دراسة إحصائية للعوامل المؤثرة في مرض سرطان الدم
(دراسة تطبيقية)

أ.د/ياسر محمد العدل عبد اللطيف **د/أحمد محمد عبد المنعم رياض**
أستاذ الإحصاء **مدرس الإحصاء**
كلية التجارة_ جامعة المنصورة **كلية التجارة_ جامعة المنصورة**
إعداد الباحث
علي محمد علي

المستخلص:-

تهدف هذه الدراسة إلى استخدام نموذج إحصائي مناسب يساعد في تحديد المتغيرات المؤثرة في مرض سرطان الدم ، من خلال استخدام الأساليب الإحصائية التالية (التحليل التمييزي- الانحدار اللوجستي – الشبكات العصبية)، حيث تُعد هذه الأساليب من أهم اساليب طرق التصنيف، وذلك لتصنيف مفردة ما إلى إحدى المجتمعات معتمداً في ذلك على متغيرات (صفات) ذات أهمية إحصائية في عملية التصنيف، وسوف نتعامل مع مسائل التصنيف بين مجتمعين أو أكثر بالاعتماد على مشاهدات متعدد المتغيرات الطبيعي.

ويمكن أن يلخص الباحث مشكلة الدراسة في التساؤلات التالية:-

هو كيف يمكن تحديد العوامل التي لها الدور الفاعل والمؤثر في الإصابة بمرض سرطان الدم قبل أن يستفحل المرض ولا يمكن علاجه عندئذ.؟

وكذلك كيف يمكن تطبيق الأساليب الإحصائية (التحليل التمييزي- الانحدار اللوجستي – الشبكات العصبية) على بيانات الأشخاص المراجعين لمستشفيات بغداد للحد من هذا المرض.؟

وعند استخدام الأساليب الإحصائية الثلاثة تبين أن الأفضلية هي لأنموذج الانحدار اللوجستي لانه قد إجتاز جميع معايير الجودة المطلوبة، حيث تمثلت المعايير في:-

$(R^2, \chi^2, Box-M, Hosmer\ \&\ Lamsho, MSE, RMSE)$

Abstract:

This study aims to estimate a suitable statistical model to help to determine affecting the blood cancer disease, The study used the following statistical methods the variable (Disicriminant Analysis- logistic regression - neural networks), where these methods are of the most important methods of classification, so as to classification Single what to one relying on the variables (recipes) with statistical significance in the classification process, we will deal with classification issues between two or more population depending on the views multinormal.

And it summarizes the researcher can study the problem in the

following questions: - is how to identify the factors that have active and influential role in the incidence of cancer of the blood before it is too late to treat the disease can not be then؟.
As well as how it can be applied statistical methods ((Disicriminant Analysis - logistic regression analysis - neural networks) data on persons reviewers to hospitals Baghdad to curb the disease؟.
When using the three statistical methods show that the preference is for the logistic regression model because he has passed all the required quality standards, was where standards(R^2,X^2,Box-M,Hosmer & Lamsho,MSE,RMSE).

١ـ المقدمة:

إن الاصابة بمرض سرطان الدم يعتبر واحد من أكبر وأهم المشكلات الصحية في العالم أجمع، ويتزايد معدل الإصابة بأمراض السرطان على مستوى العالم نظراً لزيادة وتطور مسببات هذا المرض، وخطورة هذا المرض ليس بسبب انتشاره فقط، وانما ترجع خطورته إلى ارتفاع معدل الوفيات نتيجة الإصابة بهذا المرض لقد تم استخدام الاساليب الإحصائية (التحليل التمييزي-الانحدار اللوجستي-الشبكات العصبية) لغرض الوصول إلى أهم المتغيرات المعنوية المؤثرة في الاصابة بمرض سرطان الدم.

٢ـ مشكلة البحث:

تتمثل مشكلة البحث في تحديد أهم العوامل (الشخصية، الاجتماعية، الاقتصادية، الطبية) المؤثرة في مرض سرطان الدم وذلك من خلال معرفة أي من الأساليب الإحصائية المستخدمة في البحث (تحليل تمايز الانحدار اللوجستي والشبكات العصبية)، يكون قادراً على تحديد العوامل المؤثرة في الاصابة بمرض سرطان الدم.

٣ـ أهمية البحث:ـ

تنبع أهمية البحث من خلال التشخيص الدقيق لمرض السرطان ومعرفة أن ثلث حالات السرطان يمكن الوقاية منها، وأن ثلث آخر يمكن الشفاء منها إذا اكتشف المرض وعولج في مرحلة مبكرة فمعرفة أسباب هذا المرض يكون لها الأثر البالغ في خفض معدل الإصابة، وتوفير الكثير من النفقات المالية والأعباء الاجتماعية على الفرد والمجتمع، وتلك هي الخطوة الأولى في التصدي لهذا المرض، ومن هنا يظهر دور استخدام الأساليب الإحصائية في حل مثل هذه المشكلات ولاسيما الأساليب المستخدمة في هذه الدراسة التطبيقية ، وهنا

تظهر أهمية الدراسة التي يتم تحديدها في النقاط التالية:
أ- تحديد أهم المتغيرات المؤثرة في تحديد المرض مما يساعد في خفض الإصابة أو الاكتشاف المبكر للمرض وسرعة علاجه مما يخفف من معاناة المرضى وأسرهم وكذلك خفض التكاليف التي تتحملها الدولة في العلاج.
ب- تحديد المتغيرات والعوامل المؤثرة بشكل معنوي وذلك من خلال دراسة المتغيرات الديموغرافية والبيئية وليس المتغيرات الطبية فقط.

٤- الاساليب الاحصائية المستخدمة:

٤-١ التحليـــل التمييـــزي Discriminant Analysis

التحليل التمييزي هو الأسلوب الذي يستخدم في معالجة البيانات عند تحليل العلاقة بين المتغيرالتابع النوعي Categorical Response Variable (وصفي) ومجموعة من المتغيرات المفسرة أو المستقلة سواء كانت المتغيرات التفسيرية كمية أو وصفية، ومقاسة بمقياس اسمي (Nominal) أو مقياس ترتيبي (Ordinal).

النتيجة أن التحليل التمييزي وظيفته هو تصنيف الأفراد إلى واحد من مجتمعين أو أكثر بالاعتماد على قياس مجموعة من المتغيرات ذات العلاقة. على فرض أن المجتمعات ذات فروق معروفة، وكل مشاهدة ترجع إلى أحد هذه المجتمعات وفقاً إلى هذه القياسات، كذلك يمكن استخدام هذا الأسلوب لبيان المتغيرات التي تساهم في التصنيف، وهي كما في تحليل الانحدار لديها استخدامين الوصف والتنبؤ.

الفروض التي يتطلبها التحليل التمييزي:-

١) يجب أن توزع X's توزيعاً طبيعياً متعدداً (والتي تؤدي إلى حالات أخرى هي أن كل متغير X على حدا يملك توزيع طبيعي وأي دالة خطية من الـ(X's) تملك توزيع طبيعي).
٢) يجب أن تكون مصفوفة التباين والتباين المشترك في كلا المجموعتين هي متجانسة، وإذا تم الحصول على هذه الفروض فإنه يمكن إثبات أن:

$$F = \frac{n_1 n_2 (n_1 + n_2 - p - 1)}{(n_1 + n_2)(n_1 + n_2 - 2)\, p} D^2 \approx F(p, n_1 + n_2 - 1) \qquad (١)$$

والتي توفر لنا أداة جيدة لاختبار وجود فروق معنوية بين أوساط المجتمعين للمتغيرات كلها معا. وهذه الأداة ترتبط بأداة اختبار معنوية ملائمة نموذج الانحدار، خصوصا في جدول ANOVA التالي:

جدول رقم (١) جدول التباين لمتغير تابع وهمي

Source	d.f	S.S
Due to Regression	p	$SS_R = \frac{n_1 n_2}{n_1+n_2} \sum_{i=1}^{p} \hat{\beta}$jdj
Deviation from Regression	n1+n2-p-1	$SS_E = \frac{n_1 n_2}{n_1+n_2} (1 \sum_{i=1}^{p} \hat{\beta}$jdj
Total	n1+n2-1	$\frac{n_1 n_2}{n_1 + n_2}$

ونستطيع أن نبرهن عادة أن مجموع المربعات:

$$\frac{SS_P}{P} / \frac{SS_E}{n_1 + n_2 - p - 1} \quad (٢)$$

هي نفسها

$$(\frac{n_1 n_2 (n_1 + n_2 - p - 1)}{(n_1 + n_2)(n_1 + n_2 - 2)\, p} D^2)$$

في القيمة،(إن فروض التحليل التمييزي، هي معاكسة لفروض الانحدار بأن المتغير التابع يتبع التوزيع الطبيعي و (X's) هي غير عشوائية، أن اختبار (F) لجدول تحليل التباين السابقة ستستخدم لاختبار معلمات الانحدار لفرضية العدم.

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

$$D^2 = \frac{(n_1 + n_2)(n_1 + n_2 - 2)}{n_1 n_2} (\frac{R^2}{1 - R^2}) \quad (3)$$

أي أن: $R^2 = C^* D^2$

وإذا تم رفض H_0 فهذا دليل على أن هناك فروقات معنوية بين مجموعتين على الأقل وهذا سيساعدنا في التمييز بين مجموعتين بصورة صحيحة (Johnson,1998).

٤-٢ الانحدار اللوجستي Logistic Regression

يعتبر نموذج الانحدار اللوجستي من النماذج الرياضية المستخدمة في وصف العلاقة بين بعض المتغيرات التفسيرية،

والمتغير التابع الثنائي Binary Independent variable، سواء كانت المتغيرات التفسيرية كمية أو وصفية، ويتم استخدام الانحدار اللوجستي وبصفة خاصة في المجالات الاقتصادية والاجتماعية والطبية، أن الانحدار اللوجستي الثنائي(موضوع الدراسة) مبني على فرض أساسي وهو أن المتغير التابع هو متغير نوعي ثنائي(ذو فئتين) فيسمى بنموذج الانحدار اللوجستي الثنائي Binary Logistic Regression Model وعندما يكون المتغير التابع نوعي متعدد (له أكثر من فئتين) فيسمى النموذج بنموذج الانحدار اللوجستي المتعدد Multi Logistic Regression Model وهو امتداد لنموذج الانحدار اللوجستي الثنائي Hosmer (and Lemeshow,2000)، وعلى ذلك يمكن تعريف الانحدار اللوجستي الثنائي:- هو نموذج إحصائي يستخدم لبيان العلاقة بين المتغير التابع (الوصفي) ذو الناتجين والمتغيرات التفسيرية التي من الممكن أن يكون بعضها كمي Quantitative والبعض الآخر نوعي Qualitative (طارق عبدالحميد،٢٠٠١).

تحديد معادلة نموذج الانحدار اللوجستي الثنائي:

نموذج الانحدار اللوجستي الثنائي يعمل على افتراض أن المتغير التابع variable dependent يكون متغير ثنائي binary response variable له ناتجين، الأول يسمى النجاح Success ويرمز له بالقيمة واحد إذا تحقق الحدث الأول، والآخر يسمى الفشل Failure ويرمز له بالقيمة صفر إذا تحقق الحدث الثاني، وأن المتغيرات التفسيرية Independent variable يمكن أن تكون كمية Quantitative، كما يمكن أن تكون نوعية Qualitative مع وجود علاقة بين المتغير التابع والمتغيرات التفسيرية، ويمكن التعبير عن هذا النموذج في حالة ما إذا كان المتغير التابع متغيراً نوعياً ثنائياً في الصورة $Y_i = (0,1)$ ويأخذ شكل العلاقة التالية: (Wayne,1995,p.484)

$$\pi(X) = P(X) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

حيث أن $P(X)$ هو احتمال حدوث حدث معين (X) في مجموعة من المتغيرات

التفسيرية ويمكن الرمز لها بالرمز $\pi(X)$، وحيث أن الطرف الأيسر تنحصر قيمته بين $[0 \le P(X) \le 1]$، فباستخدام تحويلة اللوجيت تكون المعادلة كالتالي. (Fred,2000,p.71).

Logit(P) =

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_j X_J$$

$$\left(\frac{P}{1-P}\right) = e^{\beta_0 + \beta_j X_j}$$

$$P = e^{\beta_0 + \beta_j \sum X_i} -$$

$$P e^{\beta_0 + \beta_j X_j}$$

$$P + P e^{\beta_0 + \beta_j \sum X_i} =$$

$$e^{\beta_0 + \beta_j X_j}$$

$$P\,(1 + e^{\beta_0 + \beta_j \sum X_i}) =$$

$$e^{\beta_0 + \beta_j X_j}$$

والتي نصل منها إلى صيغة الانحدار اللوجستي التالية:-

$$P = \frac{e^{\beta_0 + \beta_j X_j}}{1 + e^{\beta_0 + \beta_j X_j}} \quad (4)$$

فروض نموذج الانحدار اللوجستي:-

أ- أن المتغير التابع متغير وصفي ذو وجهين أو أكثر والمتوسط الشرطي لهذا المتغير $E(Y \mid X)$ عبارة عن متغير محدود بالقيم (0،1). أما المتغيرات التفسيرية فيمكن أن تكون كمية (مستمرة أو متقطعة) أو وصفية (ثنائية أو متعددة) كما يفترض أن المتغيرات التفسيرية تقاس بدون أخطاء.

ب- العلاقة بين المتغير التابع والمتغيرات التفسيرية تكون علاقة دالية غير خطية.

ج- حد الخطأ يتبع توزيع ذو الحدين Binomial Distribution بتوقع صفر وبتباين $P(X)[1-P(X)]$.

د- لا يوجد ارتباط ذاتي Auto correlation بين البواقي (حدود الأخطاء العشوائية)، أي أن $E(e_i, e_j) = 0$.

الانحدار اللوجستي المتدرج Stepwise Logistic Regression:-

من أكثر الأساليب المستخدمة في الانحدار الخطي هو أسلوب الاختيار المتدرج للمتغيرات، حيث أن كل البرامج الجاهزة تقدم اختيارا لإجراء الانحدار اللوجستي المتدرج، وكذلك يُعد من الأساليب المهمة في بناء النموذج، بسبب أنه وسيلة مفيدة وفعالة في تحليل البيانات، وبصفة خاصة إذا ما واجهنا متغير تابع يكون جديد نسبيا، بالإضافة إلى المتغيرات التفسيرية التي قد لا تكون معروفة وارتباطها مع المتغير التابع غير مفهوم، في مثل هذه الحالات تقوم معظم الدراسات بتجميع العديد من المتغيرات الممكنة وتقوم بعد ذلك بتحديد معنوية المتغيرات ذات الارتباط بمتغير الاستجابة. (Hosmer and Lemeshow,2000.p.116)

إن الذي يستطيع أن يوفر لنا وسيلة سريعة وفعالة في تحديد عدد كبير من المتغيرات وفي توفيق عدد من معادلات الانحدار اللوجستي في آن واحد هو أسلوب الانحدار المتدرج، والذي يفيدنا في أنشاء نماذج بشكل متدرج ويسمح بفحص مجموعة من النماذج لا تفحص بأي أسلوب أخر، إن أي أسلوب متدرج لا بد أن يعتمد على مجموعة من الخطوات الإحصائية StatisticalAlgorithm والتي تختبر أهمية المتغيرات وتحدد الإبقاء عليها أو استبعادها على أساس قاعدة قرار ثابتة، تقاس أهمية المتغير من خلال المعنوية الإحصائية لمعامل المتغير، والإحصاء المستخدم يعتمد على فروض النموذج، أن الأخطاء في الانحدار الخطي المتدرج يُفترض أنها تتبع التوزيع الطبيعي، ولهذا يتم استخدام اختبار(F)، بينما الأخطاء في الانحدار اللوجستي تتبع توزيع ذو الحدين، وبالتالي يتم تحديد المعنوية عن طريق اختبار χ^2 لنسبة الإمكان (Likelihood ratio chi-squared test) بحيث أنه عند أي خطوة في الأسلوب نجد أن المتغير الأكثر أهمية هو الذي ينتج عنه أكبر تغيير في دالة الإمكان مقارنة بالنموذج الذي لا يحتوي ذلك المتغير (بمعنى هو ذلك المتغير الذي ينتج عنه أكبر قيمة لإحصاء نسبة الإمكان G).

٤-٣ الشبكات العصبية الاصطناعية Artificial Neural Network

هي نظام لتشغيل المعلومات له خصائص أداء معينة تشبه الشبكات العصبية الحيوية بناء على الافتراضات التالية (**Fausett**،1994).

أ- تتم معالجة المعلومات في عناصر معالجة بسيطة تسمى نيورونات (Neurons).

ب-الإشارات Signals تمر بين النيورونات على روابط اتصال Connection Links

ج- كل رابطة اتصال لها وزن Weight خاص بها يتم ضربه في الإشارة المنقولة.

د-كل نيورون يطبق دالة تحويل Transfer Function (عادة غير خطية) على مدخلاته input (مجموع الإشارات الداخلة المرجحة بالأوزان) لتحديد إشاراته الخارجة.

الشبكات العصبية الاصطناعية مهمتها معالجة المعلومات وهي تشبه في عملها الشبكات العصبية الحيوية (عيسى،٢٠١٢)، ولقد تبين للباحثون إن لمعالجة البيانات في الشبكات العصبية الاصطناعية بدأت تشبه إلى حد كبير طرق المعالجة في الشبكات الحيوية وبذلك تم استخدامها في مجالات عديدة منها مجالات الطب والتحكم والإعمال وغيرها وكذلك تستطيع الشبكات العصبية الاصطناعية إن تعالج العديد من البيانات كالبيانات الخطية Linear وغير الخطية Non Linear بالإضافة إلى البيانات الناقصة Incomplete والمشوشة Noisy.

وصف الشبكة العصبية الاصطناعية:-

أ- نمط أو طريقة الاتصالات بين النيورونات (تسمى معمارية أو بنية أو هيكل الشبكة Architecture).

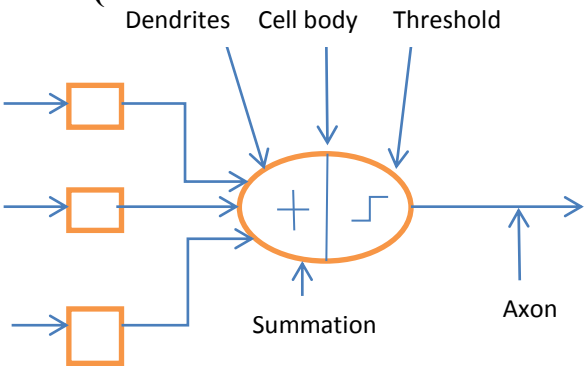
ب- طريقة تحديد الأوزان على الوصلات (تسمى خوارزم التدريب أو التعلم Training or Learning).

ج- دالة التحويل Transfer Function أوتسمى دالة الاستثارة Activation.

الشبكات العصبية تتألف من عدد كبير من عناصر تشغيل بسيطة اسمها نيورونات وهذه نيورون متصلة بعضها ببعض من خلال روابط اتصال موجه ومرافقة لها وزن خاص بها، هذه الأوزان هي المعلومات المستخدمة في الشبكة لحل المشاكل، لكل نيورون حالة داخلية خاصة به هي مستوى الاستثارة وهو دالة في المدخلات التي يستقبلها

النيورون، ويرسل مستوى الاستثارة من قبل النيورون نفسه كإشارة لنيورونات أخرى، النيورون بإمكانه إرسال إشارة واحدة فقط وهذه الإشارة توزع على الكثير من النيورونات الأخرى.

ويمكن توضيح كيفية عمل النيورون من خلال الشكل التالي (Shankar 2008):



شكل رقم (٣) نيورون صناعي بسيط

وبالتالي يتألف النيورون من:

أ- المدخلات Inputs ($Z_1,Z_2,\ldots,Z_n$).

ب- الأوزان المرجحة Weights ($w_1,w_2,\ldots,w_n$)

ج- عنصر المعالجة processing Elements

ويقسم هذا العنصر إلى جزئين هما:

- **دالة التجميع Summation) (Function**

مهمتها تجميع كل الإشارات الداخلة في إشارة واحدة داخل العقد أو النيورونات وبالنتيجة مجموع الإشارات الداخلة المجمعة y هو مجموع الإشارات المرجحة بالأوزان كالتالي.

$$y - in = W_1 Z_1 + W_2 Z_2 + \ldots\ldots + W_n Z_n$$

حيث:y-in: مجموع إشارات المدخلات المرجحة بالأوزان.

- **دالة التنشيط Activation) (Function**

وهي دالة غير خطية ذات عتبة معينة Threshold مهمة هذه الدالة تنشيط (نشر) القيمة الداخلة إليها واستنادا إلى نوع الدالة المستخدمة ووفق قيمة حد القيمة المحددة (أي الحد الذي يعمل النيورون على أساسه معطيا قيمة (1) إذا كان المجموع الموزون لقيم الداخلي أكبر من قيمة معينة تدعى العتبة و(0) إذا كان المجموع الموزون أقل من العتبة) ويطلق عليها أيضاً دالة التحويل Transfer Function .

$$y = f(y - in) \quad (5)$$

٥- الدراسات السابقة:-

١- دراسة (1989) Wu,Kg
تهدف هذة الدراسة الى تحديد عوامل الخطر المسببة للإصابة بسرطان الرئة Lung Cancer بين العاملين في منجم Dachang tin بالصين، حيث شملت العينة ٦٩ فرداً من العاملين بالمنجم، تم تجميع البيانات عن المتغيرات التفسيرية وفي نهاية الدراسة وجد أن ٥٥ حالة مصابة، ١٤ حالة سليمة، وثبت معنوية أثر التدخين، والوقت الذي يقضيه العامل داخل المنجم (الذي يسمى وقت الشم) كما وجد أن العمر عند بداية العمل بالمنجم من العوامل المؤثرة على الإصابة بالمرض، وقد استخدم الانحدار اللوجستي الثنائي في هذه الدراسة،ويعاب على هذه الدراسة أنها استعانت بعينة صغيرة، وكذلك عدد صغير من المتغيرات التفسيرية.

٢- دراسة بدر الياسين (٢٠٠٩) أن البحث يتضمن تشخيص بعض أنواع سرطان الدم (اللوكيميا) باستخدام طرائق تمييز حصينة تعتبر مهمة لقلة تناولها. وتم التركيز في هذا البحث على إيجاد مقدرات حصينة لمعلمتي $\underline{M}$ (موجه المتوسط) والقياس Σ (مصفوفة التباين والتباين المشترك) من خلال استخدام طرائق تقدير حصينة منها طريقة مقدر اصغر محدد تباين مشترك (MCD)، وطريقة مقدر اصغر محدد تباين مشترك معاد الأوزان (RMCD)، وطريقة مقدر اصغر حجم مجسم قطع بيضوي (MVE) وبالتالي استخدام دوال تمييز حصينة لما لها من دور كبير في التحصين ضد وجود الشواذ في البيانات ومقارنتها مع دوال التمييز التقليدية للوصول إلى أفضل دالة تمييز لتشخيص بعض أنواع مرض سرطان الدم (اللوكيميا) اعتمد الباحث على بعض أنواع المرض والصفات والمتغيرات المصاحبة للمرض في بناء نموذج احتمالي للتمييز بين نوعين من سرطان الدم (اللوكيميا) هما سرطان الدم النخاعي الحاد وسرطان الدم اللمفاوي الحاد.

٣- دراسة (٢٠١٠) Turgay Ayar et al. تم وضع نماذج حاسوبية في التشخيص الطبي لمساعدة الأطباء على التفريق بين الأشخاص الأصحاء والأشخاص الذين يعانون من المرض ويمكن لهذه النماذج إن تساعد في اتخاذ القرار الناجح من خلال السماح حساب احتمال المرض على أساس خصائص المريض المعروفة ونتائج اختبارات السريرية، اثنين من

نمـــاذج الكمبيـــوتر الأكثـــر استخداماً فـي تقـدير المخـاطر الســـريرية هـــي الانحـــدار اللوجسـتي والشـبكات العصـبية الاصطناعية.

وقـد أجريـت هـذه الدراسـة لمقارنـــة ومراجعـــة هـــذين النمـــوذجين وتوضـــيح مزايـــا وعيـوب كـل منهمـا وتـوفير معايير لاختيار النموذج استخدم النموذجين لتقدير خطر الإصابة بسـرطان الثـدي علـى أسـاس أصـفات ثـدي شـعاعي وعوامـل الخطر الديموغرافي على الرغم مـن أنهـا أظهـرت أداء ممـاثلاً ونمـاذج لهـا نقـاط القـوة فريـدة الخصـائص فضـلاً عـن القيـود التي يجب أخذها بعين الاعتبار وربما تكون مكملة في المساهمة فـي تحسين عمليـة صـنع القرار السريري.

٤- دراسة عبد اللطيف و الزوبعي (٢٠١١) هدف هذه الدراسة إلى تطبيق طريقة إحصائية علمية متمثلة في التحليل التمييزي المتعدد لتصنيف مراحل سرطان الثدي معتمدين على جملة من المتغيرات ويتم ذلك ببناء دوال تمييزية تساعد الطبيب المختص في التشخيص الأولي وتم استخدام طرق مختلفة للتمييز بين المجموعات والمقارنة بينها لتحديد الطريقة الأفضل ثم استخدام هذه الدوال التمييزية في تصنيف الحالات الجديدة إلى مجموعتها الصحيحة. وتم تقسيم بيانات البحث التي جمعت في مستشفى الذرة بالخرطوم إلى أربعة أقسام بالقسم الأول خمس مجموعات، والقسم الثاني يحتوي على ثمانية مجموعات، أما القسم الثالث يحتوي على بيانات المرحلة الثانية لسرطان الثدي والقسم الأخير يحتوي على بيانات المرحلة الثالثة لسرطان الثدي. وتم الحصول على وصف لمتغيرات البحث لكل مجموعة على حدة وعمل مقارنة بين المجموعات في الأقسام الأربعة، وبعد ذلك تم بناء الدوال التمييزية بطريقة المراحل والتمييز القويم.

٦- الجانب التطبيقي:

٦-١التحليـــل التمييـــزي Discriminant Analysis

التحليل التمييزي هو اسلوب احصائى مبنـى علـى تحديد دالـة بصـــورة معينـــة بغـــرض اسـتخدامها فـى تحديـد انتمـاء مشـاهدة معينـة لمجتمـع مـا مـن بـين عـدة مجتمعـات بديلـة لا تتقـاطع فيمـا بينهـا (متنافيـة بالتبـادل) مـن خـلال بعـض الخـواص (المتغيـرات) التـى يمكـن عـن طريقهـا تحديـد

المجتمع التى تنتمى اليها هذه المشاهدة بصورة جيدةويهدف تحليل التمايز لدراسة الفروق بينن مجموعتين أو اكثر بالنسبة لعدد المتغيرات أنيا،وهو اسلوب احصائى متقدم يستخدم لدراسة متغير مقسم الى مجموعات مع مجموعة من المتغيرات المستقلة، وهو من الاساليب متعددة المتغيرات هدفه التوصل الى تصنيف عدة مفردات الى مجتمع معين من عدة مجتمعات وبناء قاعدة يمكن من خلالها المساعدة فى تحديد المجتمع التى تنتمى اليها مفردات جديدة مستقبلا ويفضل استخدام تحليل التمايز عندما يكون المتغير التابع وصفى ذو وجهين أو أكثر.

معايير اختيار النموذج:-

توجد عدة معايير للحكم على مدى جودة النموذج منها درجة الدقة – والصلاحية الكلية والقوة التمييزية والمعنوية الاحصائية، ويمكن الحكم على الدالة من خلال الارتباط التوافقى(canonical correlation) فهو يقيس درجة التباين المفسر بدلالة التمايز الى التباين الكلى او من خلال ويلكس لامدا(wilks lamda) وكلما كانت قيمتها صغيره كان ذلك افضل ويمكن اختبارها عن طريق مربع كاى.

وتكون قيمة لامدا واحد صحيح عندما تكون المتوسطات متساوية وتقترب من الصفر عندما يكون هناك اختلاف بين المتوسطات، واذا كان كل زوج من متوسطات المجتمعين متساويين قل الأمل فى العثور على دالة تمايز بين المجتمعين ذات كفاءة عالية ومن ثم يجب العناية باختيار المتغيرات التى تستخدم فى دوال التمايز بحيث يكون هناك اختلاف وفروق بين المتوسطات.

ويوضح الجدول التالى قيم ويلكس لامدا واختبار ف ومعنويته لاختبار تساوى المجموعات فى المتوسط وقدرة المتغيرات المستقلة فى التمييز بين النوعين.

جدول رقم (١) معايير اختيار النموذج

الخطوات	النماذج	إحصائية الاختبار ف	المعنوية	اختبار تساوى المتوسطات	
				اختبار ف	المعنوية
1	RDW	130	0.000	29.6	0.000
2	NEW	84.7	0.000	9.49	0.002
3	RBC	69.4	0.000	19	0.000
4	LYM	57	0.000	33.6	0.000
5	MPV	48.4	0.000	11.6	0.001
6	MCH	42.7	0.000	130	0.000
7	EOS	37.8	0.000	10.3	0.002
اختبار بوكس ام = ٣٢٤.٧٩٩			0.000		

ويتضح من جدول رقم (١) وجود (7) نماذج مختلفة وكانت جميعها معنوية وذات دلالة احصائية وقد تم اختيار النموذج السابع لتحقق معايير الجودة كما تبين أن هناك فروق معنوية ذات دلالة احصائية بين متوسط المجموعتين لهذه المتغيرات. كما يؤكد اختبار بوكس ام على عدم تساوى مصفوفة التغاير– التباين للوجهين (المجموعتين). ويوضح جدول رقم (٢) ملخص نتائج الدالة التمييزية لاختبار الفرض الاول والذى يوضح قيمة ايجن (1.051) كما أن معامل الارتباط التوافقى (0.716) وان مربع هذه القيمة يساوى (0.512) وهى نسبة مايفسر من التباين بين النوعين بالنموذج المقترح، بمعنى أن النموذج المقترح يفسر مايقرب من (51.2%) من التغيرات التى تطرأ على الحالة الصحية للشخص كما تبين أن معامل ويلكس لامدا يساوى (0.488)، وهى نسبة التباين غير المفسره بالنموذج ومعنوية النموذج (0.000) وبالتالى يمكن الاعتماد على هذا النموذج ليميز بين المجموعتين (مريض / غير مريض) من خلال المتغيرات السبعة.

جدول رقم (٢) قيمة ايجن

قيمة ايجن	نسبة التباين	الارتباط التوافقى	ويلكس لامدا	مربع كاى	درجة الحرية	المعنوية
1.051[a]	100.0	.716	.488	182.783	7	.000

جدول رقم (٣) معاملات النموذج

المتغيرات	المعاملات المعيارية	المعاملات غير معيارية	الاحمال
NEW	.320	.013	.331
LYM	.279	.012	.187
EOS	.199	.075	.265
RBC	.496	.547	.352
MCH	.241	.042	.207
RDW	.748	.087	.691
MPV	.244	.058	.195
الثابت		-6.057	

كما يوضح جدول رقم (٣) أن احمال التمييز للمتغيرات السبعة وهى نسب التفسير لكل متغير من متغيرات النموذج حسب نسبة التفسير (NEW، LYM، EOS،RBC، MCH، RDW، MPV) تساوى (33.1%.، 18.7%، 26.5%، 35.2%، 20.7%، 69.1%، 19.5) كما يوضح الجدول القيم المعيارية وغير المعيارية للنموذج ويشير جدول رقم (٤) الى كفاءة التقسيم الصحيح (صلاحية النتائج والحكم على قدرة النموذج فى التقسيم الصحيح للنوعين (ذكر/انثى)).

جدول رقم (٤) جدول كفاءة التقسيم

مرض سرطان الدم			المتوقع		الاجمالى
			غير مريض	مريض	
المشاهد	العدد	غير مريض	20	0	20
		مريض	3	237	240
	%	غير مريض	100.0	0.0	100.0
		مريض	1.3	98.8	100.0
a. 98.8% of original grouped cases correctly classified.					

ومن الجدول السابق أن عدد (20) من الغير مرضى تم تقسيمهم تقسيما صحيحا بنسبة بلغت (100%).
عدد (237) من المرضى تم تقسيمهم تقسيما صحيحا بنسبة بلغت (98.8%).
عدد (257) من عينة البحث والدراسة تم تقسيمهم تقسيما صحيحا بنسبة بلغت (98.8%) وهى نسبة تقسيم مرتفعة جدا.

٦-٢ الانحدار اللوجيستى Logistic (Regression)

لمعرفة هل يوجد تأثير ذات دلالة احصائية للمتغيرات المستقلة والمتمثلة فى تسعة عشرة
متغير على مستوى صحة الشخص (مريض / غير مريض)، ولإختبارهذا الفرض قد استخدم الباحث اسلوب الانحدار اللوجيستى والذى يدرس اثر المتغيرات المستقلة على المتغير التابع الوصفى ذات الوجهين (مريض / غير مريض) حيث يعتبر هذا النموذج من النماذج المهمة المستخدمة لصياغة دالة التمييز والتقسيم، وفي الانحدار اللوجيستى ليس المهم تقدير المعالم بقدر استخدامها فى حساب احتمال المستوى الصحى للشخص (مريض/غير مريض)، وباستخدام طريقة الانحدار التدريجى اظهرت النتائج وجود اكثر من نموذج انحدار ذات دلالة احصائية عند مستوى خطأ ٥% وقد تم اختبار النموذج (4) فقد بلغت معنوية النموذج (احتمال الخطأ) (0.000) كما قد بلغت كفاءة التقسيم الصحيح للنموذج (99.2%) وهى نسبة تقسيم مرتفعة جدا وقد دخل للنموذج عدد (4) متغيرات مستقلة تتمثل فى (NEW، EOS، RBC، RDW).

جدول رقم (٥) جدول مقاييس جودة نموذج

مربع كاى	درجات الحرية	المعنوية	معامل التحديد	نسبة التقسيم الصحيحة
130.028	4	.000	.940	99.2

Overall Chi-square test for all i (In simple regression, i = 1)
H0:- $B_i = 0$
H1:- $B_i \neq 0$
ويعتبر جدول رقم (٥) احدى مخرجات تحليل الانحدار اللوجيستى وتوضح هذه القيم او

المؤشرات مدى أهمية المتغيرات المستقلة ومدى تفسيرها للتغيرات التى تحدث للمتغير التابع، حيث بلغت قيمة مربع كاى (130.023) بدرجة حرية (4) بمعنوية بلغت (0.000) وهي أقل من مستوى الخطأ المسموح به،ومن ثم نرفض فرض العدم ونقبل الفرض البديل القائل بأنه على الاقل يوجد متغير ذات معامل لا يساوى الصفركما بلغت نسبة التقسيم الكلية الصحيحة (99.2%) كما أن أشباه معامل التحديد تساوي (0.940) وهي نسبة تفسير مرتفعة، ومن ثم فان النموذج بشكل عام جيد حيث ان معنوية النموذج اقل من مستوى الخطأ المسموح به.

جدول رقم (٦) معالم نموذج الانحدار (مريض / غير مريض) ومدى معنويتها

		التقدير	الخطأ المعيارى	اختبار والد	درجات الحرية	المعنوية
النموذج الرابع	NEW	-.230	.099	5.376	1	.020
	EOS	-.920	.480	3.669	1	.055
	RBC	-11.615	5.599	4.304	1	.038
	RDW	-.589	.277	4.510	1	.034
	الثابت	88.857	41.889	4.500	1	.034

وقد اتضح من جدول رقم (٦) معنوية معاملات المتغيرات محل الدراسة بالنموذج عند مستوى ثقة (0.95,0.90) حيث بلغت معنوية كلا من (NEW ، EOS، RBC ، RDW) تساوى (.020،055.، 038.، 034.) ومن ثم نقبل الفرض القائل بوجود تأثير معنوى ذات دلالة احصائية للمتغيرات والتى تم اختيارها فى النموذج.

اختبار جودة توفيق النموذج

جدول رقم (٧) جدول التوافق لاختبار (Hosmer and Lemeshow)

Contingency Table for Hosmer and Lemeshow Test						
		مرض سرطان الدم = غير مريض		مرض سرطان الدم = مريض		الكلى
		المشاهد	المتوقع	المشاهد	المتوقع	
النموذج الرابع	1	20	19.983	7	7.017	27
	2	0	.017	26	25.983	26
	3	0	.000	26	26.000	26
	4	0	.000	26	26.000	26
	5	0	.000	26	26.000	26
	6	0	.000	23	23.000	23
	7	0	.000	21	21.000	21
	8	0	0.000	85	85.000	85

ويتضح من جدول رقم (٧) وهو يمثل اختبارا لا معلميا لجودة توفيق النموذج إذ يعتمد على حساب إحصاءة مربع كاى للفرق بين القيم المشاهدة والقيم المتوقعة وقد اقترح (**Hosmer andLemeshow**) باستخدام توزيع مربع كاى للكشف عن انحرافات نموذج الانحدار اللوجيستى حيث يوجد جزء مشاهد لا يستند الى نموذج نظرى والآخر متوقع محسوب من تقديرات النموذج اللوجيستى ومن ثم يحسب مربع كاى كمقياس لجودة التوفيق.

جدول رقم (٨) معنوية اختبار (Hosmer and Lemeshow)

النموذج	مربع كاى	درجة الحرية	المعنوية
4	.017	6	1.000

ونلاحظ من جدول رقم (٨) أن قيمة مربع كاى بلغت (0.017) وكانت المعنوية تساوي (1) مما يشير ذلك الى قبول فرضية العدم القائلة بأنه لا يوجد فروق ذات دلالة احصائية بين القيم الفعلية والقيم المقدرة مما يؤكد على جودة توفيق النموذج بالكامل.

جدول رقم (٩) نسبة التقسيم (التصنيف)

المشاهد			المتوقع		
النموذج الرابع	مرض سرطان الدم	غير مريض	مرض سرطان الدم	غير مريض	التقسيم الصحيح
	مريض		239	1	99.6
التقسيم الكلى					99.2
	غير مريض		1	19	95.0

ويتضح من جدول (٩) أن هناك (19) مفرده تم تقسيمها تقسيما صحيحا بنسبة بلغت (95.0) من بين (20) مفردة ممن كانت استجابتهم لمرض سرطان الدم (غير مريض)، أيضا هناك (239) مفردة ،تم تقسيمها تقسيما صحيحا بنسبة بلغت (99.6) من بين (240) مفردة كانت استجابتهم لمرض سرطان الدم (مريض) كما يتضح أن عدد (258) مفردة تم تقسيمها بشكل سليم بنسبة بلغت (99.2%) من أصل حجم العينة والتى تبلغ (260) مفردة.

نموذج معادلة الانحدار اللوجيستى المقدرة:-

– log odds = 88.857 - .230(NEW) - 0.920(EOS) - 11.615(RBC) - 0.589(RDW)

وبالتعويض عن المتغيرات بالقيمة (1)

– log odds = 88.857 - .230(1) - 0.920(1) - 11.615(1) - 0.589(1)

– log odds = 88.857 – 13.354 = 75.503

odds = exp(77.503) = 4.56168E+33

prob = 4.56E+33/ (1+4.56E+33) =1 .0

جدول رقم (١٠) اختبار حساسية النموذج

المساحة	الخطأ المعيارى	المعنوية	فترة الثقة (95%)	
			الحد الأدنى	الحد الأعلى
.999	.001	.000	.997	1.000

ولقياس حساسية النموذج فى التصنيف تم حساب المساحة تحت المنحى (Roc-curve) قد بلغت تقريبا (0.999) وهى مرتفعة الى جانب ذلك فهى معنوية عند مستوى ثقة (٠.99)، كما يتضح من جدول رقم (١٠).

حيث بلغت معنوية الاختبار (0.000) وهذا يعنى أن الانحدار اللوجيستى يصنف بطريقة اكثر معنوية وأن التصنيف لا يرجع للصدفة.

٦-٣ الشبكات العصبية الاصطناعية (Artificial NeuralNetwork)

وقد تبين من نتائج التحليل للعينة موضوع البحث والدراسة أن نسبة التصنيف للتدريب والاختبار قد بلغت (100%) وهى نسبة تقسيم تامة وهذا ما يوضحه جدول رقم (١١) التالي:

جدول رقم (١١) نسب التقسيم

العينة		المتوقع		
		غير مريض	مريض	النسبة الصحيحة
التدريب	غير مريض	13	0	100.0%
	مريض	0	163	100.0%
	النسبة	7.4%	92.6%	100.0%
الاختبار	غير مريض	7	0	100.0%
	مريض	0	77	100.0%
	النسبة	8.3%	91.7%	100.0%

كما بلغت المساحة تحت المنحنى والتى تعبر عن حساسية النموذج الواحد الصحيح للمتغيرات المعنوية والتى تم استخلاصها من نتائج التحليل اللوجيستى كما يتضح من جدول رقم (١٢).

جدول رقم (١٢) المساحة تحت المنحنى

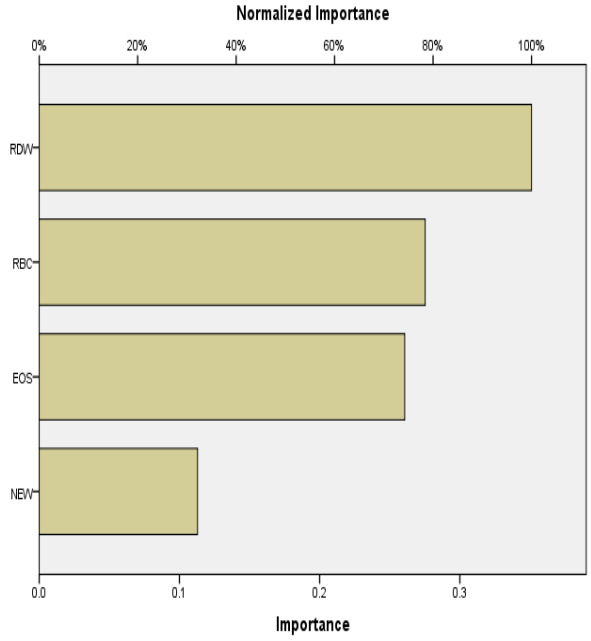
		Area
غير مريض	مرض سرطان الدم	1.000
مريض		1.000

الأهمية المعيارية للمتغيرات

جدول رقم (١٣) الاهمية المعيارية لمتغيرات النموذج

	الأهمية	الأهمية المعيارية
NEW	.113	32.2%
EOS	.261	74.2%
RBC	.275	78.4%
RDW	.351	100.0%

ومن جدول رقم (١٣) يتضح أن المتغير(RDW) هو أهم المتغيرات والأكثر تأثير على المتغير التابع بنسبة بلغت (٠.351) يلى ذلك فى المرتبة الثانية من حيث التأثير على المتغير التابع)(RBC بنسبة بلغت (٠.275) كما جاء فى الترتيب الثالث (EOS) بنسبة بلغت (٠.261) وفى الترتيب الأخير جاء (NEW) بأهمية بلغت (٠.113). كما يتضح ذلك من شكل رقم (١) التالى:-



شكل رقم (١) المتغيرات المؤثرة من حيث الاهمية

جدول رقم (١٤) مقارنة بين الانحدار اللوجيستى - التحليل التمييزي والشبكات العصبية

معايير الجودة	الانحدار اللوجيستى		التحليل التمييزي		الشبكات العصبية المدربة	
معامل التحديد	94%	√	51%	x	0.899	x
مربع كاى	130.028**	√	182.783**	√	---	---
نسبة التقسيم	99.20%	√	98.80%	x	91.7	x
اختبار Box-M	---	x	324.799**	x	---	---
اختبار Hosmer	0.017	√	---		---	---
MSE	---	---	---	---	0.001	
RMSE	---	---	---	---	0.035	

ومن جدول رقم (١٤) تبين أن معامل التحديد لنموذج الانحدار اللوجيستى بلغ (94%) وهو أكبر بكثير من معامل التحديد لنموذج الشبكات العصبية والذي بلغ (٨٩%)، وكذلك بالنسبة لتحليل التمييزي والذى بلغ (51%), كما تبين أن نسبة التقسيم للنموذجين الانحدار اللوجستي والتحليل التمييزي أكبر مما لنموذج الشبكات العصبية، وإن كانت كبيرة للنموذجين إلا أنها لنموذج الانحدار اللوجيستى أكبر من نموذج التحليل التمييزي، وكأحد معايير الجودة المطلوبة فى النموذج فنجد أن التحليل التمييزي لم يجتاز اختبار Box-m حيث أن الاختبار معنوى عند مستوى ثقة (٩٩%) وفى المقابل فى تحليل الانحدار اللوجيستى فنجد أنه قد اجتاز اختبار(Hosmer) وأن القيم المشاهدة لا تختلف عن القيم المتوقعة، وعليه فنجد أن تحليل الانحدار اللوجيستى اجتاز أكبر عدد من معايير الجودة ومن ثم فإن نموذج تحليل الانحدار اللوجيستى أفضل من نموذج التحليل التمييزي والشبكات العصبية.

٧- الاستنتاجات:-

١- إن الطرق الإحصائية الثلاث لهما القدرة على التصنيف وبنسب متفاوتة أي يمكن استخدام أي منهما لتصنيف الحالات الجديدة اعتماداً على قيم المتغيرات المستقلة لتلك الحالات.

ولكن افضل نموذج بينهم هو نموذج الانحدار اللوجستي لما امتلكه من معايير جيدة للجودة.

حيث تمثلت المعايير في $R^2, \chi^2, Box-M, Hosmer\ \&\ Lamsho, MSE, RMSE$

2 - يوجد هناك تطابق خاصة في اسلوبي الانحدار اللوجستي

والشبكات العصبية من حيث أهمية المتغيرات المستقلة المؤثرة معنوياً وغير المؤثرة في عملية تصنيف المفردات الجديدة، حيث اثبتت كلتا الطريقتين أن المتغيرات المستقلة
(RDW,RBC,EOS,NEW)
تعتبر من أهم المتغيرات لتحديد الاصابة بمرض سرطان الدم، بينما ظهر واضحا أن المتغيرات التي لها تأثير واضح في التحليل التمييزي هم (RDW، NEW، RBC ، LYM، MPV، MCH ، EOS).

٣- إن نسبة الخطأ الحقيقي في التصنيف وفي جميع الطرق المستخدمة قليلة نسبياً وهذا دليل كفاءة الطرق المستخدمة.

٤- من جدول التصنيف في التحليل التمييزي وجد أن (٢٠) مفردة (غير مريض) تم تقسيمهم تقسيما صحيحا بنسبة بلغت (100%)، وأن (٢٣٧) مفردة (مريض) تم تقسيمهم تقسيما صحيحا بنسبة بلغت (100%)، أيضا أن (٢٥٧) من عينة البحث والدراسة تم تقسيمهم تقسيما صحيحا بنسبة بلغت (98.8%) وهي نسبة تقسيم مرتفعة جدا.

٨- التوصيات:-

١- الاستفادة من الأساليب الإحصائية المتقدمة مثل النموذج اللوجستي، ونماذج التصنيف الحديثة المتمثلة في نماذج الشبكات العصبية للفصل أو التمييز بين مجموعتين أو أكثر، في جميع مجالات المعرفة إذا كانت المتغيرات خليط بين المستمرة والمتقطعة أو لا تتبع التوزيع الطبيعي.

٢- تعميم فكرة استخدام أساليب تحليل التمايز - الانحدار اللوجستي الثنائي وأسلوب الشبكات العصبية في المجالات الاجتماعية والاقتصادية وعدم تركيزها على المجالات الطبية فقط كما كان في السابق.

٣- إذا كان لدينا نموذج ذو متغير تابع نوعي (ثنائي الاستجابة)، وخليط (كمية ، نوعية) من المتغيرات المستقلة، أو إذا كان المتغير التابع أو بعض المتغيرات المستقلة لا تتبع التوزيع الطبيعي يفضل استخدام أسلوب النموذج اللوجستي الثنائي أو الشبكات العصبية الاصطناعية.

٩- المراجع العربية والاجنبية:-

٩-١ المراجع العربية:

١- عيسى، مريم كرسوع (٢٠١٢)، **"مرض السرطان في قطاع غزة"**، رسالة ماجستير، كلية الآداب، الجامعة الإسلامية، غزة.

٢- البيلي، منى مصطفى إبراهيم (١٩٩٢) ، " دراسة جزئية وكلية لوفيات الرضع في مصر من حيث (المستوى- الاتجاه – المحددات")،رسالة ماجستير، جامعة المنصورة.

٢-٩ المراجع الاجنبية:

1- Agrest, Alan. et al, (1997).“ **Statistical Methods for the Social Science”**, Third Edition, upper Saddle River, New Jersey, P.582.
2- Fausett , L. (1994). “**Fundamentals of Neural Networks Architectures, Algorithm and Application”.** New York: prentice- Hall, Inc.
3- Hosmer, D.W &Lemeshow, S. (1998). **“Applied Logistic Regression”**. New York. John Wiley.
4- Johnson, R. A. &Wichern, D.W. (1998). “ **AppliedMultivaraite Statistical Analysis “**, Prentice- Hall, INC, Englewood cliffs, New Jersey,p.594.
5- Pample, Fred C. (2000). “ **Logistic Regression A primer “** ,SAGE, publication, London, p.71.
6- Shankar, T.N.(2008). **“Neural Networks"**. UNIVERSITY SCIENCE PRESS.
7- Wayne, W. (1995). “ **BIOSTATISTICS: A foundation for Analysis in the Health Science “**. SIX

Edition, John Wiley.