللوجستي متعدد الاستجابة وسيكون تركيزنا على النوع الثاني .

1.12.2 : مميزات الانموذج اللوجستي[43]

1. يتطلب القليل من الافتراضات حول التوزيع، فهو لا يفترض وجود علاقة خطية بين المتغير التابع والمتغيرات التوضيحية.
2. سهولة استخدامه، فعندما يتم تقدير معالم الانموذج فان تحديد موقع مفردة جديدة يتطلب حساب دالة خطية واحدة فقط.
3. تكون المعالم المقدرة من الانموذج ذات تفسير مباشر كدالة خطية للوغاريتم نسبة الارجحية .
4. يجب ان تكون الفئات محددة وشاملة بحيث ان كل مفردة تنتمي إلى فئة واحدة فقط .

5. الانموذج اللوجستي لا يشترط ان تكون المتغيرات التوضيحية من النوع المستمر ولا تتبع التوزيع الطبيعي ولا ان تكون العلاقة بين المتغير التابع والمتغيرات التوضيحية خطية ولا يفترض تساوي التباين ضمن كل فئة . وهذا يجعل الانموذج اللوجستي اكثر مرونة من بقية نماذج التنبؤ والتصنيف .
6. يجب ان يكون حجم العينة المستخدم في الانموذج اللوجستي اكبر من حجم العينة المستخدم في الانموذج الخطي، لأن معاملات انموذج اللوجستي يتم تقديرها باستخدام طريقة الامكان الاعظم وهي طريقة تحتاج إلى عينة كبيرة الحجم نسبياً، وهناك طرق اخرى مثل طريقة المربعات الصغرى الموزونة.

2.12.2 : الانموذج اللوجستي ثنائي الاستجابة[49],[66]

يطبق هذا الانموذج عندما يأخذ المتغير العشوائي **Y** قيمتين فقط قد تدل على نمط الاستجابة لمؤثر ما (كالأدوية والمنشطات والسموم) . في هذه الحالة تأخذ **Y** القيمتين 0 و1 تشير إلى عدم حدوث وحدوث الاستجابة على التوالي . كذلك يستعمل في مسائل التشخيص الطبي والتمييز بين أنواع الإمراض. فالأورام السرطانية مثلا قد تكون خبيثة أو حميدة.

تعد هذه الطريقة من الطرائق المهمة وذلك لأنها تعالج مشاكل التمييز في حالة ان تكون المتغيرات مستمرة أو متقطعة أو مختلطة فضلا عن سهولتها في الاستعمال إذ إن عدد حدودها يساوي عدد حدود الدالة الخطية، يتكون انموذج اللوجستي من متغير الاستجابة Y_i يتأثر بمجموعة من المتغيرات الموضحة $(X_i's)$ (*Explanatory variables*) على وفق علاقة تحتوي على مجموعة من المعالم $(b_i's)$ إذ أن متغير الاستجابة Y_i في التجارب الحياتية ثنائية الاستجابة الذي يعد انموذج اللوجستي احدها يتوزع حسب توزيع برنولي اي ان له مستويان هما الصفر والواحد.

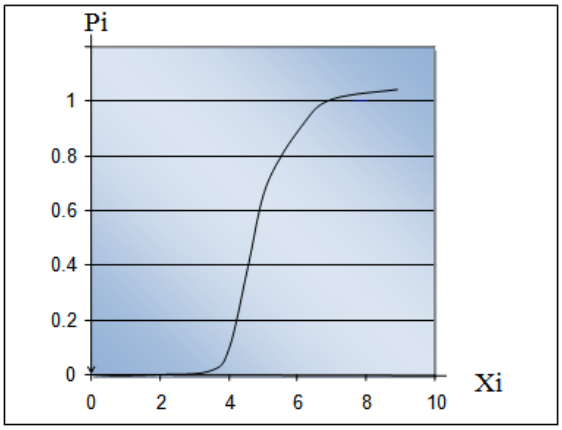
$$Y_i \sim b(1, P_i)$$

يعتمد P_i على مجموعة من المتغيرات التوضيحية $(X_i's)$ على وفق الصيغة الآتية التي يطلق عليه االنموذج دالة الاستجابة اللوجستية :

$$P_i = \frac{e^{(B_0 + B_1 X_1, \dots B_P X_P)}}{1 + e^{(B_0 + B_1 X_1, \dots B_P X_P)}} \qquad (2-67)$$

$$1 - P_i = \frac{1}{1 + e^{(B_0 + B_1 X_1, \dots B_P X_P)}} \qquad (2-68)$$

وان الصيغة (67 – 2) تسمى دالة الاستجابة اللوجستية ولها مخطط في الشكل (1 – 2) :

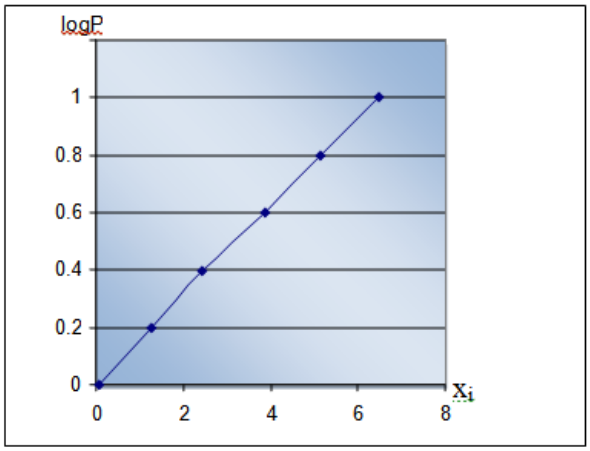


الشكل (1 – 2) يوضح العلاقة بين احتمال الاستجابة (P_i) والمتغير الموضح (x_i)

من **الشكل (1 – 2)** نلاحظ عدم خطية العلاقة بين المتغير الموضح x_i واحتمال الاستجابة P_i، ولغرض اجراء التحويل الخطي للدالة اللوجستية تمكن الباحث (***Berkson***) عام (1944) من تحويل العلاقة بين المتغيرات (x_i) ومتغير الاستجابة (P_i) إلى علاقة خطية وذلك برسم Log P بدلاً من Probit P مقابل (X) كالآتي [38].

$$L_i = ln\left(\frac{P_i}{1-P_i}\right) = log\ P_i = x_i b \qquad (2-69)$$

والصيغة (69 – 2) تبين التحويل الخطي (log) كما في الشكل (2 – 2) :



الشكل (2 – 2) يوضح العلاقة الخطية بين x_i و $logP_i$ [13]

$$L_i = ln\left(\frac{P_i}{1-P_i}\right) = b_0 + b_1 x_1 + \ldots\ldots\ldots + b_P x_P \qquad (2-70)$$

اذ تكون $\underline{b}$ موجه عمودي من رتبة [(P+1)×1] للمعلمات المطلوب تقديرها

$$\underline{b} = \begin{pmatrix} b_0 \\ b_1 \\ . \\ . \\ b_P \end{pmatrix}$$

X مصفوفة من رتبة [g×(p+1)] للمتغيرات التوضيحية تكون :

$$X = \begin{pmatrix} 1 & X_{11} \cdots & X_{1P} \\ 1 & X_{21} \cdots & X_{2P} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 1 & X_{g1} \cdots & X_{gP} \end{pmatrix}$$

وبما ان التوزيع هنا هو برنولي ولحالتين فقط عندئذ يكون [52]:

$$\underline{L_i} = \underline{P_i} = \begin{pmatrix} 0 \\ 0 \\ . \\ . \\ 0 \\ 1 \\ . \\ . \\ 1 \end{pmatrix}$$

$$\underline{L_i^*} = ln\left(\frac{P_i + \frac{1}{2n}}{(1-P_i) + \frac{1}{2n}}\right) \qquad (2-71)$$

إذ استعمل هنا معامل تصحيح الاستمرارية $\frac{1}{2n}$ حيث اعطى نتائج اكثر دقة وبخطأ اقل [30].

ونستعمل بعد ذلك طريقة المربعات الصغرى العامة *Generalized Least Squares* (GLS) التقديرية وهي:

$$\underline{b^*} = (x'x)^{-1} X' \underline{L_i^*} \qquad (2-72)$$

$$\underline{\hat{L}_i} = \hat{b}_0 + \hat{b}_1 X_1 + \ \dots\dots\dots + \hat{b}_P X_P \qquad (2-73)$$

وسوف يتم تصنيف المشاهدات الجديدة بأنها تعود إلى المجموعة الأولى إذا كانت $\underline{\hat{L}_i}$ اكبر من الصفر وإلى المجموعة الثانية اذا كانت $\underline{\hat{L}_i}$ اصغر من الصفر وعشوائياً عدا ذلك.

3.12.2 : الانموذج اللوجستي متعدد الاستجابة

يعد انموذج اللوجستي متعدد الاستجابة امتداداً طبيعياً لأنموذج ثنائي الاستجابة . وهي الحالة التي يكون فيها المفردة أو المشاهدة محددة بواحدة من $\boldsymbol{g}$ من القيم الممكنة تدعى بفئات أو مجاميع الاستجابة . فقد تكون ذات طبيعة ترتيبية مثل درجة خطورة أو شدة مرض معين . كما يمكن أن تكون مجاميع الاستجابة ذات طبيعة اسمية كأصناف الدم مثلاً أو نوع واسطة النقل المفضلة في السفر (سيارة ، طائرة ، قطار ،.....).

نفرض وجود مجتمع كبير يحتوي على g من المجاميع المختلفة $(P_1, \dots., \mathrm{P_g})$ وسحبت عينة عشوائية منه . فأن عدد المفردات الداخلة ضمن العينة وتنتمي إلى المجموعة P_s يتبع توزيع متعدد الحدود *Multinomial Distribution* وان احتمال وقوع Y من المشاهدات ضمن العينة يكون كالآتي[57]:

$$P(Y_1 = y_1, \dots., Y_g = y_g) = \binom{n}{y} P_1^{y_1}, \dots\dots, P_g^{y_g} \qquad (2-74)$$

إذ إن $(P_1, \dots., P_g)$: تمثل نسبة المجاميع في المجتمع

$$\binom{n}{y} = \frac{n!}{y_1!\ y_2! \dots. y_p!} \qquad (2-75)$$

نستنتج من ذلك إن التوزيع متعدد الحدود يتولد كنتيجة لأسلوب المعاينة. فإذا رمزنا لاحتمال حدوث الاستجابة j بـ $P_j(j = 1, \dots, g)$ فعند قيم المشاهدة i ذات متجه الخصائص $= x_{\mathrm{i}}$ $(x_{i1}, \dots, x_{iP})$ يكون الانموذج بالشكل الآتي:

$$P_j(x_i) = \frac{\exp(B_j' x_i)}{\sum_{j=1}^{g} \exp(B_j' x_i)} \quad , j = 1, \dots., g-1 \qquad (2-76)$$

ويكون $b_g = 0$ إذ يمثل g مجموعة الأساس و b_j يمثل متجه المعالم الخاصة بالمجموعة j إذ يوجدg-1 من المتجهات المختلفة ضمن الانموذج . وبذلك فان الانموذج يحتوي على (g-1)(p+1) من المعالم المجهولة و يمكن كتابة الانموذج بصيغة لوغاريتم نسبة الإمكان الأعظم كما يلي :

$$logt\left(\frac{P}{1-P}\right) = B_0 + \sum B_j X_i \quad , j = 1, \dots., g-1 \qquad (2-77)$$

4.12.2 : تقدير معلمات الانموذج اللوجستي[35],[57]

لاحظ ***Anderson* (1972)** إن تطبيق الانموذج اللوجستي يخضع للمعادلة الآتية:

$$logt\left(\frac{P_j(x_i)}{P_g(x_i)}\right) = B_0 + B'x \qquad (2-78)$$

وهو تركيب خطي للمتغير x يناسب المتغيرات الثنائية والثلاثية عندما x يأخذ القيم (0 ، 1 ، 2) ، بافتراض إن نسبة احتمال اللوغاريتمي هو تركيب خطي للمتغير x .

من اجل تطبيق الانموذج اللوجستي يجب تحويل المتغير x إلى متغيرين ثنائيين x_1 و x_2 ،

إذ إن $x = 0$ إذا وفقط إذا كان $x_1 = 0$ و $x_2 = 0$ ؛ $x = 1$ إذا وفقط إذا كان $x_1 = 1$ و $x_2 = 0$؛ و $x = 2$ إذا وفقط إذا $x_1 = 0$ و $x_2 = 1$. على نحو مماثل، يمكن استبدال كل متغير مع أكثر من $r \geq 2$ مستويات من المتغيرات الثنائية $r - 1$.

وان الانموذج اللوجستي في تحليل التمايز يكون :

$$P_1(x) = \frac{e^{(B_0 + B'x)}}{1 + e^{(B_0 + B'x)}} \qquad (2-79)$$

إذ إن $P_1(x)$ تمثل دالة الانحدار اللوجستي (احتمال الاستجابة) من خلال نمذجة العلاقة بين متغير الاستجابة الفئوية الذي نرمز له بالرمز G وبين متجه المتغيرات التوضيحية X .

علماً أن متغير الاستجابة G هو متجه من المجموعات أو الفئات المختلفة $(G_1, \ldots., G_g)$.

وعليه يكون انموذج لوغاريتمي خطي للاحتمالات اللاحقة للمتغير مستقل مع المتغير x مقارنة مع متغير أساس x_0 هو :

$$log\left[\frac{P_1(x)}{P_2(x)} \Big/ \frac{P_1(x_0)}{P_2(x_0)}\right] = B'(x - x_0) \qquad (2-80)$$

في كثير من الأحيان المرض قيد الدراسة يكون نادراً إلى حد ما ، بحيث $\mathrm{P}_2(x)$ و $\mathrm{P}_2(x_0)$ قريبة من الواحد ، وبالتالي نسبة الفرق تقريبا تكون :

$$P_1(x)/P_1(x_0) \qquad (2-81)$$

ومن خلال التحويل اللوجستي نحصل على :

$$logit[P_1(x)] = log[P_1(x)/P_2(x)]$$

$$= B_0 + B'x \qquad (2-82)$$

ولربط x بالاحتمال اللاحق $P_1(x)$ يكون التمييز الاحتمالي بالصيغة الآتية :

$$probit[P_1(x)] = \Phi^{-1}[P_1(x)] \qquad (2-83)$$

إذ إن Φ هي دالة التوزيع الطبيعي القياسي .

عملياً يمكن استعمال المعادلة الآتية للنماذج اللوجستية والاحتمالية وهي :

$$logit[P_1(x)] \cong c\, \Phi^{-1}[P_1(x)] \qquad (2-84)$$

إذ إن $c = \sqrt{8/\pi} = 1.6$.

13.2 : دالة الجار الاقرب *Nearest Neighbor Function*

تعد طريقة الجار الأقرب من الطرق المبنية على الذاكرة، بمعنى انها لا تتطلب توفيق انموذجي 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[78].

و الفكرة الأساسية لطريقة الجار الأقرب 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.

ولسهولة النهج التي تتعامل به طريقة الجار الأقرب فإنها لاقت على مر السنين اهتماما كبيرًا من قبل معظم الباحثين، إذ إن خوارزمية الجار الأقرب تتطلب مجموعة من القوالب المصنفة والتي يتم استعمالها في التعرف على مجموعة من أنماط مختارة، وذلك من خلال حساب المسافة بين النمط المراد تصنيفه مع بقية أنماط المجموعة، واختيار الأنماط التي عددها k الأقرب لاتخاذ القرار[39].

1.13.2 : خوارزمية الجار الأقرب للتصنيف[70]

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

$$\theta = \int_{v(x)} P(x)\, dx \qquad (2-85)$$

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

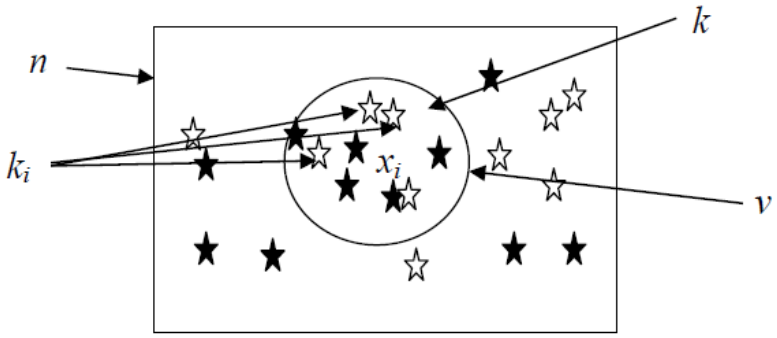
$$\theta \sim P(x)v \qquad (2-86)$$

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 ، فإن:

$$\theta \sim \frac{k}{n} \qquad (2-87)$$

ومن الصيغتين $(2-86)$ و $(2-87)$ نحصل على تقدير دالة الكثافة لـ x كما يأتي :

$$\hat{P}(x) = \frac{k}{nv} \qquad (2-88)$$



الشكل $(2-3)$ يوضح خوارزمية الجار الاقرب

2.13.2 : قاعدة قرار بيز[70] *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 $\hat{P}(\omega|x)$، وعليه يمكن تأشير النمط x إلى الفئة أو الصنف w_m إذا تحقق الشرط التالي :

$$\hat{P}(\omega_m|x) > \hat{P}(\omega_i|x) \quad ,\forall\, i \in \mathrm{c} \qquad (2-89)$$

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

$$\hat{P}\big(\omega_j|x\big) = \frac{P\big(x|\omega_j\big).P(\omega_j)}{P(x)} \qquad (2-90)$$

نحصل على قاعدة القرار الخاصة بالتعرف على الأنماط :

$$P(x|\omega_m).P(\omega_m) > P\big(x|\omega_j\big).P(\omega_j) \qquad (2-91)$$

إذ إن $P(x)$ حذف من طرفي المعادلة $(90-2)$.

مما تقدم يمكن استخدام دالة الكثافة الاحتمالية في الصيغة $(88-2)$ لتحديد قاعدة للقرار للتصنيف باستعمال دالة الجار الأقرب، فلو فرضنا أن أول أقرب مجموعة من الأنماط عددها k هناك k_m تقع في الصنف أو الفئة ω_m بحيث إن :

$$\sum_{\mathrm{m}=1}^{c} k_m = k \qquad (2-92)$$

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

$$\sum_{\mathrm{m}=1}^{c} n_m = n \qquad (2-93)$$

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

$$\hat{P}(x|\omega_m) = \frac{k_m}{n_m \,.\, v} \qquad (2-94)$$

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

$$\hat{P}(\omega_m) = \frac{n_m}{n} \qquad (2-95)$$

وبتعويض كل من المعادلتين $(2-94)$ و $(2-95)$ في قاعدة قرار بيز $(2-91)$ نحصل على الآتي :

$$\frac{k_m}{n_m \,.\, v}\,\frac{n_m}{n} > \frac{k_i}{n_i \,.\, v}\,\frac{n_i}{n} \qquad ,\forall\, i \in \mathrm{c} \qquad (2-96)$$

وعلية سيتم تحديد النمط x إلى الفئة أو الصنف ω_m إذا تحقق الشرط الآتي:

$$k_m > k_i \quad ,\forall\, i \in \mathrm{c} \qquad (2-97)$$

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 أن:

$$h(x) = \frac{k_i}{k} \qquad ,\forall\, i \in \mathrm{c} \qquad (2-98)$$

3.13.2 :مقاييس التشابه والمسافة [39] *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* التي تعود إلى العالم الرياضي الاسكندينافي اليوناني أقليديس ، وكما في الصيغة الآتية :

$$D(x_i, x_j) = \sqrt{\sum_{t=1}^{n} (x_{it} - x_{jt})^2} \qquad (2-99)$$

أما شروط المسافة فهي :

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

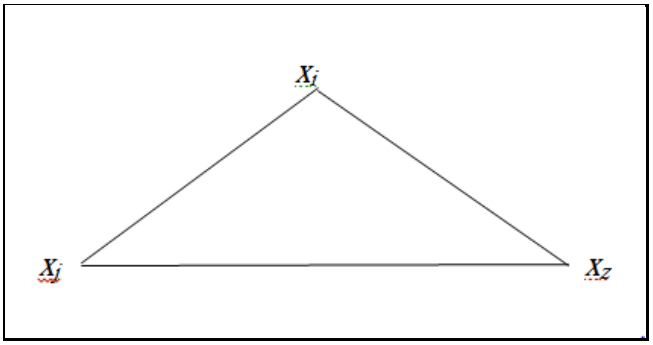
2. المسافة بين النقطتين x_i و x_j هي نفسها المسافة بين النقطتين x_j و x_i اي ان :

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

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

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

ويدعى هذا الشرط بمتباينة المثلث، كما في الشكل $(\mathbf{2-4})$:



الشكل $(\mathbf{2-4})$ يوضح متباينة المثلث

14.2 : 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 [8] .

1.14.2 : احتمال خطأ التصنيف

يعرف خطأ التصنيف بأنه احتمال تصنيف مفردة (مشاهدة) معينة إلى المجموعة (i) بينما هي في الحقيقة تعود للمجموعة (j) وبالعكس، وهذا التصنيف غير الصحيح يحدث عند تحديد النقطة الفاصلة بين المجموعتين [45].

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.

1.1.14.2 : احتمال خطأ التصنيف في حالة مجموعتين[48]:

أولاً: احتمال خطأ التصنيف $\boldsymbol{P_{12}}$ وهو احتمال تصنيف المشاهدة إلى المجموعة الثانية وهي تعود إلى المجموعة الاولى.

$$P_{12} = \Phi\left(-\frac{\delta}{2}\right) \qquad (2-100)$$

ثانياً: احتمال خطأ التصنيف $\boldsymbol{P_{21}}$ وهو احتمال تصنيف المشاهدة إلى المجموعة الاولى وهي تعود إلى المجموعة الثانية.

$$P_{21} = \Phi\left(-\frac{\delta}{2}\right) \qquad (2-101)$$

إذن :

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

وسيكون احتمال خطأ التصنيف كالآتي:

$$P_{12} = P_{21} = \Phi\left(-\frac{D}{2}\right) \qquad (2-102)$$

جدول (1 – 2) يبين حساب نسبة التصنيف الصحيح

المجموع	التصنيف		التصنيف الصحيح (Y)
	المجموعة الثانية	المجموعة الاولى	
P_1	P_{12}	P_{11}	المجموعة الاولى
P_2	P_{22}	P_{21}	المجموعة الثانية
المجموع الكلي	P_2	P_1	المجموع

2.1.14.2 : احتمال خطأ التصنيف في حالة اكثر من مجموعتين

في حالة لدينا k من المجاميع تتوزع توزيع طبيعي بمتجهات متوسطات مختلفة ومصفوفة تباين مشتركة لكل المجاميع ∑ نستخدم :

$$D_i^2 = \left(\overline{X} - \overline{X}_i\right)' \mathrm{S}^{-1} \left(\overline{X} - \overline{X}_i\right) \qquad (2 - 103)$$

وقاعدة التصنيف ستكون بأن تصنف المشاهدة x الى المجموعة i من خلال اختيار اقل القيم للمقياس D_i^2 مقارنة بكل المجاميع.

الفصل الثالث

الجانب التطبيقي

المبحث الاول : الجانب الطبي

المبحث الثاني : الجانب الإحصائي

الفصل الثالث

الجانب التطبيقي

1.3 : مقدمة *Introduction*

تم جمع البيانات من خلال اخذ عينة عشوائية بسيطة حجمها **270** من طبلات المرضى الراقدين في مستشفى (المرجان) 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 الآتية لكل مجموعة:

(الجنس، العمر، مهنة المريض، نوع الورم، حالة خروج المريض، فترة بقاء المريض في المستشفى).

المبحث الاول: الجانب الطبي

1.1.3 : السرطان[75] *Cancer*

السرطان هو مجموعة من الأمراض التي تتميز خلاياها بالعدائية (وهو النمو والانقسام الخلوي غير المحدود)، وقدرة هذه الخلايا المنقسمة على غزو الأنسجة المجاورة وتدميرها، أو الانتقال إلى أنسجة بعيدة في عملية نطلق عليها اسم النقلية . وهذه القدرات هي صفات الورم الخبيث على عكس الورم الحميد، الذي يتميز بنمو محدد وعدم القدرة على الغزو وليس لهُ القدرة على الانتقال أو النقلية. كما يمكن أن يتطور الورم الحميد إلى سرطان خبيث في بعض الأحيان.

فهو نمو الخلايا بشكل غير طبيعي لنسيج من انسجة الجسم فهو يصيب أنواعا مختلفة من الأعضاء وتختلف الأعراض عادة بإختلاف العضو أو النسيج المصاب . ويمكن توقي العديد من السرطان بتجنّب التعرّض لعوامل الإخطار الشائعة، مثل دخان التبغ. كما يمكن علاج نسبة كبيرة من السرطانات عن طريق الجراحة أو المعالجة الإشعاعية أو المعالجة الكيميائية، خصوصاً إذا تم الكشف عنها في مراحل مبكّرة .

يصيب السرطان كل المراحل العمرية عند الإنسان حتى الأجنة، ولكن تزيد مخاطر الإصابة به كلما تقدم الإنسان في العمر . كما يصيب السرطان الإنسان فإن أشكال منه تصيب الحيوان والنبات على حد سواء.

في الأغلب، يعزى تحول الخلايا السليمة إلى خلايا سرطانية إلى حدوث تغييرات في المادة الجينية المورثة . وقد يكون سبب هذه التغيرات عوامل مسرطنة مثل التدخين، أو الأشعة أو مواد كيميائية أو أمراض مُعدية (كالإصابة بالفيروسات). وهناك أيضا عوامل مشجعة لحدوث السرطان مثل حدوث خطأ عشوائي أو طفرة في نسخة الحمض النووي DNA عند انقسام الخلية، أو بسبب توريث هذا الخطأ أو الطفرة من الخلية الأم.

الأعراض

تنقسم أعراض السرطان إلى ثلاثة أقسام هي كالآتي:

1. **أعراض عامة :** كفقدان الوزن، تعب وإرهاق عام، فقدان الشهية، التعرق خصوصاً أثناء الليل.
2. **أعراض موضعية :** كظهور كتلة صلبة أو تغيرات في شكل سطح الجلد الخارجي.
3. **أعراض تدل على الانتشار:** كحدوث تضخم في الكبد أو في الغدد الليمفاوية المختلفة في الجسم أو ألم في العظام.

2.1.3 : أنواع السرطان

تم استعمال دالة التمييز لتشخيص ثلاث أنواع من الأورام السرطانية وهي اورام الثدي والعمود الفقري (سرطان العظم) والجاز التنفسي (سرطان الرئة) لكثرة شيوعها في مجتمعنا حاليا وهي :

1.2.1.3 : سرطان الثدي[67] *Breast Cancer*

سرطان الثدي هو نوع من أنواع السرطان يظهر في أنسجة الثدي. من علاماته تغير في شكل الثدي، وظهور كتلة في الثدي، و ظهور بقعة حمراء ذات قشور أو خروج سائل من الحلمة . في حالة انتشار المرض في الجسم.

اعراض الاصابة بسرطان الثدي

تظهر العلامات التالية: ألم في العظام، ضيق في التنفس أو اصفرار في الجلد، انتفاخ في الغدد الليمفاوية.

العلاج

توفر اليوم تشكيلة منوعة من العلاجات لكل مرحلة من مراحل المرض . غالبية النساء تخضع لعمليات جراحية لاستئصال الثدي، بالإضافة إلى العلاج الكيماوي، العلاج الهرموني أو الإشعاعي . كما إن هناك أيضا مجموعة متنوعة من العلاجات التجريبية لهذا النوع من السرطان.

2.2.1.3 : سرطان العظم[77] *Bone cancer*

سرطان العظام هو سرطان يبدأ في العظم نفسه (سرطان عظم أولي) وهو سرطان نادر الحدوث. هذه السرطانات تصيب الاطفال أكثر من الكبار.

أكثر أنواع سرطان العظام الأولية شيوعا هي:

1. **ساركومة عظمية *Osteosarcoma* :** وهي أصابة تحدث أساسا ضمن انسجه العظام النامية.
2. **ساركومة يوينغ *Ewing's sarcoma* :** وهي إصابة تنشأ في الأنسجة غير الناضجة في نخاع العظام.
3. **ساركومة غضروفية *Chondro sarcoma* :** وهي أصابة تحدث ضمن الغضروف.

اعراض الاصابة بسرطان العظم

الألم هو أكثر أعراض سرطان العظم شيوعا. يمكن لسرطان العظم ان ينشأ في اي من عظام الجسم، إلا أنه في معظم الأحيان يبدأ في العظام الطويلة في الذراعين والساقين.

العلاج

تعد الأدوية أحد الاهتمامات الرئيسية هي كثافة خسارة العظام، البايفوسفونيت غير الهرمونية *Non-hormonal bisphosphonates* تزيد من قوة العظام وهي متاحة كوصفة حبوب مرة أسبوعياً *Metastron* .المعروف ب *strontium-89 chloride* هو دواء عن طريق الوريد يعطى للتخفيف من الألم كل ثلاثة أشهر، ويتوقف علاج سرطان العظم على موقع الإصابة وحجمها.

3.2.1.3 : سرطان الرئة[78] *Lung Cancer*

يعد سرطان الرئة أحد أمراض الرئة التي تتميز بحدوث انقسامات خلوية غير مضبوطة للخلايا الحية، وقدرة هذه الخلايا المنقسمة على غزو الانسجة الأخرى للرئة والانتشار فيها ،إما عن طريق نمو مباشر باتجاه نسيج مجاور أو الانتقال وغزو أنسجة بعيدة في عملية يدعى باسم النقيلة. معظم انواع السرطان التي تبدأ في الرئة والمعروفة باسم سرطان الرئة الاولية هي التي تستمد من سرطان الخلايا الظهارية وتسمى أيضا سرطان خلية الشوفان والسبب الرئيس هو التعرض لدخان التبغ الذي يمثل تسعون بالمائة من حالات سرطان الرئة وهذه الحالات تنتسب غالبا إلى العوامل الوراثية أو غاز الرادون أو الاسبستوس

ويعد سرطان الرئة هو أكثر السرطانات شيوعاً في العالم (بعد سرطان الجلد، سرطان الثدي عند النساء، وسرطان البروستات عند الرجال) وهو المسبّب الأكثر للوفاة بأمراض السرطان.

اعراض الاصابة بسرطان الرئة

تشمل الأعراض ظهور آلام صدرية، تغير طبيعة سعال المدخن المزمن أو الشكوى من سعال جديد، نفث دم، ضيق تنفس، التهاب رئوي، خاصة المتكرر في نفس المكان وما يرافقه من حرارة وضيق تنفس، وأخيرا نقص في الوزن.

العلاج

علاج سرطان الرئة يعتمد على نوع الخلية السرطانية، مدى انتشاره وأداء المريض. العلاجات الشائعة تتضمّن الرعاية لتخفيف الألم، الجراحة، العلاج الكيميائي والعلاج الإشعاعي .

المبحث الثاني : الجانب الإحصائي

1.2.3 : تـعريـف 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، (تم احتساب هذا المتغير من خلال حساب تاريخ دخول المريض وتاريخ خروجه من المستشفى)

2.2.3 : بعض الاختبارات المهمة *Some important tests*

قبل أن نبدأ بتطبيق التحليل يجب إجراء بعض الاختبارات لنتحقق من شروط تطبيق التحليل التمييزي وكالاتي :

1.2.2.3 : اختبار التوزيع الطبيعي لكل مجتمع

نختبر البيانات لمعرفتها فيما إن كانت المتغيرات التوضيحية للمجموعات الثلاث لإمراض السرطان تتوزع طبيعي أم لا باستخدام (***Kolmogorov-Smirnov***) وبموجب البرنامج الإحصائي SPSS وحسب الفرضية التالية:

H_0: البيانات تتبع التوزيع الطبيعي

H_1: البيانات لا تتبع التوزيع الطبيعي

الجدول (1 – 3) نتائج اختبار البيانات للتوزيع الطبيعي

Variables	Kolmogorov-Smirnov	
	statistic	Sig.
الجنس (X_1)	**0.40**	**.000**
العمر (X_2)	**0.14**	**.000**
المهنة(X_3)	**0.34**	**.000**
حالة خروج المريض (X_4)	**0.48**	**.000**
فترة بقاء المريض (X_5)	**0.17**	**.000**

في **الجدول(1 – 3)** أظهرت نتائج قيم اختبار **(*Kolmogorov-Smirnov*)** إن مستوى المعنوية لجميع المتغيرات اقل من **0.05** المستوى المعتمد في هذه الدراسة وعليه نرفض فرضية العدم القائلة إن البيانات تتبع التوزيع الطبيعي. ولكون حجم البيانات **270** مشاهدة يمكن إن نعّد إن البيانات تقترب من التوزيع الطبيعي حسب نظرية (الغاية المركزية)[27].

2.2.2.3 : التأكد من عدم وجود ازدواج خطي بين المتغيرات التوضيحية

نتأكد من عدم وجود ارتباط بين المتغيرات التوضيحية الذي يؤثر وجوده في درجة دقة النتائج ، فإننا نقوم بحساب معامل التسامح حيث إن :

$$Tolerance = 1 - R^2_{xi\ ,\ others}$$

$R^2_{xi\ ,\ others}$: يمثل مربع معامل الارتباط المتعدد بين المتغير التابع وبقية المتغيرات التوضيحية.

ثم نستخرج قيمة VIF (Variance Inflation Factor) لكل من المتغيرات التوضيحية اذ إن :

VIF=1/Tolerance

الجدول (2 – 3) اختبار معامل الارتباط

Variables	Collinearity Statistics	
	Tolerance	VIF
الجنس (X_1)	.918	1.089
العمر (X_2)	.818	1.222
المهنة(X_3)	.815	1.227
حالة خروج المريض (X_4)	.916	1.092
فترة بقاء المريض (X_5)	.971	1.030

يتضح من **الجدول (2 – 3)** إن قيمة VIF لكل المتغيرات اقل من 5.00 يمكن أن الاستنتاج بأنه لا يوجد مشكلة ارتباط متعدد (ازدواج خطي) بين المتغيرات المستقلة .

3.2.2.3 : التأكد من عدم وجود قيم شاذة

في ما يتعلق بعملية التأكد من عدم وجود قيم شاذة للبيانات في كافة المتغيرات التوضيحية فانة يمكن اجراء اختبار ***Mahalanobis*** تحت الفرضية الآتية :

H_0: عدم وجود قيم شاذة

H_1: وجود قيم شاذة

الجدول (3 – 3) يبين نتائج قيم مهانلوبس

MAH	MAH	MAH	MAH	MAH	MAH	MAH	MAH	MAH	MAH
1.60	3.62	4.50	1.33	1.21	8.46	6.96	6.30	4.04	8.38
2.96	7.70	3.48	3.26	1.45	2.30	4.92	10.61	6.93	1.76
3.26	3.82	1.07	4.86	1.60	3.48	6.51	4.92	2.60	1.42
3.22	3.82	3.58	3.26	1.32	8.24	5.34	8.49	4.13	14.32
1.27	5.62	3.82	7.12	1.84	8.54	7.50	7.86	4.59	15.23
4.17	3.80	7.50	5.55	14.96	8.46	4.52	10.22	3.58	98.
2.30	3.89	2.40	4.20	1.42	8.91	6.51	3.53	11.64	1.34
8.82	4.51	2.15	6.39	1.09	2.17	5.19	4.08	3.65	1.60
4.72	1.10	20.43	2.72	2.96	1.90	4.52	4.16	4.22	1.85
2.24	3.88	4.89	3.58	1.04	2.11	6.33	3.91	3.36	7.87
5.05	3.94	6.67	3.00	1.04	2.13	7.04	4.16	7.79	7.52
3.88	3.46	5.20	3.54	1.04	6.56	7.81	6.75	5.46	2.57
4.66	2.30	4.06	2.80	4.64	1.42	6.21	6.75	3.95	1.60
3.75	1.36	2.73	2.92	2.17	2.31	5.25	8.91	4.22	8.38
15.75	1.37	3.78	6.97	1.94	2.52	6.21	6.75	10.87	14.27
7.48	1.45	10.76	5.19	2.79	1.65	8.88	7.68	14.06	3.91
5.21	1.01	3.96	3.89	7.16	2.04	3.91	6.78	2.96	8.01
2.61	1.04	4.30	8.51	1.17	1.00	4.55	6.96	8.29	10.37
6.34	2.08	4.60	1.55	99.	5.06	4.68	6.63	4.82	8.57
3.21	2.95	5.98	1.69	8.18	5.32	6.33	7.19	3.54	13.67
3.59	5.60	3.91	4.88	7.54	8.18	3.19	6.08	10.29	9.41
6.04	2.11	2.40	5.67	3.22	8.18	10.13	5.20	9.09	7.80
4.90	2.12	5.32	1.65	9.44	3.93	2.57	5.20	1.42	2.33
5.77	2.96	2.97	2.96	8.08	2.57	8.01	5.08	4.58	1.55
4.63	3.70	8.85	2.94	5.32	1.09	1.07	5.25	4.55	5.01
16.23	3.56	10.57	4.36	3.14	3.22	1.01	4.05	9.12	7.46
4.60	3.40	6.86	6.17	2.30	4.34	1.27	7.14	4.98	4.69

بمراجعة كافة القيم نجد ان كل من هذه القيم اقل من القيمة الجدولية لـ χ^2 عند درجة حرية $(1-3)(1-5)$ وبمستوى معنوية 0.995 والتي تساوي (21.95) وبناءً عليه لا نرفض فرضية العدم القائلة بعدم وجود قيم شاذة بين كل البيانات المتعلقة بكافة المتغيرات التوضيحية.

4.2.2.3 : اختبار شرط تجانس التباين

لمعرفة مدى تجانس أفراد المجموعات نختبر الفرضية الآتية:

$\mathbf{H_0}: \Sigma_1 = \Sigma_2 = \Sigma_3$
$\mathbf{H_1}: \Sigma_1 \neq \Sigma_2 \neq \Sigma_3$

ويمكن الاستعانة باختبار (Box's M)

الجدول (4 – 3) يبين معامل التحديد

Log Determinants		
Y	**Rank**	**Log Determinant**
سرطان الثدي	**2**	**1.287**
سرطان العظم	**2**	**3.297**
سرطان الرئة	**2**	**3.545**
Pooled within-groups	**2**	**3.094**

الجدول (5 – 3) يبين اختبار تجانس التباينات بين المجاميع

Box's M		98.14
F	Approx.	16.150
	df1	6
	df2	342154.313
	Sig.	.000

وفقاً **لجدول (4 – 3)** أشارت النتائج إلى إن قيم log Determinant تقريبا متساوية للمجاميع ومن اختبار تساوي التباينات في **الجدول (5 – 3)** وجد إن قيمة (Box's M=98.14) وبمستوى معنوية (0.000) وهذا يدل على عدم تجانس التباين بين المجموعات، وهذا يحدث في حالة البيانات الكبيرة (ففي البحث 270 مشاهدة) هذا الاختبار حساس جداً لاختبار فرضية تجانس التباين ولهذا يمكن التعامل مع قيم Log Determinant لتفسير نتائج اختبار تجانس التباين والتباين المشترك حيث كانت قيمه متقاربة ومحصورة بين (3.545 - 1.287) لذلك نسبياً يفترض تجانس المصفوفات للتباينات المشتركة[26].

3.2.3 : خطوات اجراء تحليل التمايز

1.3.2.3 : اختيار المتغيرات المكونة للمعادلة التمييزية

تم اختبار معنوية المتغيرات لمعرفة أهمية كل متغير بشكل منفرد ومدى تأثيره في بناء الدالة التمييزية الخطية وكانت النتائج كما في **الجدول (6 – 3) :**

الجدول (6 – 3) يبين اختبار معنوية متغيرات الدالة التمييزية

	Wilks' Lambda	F	df_1	df_2	Sig.
الجنس (X_1)	.728	49.99	2	267	.000
العمر (X_2)	.251	398.88	2	267	.000
مهنة المريض (X_3)	.877	18.74	2	267	.000
حالة خروج المريض (X_4)	.936	9.18	2	267	.000
فترة بقاء المريض (X_5)	.971	4.03	2	267	.019

من خلال **الجدول (6 – 3)** يتضح إن قيمة (**P-Value**=0.000) اقل من 0.05 نستنتج ان جميع متغيرات الدراسة لها تأثير واهمية في تكوين وبناء الدالة التميزية .

2.3.2.3 : المعاملات التمييزية المعيارية

الجدول (7 – 3) يبين المعاملات دالة التمييز المعيارية

Standardized Canonical Discriminate Function Coefficients		
	Function	
	الدالة الاولى	الدالة الثانية
الجنس (X_1)	.180	.995
العمر (X_2)	1.011	-.032-

الجدول (7 – 3) يبين الأهمية النسبية للمتغيرات التوضيحية في تقدير المتغير المعتمد ونلاحظ إن متغير العمر**(X_2)** التي بلغت قيمته المعيارية (1.011) له التأثير الأكبر وله زيادة قوة التمييز بين المجاميع في الدالة التمييزية الأولى، أما بالنسبة للدالة التمييزية الثانية فقد كان متغير الجنس**(X_1)** المساهمة الكبيرة في تكوين تلك الدالة وبقيمة معيارية مقدارها (0.995).

3.3.2.3 : المعاملات التمييزية غير المعيارية

الجدول (8 – 3) يبين المعاملات دالة التمييز غير المعيارية

Canonical Discriminate Function Coefficients		
	Function	
	الدالة الاولى	الدالة الثانية
الجنس (X_1)	.432	2.383
العمر (X_2)	.089	-.003-
(Constant)	-5.082-	-3.711-

يوضح **الجدول (8 – 3)** المعاملات التمييزية غير المعيارية للارتباط بين كل متغير من المتغيرات المنبئة الداخلة في التحليل وبين الدالة التمييزية وتحسب الدرجة التمييزية من خلال ضرب المعاملات التمييزية غير المعيارية في قيم المتغيرات المدخلة ثم جمع الناتج وإضافته إلى القيمة الثابتة (5.082 -) و(3.711-) .

4.2.3 : اختبار قدرة الدالة التمييزية على التمييز

1.4.2.3 : مقياس ويلكس *Wilks' Lambda*

يستعمل هذا المقياس لاختبار وجود علاقة خطية بين المتغيرات وذلك بالاعتماد على الفرضية الآتية:

$\mathbf{H_0}$: لا توجد علاقة خطية بين المتغيرات

$\mathbf{H_1}$: توجد علاقة خطية بين المتغيرات

تكون صيغة مقياس ***(Wilks' Lambda)*** كما في المعادلة $(3-2)$ وهي تتوزع تقريبا χ^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** $(9-3)$:

الجدول $(9-3)$ اختبار معنوية الدالة التمييزية

Test of Function(s)	Wilks' Lambda(Λ)	Chi-square	df	Sig.
1 through 2	**.179**	**458.572**	**4**	**.000**
2	**.729**	**84.243**	**1**	**.000**

أظهرت نتائج الاختبار من خلال **الجدول** $(9-3)$ وجود فروق معنوية بين المتوسطات الثلاث إذ إن قيمة P-value لـ(χ^2) اقل من (0.05) أي نرفض فرضية العدم ونرجح الفرضية البديلة القائلة عدم تساوي المتوسطات الثلاث 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 الثلاث.

2.4.2.3 : اختبار F

بالنسبة لإحصائية اختبار F فنستخدم المعادلة $(6-2)$ وتكون النتيجة كالآتي:

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

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

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

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

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

بدرجات حرية

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

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 .

3.4.2.3 : القيم الذاتية

الجدول (10 − 3) يبين القيم الذاتية

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 (10 − 3) يبين قيمة ***Eigen values*** للدالة التمييزية اذ 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** (**11 – 3**):

الجدول (**11 – 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 الآتي:

الجدول (**12 – 3**) 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** (**12 – 3**) 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 (13 – 3) يبين المتغيرات غير الداخلة 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	.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 (13 – 3)** 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 وهي مهنة المريض**(X_3)**، حالة خروج المريض**(X_4)**، فترة بقاء المريض**(X_5)**.

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

الجدول (14 – 3) 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	.000
2	2	.179	2	2	267	181.41	4	532.00	.000

في **الجدول (14 – 3)** كانت Wilks' Lambda 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 مساوية الى (0.000).

اما المصفوفة الهيكلية فهي كالتالي:

الجدول (15 – 3) يبين المصفوفة الهيكلية

Structure Matrix		
	Function	
	الدالة الاولى	الدالة الثانية
العمر (X_2)	.984*	-.179-
حالة خروج المريض (X_4[b])	-.139-*	.015
فترة بقاء المريض (X_5[b])	.031	1.000*
الجنس (X_1)	.055	-.282-*
مهنة المريض (X_3[b])	-.032-	-.038-*

يبين **الجدول (15 – 3)** الارتباط داخل المجموعات بين كل متغير من المتغيرات المنبئة الداخلة في التحليل وقيمة الدالة التمييزية وقد كان معامل الارتباط مع متغير العمر **(X_2)** أقواها إذ بلغ (0.984) ثم الارتباط مع متغير الجنس **(X_1)** الذي بلغ (0.055) أما بقية المتغيرات فقد تم استبعادها من التحليل ووضع رمز مشير إليها. وتعني الإشارة السالبة وجود علاقة عكسية بين المتغير المعتمد تشخيص المرض (مرض السرطان) وبين المتغيرات التوضيحية.

ويمكن صياغة دوال التمييز بالشكل الآتي:

$$Y_1 = 0.984 - 0.139 + 0.031 + 0.055 - 0.032$$

$$Y_2 = -0.79 + 0.015 + 1 - 0.282 - 0.038$$

وقيم المتوسطات التي تميز المجاميع لدالة التمييز الاولى والثانية كما يأتي:

الجدول (16 – 3) يبين الدالة التمييزية ومتوسطات المجموعات

Functions at Group Centroids		
Y	Function	
	الدالة الاولى	الدالة الثانية
سرطان الثدي	.372	.833
سرطان العظم	-3.408-	-.257-
سرطان الرئة	1.208	-.506-

من **الجدول (16 – 3)** نلاحظ قيم المتوسطات والتي تميز بين المجاميع بالنسبة للدالة التمييزية الاولى قد بلغت (0.372، 3.408- ، 1.208). وهذا يعني ان قيمة الدالة التمييزية لإحدى المشاهدات بالنسبة للدالة الاولى اذا كانت سالبة ، فان المشاهدة تصنف ضمن المجموعة الثانية.

اما اذا كانت قيمة الدالة التمييزية موجبة ، فإنها تصنف ضمن المجموعة الاولى او الثالثة ويتم الانتقال إلى الدالة التمييزية الثانية لتصنيف تلك المشاهدة بشكل ادق ضمن احدى هذه المجاميع .

5.2.3 : الاحتمال للتصنيف 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 دوال التمييز الأربعة.

1.5.2.3 : تصنيف دالة التمييز الخطية

الجدول (17 – 3) 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 (17 – 3)** إلى مدى دقة النتائج النهائية لتصنيف الدالة الخطية 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 وبناءً عليه فإن 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 وبناءً عليه فإن 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) .

2.5.2.3 : تصنيف دالة التمييز التربيعية

الجدول (18 – 3) يبين تصنيف الدالة التربيعية

التصنيف الصحيح (Y)	التصنيف			المجموع
	المجموعة الاولى	المجموعة الثانية	المجموعة الثالثة	
المجموعة الاولى	87	0	5	92
	94.57	0.00	5.43	%100
المجموعة الثانية	1	52	1	54
	1.85	96.30	1.85	%100
المجموعة الثالثة	47	2	75	124
	37.90	1.61	60.48	%100
المجموع	135	54	81	270
	50.00	20.00	30.00	

يبين **الجدول (18 – 3)** مدى دقة النتائج النهائية لتصنيف الدالة التربيعية إذ يتبين إن (87) حالة من المجموعة الأولى وبنسبة 94.57 % قد تم تصنيفها بشكل صحيح وإن (5) حالات وبنسبة 5.43% 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 وبناءً عليه فإن 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 (75) حالة وبنسبة 60.81% تم تصنيفها بشكل صحيح وبناءً عليه فإن هناك (47) حالة ونسبة 37.90% 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 (214) حالة أي ما نسبته 79.26% من حالات العينة البالغة (270) .

3.5.2.3 : تصنيف دالة التمييز اللوجستية

الجدول (19 – 3) يبين تصنيف الدالة اللوجستية

التصنيف الصحيح (Y)	التصنيف			المجموع
	المجموعة الاولى	المجموعة الثانية	المجموعة الثالثة	
المجموعة الاولى	83	1	8	92
	90.22	1.09	8.70	%100
المجموعة الثانية	1	52	1	54
	1.85	96.30	1.85	%100
المجموعة الثالثة	42	2	80	124
	33.87	1.61	64.52	%100
المجموع	126	55	89	270
	46.67	20.37	33.96	

يبين **الجدول (19 – 3)** إلى مدى دقة النتائج النهائية لتصنيف الدالة اللوجستية إذ يتبين إن (83) حالة من المجموعة الأولى وبنسبة 90.22 % قد تم تصنيفها بشكل صحيح وبناءً عليه فإن هناك حالة واحدة وبنسبة 1.09% تم تصنيفها بشكل خاطئ إذ صنفت ضمن المجموعة الأولى وهي تعود إلى المجموعة الثانية وكذلك يتبين إن هناك (8) حالات وبنسبة 8.70% 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 وبناءً عليه فإن 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 (80) حالة وبنسبة 64.52% تم تصنيفها بشكل صحيح وبناءً عليه فإن هناك (42) حالة ونسبة 33.87% 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 (215) حالة أي ما نسبته 79.62% من حالات العينة البالغة (270).

4.5.2.3 : تصنيف دالة الجار الأقرب

الجدول (20 – 3) 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 (20 – 3)** مدى دقة النتائج النهائية لتصنيف دالة الجار الأقرب 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 وبناءً عليه فإن 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.

وفي الجدول الآتي تلخيص لحالات التصنيف الصحيح للنماذج الأربعة:

الجدول (21 – 3) يبين نسبة التصنيف الصحيح

دالة التمييز	التصنيف الصحيح	نسبة التصنيف الصحيح %
دالة التمييز الخطية	**218**	**80.74**
دالة التمييز التربيعية	**214**	**79.26**
دالة التمييز اللوجستية	**215**	**79.62**
دالة الجار الأقرب	**260**	**96.30**

يبين **الجدول (21 – 3)** المقارنة بين النماذج الأربعة (**دالة التمييز الخطية ، دالة التمييز التربيعية ، دالة التمييز اللوجستية ، دالة الجار الأقرب**) وتبين إن نموذج دالة الجار الأقرب أفضل النماذج إذ يحوي على أعلى نسبة احتمال التصنيف الصحيح أي اقل خطأ تصنيف ممكن .

أي أن النموذج اللامعلمي كان هو الافضل في تمثيل بيانات الدراسة كون البيانات كانت تعاني من مشاكل في التوزيع الطبيعي وتجانس التباينات .

الفصل الرابع

الاستنتاجات والتوصيات

الفصل الرابع

الاستنتاجات والتوصيات

يتضمن هذا الفصل أهم الاستنتاجات التي تم الوصول إليها اعتماداً على ما تم عرضه في الجانب النظري ونتائج العمل التي تم الحصول عليها في الجانب التطبيقي من خلال التحليل الإحصائي ، فضلاً عن التوصيات التي تم الوصل إليها.

4.2 : الاستنتاجات *Conclusions*

1. لوحظ ان المتغيرين الافضلين في تمييز المرضى المصابين بالأورام السرطانية الثلاثة كانا متغير العمر ومتغير الجنس فقد كان معامل الارتباط بين الدالة التمييزية ومتغير العمر(X_2) ارتباط طردي قوي جداً، ثم الارتباط مع متغير الجنس (X_1) ارتباط طردي ضعيف أما بقية المتغيرات فقد تم استبعادها من التحليل، وتبين وجود علاقة ارتباط بين مهنة المريض وجنسه.

2. تبين من 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 وعند اختبار معنوية المتغيرات التوضيحية ظهرت جميع المتغيرات لها تأثير واهمية في تكوين وبناء الدالة التميزية.

3. عند اختبار القيم الذاتية تبين ان قيمة ***Eigen values*** للدالة التمييزية قدرة عالية على التمييز حيث ان قيمة ***Eigen values*** اكبر من الواحد الصحيح . اما في ما يتعلق بالارتباط القانوني ***Canonical Correlation*** فقيمته تدل 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، ويليه معامل الارتباط بين متغير الجنس(X_1) ومتغير العمر(X_2)، ويليه معامل الارتباط بين متغير العمر(X_2) ومتغير حالة خروج المريض(X_4) .

5. عند اختبار نسبة التصنيف الصحيح تبين ان دالة الجار الاقرب افضل النماذج اذ كانت اعلى نسبة، تليها دالة التمييز الخطية ثم دالة التمييز اللوجستية واخيراً دالة التمييز التربيعية.

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 .

3.4 : التوصيات *Recommendations*

اعتمادا على الاستنتاجات التي توصل اليها الباحث يدرج ادناه التوصيات التي نعتقد بانها ضرورية وعملية في التطبيق.

1. ضرورة تطبيق التصنيف على وفق الصيغ الأربع (دالة التمييز الخطية، دالة التمييز التربيعية، دالة التمييز اللوجستية، دالة الجار الاقرب) في مجالات أخرى غير الطبية، وكلك 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.

2. ينبغي أستخدام الدوال التمييزية الأخرى (دالة كيرنل، دالة الرتبة، الدالة اللوجستية التربيعية، نايف بايز للتصنيف).

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

4. ضرورة تطوير مراكز الإحصاء في المستشفيات فضلاً عن معلومات جديدة للمرضى المصابين.

5. ضرورة حث المرضى لمراجعة المستشفيات والمستوصفات لتشخيص المرض وإجراء العمليات الجراحية اللازمة في وقت مبكر.

المصادر

- القرآن الكريم

أولاً: المصادر العربية *Arabic Reference*

1. أبو علام، رجاء ، **"التحليل الإحصائي للبيانات بأستخدام برنامج SPSS"** ، دار النشر للجامعات، القاهرة ، ص224، (2003).

2. إحسان، نازك، **"استخدام الدالة المميزة للتنبؤ بنتيجة الطالب"** ، بحث منشور ،المجلة العراقية لبحوث السوق وحماية المستهلك ،مجلد(3) ، العدد(5) ، معهد الإدارة /الرصافة، (2011).

3. احمد ، فارس غانم، **" تسوس الأسنان لدى الأطفال (دراسة إحصائية تمييزية)"**، مجلة تنمية الرافدين العدد (51) ، (1997).

4. البكري، رباب عبد الرضا، **"مقارنة بعض طرق التقدير للدالة المميزة"**، رسالة ماجستير، كلية الإدارة والاقتصاد، جامعة بغداد ، (1997).

5. البكري، رباب عبد الرضا و رضا، صباح منفي و الليف، عادلة عبد، **"مقارنة بين انموذج اللوجستي وانموذج التحليل المميز الخطي بأستعمال المركبات الرئيسية لبيانات البطالة لمحافظة بغداد"** بحث منشور، مجلة العلوم الاقتصادية والادارية، المجلد (23)، العدد(95)، كلية الادارة والاقتصاد، جامعة بغداد،(2017).

6. البلداوي، تسنيم حسن كاظم، **"مقارنة تحليلية بين نماذج اللوجستك ونماذج الدوال التمييزية"** ، أطروحة دكتوراه فلسفة في الإحصاء كلية الإدارة الاقتصاد، جامعة بغداد، (1996).

7. الجاعوني، فريد و غانم، عدنان، **"التحليل الإحصائي متعدد المتغيرات (التحليل التمييزي) في توصيف وتوزيع الأسر داخل الهيكل الاقتصادي الاجتماعي في المجتمع"** ، بحث منشور ، مجلة جامعة دمشق للعلوم الاقتصادية والقانونية، المجلد الثالث والعشرون، العدد الثاني، (2007).

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

9. الحنيطي، دوخي وآخرون ، **" تمييز الأسر الفقيرة من غير الفقيرة في المناطق النائية التابعة لإقليم جنوب الأردن"** ، مجلة التنمية والسياسات الاقتصادية، المجلد السابع، العدد الأول ،(2004).

10. الزوبعي، عبيد محمود و المشهداني، كمال علوان ،**"اقتراح انموذج إحصائي لتصنيف الطلبة إلى الفرعين العلمي والأدبي"** ، بحث منشور في مجلة العلوم الاقتصادية والإدارية العدد الثاني، (1988).

11. السليماني، مؤيد سلمان عباس ، **"استخدام الدالة المميزة لتشخيص حالات التهاب الأمعاء عند الأطفال الرضع"** ، رسالة ماجستير في الإحصاء، كلية الإدارة والاقتصاد، جامعة بغداد، (1998).

12. الكاتب، محمد أسامة، **"الدالة المميزة الخطية في حالة أكثر من مجموعتين"** ، بحث منشور، مجلة التربية والعلم ، المجلد(23) ، العدد (4)، قسم أنظمة الحاسبات، المعهد التقني، نينوى، (2010).

13. بيثون، نغم نافع متي، **"خواص قوة الاختبار وحدود الثقة لمعاملات انموذج اللوجستيك الخطي دراسة مقارنة"** ، رسالة ماجستير في الإحصاء، كلية الإدارة والاقتصاد، جامعة بغداد، (1992).

14. جبارة، أزهار كاظم، **"تحليل البيانات متعددة الاستجابة لتشخيص أمراض العيون بأستخدام الدالة التمييزية والانحدار اللوجستي (دراسة مقارنة)"**، رسالة ماجستير، كلية الإدارة والاقتصاد ، الجامعة المستنصرية ، (2014).

15. حسين، محمد رمضان عتاب، **"استخدام الدالة المميزة التربيعية في تمييز أنماط الأرقام العربية"** ،اطروحة دكتوراه، كلية الإدارة والاقتصاد ، جامعة بغداد ، (1999).

16. حمودات، الاء عبد الستار، **"الدالة التمييزية وطرق تحديد متغيراتها"** رساله ماجستير في علوم رياضيات، جامعه الموصل، (2005).

17. حميد، رند سليم ، **"استخدام الدالة المميزة في تشخيص بعض الأورام السرطانية"** ، رسالة ماجستير في الإحصاء، كلية الإدارة والاقتصاد، جامعة بغداد ، (1991).

18. رشيد، ظافر حسين، **"انموذج احتمالي لتشخيص الحالة الصحية للطفل الخديج"**، بحث منشور في مجلة العلوم الاقتصادية والإدارية المجلد الثاني، العدد الثالث ، (1995).

19. رشيد، ظافر حسين و خمو، خلود يوسف و حميد، رند سليم ، **"تشخيص أمراض قرحة المعدة باستخدام دالة التمييز اللوجستية"**، مجلة كلية الإدارة والاقتصاد، جامعة بغداد، المجلد0 ،عدد15-10 ،ص 45، (2001).

20. سليمان، علي ابشر فضل المول ، **"المقارنة بين التحليل التمييزي والانموذج اللوجستي الثنائي ونماذج الشبكات العصبية في تصنيف المشاهدات"**، أطروحة دكتوراه في علوم الإحصاء ، كلية الدراسات العليا ، جامعة السودان للعلوم التكنولوجيا، (2015).

21. شومان، عبد اللطيف حسن ، **"التحليل المميز وتطبيقه في تصنيف طلاب المدرسة الثانوية"** ، رسالة ماجستير، جامعة الموصل، (1977).

22. صالح، سميرة محمد ، **"استخدام التحليل المميز (التصنيفي) لتحديد أهم العوامل المؤثرة في تسرب ورسوب الطلبة في جميع المراحل الدراسية "**،رسالة ماجستير في الإحصاء ،كليه الإدارة والاقتصاد ،جامعه السليمانية ، (2003).

23. عباس، لازم محمد و خلف ، علي عطشان و طنيش ، مشرق عزيز، **"التصنيف وفقاً لبعض القدرات البدنية الخاصة والأداء المهاري المركب والقياسات الجسمية للاعبي كرة السلة"** ، بحث منشور ، كلية التربية البدنية وعلوم الرياضة ، جامعة القادسية، (2016).

24. عبد الرزاق، أسيل ، **"استخدام الدالة التمييز الخطية لتصنيف مرضى الثلاسيميا في مستشفى كركوك العام"** ، بحث منشور، مجلة الإدارة والاقتصاد، العدد الثامن والسبعون، الجامعة المستنصرية، (2009).

25. عبد الكريم، أنوار ضياء ، **"استخدام الطرائق التمييزية الإحصائية لتشخيص بعض أمراض القلب"** ، بحث منشور، مجلة جامعة كركوك الدراسات العلمية، المجلد(1)، العدد(2)، كلية العلوم ، جامعة كركوك ، (2006).

26. نجيب، حسين والرفاعي، غالب، **"تحليل ونمذجة البيانات باستخدام الحاسوب تطبيق شامل للحزمة SPSS"** ، دار الأهلية للنشر والتوزيع، عمان، ص 445، (2006).

27. هرمز، امير حنه، **" الاحصاء الرياضي"**، قسم الاحصاء، كلية الادارة والاقتصاد، جامعة الموصل، ص 482، (1990).

ثانياً: المصادر الأجنبية *Foreign Reference*

28. Abbas, F. M. Azhar, M. E , **"Comparing Discriminant Analysis and Logistic Regression Model as a Statistical Assessment Tools of Arsenic and Heavy Metal Contents in Cockles"** School of Industrial Technology, Environmental Technology Division Universiti Sains Malaysia, 11800 Penang, Malaysia , (2008).

29. Afifi A.A. and Clark V., **"Computer Aided Multivariate Analysis"**, Life time Learing Publicatons, Belmont, Catlifornia, U.S.A, (1984).

30. Ager, J.W, JR and Brent, S.B, **"An index of agreement between ahypothesized partial order and an empirical Rank order"**, American statistical Association, volume 73, number 364, Theory and method section, (1978).

31. Al-Iathary, I., A., **"Comparison of discriminant techniqus to a Binary set of data of PSYCHOLGICALLY disturbed patients"**, Tanmlyat Al-Rafiduin (42), (1994).\

32. Alkhatib, Khalid, Najadat, Hassan, Hmeidi, Ismail & Saharawi Mohammed **," Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm"**, International Journal of Business, Humanities and Technology, Vol. 3 No. 3 ,pp.32-44, (2013).

33. Altman, E.I., **"Financial Ratios Discriminant analysis and the prediction of corporate Bankruptcy"**, Journal of finance, 1968.

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

35. Anderson, J.A., **"Separate Sample Logistic Discrimination"**, Biometrika, 59, 19-35, (1972).

36. Bartlett, M. S., **"Properties of Sufficiency and Statistical tests"**, Proceedings of the Royal Statistical Society, Series A, 160, 268–282, (1937).

37. Begg, C.B. and Gray, R., **"Calculation of Polychotomous Logistic Regression Estimates using Individualized Regressions"**, Biometrika, 71, 11-18, (1984).

38. Berkson J., **"Application of the Logistic function to Bioassay"**, JASA, vol. pp.357-365, (1944).

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

40. Broffit J. D., Randles R. H. and Hogg R. V., **"Distribution-Free pratial Discriminant Analysis"** .JASA.Vol.17, No: 356pp.934-939, (1976).

41. Brpffitt J. D., Randls R. H., Ramberg, J.S., and Hogg R.V., **"Generalized Linear and Quadratic Discriminate Functions Using Robust Estimates"**, JASA Vol.73 No 363, pp. 564-568, (1978).

42. Bull, S. B. and Donner, A., **"Derivation of large sample efficiency of multinomial logistic regression compared to multiple group**

discriminant analysis", Festschrift for V.M. Joshi, Volume 5, Biostatistics (I.B. MacNeill and G.J. Umphrey, eds.), Reidel, Boston, 177-197, (1987a).

43. Burns, Robert and Burns, Richard, **"Business Research Methods and Statistics using SPSS"**, Five extra advanced chapters, chapter 24 Logistic Regression: pp 568 -575, (2008).

44. Constanza, M.C. and Afifi, A.A., **"Comparison of Stopping Rules in Forward Stepwise Discriminant Analysis"**, JASA. Vol. 74, No.368, pp 777 – 785, (1979).

45. David G. and Kupper, L., **"Applied Regression Analysis and other Multivariate Methods"**, The University of North Carolind and Chapel Hill, (1978).

46. Hair, J. F, Black, W.C, Babin, B.J and Anderson, R.E, **"Multivariate Data Analysis"**, 7th Edition. Prentice Hall, (2009).

47. Hardle W., **"Multivariate statistics"**, Printed on acid – free paper, p.227, (2007).

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

49. He, X. and Fung, W.K., **"High breakdown estimation for multiple with applications to discriminant analysis"**, Journal of Multivariate analysis, 72, 151-162, (2000).

50. Huberty, C. J., "**Applied discriminant analysis"**, New York, John Wiley & Sons, Inc, (1994).

51. Izenman, A.J., **"Modern Multivariate Statistical Techniques, Regression, Classification, and Manifold Learning"**, New York, Springer Science, Business Media, LLC, (2008).

52. Johnson, R.A. and Wichern, D.W., **"Applied Multivariate Statistical Analysis"**, Sixth edition, New Jersey: Pearson Education, Inc, (2007).

53. Kiang, M.Y., **"A Comparative Assessment of Classification Methods"** .Decision Support Systems, 35, 441-454, (2003).

54. Krieng kitbumrungrat, **"Comparison logistic Regression and Discriminant Analysis in Classification groups for breast Cancer"**, Faculty of Information Technology, Rangsit University, Thailand, (2012).

55. Kshlr Sagar M.A., **"Multivariate Analysis"**, Marcel Dekker, Inc, New York, (1972).

56. 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).

57. McLachlan, G. J., **"Discriminant Analysis and Statistical Pattern Recognition"**, New York: Wiley, (1992**).**

58. Mirakabad, M., Yazdy, Fatimah & Xia, Feng **"A Novel Neighbor Selection Approach for KNN: Physiological Status Prediction Case Study"**, Sharif University of Technology. [Cited:12 05, 2014] , (2011).

59. Morrison, Donald F., **"Multivariate statistical Method",** second Edition, Tokyo, London. Mexico, New Delhi, (1976).

60. NG, T.H, **"Rank procedure in discriminant analysis for two population"**, وزارة الصناعة والمعادن/مركز المعلومات, (1980).

61. Nie, Norman H., et al., **"Statistical Package for the Social Science"**, 2nd ed. (N.Y., McGraw-Hill Book Company), (1975).

62. Nussbaumer, H. J., **"Fast Fourier Transform and Convolution Algorithms"** Springer – Verlag Berlin Heidelberg, (1982).

63. Rausch, J. R., **"A comparison of Linear, quadratic, and mixture models for Discriminant Analysis under non Normal Continuous Predictors"**, http://www.stat.notre dame.edu/www/research/reports, (2005).

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

65. Sabine L., Brian S.E, **"A Handbook of Statistical Analysis using SPSS"**, Chapman & Hall/CRC press LLC, (2004).

66. Schlotzhauer, D.C, **"Some issues in using proc logistic for Binary logistic regression",** the technical for SAS software users vol. 2. No. 4. (1993).

67. Theodoridis S. and Koutroumbas K, "**Pattern Recognition**", 2nd ed., Elsevier (USA), (2003).

68. Therrien, C.W., **"Decision Estimation and classification"**, John wiley & Sons New York, (1989).

69. Titterington, D. M., Murray, G. D. Murray, L. S. Spiegelhalter, D. J. Skene, A. M. Habbema, J. D. and Gelpke G. J. ,**"Comparison of discriminant techniques applied to a complex data set of head injured patients"**. J. R. Statist. Soc., A144, 145-75, (1981).

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

71. Wijesinha, A., Begg, C. B. Funkenstein, H. H., and McNeil, B. J., **"Methodology for the differential diagnosis of a complex"**, data-set. Med. Decision Making, 3, 133-154, (1983).

72. Young, T.Y., **"Classification, Estimation and Pattern Recognition"** American Elsevier Publishing Company Inc, New York, (1974).

73. Zhang, M.Q., **"Discriminant Analysis and its Application in DNA Sequence Motif Recognition"**, http://www.nslij genetics.gene /Zhang, (2000).

ثالثاً : مواقع الانترنت *Internet sites*

74. https://ar.wikipedia.org/wiki/%D8%B3%D8%B1%D8%B7%D8%A7%D9%86

75. https://ar.wikipedia.org/wiki/%D8%B3%D8%B1%D8%B7%D8%A7%D9%86_%D8%A7%D9%84%D8%AB%D8%AF%D9%8A

76. https://ar.wikipedia.org/wiki/%D8%B3%D8%B1%D8%B7%D8%A7%D9%86_%D8%A7%D9%84%D8%B9%D8%B8%D9%85

77. https://ar.wikipedia.org/wiki/%D8%B3%D8%B1%D8%B7%D8%A7%D9%86_%D8%A7%D9%84%D8%B1%D8%A6%D8%A9

78. Rajender Parsad and Cini Verghese. NEARESTNEIGHBOURHOOD DESIGNS http://www.iasri.res.in/iasriwebsite/DESIGNOFEXPAPPLICATION/Electronic-Book/module5/1

Abstract

The discriminate analysis is a statistical method based on a sample of observations taken from known societies to build a base that can help in identifying the society to which the new observations belong based on variables with discriminatory characteristics and applied to diagnose three types of tumors, namely breast, vertebral column (*bone*) *and the respiratory system* (*lung*) *tumors as their prevalent in our society is abundance nowadays.*

For the purpose of studying this subject, the values of observations were recorded for five variables (*sex, age, occupation, prognosis and duration of hospitalization*) *for* ***270*** *patients* (*random samples*) *and they were studied in four statistical models. The models were* (*linear discrimination function*) *and nonlinear* (*quadratic discrimination function and logistic discrimination function*) *and non-parametric model* (*nearest neighbor function*).

The statistical results were derived from the statistical program ***Stata*** *as well as the statistical package* ***SPSS*** *to classify the tumors to which group they belong. The classification error criterion was used as a criterion for comparison between the four models, to the best model with the least error of invention in possible.*

It was found that the nearest neighbor model was superior to the rest of the models in terms of lack of classification of the wrong as compared to other models.

Republic of Iraq
Ministry of Higher Education and Scientific Research
Karbala University
College of Administration and Economics
Department of Statistics

An Applied Study to Analyze Discriminant in Some Statistical Models

A thesis Submitted to

Council of the Faculty of Management and Economics at the University of Karbala As part of the requirements for a Master's Degree in Statistics Science

By

Maryam M. A. Al-khazali

Supervised By

Asis. Prof. Dr. Shrook A.S.AL-Sabbah

2018 Ad **1439 Ah**