استخدام الانحدار اللوجستي والتحليل التمييزي لدراسة حالات الاصابة بمرض الاسهال لدى الاطفال في العراق (دراسة تطبيقية)

أ. د/ البيومي عوض عوض طاقية د/ هشام محمد رجب المنجى

الباحث/ كريم خلف عزر العتابي

المستخلص:

من أوسع الأساليب الإحصائية استخداماً في مجال تحليل البيانات المصنفة هما اسلوبي الانحدار اللوجستي والتحليل التمييزي واللذان يمكن استخدامهما كنماذج خطية او غير خطية في مجال تصنيف البيانات.

ان استخدام التحليل التمييزي يشترط فيه توافر عدد من الافتراضات من أهمها ان تكون بيانات المتغيرات المستقلة مستمرة وتتبع التوزيع الطبيعي، وكذلك تساوي مصفوفتي التباين والتغاير لمجتمعين، اما في الانحدار اللوجستي لا توجد شروط تخص المتغيرات المستقلة اذ يعد نسبيا أكثر حصانه من التحليل التمييزي وكذلك يتسم بالمرونة والبساطة ويعطي تفسيرا واضحاً وذو دلالة ومعنى لوصف العلاقة بين المتغير التابع والمتغيرات المستقلة.

تهدف هذه الدراسة إلى تصنيف البيانات على وفق ثلاث نماذج الأول نموذج الدالة الاحتمالية لدالة الانحدار اللوجستي المتعدد الاستجابة، والثاني نموذج دالة التمييز الخطي والثالث نموذج احتمالات الاستجابة لدالة التمييز الخطي، حيث استخدم الباحث اسلوبين من أساليب التصنيف والتنبؤ هما التحليل التمييزي والانحدار اللوجستي وعمل مقارنه بينهما لتحديد اهم العوامل الذي توثر على طبيعية المرض وقد تكون هذه العوامل اقتصادية وديموغرافية وطبية وغيرها والمتمثلة بالمتغيرات المستقلة وما مدى تأثيرها على المتغير التابع ومعرفة أي من النماذج الثلاثة افضل وأدق في توفيق البيانات وأيهما يظهر باقل نسبة خطأ تصنيف ممكن.

ABSTRACT

Among the most widely used statistical methods in classifyieg data are logistic regression, and discriminatory analyzed, which can be used as

linear models in data classification.

There should be a number of assumptions for using analysis discriminatory, the most important among them is that the data of the independent variables are continuous and follow the normal distribution, and the population have common variance matrix. While in logistic regression, there are no conditions concerning the independent variables, as it is encountered to be more accurate than analyzed discriminatory, and it is also simple and flexible, and gives a clear significant explanation and a meaning for describing the relation between both dependent and independent variations.

This study also aims at classifying the data based on three models: the first model is the probability function for the multi-responsive logistic regression. The second model is linear discriminatory function. The third model is the prospected discriminatory function response; The researcher used two models of classification and prediction, which are analyzed discriminatory and logistic regression and made a comparison between them to identify the most important factors that affect the nature of the disease, these factors might be economic, demographic, medical, etc. which are represented in dependent variables, and how it has an effect on the independent variables , besides recognizing which of the two models is better and more accurate in results, and which of them shows the least percentage of fault in classification.

(١-١) مقـدمـــــة:

يعد مرض الاسهال من الامراض المنتشرة والخطيرة التي يعاني منها المجتمع عامة والأطفال على وجه الخصوص، ويعتبر هذا المرض من اكثر الاسباب الكامنة وراء وفيات الاطفال والحالات المرضية التي تصيبهم في مختلف انحاء العالم وهو يحدث أساسا جراء الاغذية ومصادر المياه الملوثة والبيئة غير الجيدة وغيرها، وفي عام 2014 كان مرض الاسهال ثالث اهم اسباب الوفاة في البلدان منخفضة الدخل حيث تسبب في وقوع 6.9% من مجموع الوفيات كما يمثل هذا المرض ثاني اهم اسباب الوفاة الاطفال دون سن الخامسة من عمرهم بعد الالتهاب الرئوي، والملاحظ ان 80% من مجموع الاطفال الذي قضوا نحبهم بسبب هذا المرض اللعين في عام 2014 كانوا دون سن الثانية من ربيع عمرهم ومن هنا لابد ان يكون دور هام وفعال لعلم الاحصاء كونه العلم الذي

يهتم في جميع مجلات العلوم الاخرى قد تكون اقتصاديه او صحية او بيئة او اجتماعيه وغيرها لتحديد اهم العوامل المؤثرة على مرض الاسهال وكيف يتم الحد منها حيث تطرق الباحث في دراسته التطبيقية باستخدام اسلوبين من اساليب التصنيف والتنبؤ هما التحليل التميزي والانحدار اللوجستي وعمل مقارنة بينهما ومعرفة ايهما أفضل من خلال دراسته التطبيقية على بيانات الاطفال المصابين بمرض الاسهال.

مشكلة البحث:

يعتبر علم الاحصاء أحد الوسائل المهمة في البحث العلمي من خلال استخدام قواعدهُ وقوانينهُ وطرقهُ في جمع البيانات والمعلومات اللازمة للبحث العلمي وتحليل هذه البيانات والمعلومات من اجل الوصول الى النتائج التي يهدف لها البحث، ان استخدام الاسلوب الاحصائي في البحث العلمي عن الظاهرة ودراستها في ذلك البحث و وضع الخطط المستقبلية عن طريق التنبؤ بالنتائج والوصول الى اهم العوامل المؤثرة على المتغير التابع، ولهذا العلم دور هام وفعال في كافة القطاعات سواء كانت انتاجيه ام خدميه وحيث ان الاحصاء بحد ذاته يعتبر وسيلة وليس غاية فذلك يعني امكانية استخدامه اين ما وجد البحث العلمي ويدخل هذا العلم في جمع المجالات ان كانت زراعية، صناعية، نفسية، تربوية، رياضية، طبية، اقتصادية، إدارية، هندسية وغيرها، في هذه الدراسة تم التطرق الى اسلوبين من أساليب التنبؤ والتصنيف هما الانحدار اللوجستي وتحليل التمييزي لتحديد اهم العوامل الذي تؤثر على المتغير التابع (مرض الاسهال) وفي هذه الدراسة نبين التالي:-

١-تحديد المتغيرات المستقلة التي تؤثر على المتغير التابع (مرض الاسهال).

٢-ابعاد المتغيرات الضعيفة والزائدة لان وجودها يعقد التحليل ويجعل أخطاء التصنيف أكثر.

٣-اختيار المتغيرات المستقلة ذات التأثير الواضح على المتغير التابع وذلك من خلال مقاييس خاصة بكل أسلوب.

هدف البحث:

تهدف هذه الدراسة إلى تصنيف البيانات وفق ثلاثة نماذج الأول نموذج الدالة الاحتمالية لدالة الانحدار اللوجستي المتعدد الاستجابة، والثاني نموذج دالة التمييز الخطي والثالث نموذج احتمالات الاستجابة لدالة التمييز الخطي، والمقارنة بينهما على أساس احتمال خطأ التصنيف من أجل الوصول إلى أفضل نموذج خطي لتشخيص مرض الإسهال لدى الأطفال وتحديد أهم العوامل المؤثرة على الإصابة به.

اهـــمية الـبـحـــث:

كخطوة مساهمة في خدمة القطاع الصحي والرفع من مستواه تناول الباحث في دراسته التطبيقية بيانات الأطفال المصابين بمرض الإسهال التي تؤثر على طبيعة هذا المرض اللعين الذي يعانيه الأطفال، قد تكون عوامل بيئية، صحية، اقتصادية وغيرها من العوامل التي تؤثر على هذا المرض.

واتبع ما يلي: -

١-عمل مقارنة بين أسلوبين من أساليب التصنيف والتنبؤ هما: التحليل التمييزي، والانحدار اللوجستي، ومعرفة أهم العوامل

٢-معرفة أي من الأسلوبين أفضل وأدق في اظهار النتائج، وأيهما يظهر بأقل نسبة خطأ تصنيف ممكن.

متغيرات الدراسة المقترحة: -

المتغير التابع: مرض الإسهال.

أما المتغيرات المستقلة وهي: -اقتصادية، اجتماعية، صحية، ديموغرافية، ومتغيرات خاصة بالأغذية وغيرها

حجم ومجتمع الدراسة:

تم عمل استمارة استبيان وذلك بالتعاون مع بعض الأطباء المختصين بمستشفى الطفل العربي في بغداد، للوصول الى اهم العوامل المؤثر على طبيعة المرض وقد شملت الاستبانة ٣٣ سؤال وتم تقسيمهم الى أربع مجاميع وهي (ديموغرافية، وصحية، اقتصادية، مجموعة من الأسئلة خاصة بالتغذية)، بينما توصلنا احصائيا ومن خلال المعادلة التالية: -

$$n = \frac{z^2 p(1-p)}{E^2}$$

الى عمل ٣٨٤ استمارة استبيان وتم توزيعها على أولياء أمور الأطفال المصابين والاصحاء الذين يعيشون تحت نفس الظروف.

P: نفرض ان الظاهرة محل الدراسة تحقق في المجتمع بنسبة ٥٠%

1-P : نسبة عدم تحققها٥٠%

وعند معامل ثقة ٩٥%

وقيمة Zالمعيارية= ١.٩٦

E الخطأ =٥٠%

حيث ان:

$$n = \frac{(1.96)^2 0.5(0.5)}{(0.05)^2} = 384.16$$

الأســالــيب الاحصــائيــة المستخدمة:

١- <u>الانحدار اللوجستـي</u>:

يعتـــر نمـــوذج الانحـــدار اللوجستـــي مـــن اهـــم النمـــاذج المستخدمـــة لصيـــاغة دالـــة التمـــيز، ويتميـــز هـــذا النمـــوذج بملائمتـــه للعديـــد مـــن الاستخـــدامات، ويســـتخدم الانحـــدار اللـــوجستـــي بصفـــة عـــامة لتحليـــل العـلاقـــة بيـــن مـــا حـــدث فـــي المـــاضي ومـــا يمكـــن ان يحـــدث فـــي المستقبـــل، وينشـــا الانحـــدار اللوجيستـــي عن اســـتخدام تحويلـــه اللوجيـــت عنـــدما يكـــون المتغيـــر التـــابع (Y) يعبـــر عـــن احتـــال حـــدوث حـــدث معيـــن (θ)، وهنـــاك مجمـــوعة مـــن المتغيـــرات التفسيـــرية $X = (X_1, X_2, \ldots, X_P)$ عـــددها P متغيـــر، وهنـــاك عـــلاقة بيـــن الاحتمـــال (θ) والمتغيـــرات التفسيـــرية علـــى الصـــورة الخطيـــة التاليـــة:

$$\theta_k = B_0 + B_1 x_{1k} + B_2 x_{2k} + \ldots\ldots + B_p x_{pk} + ei$$

حيث ان الطرف الايسر محدود ($0 < \theta_k < 1$) بينما الطرف الايمن من المعادلة غير محدود. وبإستخدام تحويلة اللوجيت نصل الى ان:

$$P_r = (Y_K = 1/X) = \frac{\exp(BX)}{1+\exp(BX)} = P_r(X)$$

حيث تعرف تحويله اللوجيت والتي تعتبر الاساس في هذا النموذج وكما يلي:

$$g(x) = Ln\left(\frac{P_r(x)}{1-p_r(x)}\right) = BX$$

التحليل التمييزي:

يعد تحليل التمايز Discriminant Analysis أحد الأساليب المهمة في التحليل الإحصائي متعدد المتغيرات Multivariate Statistical Analysis، والتي تهدف الى تكوين نموذج إحصائي يمثل العلاقة المتبادلة بين عدة متغيرات مستقلة من حيث قدرتها على التنبؤ بمتغير تابع نوعى (ثنائي، متعدد)، حيث انه يهتم بالتمايز بين مجموعتين أو أكثر، والتفرقة بين المجتمعات المتداخلة.

وفي عام ١٩٣٦ قام الباحث فيشر Fisher باقتراح الدالة المميزة الخطية للوصول إلى أعلى مجال للفصل بين مجموعتين تابعتين إلى نفس المجتمع وذلك باستعمال عدد من المتغيرات المرتبطة بتصنيف المفردات حسب مجتمعاتها وهذه أول عملية تمييز حقيقية لتصنيف المفردة

إلى أحد المجموعتين وعرف الدالة التميزية الخطية بالشكل التالي: -

$$Y = B_1 X_1 + B_2 X_2 + ... + B_P X_P$$

$$Y = \underline{B}\,\underline{X}$$

حيث ان $\underline{B}$ متجه المعالم، $\underline{X}$ مصفوفة البيانات

أنواع الدوال التمييزية:

هناك عدة أنواع من دوال التمييزي والتي يمكن استخدامها بالاعتماد على البيانات المدروسة وخصائصها ومن هذه الدوال: -

دالة التمييز الخطية. الذى يستخدم لدراسة مدى تداخل المجموعات.

دالة التمييز التربيعية. (٢) يهدف إلى تكوين قاعدة للفصل بين المجموعات محل البحث والدراسة بناءً على عينة من المفردات يتم سحبها من هذه المجموعات.

دالة التمييز اللوجستية.

وتتمثل خصائص أسلوب التحليل التمييزي فيما يلي:

(١) يعتبر أحد الاساليب الإحصائية متعددة المتغيرات

انواع التحليل التمييزي:

ينقسم تحليل التمييزى إلى

١- التحليل التمييزي الخطى.

٢- التحليل التمييزي المختلط.

تحليل التمييزى الخطي

يعتبر تحليل التمييزي الخطى من اكثر انواع تحليل التمييزي شيوعا، وتقوم فكرته على تعظيم التباين بين المجموعات نسبة الى التباين داخل المجموعات (Fisher ,1936).

والهدف من ذلك هو عمل اساس لتصنيف المشاهدات المستقبلية بشكل دقيق الى المجموعات المستقلة التي تنتمى اليها، وتصل درجة التصنيف وفقا لهذا الاسلوب الى حدها الاقصى في ظل تحقق الفروض التالية: -

i. تتبع المتغيرات العشوائية التوزيع الطبيعي المتعدد.

ii. تساوى مصفوفة التغاير لجميع المجموعات.

ولتوضيح كيفية تصنيف البيانات نفترض ان المجتمع الأصلي للبيانات ينقسم الى مجموعتين G_1 و G_2، وأن كلاهما يتبعان التوزيع الطبيعي متعدد المتغيرات بمصفوفة تغاير متساوية " مشتركة " للمجموعتين ومتوسطات و μ_1, μ_2 ، وبافتراض أن لدينا عينة عشوائية حجمها (n) من المفردات.

ونرغب في تصنيف مفردات هذه العينة إلي المجموعات التي تنتمى اليها فان ذلك يتم عن طريق ما يسمى (بمعادلة الاحتمالات البعدية) والتي على اساسها يتم تصنيف المفردات في المجموعتين، حيث تعطى المعادلة احتمال ان تكون المفردة ضمن المجموعة الاولى ويكون الاحتمال المكمل هو احتمال ان تكون المفردة ضمن المجموعة الثانية, وبناءً على ذلك يتم تصنيف المفردات الى المجموعة الاولى اذا كانت معادلة الاحتمال البعدي اكبر من $\frac{1}{2}$ بينما تصنف المفردات إلي المجموعة الثانية فيما عدا ذلك، ويتم حساب الاحتمال البعدي بالشكل التالي:-

$$\pi_1 = \frac{p_1 \phi(x \mid \mu_1, \Sigma)}{p_1 \phi(x \mid \mu_1, \Sigma) + p_2 \phi(x \mid \mu_2, \Sigma)}$$

حيث ان:

X: هو متجه المتغيرات العشوائية q

$\varphi(x/u_g, \sum)$:-هو داله كثافة التوزيع الطبيعي متعدد المتغيرات x بمتوسط μ_g ومصفوفة تغاير $\sum$.

p_1, p_2: هي نسب المجموعات ٢،١ في المجتمع.

وتعتبر كلا من($\sum_g, u_g$) مجهولتان وتقدران من البيانات باستخدام طريقه المربعات الصغرى، اما الاحتمال البعدي للمجموعة (٢) يتم حسابه كالاتي:-

$$\pi_2 = 1 - \pi_1$$

ومنذ تقديم طريقه تحليل التمييزي الخطى، قام العديد من الباحثين بإدخال تعديلات على تحليل التمييزي الخطى التقليدي وذلك لتحقيق أداء تصنيفي أفضل. ومن الطرق المنبثقة عن تحليل التمييزي الخطى ما يلي

تحليل التمييزي المختلط Mixture Discriminant Analysis (MDA)

يعتبر تحليل التمييزي المختلط امتداداً لتحليل التمييزي الخطى والذي يقيس كثافة المتغيرات داخل المجموعة من خلال مزيج من التوزيعات الطبيعية متعددة المتغيرات ويستخدم هذا النوع من تحليل التمييزي بشكل أساسي عندما تكون العلاقات بين المتغيرات داخل كل مجموعة غير خطية، حيث يتيح تحليل التمييزي المختلط نموذجة تلك العلاقات مما يزيد من دقة التصنيف بالإضافة الى امكانية تحديد العوامل الكامنة المؤثرة في تصنيف المشاهدات داخل كل مجموعة ان وجدت.

ويتم تصنيف البيانات بتحليل التمييزي المختلط اعتمادا على معادلة الاحتمال البعدي، ويتم حساب الاحتمال البعدي للمجموعة (١) كالاتي: -

$$\pi_1 = \frac{p_1 \sum_{c=1}^{C_1} \tau_{c1} \phi(x \mid \mu_{c1}, \Sigma_{c1})}{p_1 \sum_{c=1}^{C_1} \tau_{c1} \phi(x \mid \mu_{c1}, \Sigma_{c1}) + p_1 \sum_{c=1}^{C_2} \tau_{c2} \phi(x \mid \mu_{c2}, \Sigma_{c2})}$$

حيث ان:

t_{cg}: هي مزيج من نسب المكون c في المجموعة g ويعرف المكون c كتوزيع طبيعي متعدد المتغيرات يستخدم لقياس كثافه المتغيرات العشوائية داخل المجموعات. تشير إلى داله الكثافة.

أما الاحتمال البعدي للمجموعة(٢) يتم حسابه كالاتي:

$$\pi_2 = 1 - \pi_1$$

وغالبا ما يتم تقدير المعالم في دالة الاحتمال البعدي بطريقة الامكان الاعظمMLE وذلك باستخدام خوارزمية Expectation (EM) Maximizationلإيجاد

تقديرات دالة الإمكان الأعظم MLE لمعالم نموذج تحليل التمييزي المختلط.

الدراسات السابقة

- هدفت دراسة النقاش، عائدة صالح هادي(2008) بعنوان (استخدام التحليل المميز لتشخيص بعض امراض العيون) إلى تشخيص العوامل المؤثرة في أمراض العيون ومعرفة تأثير كل عامل من هذه العوامل من خلال استخدام الدالة التمييزية الخطية، تم استخدام التحليل المميز لتصنيف بعض أمراض العيون إلى ثلاث مجموعات اعتماداً على متغيرات ذات صفات تمييزية، تضمنت الدراسة سحب عينة عشوائية لكل مجموعة على حدة، وباستخدام الدالة المميزة الخطية تم تصنيف بعض أمراض العيون على أساس مجموعة من المتغيرات المصاحبة للمرض.

• أجريت دراسة عبد الماجد، أنور إدريس حسين (2009) بعنوان: (استخدام النموذج اللوجستي المتعدد لتحديد العوامل المؤثرة في مرض الجل كوما) بهدف تحديد عوامل الخطر على مرضى الجل كوما، التقنية التي استخدمت في التحليل هي نموذج الانحدار اللوجستي المتعدد.

تم استخدام خمسة متغيرات كعوالم خطر محتملة، إلا أن اثنين منهما فقط هما ضغط العين والعامل الوراثي تبين أن لهما علاقة معنوية بالمرض وبالتالي يمكن اعتبارها عوامل خطر، بصفة خاصة وجد الباحث أن خطر الإصابة بالجل كوما للأشخاص الذين يعانون من ضغط عين مرتفع حوالي (1.8) مرة ذلك الذي للأشخاص بضغط عين طبيعي، كذلك فإن الأشخاص الذين يوجد في أسرتهم من عانى مرض الجل كوما يكون الخطر حوالي (31.9) مرة مقارنة بالأشخاص الذين ليس في أسرهم من يعاني من المرض.

• **بينت Abbas, F. M. Azhar, M.E (2008)** في هذه الدراسة إلى مقارنة النماذج الخطية والتربيعية والمختلطة للتحليل التميزي تحت متغيرات مستمرة غير الحالة الطبيعية فيما يتعلق بدقة التصنيف من خلال محاكاة مونت كارلو وتوصل إلى أن الطرق الخطية

للتحليل التميزي والتحليل التميزي التربيعي يمكن أن تكون شبه مثلى في التوزيعات غير المتماثلة (الملتوية) ، والتمييز اللوجستي التربيعي يتم توضيحه بأنهُ يحلل على نحو ملائم حالة عد م الالتواء فقط في الحالات التي توجد فيها المتغيرات غير المتماثلة (الملتوية) بنفس الاتجاه عبر المجاميع وأحجام العينة الكبيرة جداً على نحو كاف والتحليل التميزي المختلط ذو العدد الصغير نسبياً من المركبات في كل مجموعة إذ تم التوصل بأنهُ مدخل واعد للتعامل مع المتغيران المستمرة غير الحالة الطبيعية في سياق التحليل التميزي.

الجانب التطبيقي:

التأكد من صلاحية أدوات الدراسة

معامل الثبات Reliability Coefficient

صدق الاتساق الداخلي لفقرات الاستبانة

والمقصود بثبات أداة القياس أن يعطي النتائج نفسها إذا أعيد تطبيق الاستبانة على نفس العينة في نفس الظروف ويتم قياسه بثلاث طرق:

الطريقة الأولى: الاختبار وإعادة الاختبار

يتم في هذه الطريقة تطبيق الاستبانة على عينة استطلاعية مرتين بينهما فارق زمني مدته أسبوعان ثم حساب معامل الارتباط بين إجابات المفحوصين في المرتين، فإذا كانت معامل الارتباط مرتفعا فان هذا يكون مؤشرا على ثبات الاستبانة وبالتالي على صلاحية وملائمة هذه الاستبانة لأغراض الدراسة.

الطريقة الثانية: الثبات عن طريق التجزئة النصفية:

حيث يتم تجزئة فقرات الاستبانة إلى جزأين، الجزء الأول يمثل الأسئلة الفردية والجزء الثاني يمثل الأسئلة الزوجية ثم يحسب معامل الارتباط (r) بين درجات الأسئلة الفردية ودرجات الأسئلة الزوجية ثم تصحيح معامل الارتباط بمعادلة بيرسون براون كالتالي:

$$\text{Reliability Coefficient} = \frac{2\ r}{1\ +\ r}$$

الطريق الثالثة: معامل ثبات كرو نباخ الفا

يتم حساب معامل ثبات كرو نباخ الفا باستخدام برنامج SPSS والذي من خلاله نحسب معامل التمييز لكل سؤال حيث يتم حذف السؤال الذي معامل تمييزه ضعيف أو سالب يقصد بالاتساق الداخلي لأسئلة الاستبانة هي قوة الارتباط بين درجات كل مجال ودرجات أسئلة الاستبانة الكلية، والصدق ببساطة هو أن تقيس أسئلة الاستبانة أو الاختبار ما وضعت لقياسه أي يقيس فعلا الوظيفة التي يفترض انه يقيسها.

وسوف نعرض النتائج التي طبقها الباحث على اسئلة الاستبيان.

وقد اشتمل كل مجال على عدد من الفقرات.

تحقق من صدق الاداة باستخدام نوعين من الصدق وهما: الصدق الظاهري وذلك بتوزيع الأداة على عدد من المفحوصين الذين استجابوا عليها بسهولة وبيسر حيث كانت الأسئلة والفقرات واضحة بالنسبة لهم وقد كان ذلك مؤشرا على صدق الأداة الظاهري، أما النوع الثاني من أنواع الصدق فهو صدق المحكمين حيث تم عرض الأداة على عدد من الأساتذة المختصين في مجال البيئة والاحصاء، حيث طلب منهم تحكيم الاستبانة في ضوء اهداف الدراسة، وقد ابدى المحكمون ملاحظاتهم، وقد صيغت الاداة وفق ملاحظاتهم واعتبرت هذه الاجراءات دلالات صدق ظاهري للأداة.

لإيجاد معامل الثبات بطريقة التجزئة النصفية نوجد معامل الارتباط بين المتغيرين " av_odd" , "av_even " وتكون النتائج كالتالي:

من النتائج يتبين أن معامل الارتباط يساوي ٠.٧٦٣

وبحساب تصحيح معامل الارتباط باستخدام معادلة سيبرمان براون نجد أن معامل الثبات يساوي

$$\text{معامل الثبات} = \frac{2 \times 0.763}{1 + 0.763} = 0.866$$

هو معامل ثبات مقبول ودال إحصائياً

الإحصاءات الوصفية

اذ يوضح الجدول الاول أهم تلك الاحصاءات الوصفية وهي نسبة الإصابة بمرض الاسهال لعينة البحث (70.3) وهي نسبة مرتفعة إلى حدٍ ما اذ تشكل ما يقرب من ثلاث ارباع عينة البحث

جدول(١)

القيم التكرارية لتصنيف المتغير التابع

نوع الاستجابة	عدد الحالات	النسبة المئوية %	المتوسط	الانحراف المعياري
مصاب بالإسهال	٢٧٩	70.3	٠.7	٠.458
غير مصاب بالإسهال	١١٨	29.7		
المجموع	٣٩٧	100		

اما جدول (٢) فيعرض الاحصاءات الوصفية للمتغيرات التوضيحية المستقلة، من متوسطات ونسب مئوية كما يعرض الجدول الانحرافات المعيارية لمتغيرات البحث

جدول (٢)
الإحصاءات الوصفية للمتغيرات المستقلة

المتغيرات	عدد الحالات	النسبة المئوية	المتوسط	الانحراف المعياري
نوع الطفل				
ذكر	٢٢٨	57.4	1.43	0.495
انثى	١٦٩	42.6		
عمر الطفل				
اقل من سنتين	١٨٩	47.6	1.52	0.5
أكبر من سنتين	٢٠٨	52.4		
السكن				
مدينة	٢٣٨	59.9	1.4	0.491
ريف	١٥٩	40.1		
تعليم الام				
امية	٥٣	13.4	3.28	1.51
ابتدائي	٩٢	23.2		
متوسط	٨٠	20.2		

أعدادي	٦٠	15.1		
جامعي	٨٨	22.2		
دراسات عليا	٢٤	6.0		
تعليم الاب				
امي	٣٧	9.3		
ابتدائي	٥٤	13.6		
متوسط	٨٣	20.9	3.78	1.534
أعدادي	٥٨	14.6		
جامعي	١١٤	28.7		
دراسات عليا	٥١	12.8		
حالة الام				
الاجتماعية	٤٧	11.8		
مطلقة	٤٨	12.1	2.64	0.684
ارملة	٣٠٢	76.1		
غير ذلك				
عمر الام				
١٥-٢٥ سنة	٦٦	16.6		
٢٥ – ٣٥ سنة	٢٠٩	52.6	2.14	0.674
٣٥ – ٤٥ سنة	١٢٢	30.7		
نوع السكن				
ملك	٢٨٠	70.5	1.29	0.456
ايجار	١١٧	29.5		
نوع الرضاعة				
طبيعية	٢٧٥	69.3	1.31	0.462
صناعية	١٢٢	30.7		
نوع مياه الشرب				
معدنية	١٥٦	39.3	1.61	0.489
مصفاة	٢٤١	60.7		

التحليل التمييزي

يستخدم التحليل التمييزي عندما تكون العينة مختارة من مجتمع يتبع التوزيع الطبيعي وفى حالة عدم تحقق هذه الفرضية فأننا نستخدم الانحدار اللوجستي والذي لا يتطلب اختبارات حول التوزيع الطبيعي.

اختبار ان العينة المختارة من مجتمع يتبع التوزيع الطبيعي

جدول (٣)
اختبار كولومجروف سيمرنوف حول التوزيع الطبيعي للبيانات

المتغيرات	احصاء الاختبار	مستوى المعنوية
الاصابة بالإسهال	**2.3**	**.2**
النوع للطفل	**0.157**	**.17**
العمر للطفل	**1.7**	**.013**
السكن	**0.185**	**1.7**
تعليم الام	**1.4**	**.952**
تعليم الاب	**0.437**	**1.6**
الحالة الاجتماعية	**.315**	**.452**
عمر الام	**1.67**	**.005**
نوع السكن	**1.92**	**1.2**
نوع الرضاعة	**.831**	**1.1**
مياه الشرب	**.357**	**.638**

تبين نتائج جدول (٣) ان غالبية المتغيرات تتبع التوزيع الطبيعي وذلك اعتمادا على احصاء اختبار كولومجروف سيمرنوف ومستوى المعنوية حيث يتضح ان مستوى المعنوية لجميع المتغيرات أكبر من 0.05 فيما عدا متغير العمر للطفل وعمر الام وحيث ان حجم البيانات هو ٣٩٧ وهذا يتجاوز العدد ٣٠ مشاهدة فانه يمكن اعتبار هده البيانات تتوزع بالتقريب حسب التوزيع الطبيعي طبقا لنظرية النهاية المركزية.

اختبار تجانس تباين المجموعتين: -

تم حساب احصاء الاختبار Box's M في جدول (٤) وكانت قيمة مستوى المعنوية هي 0.213 وهي أكبر من 0.05 وهذا يشير الى قبول الفرض العدمي الذي يدل على

تجـانس التباينـات للمجمـوعتين الدالـــــــــة التمييزيـــــــــة وهـذا يـدل علـى تحقـق شـروط

جدول (٤)

اختبار تجانس المجموعتين

Box's M	42.417
F approx.	1.41
Df1	62
Sig.	.213

اختبـــار معنويـــة المتغيـــرات المســـتقلة باســـتخدام اختبـــار F للحصـول علـى الدالـة التمييزيـة المثلى بين المجموعتين لمعرفة اهميـة كـل متغيـر ومـدى تـأثيره على المتغير التابع وهو الاصابة بمرض الاسهال

ومن نتائج جدول (٥) نجد ان المتغيرات عمر الطفل والسكن والمستوى التعليمي للام ونوع الرضاعة ونوع مياه الشرب لها تأثيرات كبيرة في عملية التصنيف اما بقية المتغيرات فكانت ذو تأثير ضعيف

جدول (٥)
اختبار معنوية متغيرات الدالة التمييزية

المتغيرات	Wilks' Lamda	F	Df1	Df2	مستوى المعنوية
النوع للطفل	٠.943	1.4	١	٣٩٥	٠.23
العمر للطفل	٠.982	8.2	١	٣٩٥	٠.001
السكن	0.931	7.12	١	٣٩٥	0.0
تعليم الام	٠.927	5.77	١	٣٩٥	٠.008
تعليم الاب	٠.997	6.3	١	٣٩٥	0
الحالة الاجتماعية	1.0	.981	١	٣٩٥	٠.41
عمر الام	0.993	1.8	١	٣٩٥	٠.13
نوع السكن	٠.962	.574	١	٣٩٥	٠.512
نوع	٠.973	6.92	١	٣٩٥	٠.009

الرضاعة					
مياه الشرب	٠.921	6.44	١	٣٩٥	٠.005

ولتقدير دالتي التمييز الخطية للمجموعتين كانت النتائج كما يلى:

جدول (٦)

الدوال التمييزية الخطية التقديرية

المتغيرات	الدالة التمييزية الاولى	الدالة التمييزية الثانية
الثابت	-55.6	-59.8
النوع للطفل (X1)	17.44	15.4
العمر للطفل (X2)	17.52	19.62
السكن (X3)	19.13	18.6
تعليم الام (X4)	17.81	18.8
تعليم الاب (X5)	19.43	18.3
الحالة الاجتماعية (X6)	8.92	8.98
عمر الام (X7)	7.82	13.65
نوع السكن (X8)	16.2	15.5
نوع الرضاعة (X9)	22.7	21.6
مياه الشرب (X10)	17.89	16.66

$$y_1 = -55.6 + 17.44X_1 + 17.52X_2 + 19.13X_3 + 17.81X_4 + 19.43X_5 + 8.92X_6 + 7.82X_7 + 16.2X_8 + 22.7X_9 + 17.89X_{10}$$

$$y_2 = -59.8 + 15.4X_1 + 19.62X_2 + 18.6X_3 + 18.8X_4 + 18.3X_5 + 8.98X_6 + 13.65X_7 + 15.5X_8 + 21.6X_9 + 16.66X_{10}$$

يلاحظ من جدول (٦) ان معاملات بعض المتغيرات تقريبا متساوية في الدالتين وهذا يعنى ان هذه المتغيرات سيكون لها تأثير ضعيف في عملية تصنيف البيانات اما البيانات التي يكون لها دور مهم في تصنيف مرض الاسهال فهي كما يلي نوع الرضاعة ومياه الشرب والسكن وعمر الطفل والمستوى التعليمي لكل من الام والاب.

تصنيف مفردات (مشاهدات) المجموعتين حسب دوال التمييز التقديرية: قد يحدث ان يتم تصنيف مشاهدة معينة الى المجموعة الاولى بينما هي في الحقيقة تعود للمجموعة الثانية

او بالعكس وهذا ما يسمى خطأ التحليل لخطأ التصنيف.
التصنيف وفيما يلي نتائج هذا

جدول (٧)
تصنيف المشاهدات بحسب دالة التمييز الخطي لمجموعتين

الحالة	التصنيف		نسبة التصنيف الصحيح
	عائد للمجموعة الاولى	عائد للمجموعة الثانية	
المريض حسب المجموعة الاولى	١٣٤	٧٥	67.7
المريض حسب المجموعة الثانية	٤٧	١٤١	71.2
نسبة التصنيف الكلى	45.6	54.4	69.45

يلاحظ من جدول (٧) ان احتمال التصنيف الصحيح لمريض بالإسهال يعود للمجموعة الاولى هو 67.7% ويعود للمجموعة الثانية هو 71.2% في حين احتمال خطأ التصنيف للمجموعة الاولى 33.3%وللمجموعة الثانية 28.8%وبلغت نسبة التصنيف الصحيح الكلية 69.45% ونسبة التصنيف الخاطئ الكلية 30.55%

وكذلك جرى التصنيف باستخدام الصيغ الاحتمالية التقديرية كطرق بديلة للتصنيف وكانت النتائج كما يلى

جدول(٨)
تصنيف المشاهدات بحسب الصيغة الاحتمالية لدالة التمييز الخطي

الحالة	التصنيف		نسبة التصنيف الصحيح
	عائد للمجموعة الاولى	عائد للمجموعة الثانية	
المريض حسب المجموعة الاولى	١٢٩	٧٣	65.2
المريض حسب المجموعة الثانية	٤٨	١٤٧	74.2

نسبة التصنيف الكلي	44.6	55.4	69.7

بينت نتائج جدول (٨) ان احتمال التصنيف الصحيح لمريض يعود الى المجموعة الاولى كانت 65.2%، وللمجموعة الثانية كانت 74.2%في حين خطأ التصنيف للمجموعة الاولى 34.8% وللمجموعة الثانية 25.8% اما نسبة التصنيف الكلية الصحيحة هي 69.7%ونسبة التصنيف الكلية الخاطئة هي 30.3% وذلك وفق الصيغة الاحتمالية لتصنيف المشاهدات.

الانحدار اللوجستي:-

نتائج تقدير واختبار النموذج اللوجستي لعينة البحث:

وفيما يلي عرض ومناقشة للنتائج التي تم الحصول عليها في البداية يتم تضمين الثابت فقط في النموذج ثم بعد تلك الخطوة تضاف جميع المتغيرات التوضيحية (المستقلة) للاستدلال على أثرها مجتمعة في متغير الاستجابة (التابع) لتحديد كفاءة النموذج ككل.

جدول (٩) يظهر كل النتائج الخاصة بالنموذج البدائي الذي يحتوي على الثابت فقط .

جدول (٩)
جودة توفيق النموذج

Iteration		-2 Log likelihood	Coefficients Constant
Step 0	1	478.091	.806
	2	477.896	.854
	3	477.896	.855

إذ أوضحت هذه النتائج التوقف عند المحاولة (٣)، كما جاء تقدير احصاءة الإمكان الأعظم (-2 log likelihood) هو (٤٧٧.٨٩٦) مساويا إلى المحاولة السابقة، وقيمة

الثابت(.855٠)، واحصاءة Wald لهذا الثابت (٥٩.٩٤٥)، وقيمة مستوى المعنوية الخاص بهذه الاحصاءة (Sig = 0.000) وهي اقل من مستوى المعنوية المحدد من قبل الباحث (α = 0.05) لذا نرفض فرضية العدم ونقبل البديلة، أي نقبل معنوية هذا الثابت كما بينها جدول(١٠)

جدول (١٠)

تقدير معالم نموذج الانحدار اللوجستي

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 0 Constant	0.855	0.110	59.945	1	.000	2.350

وبدراسة جودة التوفيق للنموذج تم استخدام اختبار نسبة الامكان واختبار مربع كاي حيث بينت قيمة كا^{2} = 10.582وبمستوى معنوية 227. وهذه القيمة اكبر من 0.05مما يؤكد ملائمة النموذج للبيانات كما يتضح من الجدولي (١١) (١٢).

جدول(١١)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	352.774[a]	.273	.388

جدول(١٢)

جودة توفيق النموذج

Step	Chi-square	df	Sig.
1	10.582	8	.227

النتائج والتوصيات

اولاً: النتائج: -

• تم صول الى ان نموذج الانحدار اللوجستي من خلال مشتقة ضعف دالة الإمكان الأعظم حيث بلغت نسبتها(1.018) وهي اعلى وفق نتائج جدول (١٥)

•تبين لنا من تصنيف البيانات على وفق صيغة دالة التمييز الخطية وصيغة احتمال الاستجابة لدالة التمييز الخطية وصيغة الاحتمال لنموذج الانحدار اللوجستي وتم الانحدار اللوجستي تعطي اقل احتمال خطأ تصنيف حيث بلغت نسبة التصنيف الصحيح الكلي 69.7 حسب جدول (٨) في حين نسبة التصنيف الخاطئ بلغت (30.3)

والخلاصة من ذلك أن الدراسة أوضحت أن هناك توزيع جغرافي واضح سواء في نسب الحدوث أو تأثير المتغيرات الديموغرافية للأسرة لحدوث مرض الإسهال في الأطفال وأن التوزيع يرتبط بالمستوى البيئي والاجتماعي الذي يختلف بين المدن والقرى والمستوى التعليمي للام والاب وغيرها من النتائج التي اظهرتها الدراسة.

الوصول الى ان طريقة التصنيف بحسب الصيغة الاحتمالية لدالة التمييز الخطي أعطت نتائج أكثر احتمال خطا التصنيف

•أظهرت النتائج عند تصنيف البيانات على صيغة دالة التمييزي الخطية والصيغة الاحتمالية لنموذج الانحدار اللوجستي اذ تم التوصل الى ان التصنيف على وفق صيغة احتمالات الاستجابة لنموذج كما ان الدراسة بينت ان تصنيف البيانات بطريقة الصيغة الاحتمالية لنموذج الانحدار اللوجستي تفوقت على دالة التمييز الخطى لها وذلك بإعطائها اقل احتمال لخطأ التصنيف ايضا تبين ان استخدام تحليل الانحدار اللوجستي كنموذج تصنيف خطى للفصل بين المجموعتين من المشاهدات ذات الاستجابة الثنائية وذلك في حالة عدم تحقق الفروض الاساسية للدالة التمييزية الخطية تم الحصول على

ثانياً: التوصيات:

١-يوصي الباحث بتطبيق التصنيف البيانات وفق ثلاث نماذج الأول نموذج الدالة

الاحتمالية لدالة الانحدار اللوجستي المتعدد الاستجابة. والثاني نموذج دالة التمييز الخطي والثالث نموذج احتمالات الاستجابة لدالة التمييز الخطي

٢-يوصي الباحث باستخدام الدوال التمييزية الأخرى (التربيعية-اللوجستي الخطي-اللوجستي التربيعي) في تصنيف البيانات ذات الاستجابة المتعددة.

المراجع:

١- النداوي، سرى صباح كتيب (٢٠٠٨) **"مقارنة بعض المقدرات الحصينة في الدوال التمييزية مع تطبيق عملي"** رسالة ماجستير، كلية الإدارة والاقتصاد، جامعة بغداد.

٢- حمودات، الاء عبد الستار (٢٠٠٥) **" الدالة التميزية وطرق تحديد متغيراته"** رسالة ماجستير، كلية التجارة، جامعة المنصورة.

٣- أنوار ضياء عبد الكريم (2006) **"استخدام الطرق التميزية الإحصائية لتشخيص بعض أمراض القلب"** رسائل المنجى، هشام محمد رجب والاقتصاد، الجامعة المستنصرية.

٤- ماجستير في علوم الرياضيات، جامعة الموصل.

٥- عباس، علي خضير (2012) **" استخدام نموذج الانحدار اللوجستي التنبؤ بالدوال ذات المتغيرات الاقتصادية التابعة النوعية"** مجلة جامعة كركوك للعلوم الادارية والاقتصادية، (2)، (٢)، تكريت. العراق.

6- Al- Afifi, R.M (2010)"**The Use of Multinomial Logistic Regression Model on Physical Violence Data**" degree of Master of Applied Statistics, Al- Azhar University-Gaza.

7- Cramer, J. S. (2002) **"The Origins of Logistic Regression"** Tinbergen Institute Discussion Paper TI 2002-119/4, available at:

http://www.tinbergen.nl/discussionpapers/02119.pdf

8-Jordan, A. (2002)" **On discriminative vs . generative classifiers A comparison of logistic regression and naïve Bayes**" Advances in neural information processing systems, 14, 841.

9-Kemp, G.C.R. (2000): "**Semi-Parametric Estimation of a Logit Model**", University of Essex. http://www.econometricsociety.org/meetings/wcoo/pdf 0879.pdf