مجلة جامعة الأزهر – غزة، سلسلة العلوم الإنسانية 2018، المجلد 20، عدد خاص (A) ص 231-260

استخدام الانحدار اللوجستي لتحديد اهم عوامل الخطر على مرضى السكري في قطاع غزة

هارون موسى بهار عبدالهادي خليل أبو سعدة

جامعة الأزهر – غزة

تاريخ الاستلام 2018/10/15 تاريخ القبول 2019/1/29

ملخص:

لقد قمنا في هذه الدراسة بتحليل بيانات تتعلق بمرضى السكري، وعينة الدراسة عبارة 232 شخص مقسمين الى 172مريض سكري و60 شخص غير مريض بالسكري، حيث تم جمع هذه البيانات داخل المراكز الصحية والعيادات سواء كانت حكومية أو خاصة، وكانت العينة موزعة على الجنسين وتحديدا الأشخاص البالغين التي تزيد أعمارهم عن 25سنة. كما تم استخدام التحليل اللوجستي لبناء نموذج احصائي لتصنيف البيانات. والمعروف إن المستوى السليم للسكر في الدم بعد صوم ثماني ساعات يجب أن يكون أقلّ من 108 ملغم/ دل، بينما المستوى الحدودي هو 126 ملغم/ دل. أما إذا كان مستوى الجلوكوز في الدم لدى شخص ما 126 ملغم/ دل فما فوق، في فحصين أو أكثر، فعندئذ يتم تشخيص إصابة هذا الشخص بمريض السكري. وتأتي أهمية هذا البحث من أنه تمكن من بناء نموذج إحصائي يستطيع التنبؤ بحالات الاصابة بمرض السكري والذي من خلاله نستطيع تحديد أهم عوامل الخطر التي إذا ما توفرت لدى اي شخص كان احتمال إصابته بهذا المرض أكبر بكثير من غيره. وكانت خلاصة التحليل أننا حصلنا على نموذج بلغت دقته وفق جدول التصنيف الى (92.8%). واهم عوامل الخطر التي ظهرت في النموذج هي: الضغط النفسي المرتفع، كمية الفواكه والخضروات التي يتناولها أسبوعيا، كمية اللحم المستهلكة أسبوعا، المؤهل العلمي. وفي هذه الدراسة تم اختيار بعض أدوات التقييم الإحصائي لتقييم النموذج الذي تم الحصول علية ولمعرفة مدى جودة التصنيف التي يحققها هذا النموذج، وهذه الطرق هي (leave-one-out cross validation, classification table and ROC curve) وقد أجريت جميع التحليلات باستخدام البرمجيات R ، SPSS.

الكلمات المفتاحية: التحليل اللوجستي، جدول التصنيف، منحنى ROC، نسبة الأرجحية، مرض السكري.

Abstract:

In this study, the two researchers analyzed data related to diabetes. The study sample consisted of 232 people divided into 172 diabetes patients and 60 non-diabetic patients. These data were collected at health centers and clinics, (governmental or private). The sample was distributed among male & female, particularly adults over 25 years. The logistic regression was used to construct the statistical model for data classification.

It is known that the proper level of sugar in the blood after fasting of eight hours should be less than 108 mg /dl, while the border level is 126 mg /dl. If a person's blood glucose level is 126 mg /dl or above, in two or more tests, then, he or she is diagnosed with diabetes.

The importance of this research is that, it was able to build a statistical model can predict the incidence of diabetes, which through we can identify the most important risk factors that if they are found in any person, the ability to affected in diabetes will be much higher than others. The summary of the analysis was the two researchers obtained a model its accuracy according to the classification table reached 92.8%.

The main risk factors in the model are: high psychological stress, weekly amount of meat, fruits and vegetables consumed per a week, scientific qualification. In this study, some statistical evaluation tools were selected to evaluate the model obtained and to determine the quality of the classification achieved by this model, and these methods are (leave-one-out cross validation, classification table and ROC curve)

All the analyzes done by using the software R, SPSS.

Keywords: Logistic regression, classification table, ROC curve, odds ratio, diabetic patients

مقدمة:

الانحدار بشكل عام هو التحليل الذي يختص بدراسة اعتماد متغير واحد يعرف بالمتغير التابع على متغير أو أكثر يعرف بالمتغير المستقل أو المتغيرات المستقلة، وذلك بغرض التصنيف أو التنبؤ. وبناء على ذلك فإن أسلوب الانحدار يستخدم للتوصل إلى نموذج رياضي يصف العلاقة الكمية بين المتغير التابع والمتغيرات المستقلة. ويعتبر الانحدار الخطي Liner Regression من ابسط واهم تطبيقات الانحدار، حيث يشترط تحقيق بعض الافتراضات من أجل صحة تطبيق النموذج الخطي (Menard 2002) ومن ضمن الافتراضات ان يكون المتغير التابع متصلا وغير محدود. لكن هناك العديد من الأوضاع والظروف في علم النفس والتربية والعلوم الاجتماعية وغيرها التي يكون فيها المتغير التابع نوعيا. ففي هذه الحالة يتم استخدام طرق أخرى للانحدار وهي الانحدار اللوجستي ، والتحليل التمييزي وغيرها ، وتكمن أهمية تحليل الانحدار اللوجستي عند مقارنته بالأساليب الاحصائية الاخرى (الانحدار الخطي والتحليل التمييزي)، في أن الانحدار اللوجستي هو

أداة اكثر قوة لأنه يقدم إختبارا لمعنوية المعاملات، كما أنه يعطي الباحث فكرة عن مقدار تأثير المتغير المستقل في المتغير التابع النوعي ثنائي القيمة ، بالإضافة إلى ذلك ، فإن الانحدار اللوجستي يرتب تأثير المتغيرات المستقلة ، مما يسمح للباحث بالاستنتاج بأن متغيرا ما يعتبر أقوى من المتغير الآخر في فهم ظهور النتيجة، كما أن تحليل الانحدار اللوجستي يمكنه أن يتضمن المتغيرات المستقلة النوعية، كما أن من مزايا استخدام الانحدار اللوجستي هو أنه أقل حساسية تجاه الانحرافات عن التوزيع الطبيعي لمتغيرات الدراسة، كما أن الانحدار اللوجستي يستطيع أن يتجاوز العديد من الافتراضات المقيدة، لاستخدام طريقة المربعات الصغرى OLS في الانحدار الخطي، الأمر الذي يجعل في نهاية المطاف تحليل الانحدار اللوجستي الأسلوب الأفضل في حالة المتغير التابع الفئوي ثنائي القيمة.(www.exeter.ac.uk). والانحدار اللوجستي عدة أنواع، إلا أن أكثرها شيوعا هو تحليل الانحدار اللوجستي الثنائي Binary Logistic Regression الذي سنستخدمه في بحثنا هذا . ويوجد هناك نوع آخر من الانحدار اللوجستي وهو الانحدار اللوجستي المتعدد Multinomial Logistic Regression المستخدم في حالة كان المتغير التابع النوعي لديه أكثر من قمتين، كما يوجد نوع ثالث من الانحدار اللوجستي وهو الانحدار اللوجستي الترتيبي Ordinal Logistic Regression والذي يستخدم في الحالات التي يكون فيها المتغير التابع ترتيبي.

مشكلة الدراسة:

ازداد معدل الإصابة بمرض السكري في الآونة الأخيرة بشكل ملفت للنظر على مستوى العالم، وتعد فلسطين وتحديدا قطاع غزة أحد المناطق العربية التي تعاني من انتشار المرض المزمن بين مواطنيها بحيث أظهرت بعض الاحصائيات في قطاع غزة أن هناك ارتفاع ملحوظ بالإصابة بمرض السكري(وفقاً للمؤتمر الدولي حول اللاجئين ومرض السكري). ونظراً لانعدام الدراسات العربية التي تناولت الوقاية من السكري فإن هذه الدراسة تأتي لتحديد اهم عوامل الخطر للإصابة بمرض السكري.

منهجية الدراسة:

تم جمع البيانات من عينة عشوائية من المرضى الذين يعانون من أمراض السكري. وعينة أخرى من الأشخاص الذين لا يعانون من المرض، حيث تم جمع البيانات من خلال سجلات المرضى بالإضافي إلى استخدام الاستبانات. تم الاعتماد على نموذج الانحدار اللوجستي الثنائي كون ان المتغير التابع يتكون من فئتين فقط، وذلك بهدف تحديد أهم عوامل الخطر للاصابة بمرض السكري، ثم تم تطبيق بعض تقنيات التقييم من أجل معرفة مدى دقة النموذج الذي تم الحصول عليه، وهذه الطرق هي: (leave-one-out cross validation, classification table and ROC curve) وقد أجريت جميع التحليلات باستخدام البرمجيات R

الدراسات السابقة:

دراسة (غانم و الجاعوني :2011) يتضمن هذا البحث دراسة أهم محددات لكفاية دخل الأسرة وتحليلها من وجهة نظر إحصائية، والتعرف على أهم المحددات المعنوية المؤثرة significant باستخدام تقنية الانحدار اللوجستي ثنائي الاستجابة Regression logistic وقد شملت عينة الدراسة 161 أسرة.

دراسة (Bhar :2013) تهدف هذه الدراسة إلى اختيار أفضل نموذج إحصائي لتحديد أهم عوامل الخطر على مرضى شرايين القلب في قطاع غزة، حيث تم تحليل بيانات تتعلق بمرضي شرايين القلب في قطاع غزة ، وتم استخدام ثلاث طرق للتصنيف وهي: الانحدار اللوجستي، وتحليل التمايز، وتحليل الشبكات العصبية ، وكان أهم النتائج أن عدد مرات الزواج يقلل من احتمال الاصابة بمرض شرايين القلب، ثم ارتفاع ضغط الدم ثم التدخين ، وكانت هذه أهم العوامل التي تؤثر على مرض الشرايين.

دراسة (Okasha and Abu Samra:2013) تهدف هذه الدراسة إلى اختيار أفضل نموذج إحصائي للمرضى الذين يعانون من ارتفاع ضغط الدم في فلسطين ، من خلال المقارنة بين الشبكات العصبية وتحليل شجرة القرارات ونموذج الانحدار اللوجستي على مجموعة البيانات الحقيقية. تشير نتائج الدراسة إلى أن الشبكات العصبية الاصطناعية (ANN) قد تؤدي أداءً أفضل من الطرق التقليدية باستخدام الانحدار اللوجستي أو شجرة القرارات، خاصة في حالة البيانات غير الخطية.

دراسة(Frank & other:2001) وهي بعنوان الغذاء ونظام الحياه وخطر الإصابة بمرض السكري لدى النساء، وهدفت هذه الدراسة إلى التعرف إلى مدى ارتباط مرض السكري من النوع الثاني بالنظام الغذائي لدى أفراد الدراسة، وتوصلت الدراسة إلى أن زيادة الوزن هو أحد أهم المؤشرات في حدوث المرض بالإضافة إلى قلة التمارين الرياضية وسوء التغذية والتدخين واستخدام الكحول.

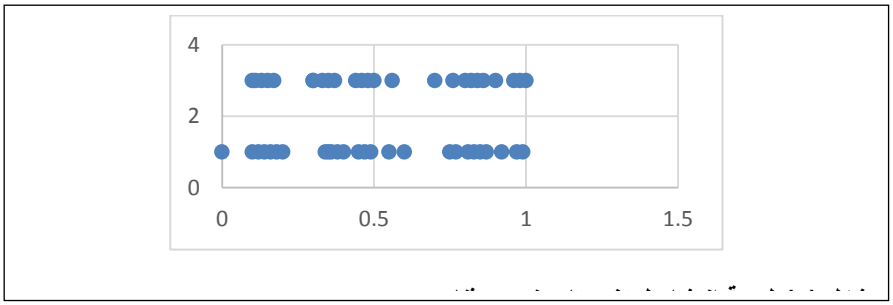
دراسة(Abo zeid & Kassb:1992) وهي بعنوان مدى انتشار مرض السكري في المجتمعات القروية في جنوب السعودية وعلاقته بالرعاية الصحية المقدمة، وهدفت هذه الدراسة الى تقدير معدلات انتشار مرض السكري والتحكم فيه، وتوصلت الدراسة إلى تشخيص (57) حاله مرضية من مجتمع مكون من 1250 فردا مع أنهم لم يشملوا كلهم بالدراسة الفعلية للتعرف على مستوى السكر في الدم لديهم، كذلك وجد فروق بين السن والجنس فيما يخص الإصابة بمرض السكري.

مفهوم الانحدار اللوجستي: يعرف الانحدار اللوجستي بشكل عام، بأنه التحليل الذي يختص بدراسة العلاقة بين متغير واحد يعرف بالمتغير التابع (في حال كان المتغير التابع نوعي مكون من فئتين

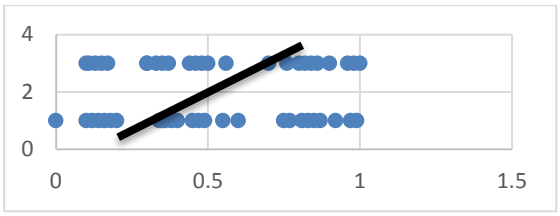
Dichotomous – في هذه الحالة يسمى الانحدار اللوجستي الثنائي) ومتغير واحد أو أكثر يعرف بالمتغير المستقل او المتغيرات المستقلة (المفسرة) وذلك بغرض التقدير، أو التنبؤ.

ويقوم نموذج الانحدار اللوجستي على فرض أساسي وهو أن المتغير التابع Y والذي نهتم بدراسته وهو متغير ثنائي يأخذ القيمة (1) باحتمال (P) والقيمة (0) باحتمال (1−P).

ونلاحظ أن لوحة الانتشار للمتغير التابع عندما يكون نوعياً ثنائياً عبارة عن مجموعتين من النقاط المتوازية كما في الشكل(1) :

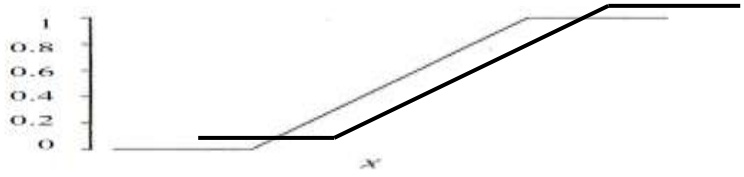


في هذه الحالة وعند القيام برسم أفضل خط مستقيم لتوفيق البيانات سيكون غير ملائم، والسبب هو ان الخط المستقيم سوف يأخذ قيمة أكبر من الواحد الصحيح وأقل من الصفر، إلا إذا كان الميل صفر.



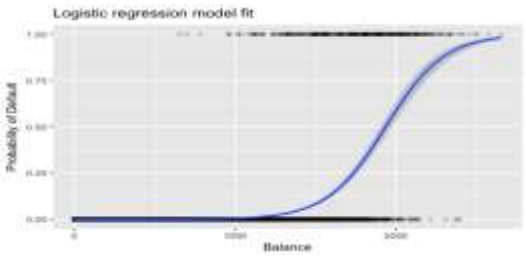
شكل (2) رسم افضل خط مستقيم لتوفيق البيانات

ويرى (Pample,2000) بأن أحد الحلول لهذه المشكلة هي: اعتماد صيغة القمة والقاع ووفقا لهذا المبدأ فان هناك حدود للقيم المتنبئ بها بحيث ألا تتجاوز الواحد الصحيح، ولا تقل عن الصفر كما في الشكل: (3)



شكل(3) اعتماد مبدأ القمة القاع لتحديد اعلى قيمة واقل قيمة

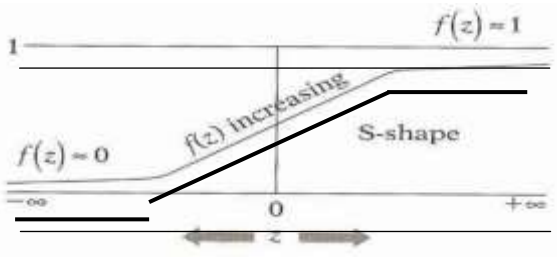
وبناء على ذلك فان توفيق البيانات في حالة المتغير التابع الثنائي لن يكون من خلال استخدام افضل خط مستقيم ولكن من خلال المنحنى اللوجستي والذي تقع قيمه بين الصفر والواحد والذي يأخذ شكل حرف ال S هو الأنسب لتوفيق هذه البيانات (Walker, 1998)، بواسطة (بابطين:1429هـ). وكذلك قال (Schmidt,2000) إن العلاقة الغير خطية الأكثر ملائمة هي المشابهة لحرف S ، بحيث تكون مستويان المنحنى محصورة بين الصفر والواحد ، وبحيث يمكن تقريبة على أنه مجموعة من الخطوط ذات الميول المخالفة وبحيث تكون الخطوط القريبة من الصفر والواحد الصحيح ذات ميول صغيرة جدا، في حين أن الخطوط في المنطقة الوسطى من المنحنى ذات ميول كبيرة، عند تمثيل تلك الخطوط بخط ناعم نحصل على منحنى على شكل حرف "S" كما في الشكل (4).



شكل(4) الحصول على منحنى على شكل حرف S

إن اعتماد الانحدار في الشكل السابق على المنحنى اللوجستي $P = \frac{e^{a+bx}}{1+e^{a+bx}}$ بدلا من الخط المستقيم $Y = a + b * x$ ، وهو الذي جعل هذا النوع من الانحدار يسمى بالانحدار

اللوجستي (Walker,:1998). كما نلاحظ أن الشكل السابق يتحدد بقيم القمة والقاع ولا يتمثل بخط مستقيم، ويفرض تحديات جديدة يجب أخذها بعين الاعتبار، فمن تلك التحديات والمشكلات الجديدة التي تظهر طبيعة العلاقة غير الخطية بين المتغير التابع والمتغيرات المستقلة، فنلاحظ في الشكل (5):*(Kleibaum & Klein , 2002,p.6)*

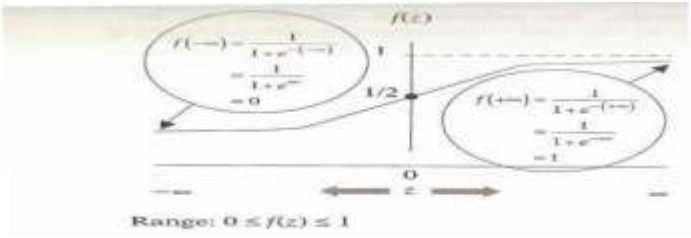


شكل(5) العلاقة غير الخطية بين المتغير التابع والمتغيرات المستقلة

إن تأثير التغير بمقدار الوحدة الواحدة في المتغير المستقل على الاحتمال المتوقع للمتغير التابع وهو ما يسمى بميل المنحنى – سوف يكون عند اقتراب المنحنى من القمة أو القاع أقل منها في حالة الجزء الأوسط من المنحنى (*Walker:1998*) وبناءا على ما سبق فان العلاقة غير الخطية بين المتغير التابع والمستقل في منطقة الوسط يمكن اعتبارها تقريبا للعلاقة الخطية لكن بدلا من أن تستمر الى اعلى او الى ادني بشكل لا نهائي فإن هذه العلاقة غير الخطية تبدأ تدريجيا في البطء كلما اقتربنا من الصفر أو الواحد الصحيح ،ونلاحظ أنه كلما اقتربت القيم أكثر فأكثر من الصفر أو الواحد ، فان الحاجة لإحداث تغير في قيمة *P* يتطلب تغيراً كبيراً في المتغير المستقل *X* مقارنة بما يحتاج اليه نفس القدر من التغير في *P* في منطقة وسط المنحنى (*Aldrich & Nelson :1984,p.32*).

فعلى سبيل المثال لإحداث تغير في احتمال حدوث حدث ما من 0.95 الى 0.96 يتطلب الامر تغيراً في المتغير المستقل أكبر بكثير مما يتطلبه تغير الاحتمال من 0.45 إلى 0.46 ويكون المبدأ ببساطه هو أن نفس الإضافة التي تحدث في المتغير المستقل يكون لها أثر أقل على المتغير التابع في المنطقة القريبة من القمة أو القاع، وللحصول على قدر من الأثر بالقرب من منطقتي القمه والقاع مقارباً للأثر الحادث في منطقة الوسط من المنحنى، فإنه لابد من حدوث تغير أكبر في قيم المتغير أو المتغيرات المستقلة (*Pample, 2000*). كما أن هناك مشكلة أخرى تنشأ من حدود

القمة والقاع عند محاولة توفيق البيانات وهي مشكلة خاصية الإضافة (*additivity*) (*Lea,1997*) فكما يرى (*CMH 2002*) بأن هناك نوعين شائعين من النماذج الإحصائية وهما: النماذج الإضافية (*additive*)، والنماذج الضربية (*multiplicative*) ففي النموذج الجمعي يفترض أن المعاملات تؤثر في المتغير التابع عن طريق تغييره من خلال الإضافة أو الطرح ، ومثال ذلك زيادة الإنتاج بمقدار معين ، لكل ساعة عمل إضافية ، أما في النموذج الضربي ، فيفترض أن المعاملات تؤثر في المتغير التابع من خلال الضرب أو القسمة، ومثال ذلك زيادة عدد السكان بشكل مضاعف كلما زاد عدد السنوات مثال آخر: وهو زيادة الأرباح بشكل مضاعف كلما زاد رأس المال، إذ أن في النماذج الجمعية تكون التغيرات التي نراها في المتغير التابع بنفس المقدار والحجم، وبغض النظر عن مستوى المتغير المستقل، أما في النماذج الضربية فإن التغيرات المشاهدة في المتغير التابع يعتمد حجمها ومقدراها على مستوى المتغير المستقل، بحيث يكون تأثير المتغير المستقل على المتغير التابع ليس ثابتا لجميع مستويات المتغير التابع ، بل يكون عالي التأثير في المستويات المتوسطة من قيم المتغير التابع ، وضئيل التأثير عند طرفي قيم المتغير التابع (*Pezzullo,2003*).



شكل(6) تأثير التغير في المتغيرات المستقلة على المتغير التابع

(Kleibaum & Klein, 2002, p.6)

إن جميع النماذج الخطية هي نماذج جمعية (*Pezzullo,2003*) . وإذا تم تمثيل الدالة الخطية جبريا على النحو التالي : $y = a + b * x$ ، فإن هذا يعني أنه لكل زيادة بمقدار وحدة واحدة في X ، سوف تزداد قيمة ال Y بمقدار b من وحدة Y ، وذلك أياً كانت مستويات X أو Y (*Pezzullo, 2002*)، لكن هذا النموذج الجمعي لا ينطبق في حال المتغيرات التابعة الثنائية، والسبب في ذلك أن المتغير التابع له حدود عليا ودنيا، وأن تأثير المتغير المستقل على المتغير التابع يقل جدا كلما اقتربت قيم المتغير التابع Y من الحد الأعلى وهو الواحد الصحيح أو من الحد الأدنى وهو الصفر، كما هو موضح من الشكل (6). أما المشكلة الإحصائية لاستخدام تحليل

انحدار المربعات الدنيا في حالة توفيق البيانات ذات المتغيرات التابعة ثنائية القيمة، فهي تكمن في انتهاك افتراضات تحليل الانحدار الخطي واهمها افتراض: اعتدالية التوزيع (*normality*) وتجانس البيانات (*homescedasticity*) ، حيث تنشأ هاتان المشكلتان بسبب الطبيعة الثنائية للمتغير التابع (*Fraas & Newman, 2003*) (*Cizek & Fitzgerald, 1999*). إن المتغير التابع Y في حالة المتغيرات التابعة ثنائية القيمة له قيمتان فقط، وبذلك فإن هناك قيمتين فقط للبواقي عند كل مستوى من مستويات X ، وإذا كان عند أي قيمة أو مستوى من مستويات x_i ، يكون الاحتمال المتوقع يساوي $b_0 + b_1 x_i$ ، فإن قيم البواقي ستكون على النحو التالي:

$1 - (b_0 + b_1 x_i)$ وذلك عندما تكون y_i تساوي الواحد الصحيح

$0 - (b_0 + b_1 x_i)$ وذلك عندما تكون y_i تساوي صفر .

حتى في المجتمع الاصلي، فإن توزيع الأخطاء لن يكون طبيعيا عند أي مستوى من مستويات X عندما يكون للتوزيع قيمتين فقط، وكما يرى (Newsom , 2003) فإن توزيع الأخطاء (البواقي) في مثل هذه الأوضاع من المحتمل، وبشكل كبير أن تتبع التوزيع اللوجستي (logistic Distribution) وليس التوزيع الطبيعي (Normal Distribution) .

معامل الترجيح Odds: أن معامل الترجيح odds هو عبارة عن طريقة للتعبير عن مدى احتمال حدوث شيء ما مقارنة باحتمال عدم حدوثة، (Pample, 2000,p.12)

الصيغة الرياضية لمعامل الترجيح يعطي فرصة أكبر لفهم العلاقة بين معامل الترجيح والاحتمالات:

$$O = \frac{p}{1-p}$$

حيث O هي معامل الترجيح ، P هي احتمال وقوع الحدث ، (1−P) هي احتمال عدم وقوع الحدث.

تقدير وتفسير معاملات الانحدار اللوجستي:

من أجل تقدير معاملات الانحدار اللوجستي يتم اللجوء إلى طريقة الاحتمال الأعظم Method Likelihood Maximum ، التي تعتبر الطريقة الأكثر ملائمة لكافة النماذج الخطية وغير الخطية ، وتعرف طريقة الاحتمال الأعظم بأنها طريقة تكرا رية Iterative تعتمد على تكرار العمليات الحسابية عدة مرات، حتى يتم الوصول إلى أفضل تقدير للمعاملات، والتي من خلالها يمكن تفسير البيانات المشاهدة . (Newsom, 2003,65) .

1) طرق تقيم النموذج :

1.1) جدول التصنيف Classification Tables: إن استخدام جدول التصنيف يعتبر احدى طرق فحص جودة مطابقة النموذج للبيانات ، وتعتمد هذه الطريقة على انشاء جدول يوضح عدد الحالات التي تمتلك الصفة المرغوب فيها أو الحالات التي لا تمتلك الصفة المرغوب فيها والتي تم تصنيفها بطريقة صحيحة أو بطريقة خاطئة (Soderstrom & Leitner, 1997) ، وتتطلب هذه

الطريقة الحصول على متغير تابع مشتق من النموذج من خلال تحديد نقطة قطع C ، ثم مقارنة الاحتمالات المتوقعة بتلك النقطة، بحيث إذا تجاوزت الاحتمالات المتوقعة نقطة القطع C أعطيت تلك الحالة تصنيفاً متوقعاً يساوي واحد، وما عدا ذلك فإن الحالة يعطى لها تصنيف متوقع يساوي الصفر، علما بأنه غالبا ما تكون نقطة القطع C تساوي 0.5 (Frass & Newman, 2003) وتعتمد فكرة استخدام هذا التحليل على أن النموذج إذا قام بتوقع تصنيف الحالات بشكل صحيح اعتماد على معيار ما ، فإن ذلك يعطي برهانا بأن النموذج يطابق البيانات المشاهدة (Ferrer & Wang, 1999) .

جدول(1) جدول التصنيف

التصنيف		التوقع		المجموع
		الموجب	السالب	
المشاهد	الموجب P	الموجب الصحيح TP	السالب الخاطئ FN	P= TP+FN
	السالب N	الموجب الخاطئ FP	السالب الصحيح TN	P'= FP+TN
المجموع		Q= TP+FP	Q'= FN+TN	N=TP+FN+FP+TN

الحساسية Sensitivity: ويرمز لها بالرمز SE ، وتعرف بأنها قيمة الاحتمال، بأن يكون التصنيف المتوقع موجباً للحالة التي تكون فعلاً موجبة وتحسب حسب المعادلة :

$$SE = \frac{TP}{TP+FN} = \frac{TP}{P}$$

الخصوصية Specificity : ويرمز له بالرمز SP، وتعرف بأنها قيمة احتمال، أو يكون التصنيف المتوقع سالبا للحالة التي تكون فعلاً سالبة، وتعطى حسب المعادلة

$$SP = \frac{TN}{FP+TN} = \frac{TN}{P'}$$

الدقة (accuracy): تعني قرب القيمة المقاسة من القيمة القياسية أو القيمة الفعلية، على سبيل المثال ، إذا كنت في المختبر تحصل على قياس للوزن يبلغ 3.2 كغ لمادة معينة ، لكن الوزن الفعلي أو المعروف هو 10 كغم ، فإن قياسك غير دقيق (not accurate). في هذه الحالة، لا يكون قياسك بالقرب من القيمة المعروفة.

ويمكن قياس الدقة (accuracy) من خلال العلاقة التالية: $\text{Accuracy} = \frac{TN+TP}{N}$

الضبط (Precision): تشير إلى قرب قياسات اثنين أو أكثر من بعضها البعض. باستخدام المثال أعلاه، إذا كنت تزن مادة معينة خمس مرات ، وتحصل على 3.2 كجم في كل مرة ،يكون القياس دقيقًا للغاية (very precise). الدقة أو الإحكام(Precision) مستقلة عن الدقة (accuracy).

يمكنك أن تكون دقيقًا جدًا (very precise) ولكن غير دقيق (accuracy). ، كما هو موضح أعلاه. يمكنك أيضًا أن تكون دقيقًا (accuracy) ولكن غير محكم (دقيق) (imprecise) .

والدقة أو الاحكام (Precision) يتم حسابها كالتالي: Precision= $\frac{TP}{TP+FP}$

يتم احتساب معدل الخطأ(error rate) لكل التصنيفات حسب الصيغة التالية : Error Rate = $\frac{FN+FP}{N}$

معدل الخطأ الموجب (The false positive rate) يتم حسابة كم خلال الصيغة: The false positive rate= $\frac{FP}{FP+TP}$

نسبة التصنيف الصحيح Hit Ratio : وتعرف بأنها قيمة احتمال التصنيف الصحيح، كما أنها تعرف أيضا بنسبة الكفاءة، واذا كانت الكفاءة Efficiency ، والتي يرمز لها بالرمز EF وتعرف بانها: $EF = \mathrm{TP} + \mathrm{TN}$

فان نسبة التصنيف الصحيح أو ما يعرف بنسبة الكفاءة تساوي

$$HitRatio = \frac{EF}{Total} = \frac{(TP+TN)}{P+P'} = \frac{(TP+TN)}{(Q+Q')}$$

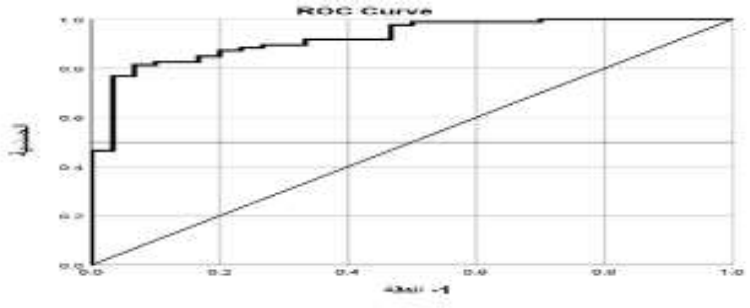
علماً بأن جميع هذه المقاييس تتطلب قاعدة للتقرير واتخاذ القرار(أو ما يسمى ب threshold) لتصنيف نتائج الاختبار إما موجباً أو سالبا (Ferrer& Wang,1999, Obuchowski,2005).

لكن كما يقول (Hosmer & Lemshow, 2000, p. 156) بأنه لسوء الحظ فان هذا الوضع وهو استخدام هذا التحليل مبرهنا ًعلى مطابقة النموذج للبيانات قد لا يعمل دائما ، حيث من السهل تصميم وضعية يكون فيها تحليل الانحدار اللوجستي صحيحاً ويطابق البيانات المشاهدة ، ومع ذلك يعطي تحليل جداول التصنيف نتائج سيئة وضعيفة ، إن سبب ذلك يعود إلى أن صحة التصنيف تعتمد على التوزيع الأصلي للمتغير التابع وحساسية التحليل لنسب أحجام مجموعتي العينة، حيث دائما ما يفضل التحليل تصنيف الحالات لصالح المجموعة الاكبر حجما، وهي حقيقة مستقلة عن وجود مطابقة النموذج، بمعنى أن دقة التصنيف أو عدمها لا يعكسان المعيار المفترض لجودة المطابقة ، وهو قرب أو بعد المسافات بين القيم المشاهدة والمتوقعة للنموذج، كما يلاحظ أنه إذا كانت هناك حالات كثيرة لها احتمالات تقترب من نقطة القطع فإن من المتوقع أن يكون مقدار سوء التصنيف كبير(Nichols, Orehovec & Ingold:1998; Soderstrom & Leitner:1998)

ويرى (Ferrer & Wang: 1999) أن إحدى مميزات هذه الطريقة أنها تمكن الباحث من مقارنة نتائج التحليل لأسلوبين إحصائيين مختلفين تماماً، كما هو الحال في مقارنة نتائج تحليل الانحدار

اللوجستي مع نتائج التحليل التمييزي، وذلك لأن كلا الأسلوبين يعيطان جدول التصنيف نفسه والذي يمكن من خلالهما إجراء المقارنة.

1.2) تحليل منحنى ROC: إن تمثيل (1-الدقة) في مقابل الحساسية لجميع نقاط القطع فإن ذلك يعطي شكلاً في غاية الأهمية في تقيم النموذج وهو ما يسمى منحنى خاصية تشغيل المستقبلReceiver Operation Characteristic والمعروف اختصاراً بمنحنى ROC كما في الشكل التالي (Westin,2005:Chatellier,1998)



شكل(7) تمثيل منحنى ROC للبيانات المتوقعة للنموذج اللوجستي

لقد بدأ استخدام منحنى ROC خلال الحرب العالمية الثانية اعتماداً على نظرية التقاط الإشارات، والتي توضح كيفية التقاط المشغل المستقبل إشارات الرادار عند وجود التشويشات، وهي ترسم احتمال التعرف على الإشارة الصحيحة (sensitivity) والإشارة الخاطئة (1- specificity) على المدى الكلي لنقاط القطع الممكنة(Fawcett,2005 & Ting,2004) ويبدأ منحنى ROC بالإحداثي(0,0) المقابل لنقطة القطع للحالات السالبة، أما الطرف الآخر فإن الإحداثي (1,1) يقابل نقطة القطع للقرار بأن جميع الحالات موجبة. أما الخط الذي يصل بين النقطتين (0,0) و(1,1) فيسمى بقطر الصدفة Chance Diagonal، وهو يمثل منحنى ROC لاختبار التصنيف الذي ليس له قدرة على التمييز بين الحالات الموجبة والحالات السالبة. ولكن عندما يكون المنحنى واقعا أعلى من قطر الصدفة فإن هذا يعني أن النموذج له قدرة تصنيفية وتميزية بين الحالات الموجبة والسالبة، وكلما ابتعد المنحنى عن قطر الصدفة، نحو الجهة اليسرى وإلى الأعلى كلما زادت قدرة النموذج على التصنيف(Obuchowski,2005). وتعطى المساحة تحت المنحنى ROC والتي تتراوح ما بين الصفر والواحد صحيح مقياساً لمدى قدرة النموذج على التصنيف بين الحالات التي تمتلك السمة موضع الدراسة، أو الفحص والحالات التي لا تمتلك هذه السمة، وهي أي المساحة تحت

المنحنى ROC تعتبر من أفضل مقاييس دقة التصنيف (Hosmer&Lemshow,2000; Bradle,1997).

1.3) تقنيات التحقق المتقاطع (Cross Validation Techniques):

يعد التحقق من المتقاطع (Cross Validation) أحد الإجراءات العامة المستخدمة في بناء النماذج الإحصائية. ويمكن استخدامه لاتخاذ قرار بشأن النموذج الإحصائي وذلك يشمل نماذج السلاسل الزمنية ونماذج الانحدار ونماذج توزيع المزيج ونماذج التمييز (Chernick-2008) .

تعد عملية التحقق المتقاطع طريقة إحصائية للتقييم والمقارنة، وذلك من خلال تقسيم البيانات إلى جزأين: يتم استخدام أحدهما لتقدير المعلمات أو تدريب النموذج والقسم الآخر يستخدم للتحقق من صحة النموذج واختباره. يتم إجراء التحقق من الصحة بطرق مختلفة.

في هذا البحث سوف نقوم بتطبيق (Leave-One-Out Cross-Validation)

التحقق من صحة التدفقات المتقاطعة (Leave-One-Out Cross-Validation) (LOOCV):

يعتبر التحقق من صحة التدفقات المتقاطعة (LOOCV) حالةً خاصة من التحقق المتقاطع k-fold حيث k يساوي عدد أجزاء البيانات، وبعبارة أخرى يتم استخدام جميع البيانات تقريبًا باستثناء جزء واحد للتدريب ويتم اختبار النموذج المقدر على هذا الجزء من البيانات المتبقي والذي لم يدخل في بناء النموذج. من المعروف أن دقة النموذج المقدر باستخدام LOOCV غير متحيزة تقريبًا، ولكنها ذات تباين عالٍ، مما يؤدي إلى تقديرات غير موثوقة (Efron، 1983). لا يزال يستخدم على نطاق واسع عندما تكون البيانات المتوفرة نادرة للغاية، خاصة في مجال المعلوماتية الحيوية حيث تتوفر عشرات البيانات فقط. يمكن تطبيق التحقق المتقاطع لملائمة النموذج عدد n من المرات، في كل مرة يتم استبعاد بعض المشاهدات وبعد ذلك يتم اختبار النموذج لتقدير معالم النموذج أو التنبؤ بالمشاهدات التي تم استبعادها في كل مرة، وهذا يوفر اختبارًا عادلاً من خلال اختبار مشاهدات لا تستخدم في ملائمة النموذج، كما أنه فعّال في استخدام البيانات لملائمة النموذج نظرًا لأنه في كل مرة يتم استخدام (n − 1) من المشاهدات في الملاءمة النموذج. نسبة النجاح (Hit ratio) هي النسبة المئوية للحالات (الأفراد ، المستجيبون ، الشركات ، إلخ) والتي تم تصنيفها بشكل صحيح حسب النموذج. يتم حسابه على أنه عدد الكائنات في قطر مصفوفة التصنيف مقسوماً على العدد الكلي (n). تُعرف نسبة النجاح أيضًا باسم النسبة المئوية للأجسام المصنفة بشكل صحيح (Hair et al. 2009).

2) التحليل الإحصائي للبيانات:

مقدمة : وقد تم اختيار مجموعة من الأشخاص الذين يراجعون العيادات الطبية سواء كانت هذه العيادات حكومية أو خاصة أو هي تابعة لوكالة الغوث وتشغيل اللاجئين ، تم الحصول على بعض

التحليلات التي أجراها الأشخاص المراجعين لهذه العيادات وتم تعيين استبانة ملحقة بهذه البيانات لنفس الأشخاص، حيث كان الاهتمام حول ما إذا كان هؤلاء الأشخاص مصابون أو غير مصابين بمرض السكري، هذا المتغير تم الاعتماد عليه على اعتبار أنه متغير تابع يتكون من فئتين هما مريض أو غير مريض، بحيث أن الفئة كانت تستهدف الأشخاص الذين كانت أعمارهم تزيد عن 25سنة (البالغين) وخضعوا لفحص، وكانت النتيجة إما مصاب أو غير مصاب بمرض السكري.

2.1) مجتمع الدراسة: الفئة المستهدفة هم المرضى الذين يراجعون العيادات الطبية أو المراكز الصحية سواء كانت هذه المراكز حكومية أو خاصة أو حتى تتبع لوكالة الغوث وتشغيل اللاجئين والذين تزيد أعمارهم عن 25سنة وذلك خلال العام 2017

2.2) العينة: تم أخذ عينة مكونة من232شخص مقسمين الى 172مريض سكري و60غير مريض بالسكري حيث تم جمع هذه البيانات داخل المراكز الصحية والعيادات سواء كانت حكومة أو خاصة وقد تم تقسيم المرضى إلى مجموعتين (مريض/غير مريض) وفقا لنتيجة الاختبار، فحص السكر في الدم، وكانت العينة موزعة على الجنسين وتحديداً الأشخاص البالغين التي تزيد أعمارهم عن 25سنة.

2.3) وصف البيانات: جدول (2) يحتوي على بعض الإحصاءات الوطنية للمتغيرات الكمية في الدراسة، حيث كان الوسط creatinine للمرضى يساوي 0.91 ولغير المرضى 0.80، ومتوسط Uric Acid للمرضى يساوي4.55 ولغير المرضى يساوي 3.81، ومتوسط كرات الدم الحمراء HGBيساوي 13.4ولغير المرضى يساوي 16.0 ، بينما متوسط مؤشر كتلة الجسم BMI للمرضى يساوي 33.12 ولغير المرضى 29.85، ومتوسط urea للمرضى يساوي34.1 ولغير المرضى 28.33، وأن متوسط العمر للمرضى 56.9 سنة ولغير المرضى 47.6 سنة، ومتوسط الوزن للمرضى يساوي 84.8 كجم ولغير المرضى 77.9 كجم، ومتوسط HDL للمرضى يساوي 55.7 ولغير المرضى 44.4، ومتوسط LDL للمرضى يساوي 117.16، ولغير المرضى 111.3، كما نلاحظ أن متوسط الطول للمرضى يساوي 162.2سم، ولغير المرضى 164.4 سم .

كما نجد أن متوسط Cholesterol للمرضى يساوي 199.6 ولغير المرضى فإن المتوسط 189.6، الجدول التالي يوضح هذه النتائج:

جدول(2) يوضح الإحصاءات الوصفية للمتغيرات الكمية للبيانات

المتغير	فئات المتغير	N	Mini	Max	Mean	Std. Deviation
Creatinine	مريض	172	0.68	3.10	0.91	0.31
	غير	60	0.71	1.07	0.80	0.09

	مريض					
	الكلي	232	0.68	3.10	0.88	0.27
Uric acid	مريض	172	2.90	12.00	4.55	1.49
	غير مريض	60	1.70	6.40	3.81	0.98
	الكلي	232	1.70	12.00	4.36	1.41
HGB	مريض	172	9.01	16.70	13.41	1.67
	غير مريض	16.00	16.0	16.00	16.00	16.00
	الكلي	232	9.01	16.70	13.23	1.76
BMI	مريض	172	16.6	73.60	33.12	10.70
	غير مريض	60	21.2	48.50	29.85	6.37
	الكلي	232	16.6	73.60	32.28	9.86
Urea	مريض	172	21	170	34.10	18.99
	غير مريض	60	21	43	28.33	5.10
	الكلي	232	21	170	32.61	16.73
HDL	مريض	172	36	185	55.69	24.44
	غير مريض	60	42	54	49.40	3.13
	الكلي	232	36	185	54.06	21.27
العمر	مريض	172	31	78	56.93	8.72
	غير مريض	60	35	72	47.57	9.53
	الكلي	232	31	78	54.51	9.82
الوزن	مريض	172	50	125	84.84	16.79
	غير مريض	60	50	95	77.90	12.12
	الكلي	232	50	125	83.04	15.98
LDL	مريض	172	1	252	117.16	39.41

	غير مريض	60	71	198	111.33	30.20
	الكلي	232	1	252	115.66	37.27
الطول	مريض	172	120	195	162.22	14.00
	غير مريض	60	135	185	164.40	11.72
	الكلي	232	120	195	162.78	13.46
Cholesterol	مريض	172	141	330	199.63	36.98
	غير مريض	60	147	261	189.60	27.90
	الكلي	232	141	330	197.03	35.08

الجدول رقم(3) يوضح بعض الإحصاءات الوصفية للمتغيرات الفئوية التي تم جمعها من أفراد الدراسة، حيث يوضح الجدول أن 44.2% من المرضى كن من الاناث و55.8% كانوا من الذكور، وأن 83.7% من المرضى كانوا يعانوا من الضغط النفسي المرتفع، وأن 16.3% لم يعانوا من الضغط النفسي المرتفع، 75.6% من المرضى لا يوجد لديهم خطة غذائية ، وأن 24.4% كان لديهم خطة غذائية، كما نجد أيضا أن 56.4% من المرضى يستهلكون اللحوم مرة في الأسبوع، وأن 43.6% من المرضى يستهلكون اللحوم أكثر من مرة خلال الأسبوع، 27.9% من المرضى يستهلكون الخضروات والفواكه مرة خلال الأسبوع وأن 72.1% يستهلكون الخضروات والفواكه أكثر من مرة خلال الأسبوع ، 91.9% من المرضى لا يعانوا من الفشل الكلوي، وأن 8.1% كانوا يعانوا من الفشل الكلوي، 91.9% من المرضى لا يعانون من ارتفاع ضغط الدم ، وأن 8.1% كانوا يعانون من ارتفاع ضغط الدم، 75.6% من المرض ليسوا مدخنين وأن 24.4% كانوا مدخنين، 84.9% من المرضى لا يحملون مؤهل جامعي و15.1% يحملون مؤهل جامعي.

2.3% من المرضى كانوا من محافظة خانيونس 1.2% كانوا من الوسطى، وأن 46.5% كانوا من محافظة غزة، وأن 50% كانوا من محافظة الشمال، والجدول التالي يوضح النتائج السابقة.

جدول(3) يوضح الإحصاءات الوصفية للمتغيرات الكمية للبيانات			
الاسم المتغير	الفئة	مريض	غير مريض
		n=172	n=60
الجنس	انثى	44.2	53.3
	ذكر	55.8	46.7

الضغط النفسي المرتفع	لا يوجد	16.3	56.7
	يوجد	83.7	43.3
يوجد لديه خطه غذائيه	لا يوجد	75.6	46.7
	يوجد	24.4	53.3
استهلاك اللحوم فى الأسبوع	مرة واحدة	56.4	66.7
	اكثر من مرة	43.6	33.3
كمية استهلاك الخضروات والفواكه اسبوعيا	مرة واحدة	27.9	3.3
	اكثر من مرة	72.1	96.7
فشل كلوي	لا يوجد	91.9	96.7
	يوجد	8.1	3.3
ارتفاع ضغط الدم	لا يوجد	91.9	96.7
	يوجد	8.1	3.3
مد خن	لا	75.6	80.0
	نعم	24.4	20.0
التعليم الجامعي	لا	84.9	60.0
	نعم	15.1	40.0
المحافظة	رفح	0	0
	خانيونس	2.3	43.3
	الوسطى	1.2	23.4
	غزة	46.5	13.3
	الشمال	50	20

في الاقسام التالية سوف نتطرق الى الانحدار اللوجيستي لإيجاد نموذج التنبوء الخاص بيانات مرضى السكري، ثم نقوم بعد ذلك بتقييم دقة النموذج الذي تم الحصول عليه باستخدام الانحدار اللوجيستي باستخدام ثلاث طرق للتقييم وهي جدول التصنيف، ومنحنى ROC وكذلك validation cross وذلك باستخدام برنامج الإحصاء(R) لتقييم دقة النموذج المتحصل عليه.

2.4 نموذج الانحدار اللوجيستي الثنائي لبيانات مرضى السكري

2.4.1 أهم المتغيرات في نموذج الانحدار اللوجيستي للبيانات

لقد تم الحصول على البيانات الخاصة بمرض السكري من خلال مسح إجرائى في قطاع غزة في العام 2017 على المراكز الصحية والطبية العاملة في قطاع غزة، وتحديداً المرضى المراجعين والذين يتلقون الخدمات الصحية في هذه المراكز، حيث تم التركيز على المرضى الذين يعانون من

مرض السكري، بمعنى تم تصنيف المرضى على أساس مريض سكري أو غير مريض سكري، وتم استخدام هذا المتغير كمتغير تابع في التحليل، حيث تم عرض مجموعة من المتغيرات المستقلة في الجدول (3.2 /3.1) وقد خضع كل متغير من هذه المتغيرات إلى عملية ترميز معينة لضمان ملائمة هذه المتغيرات للتحليل، فقد تم ترميز المتغير التابع إلى (1،0) أي (مريض /غير مريض) كما تم ترميز المتغيرات المستقلة الثنائية التفرع (1،0) ، بالإضافة إلى المتغيرات الرقمية التي تحتوي على النموذج ويعد الانحدار اللوجيستي بشكل عام طريقة لنمذجة الحالات التي فيها المتغير التابع ثنائي الاستجابة أي مكون من فئتين (0 ، 1)، في حين يمكن أن تكون المتغيرات المستقلة الداخلة في النموذج ثنائية أو أكثر من ثنائية وقد تكون أيضا متغيرات كمية ،وهذا ما تحقق في هذه البيانات الأمر الذي أجاز استخدام الانحدار اللوجستي لبناء النموذج الخاص، بتحديد أهم عوامل الخطر على مرضى السكري في قطاع غزة. حيث تم من خلال هذا الأسلوب الاحصائي بناء نموذج جيد في تحديد اهم عوامل الخطر التي إن وجدت يصاب المريض بداء السكري، واهم هذه العوامل كانت حسب المعادلة التالية:

$$\log it\left[p(Y=1)\right]=\ln\left[\frac{p(Y=1)}{1-P(Y=1)}\right]=-27.7+0.262X_1+2.945X_2-0.425X_3+0.167X_4-2.226X_5+0.705X_6+1.705X_7$$

جدول (4) المتغيرات الهامة في النموذج	
Y	المتغير التابع مرض السكري (مريض/ غير مريض)
X1	ضغط الدم المنخض
X2	الضغط النفسي المرتفع
X3	كمية استهلاك الخضروات والفواكه أسبوعيا
X4	العمر
X5	المؤهل العلمي
X6	استهلاك اللحوم في الأسبوع
X7	وجود خطة غذائية

3) عوامل الخطر على مرضى السكري:

تم استخدام نموذج الانحدار اللوجستي الثنائي، والذي تم شرحة في فقرة سابقة حيث كان هذا النموذج مناسبا لبيانات مرضى السكري على اعتبار أن المتغير التابع يتكون من فئتين وهما: (مريض / غير مريض). الجدول (5) يوضح معاملات النموذج مع الخطأ المعياري ونسب الأرجحية لعوامل الخطر على مرضى السكري.

جدول (5) معاملات النموذج المقدر والخطأ المعياري ونسبة الارجحية

	B	STD	Z-VALUE	SIG	EXP(B)
الثابت	-27.67	5.252	-5.269	0.00	0
ضغط الدم المنخفض	0.245	0.054	4.865	0.00	1.3
الضغط النفسي المرتفع	2.942	0.631	4.665	0.00	19.1
كمية استهلاك الخضروات والفواكه اسبوعيا	-0.425	0.114	-3.716	0.00	0.65
العمر	0.167	0.034	4.884	0.00	1.18
المؤهل	-2.226	0.643	-3.461	0.001	0.107
استهلاك اللحوم في الأسبوع	0.704	0.326	2.157	0.03	2.02
يوجد لديه خطة غذائية	1.705	0.591	-2.883	0.003	1.82

يوضح الجدول السابق، نتائج التحليل اللوجيستي الثنائي حيث احتوى النموذج على أهم عوامل الخطر، على مرضى السكري، كذلك يوضح الجدول السابق (Odds Ratio) لكل عامل من عوامل الخطر لمرضى السكري، حيث نلاحظ أنه كلما ابتعد ت قيمة (OR) عن الواحد الصحيح سواء بالزيادة أو النقصان كان هناك تأثير واضح لهذه المتغيرات المستقلة على المتغير التابع، حيث أظهر الجدول السابق أن متغير (الضغط النفسي المرتفع) كان له تأثير قوي جداً على مرضى السكري حيث يعتبر أهم عوامل الخطر على مرضى السكري حيث كانت قيمة (sig=0.0) وقيمة (OR=19.1)، وهذا يعني أن الشخص الذي يتعرض للضغط النفسي المرتفع، معرض للإصابة بمرض السكري 19 ضعفا مقارنة بالشخص الذي لا يعاني من هذا الضغط النفسي، في حين أن العامل الثاني كان (التعلم الجامعي) حيث كانت قيمة (sig=0.001) وقيمة(OR=0.107) وهذا يعني أن الشخص الذي يحمل مؤهلاً جامعياً معرض للإصابة بمرض السكري (9.3) مرة عن الشخص الذي لا يحمل مؤهل جامعي. في حين يوضح الجدول السابق أيضا أن العامل الثالث من حيث الأهمية هو (استهلاك اللحوم في الأسبوع) حيث كانت قيمة sig=0.000 وقيمة OR=2.02 وهذا يعني أن الشخص الذي يستهلك اللحوم أكثر من مرة في الأسبوع، معرض للإصابة ضعف الشخص الذي يستهلك اللحوم مرة واحدة خلال الأسبوع. في حين أن العامل الرابع كان (يوجد لديه خطة غذائية) حيث كانت قيمة sig=0.003 وقيمة OR=1.82 بمعنى أن الشخص الذي لا يوجد لديه خطة غذائية معرض للإصابة بمرض السكري بمقدار 1.8 عن الشخص الذي لديه خطة غذائية. في حين أن العامل الخامس كان (كمية استهلاك الخضروات والفواكه أسبوعيا) حيث نجد أن قيمته المعنوية sig=0.00، وقيمة OR= 0.65 بمعنى أن الشخص الذي يستهلك الخضروات

والفواكه مرة واحدة في الأسبوع معرض للإصابة بمرض السكري بمقدار(1.5) عن الشخص الذي يستهلك الخضروات والفواكه أكثر من مرة أسبوعياً. كذلك يوضح الجدول السابق أن ارتفاع ضغط الدم المنخفض كان العامل السادس من عوامل الخطر حيث كانت قيمة sig=0.00 قيمة OR=1.3. في حين أن العمر كان كذلك عامل الخطر السابع، والذي يؤثر على مرضى السكري حيث كانت قيمة sig=0.00وقيمة OR=1.18 وهو أقل عوامل الخطر تأثيراً على مرضى السكري. وبالتالي يمكن ترتيب عوامل الخطر كالتالي من حيث الأهمية كالتالي: (الضغط النفسي المرتفع، التعليم الجامعي، استهلاك اللحوم في الأسبوع، يوجد لديه خطة غذائية، كمية استهلاك الخضروات والفواكه أسبوعيا، ضغط الدم المنخفض، العمر).

4. جدول التصنيف:

4.1-جدول التصنيف للنموذج المبدئي:

جدول(6) التصنيف في ظل عدم وجود المتغيرات المستقلة (للنموذج المبدئي)			
	مريض	غير مريض	النسبة الصحيحة
مريض	172	0.0	0.0
غير مريض	60	0.0	100
النسبة الكلية			74.1

يوضح الجدول (6) نتائج جدول التصنيف وذلك في ظل عدم وجود أي من المتغيرات المستقلة في النموذج حيث نلاحظ من الجدول أن نسبة التصنيف الصحيح كانت 74.1%، وهذا يعطي معلومات عن النموذج المبدئي الذي لا يحتوي على أي من المتغيرات المستقلة، وفي ظل هذا التصنيف باننا سوف نحصل على نموذج انحدار مكون من الجزء الثابت فقط في النموذج والذي يساوي y=1.053 وهذا النموذج الذي يتم بناءاً عليه التصنيف والذي في المجمل حصل على تصنيف بنسبة 74.1%.

4.2- جدول التصنيف للنموذج النهائي:

يوضح الجدول التالي التصنيف للنموذج النهائي، ويمكن مقارنة النتائج الخاصة بهذا الجدول رقم (7) مع النتائج التي حصلت عليها من الجدول رقم (6) وذلك لإظهار نسبة التحسن التي تحققت نتيجة إضافة المتغيرات المستقلة للنموذج النهائي حيث أن هذا النموذج أظهر أن نسبة التصنيف الصحيح بلغت 90.5% هذا يمثل تحسناً كبيراً وواضحاً مقارنةً بجدول التصنيف المبدئي والذي أظهر نسبة 74.1%. حيث تمثل الأعمدة في جدول التصنيف (7) القيم المتوقعة للمتغير التابع، في حين تمثل الصفوف قيمة المشاهدة الفعلية للمتغير التابع حيث يتضح من الجدول أن 94.2% من الحالات تم تصنيفهم مرضى وهم مرضى، وأن 80% من الحالات تم تصنيفهم غير مرضى وهم غير مرضى، في حين كانت النسبة الإجمالية للتصنيف الصحيح حوالي 90.5% وهذا يعني أن النموذج الذي حصلنا عليه جيد في التنبؤ بحالات مرضى السكري حيث أظهر النموذج تحسن من

74.1% في النموذج المبدئي إلى 90.5% في النموذج النهائي الذي يحتوي على المتغيرات المستقلة في النموذج.

جدول رقم (7) التصنيف للنموذج النهائي الذي يحتوي على المتغيرات المستقلة

التصنيف	التوقع			النسبة المئوية للتصنيف الصحيح
	مريض	غير مريض	المجموع	
مريض P	**162** TP	**10** FP	**172** P	**94.50%**
غير مريض N	**12** FN	**48** TN	**60** P'	**80%**
المجموع	**174** TP+FN	**58** FP+TN	**232** N	**90.50%**

ومن النتائج التي حصلنا عليها في جدول التصنيف رقم (7) يمكن الحصول على الجدول التالي

جدول رقم (8) يوضح الحساسية، معدل الدقة ومعدل الخطأ وكذلك المساحة تحت المنحنى

النموذج	Sensitivity	Specifity	Accuracy	False Positive	Error Rate	precision	Area Roc
الانحدار اللوجيستي	**0.931**	**0.80**	**0.905**	**0.058**	**0.095**	**0.942**	**0.957**

يتضح من الجدول (8) السابق أن النموذج نجح في التنبؤ بشكل صحيح بمن يعانون من مرض السكري بنسبة 94.2% ، في حين أن أولئك الذين لا يعانون من مرض السكري كانت بنسبة 80% ، بينما كان المعدل الإجمالي للتنبؤ الصحيح في النموذج النهائي، (التصنيف الصحيح) بلغ نسبة 90.5% ، مما يؤكد أن النموذج حاز على دقة عالية ويمكن الاعتماد عليها في التصنيف ، بمعنى أن نموذج الانحدار اللوجيستي استطاع أن يصنف هذه الحالات بمعدل الدقة 90.5%، وهذا يدل على وجودة النموذج، كذلك نلاحظ أن معدل الخطأ في التصنيف بلغ كان بنسبة 9.5%، في حين بلغت حساسية النموذج اللوجيستي 94.2%، وهو ما يعني أنه عندما نقوم بإجراء اختبار تشخيصي للمرض عن طريق هذا النموذج الذي حصلنا عليه من خلال الانحدار اللوجيستي فإننا سوف نحصل على 94.2% من فرصة التنبؤ بشكل الصحيح عن حالة المريض بالسكري، بمعنى أن 94.2% من مرضى السكري سيتم تصنيفهم بمرضى السكري. في حين أن نسبة الغير مرضى بالسكري، وصنفوا

كذلك (غير المرضى) كانت حوالي 80%، في حين كان معدل الخطأ الإيجابي 5.8% بمعنى أن يكون غير المريض بالسكري، ويصنف أنه مريض وبلغت هذه النسبة حوالي 5.8%.

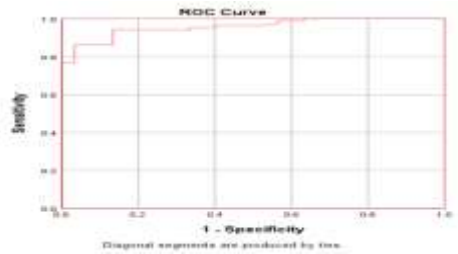
-كما يمكن ملاحظة أن المساحة تحت منحنى Roc Curve بلغت حوالي 95.7% وهي نسبة جيد جداً وتعني أن النموذج استطاع أن يصنف الحالات المرضية والغير المرضية بشكل جيد ويمكن الاعتماد عليه في التصنيف.

5-منحنى Roc Curve لتقييم النموذج:

منحنى Roc Curve هو عبارة عن تقنية لتنظيم واختيار التصنيفات بناءً على أدائها بحيث عندما يصبح التصنيف أكثر حساسية فإنه سوف يحدد نسبة أكبر من التصنيفات الإيجابية الحقيقية التي سوف نعتمد عليها في بناء النموذج اللوجيستي، وعندما يصبح النموذج أكثر دقة، أي عندما تكون عدد التصنيفات الصحيحة كبير، مقارنة بعدد التصنيفات الخاطئة، ونستخدم هذه التقنية Roc في تقييم جودة النموذج لكي نستطيع الاعتماد عليه في عملية التصنيف.

-ومن المعروف المساحة تحت المنحنى ROC تساوي الواحد الصحيح، بمعنى كلما زادت المساحة تحت المنحنى ROC والتي تتراوح ما بين الصفر والواحد صحيح كلما كان هناك قدرة أكبر للنموذج على التصنيف بين الحالات المرضية والغير مرضية.

والشكل التالي يوضح المساحة تحت منحنى ROC والذي تم بناءً على نموذج الانحدار اللوجيستي الثنائي.



الشكل (8) يوضح منحنىROC لنتائج الانحدار اللوجستي

حيث يوضح الشكل السابق أن المساحة تحت منحنى ROC للنموذج الذي تم توفيقه تساوي (95.7) عند مستوى دلالة (0.05) وهذا يعني أ ن النموذج يساعد على التنبؤ بتصنيف حالات المتغير التابع أكثر مما تفعله الصدفة.

وخلاصة ما سبق: أن نمذجة المتغيرات الخاصة بمرضى السكري وتصنيفهم إلى مرضى وغير مرضى باستخدام تحليل الانحدار اللوجيستي الثنائي أعطي نموذجاً ملائماً ومطابقاً للبيانات المشاهدة

بدرجة تفوق بكثير على الصدفة وهي 50% ، حيث يمكن الاعتماد على هذا النموذج من أجل الاستدلال الاحصائي للتعرف على احتمال أن يكون المريض مصاب أو غير مصاب بمرض السكري ومن ثم تصنيفها بناء على نتائج هذا النموذج.

6-طريقة leave-one-out cross-validation

وفقا لهذه التقنية تم توليد 1000 عينة عشوائية بدون استبدال أو إرجاع كما هو موضح لنا في الإطار النظري للدراسة، وقد أدرجت 7متغيرات في النموذج حيث كان النموذج مناسب لتصنيف ويمكن الاعتماد عليه في تصنيف الحالات المرضية، المصابة بمرضى السكري وفي النهاية تم تقييم المشاهدات المتوقعة وتم الحصول على معدل التصنيف الصحيح للمشاهدات المتوقعة وثم الحصول على جدول تصنيف النتائج وكذلك الدقة والجدول التالي يوضح التصنيف باستخدام leave-one-out cross-validation

جدول (9) التصنيف للنموذج اللوجستي باستخدام leave-one-out cross-validation

التصنيف	التوقع			النسبة المؤوية للتصنيف الصحيح
	مريض	غير مريض	المجموع	
مريض P	**711** TP	**41** FP	**752** P	**94.50%**
غير مريض N	**78** FN	**170** TN	**248** P'	**68.5**
المجموع	**789** TP+FN	**211** FP+TN	**1000** N	**88.10%**

يمكن أن نستنتج من الجدول (9) أن الدقة المقدرة لنموذج الانحدار اللوجيستي لمن يعانون من مرض السكري هو 90.1% ، في حين أن أولئك الذين لم يعانوا من مرض السكري كانت نسبة الدقة لديهم 80.6%وأن التنبؤ الصحيح للتصنيف العام والصحيح هو 88.1% ، مما يؤكد دقة التصنيف باستخدام نموذج الانحدار اللوجيستي ، كما يمكن أن نستنتج من الجدول السابق أن معدل الخطأ في التصنيف يساوي 11.9%، في حين أن حساسية النموذج اللوجيستي بلغت 90.1% وهوما يعني أنه عندما تقوم باختبار تشخيص لمرضى السكري المصابين بالمرض فإن هناك فرصة 90.1% للتنبؤ الصحيح عن حالة المريض. كما يمكن نستنتج أيضا أن الدقة الخاصة بغير المصابين

بمرضى السكري (specificity) بلغت 80.6%، وهذا يعني أن عندما نقوم بإجراء اختبار لتشخيص المرضى غير المصابين بمرض السكري فإن نسبة التنبؤ بصحة هذه الحالة بلغ 80.6% ، بينما نلاحظ أن المعدل الإيجابي the true positive ratوالذي يقيم مدى نجاح طريقة cross– validation في الكشف عن مرضى السكري المصابين بالمرض حيث بلغت هذه النسبة 90.1% والجدول التالي يوضح بعض الإحصاءات الخاصة بأداء النموذج، مثل الدقة والحساسية والتي حصلنا عليها من الجدول السابق:

جدول(10) يوضح الحساسية، معدل الدقة ومعدل الخطأ للنموذج اللوجستي باستخدام cross– validation

	sensitivity	Specificity	Accuracy	precesion	fals postive rate	error rate
LR	0.901	0.806	0.881	0.945	0.054	0.119

7–مقاييس جودة النموذج:

والآن وبعد الحصول على النموذج اللوجيستي في الفقرة السابقة، لا بد وأن نتحقق من افتراضات النموذج، ومدى ملائمة النموذج لعملية التصنيف، وبالتالي فإن إحدى الطرق للكشف عن مدى ملائمة النموذج هي من خلال اختبار likelihood–ratio–test، وبعض الاختبارات الأخرى التي من خلالها يمكن قياس مدى جودة النموذج، الذي تم الحصول عليه، حيث تقيم هذه الاختبارات حسن وملائمة النموذج الذي يصف متغير الاستجابة (التابع)، ويبين الجدول التالي نتائج اختبار (cox8snetl R square / −2logtikehood /chi square / Nagelkeke R square)

وكل هذه الاختبارات تشير إلى أنه لا يوجد دليل على عدم ملائمة النموذج، وقدرته على تصنيف فئات المتغير التابع والذي يمثل (المريض، وغير المريض) لمرضى السكري، والجدول التالي يوضح نتائج هذه الاختبارات.

جدول(11) يوضح اختبارات جودة النموذج

Step	−2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square	Chi–square	Sig.
1	106.521[a]	0.495	0.727	4.432	0.816

يوضح الجدول السابق نتائج اختبار هوسمر، وليمشو للتحقق من افتراضات النموذج ومدى ملائمة النموذج، حيث يتضح في الجدول أن قيمة اختبار (chi square) تساوي 4.432 وقيمة

sig=0.816 وهذا يعني أننا لا نستطيع رفض الفرضية الصفرية والقائلة بعدم وجود علاقة بين المتغير التابع والمتغيرات المستقلة الذي دخلت النموذج، بمعنى أن المتغيرات المستقلة كانت مهمة في النموذج، أي أن النموذج مناسب لتمثيل العلاقة بين المتغير التابع والمتغيرات المستقلة.

كما توفر قيمة Cox & Snell R Square، Nagelkerke R Square مؤشراً على مقدار التباين في المتغير التابع الذي يفسره النموذج حيث تتراوح قيمة هذين المؤشرين ما بين (0،1)وهذا يعني أن هتين القيمتين 0.495، 0.727 تشير إلى أن 49.5% و72.7%من التباين في المتغير التابع يفسر من خلال مجموعة المتغيرات المستقلة التي دخلت النموذج ، كما يوضح الجدول أن القيمة (-log=106.52) تقيس مدى قدرة النموذج على التنبؤ وكلما قلت قيمة هذا الاختبار كلما كان النموذج أكثر دقة في عملية التنبؤ.

وبشكل عام فإن المؤشرات السابق في النموذج تشير الى أن النموذج له قدرة عالية في التنبؤ بالفئة الصحيحة المرضية وغير المرضية لكل حالة.

النتائج والتوصيات:

أهم النتائج:

تبين من خلال تحليل البيانات الخاصة بمرضى السكري أن أهم عوامل الخطر على مرضى السكري كانت مرتبة حسب درجة تأثيرها على مرضى السكري وهي كالتالية مرتبة حسب الأهمية: الضغط النفسي المرتفع وحصل على المرتبة الأولى من حيث التأثير على مرضى السكري، التعلم الجامعي وحصل على المرتبة الثانية من حيث درجة التأثير، في حين كمية استهلاك اللحوم في الأسبوع حصل على المرتبة الثالثة، يليه هل يوجد خطة غذائية لدى الشخص وحصل هذا المتغير على المرتبة الرابعة، كذلك كمية استهلاك الخضروات والفواكه المستهلكة أسبوعيا حصلت على المرتبة الخامسة من ناحية التأثير على مرض السكري، في حين أن ضغط الدم المنخفض كان له أثر واضح على مرضى السكري، حيث حصل على المرتبة السادسة من حيث الأهمية، وأخيراً كان متغير العمر أحد عوامل الخطر التي تكون لها أثار على مرضى السكري. وفي المجمل استطعنا الحصول على أهم العوامل الخطر على مرضى السكري وذلك من خلال استخدام نموذج الانحدار اللوجيستي الثنائي حيث أعطى هذا النموذج نتائج جيد جداً في التنبؤ بحالة المريض المصاب بمرض السكري، كما تم استخدام ثلاث طرق إحصائية متطورة للتأكد من ملائمة هذا النموذج لتصنيف البيانات الخاصة بمرض السكري. وكانت النتائج قوية وتدعم استخدام النموذج اللوجيستي في تحليل هذه البيانات والحصول على نتائج ذات مدلولات قوية. حيث كانت المساحة تحت منحنى ROC تساوي(95.7%) وفي حين أن جدول التصنيف كانت نسبة التصنيف الصحية تساوي (90.5%)، في حين أن جدول Cross validation أعطى نسبة تصنيف صحيحة تساوي (88.1%) وهذه

المؤشرات تدعم الرأي القائل بأن الانحدار اللوجيستي جيد في تصنيف الحالات المرضية والغير مرضية، الخاصة بمرض السكري الخاصة ببيانات مرضى السكري في قطاع غزة. والنموذج النهائي الذي حصلنا عليه كان حسب المعادلة التالية.

$$\log it]p(Y=1)] = \ln\left[\frac{p(Y=1)}{1-P(Y=1)}\right] = -27.7 + 0.262X_1 + 2.945X_2 - 0.425X_3 + 0.167X_4 - 2.226X_5 + 0.705X_6 + 1.705X_7$$

Y	مرضى السكري (مريض/غير مريض) المتغير التابع
X1	ضغط الدم المنخفض
X2	الضغط النفسي المرتفع
X3	كمية استهلاك الخضروات والفواكه اسبوعيا
X4	العمر
X5	المؤهل
X6	استهلاك اللحوم في الأسبوع
X7	يوجد لديه خطة غذائية

التوصيات:

من النتائج السابقة يمكن أن نسرد بعض التوصيات الخاصة بمرضى السكري:

1. تقليل الضغط النفسي المرتفع، حيث أثبتت الدراسة أنه أحد أهم الأسباب التي تؤدي إلى مرض السكري، وقد يصنف الضغط النفسي المرتفع في المستقبل إلى السبب الرئيسي الغير مباشر للوفاة في فلسطين، نتيجة ما يترتب عليه من أعراض أو أمراض كثيرة، تؤدي في نهاية المطاف إلى الوفاة.
2. نوصي وزارة الصحة بالاهتمام بالجانب النفسي لجميع المرضى وخصوصاً مرضى السكري لما لهذا المرض من أثر فتاك وقاتل والذي قد يعتبر القاتل الرئيس في المستقبل.
3. تنظم ضغط الدم، اتخاذ كافة السبل للحيلولة دون عدم انتظامه وتحديداً ضغط الدم المنخفض لما له من أثر كبير على زيادة نسبة الإصابة بمرض السكري.
4. زيادة استهلاك الخضروات والفواكه حيث أثبتت هذه الدراسة أن الزيادة من كمية استهلاك الفواكه والخضروات تؤثر بشكل إيجابي على تقليل نسبة الإصابة بمرض السكري.
5. الحد من استهلاك اللحوم حيث أن زيادة استهلاك اللحوم يؤثر سلباً على زيادة معدل الإصابة بمرض السكري.
6. نوصي بوضع خطة غذائية مناسبة لمرضى السكري لما لذلك من أثار إيجابية على الحد من الإصابة بمرضى السكري.

7. ننصح الشباب بمتابعة دراستها الجامعية حيث أن ذلك يزيد من توعيتهم بمخاطر مرض السكري وبالتالي المحافظة على صحتهم الأمر الذي ينعكس إيجابياً على الحد من الإصابة بهذا المرض.
8. نوصي وزارة الصحة بتوفير قاعدة بيانات الكترونية جيدة لجميع المرضى والمصابين بالأمراض وكذلك المراجعين حيث يمكن استخدامها في مجال البحث، وإعطاء أهمية كبيرة لمرض السكري.
9. نوصي وزارة الصحة بتأسيس مستشفى للطبيب النفسي والاكلينيكي والتدخل المبكر لمعالجة المرضى الذين من المتوقع إصابتهم بأمراض مختلفة نتيجة الضغط النفسي لتلافي ذلك في المستقبل.
10. نوصي الباحثين بإجراء المزيد من الأبحاث حول مرض السكر للمساعدة في تحديد كل عوامل الخطورة المسببة لمرض السكر واستخدام طرق إحصائية أخرى غير الانحدار اللوجستي الثنائي وكذلك تضمين متغيرات أخرى لها علاقة بمرض السكر.

المراجع :

أولاً- المراجع العربية:

بابطين، عادل بن احمد بن حسن(1429 هـ): الانحدار اللوجستي وكيفية استخدامه في بناء نماذج التنبؤ للبيانات ذات المتغيرات التابعة الثنائية القيمة. بحيث مقدم للحصول على درجة الدكتوارة – تخصص احصاء وبحوث .

غانم، عدنان و الجاعوني، فريد خليل (2011): " استخدام تقنية الانحدار اللوجستي ثنائي الاستجابة في دراسة أهم المحددات الاقتصادية والاجتماعية لكفاية دخل الأسرة "دراسة تطبيقية على عينة عشوائية من الأسر في محافظة دمشق " ، قسم الاحصاء التطبيقي ، كلية الاقتصاد ، جامعة دمشق ، مجلة جامعة دمشق للعلوم الاقتصادية والقانونية– المجلد 27 –العدد الأول–2011

المؤتمر الدولي حول اللاجئين ومرض السكري ،(2017)، البحر الميت، الاردن https://www.unrwa.org/sites/default/files/content/resources/declaration_of_the_international_conference_on_refugees_and_diabetes_-_arabic.pdf).

ثانياً- المراجع الأجنبية:

Abu–Zeid,H.&Kassab,A.(1992). Prevalence & health care features of hypoglycemia in semiuban- rural communities in southern SaudiArabia, Journal of Diabetes Care, 15.

Aldrich, John H. & Nelson, Forrest D.(1984). Linear probability, Logit, and Probit Models. Sage University Paper series on Quantitative Applications in the Social Scienies. No.07-045. Beverly Halls, CA: Sage.

Bhar, Haroun Musa (2013). The Use of Classification Methods to Identify the Most Important Risk Factors for Cardiovascular Patients in the Gaza Strip, thesis Submitted in Partial Fulfillment of Requirements for the Degree of M.Sc. of Statistics.

Chatellier, Gilles (1998). Logistic Regression Model: Conditions Required for Stability of Prediction. Medical Informatics Department, Broussais Hospital, Paris, France.

Cizek, Gregory J & Fitzgeraid, Shawn M. (1999). Methods, Plainly Speaking: An Introduction to Logistic regression. Measurement & Evaluation in Cunseling and Development. Vol. 31, janu, 1999.

CMH(2002). The Concepts behind the Logistic Regression Model. Children's Mrecy Hospitals & Clinics. Pediatric Research. Available at: http://www.cmh.edu/stats/model/logistic-concepts.asp.

Fawcett, Tom (2005). An Introduction to ROC Analysis. Pattern Recognition Letters. No.27,2006,861-874.

Frank,B. (2001). Diet, life style & risk of the type 2 diabetes mellitus in women. The New England Journal of medicine,797-790 .(11) 345 .

Fraas, Jojan W. & Newman Isadore (2003). Ordinary Least Squares Regression, Discriminant Analysis, and Logistic Regression: Question Researchers and Practitioners Should address When Sclecting an Analytic Technique. Paper Presented at the Annual Meeting of the Education Research Association (Hilton Heal Island, GA, February 26- March 1, 2003).

Ferrer, Alvaro J. Arce & Wang Lin (1999). Comparing the Classification Accuracy among Nonparametric, Parametric Discriminant Analysis and Logistic Regression Methods. Paper Presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23,1999).

Hosmer, David W. & Lemeshow, Stanely (2000). Applied Logistic Regression. 2nd edition. New York: Johnson Wiley & Sons, Inc.

Kleibaum, David & Klein, Mitchel(2002). Logistic Regression: a Self-learning Teext. USA: Springer.

Lea, Stephen(1997). Multivariate Analysis II: Manifest Variables Analysis. Topic 4: Logistic Regression and Discriminant Analysis. University of EXETER, Department of Psychology. Revised 11th March, 1997. available at: www.exeter.ac.uk/~SEGLea/multivare2/diclogi.html.

Menard, Scott(2002). Applied Logistic Regression Analysis.2nd edition. Sage University Paper series on Quantitative Applications in the Social Sciencies. No.07-106. Beverly Halls, CA:Sage.

Newsom (2003). Data Analsis II: Logistic Regression , Fall 2003, available at: http://www.upa.pdx.edu/IOA/newsom/da2/ho_logistic.pdf.

Nichols, Jerry L.; Orehovec, Paul M.; Ingold, Scott et al (1998). Using Logistic Regression to Identify New "At-Risk" Freshmen. Journal of Marketing for Higher Education, Vol a(1) 1998. The Haworth Press, Inc. PP. 25-37.

Obuchowski, Nancy A. (2005). Fundamentals of Clinical Research for Radiologists ROC Analysis. American Roeutgen Ray Society. No.184, February 2005, 364-372.

Okasha, Mahmoud K. & Abu Samra, Ashraf I. (2013). "Classification Methods for Hypertension Patients' Data in Palestine"; Proceedings of the25th Annual International Conference on Statistics and Modeling in Humanand Social Sciences; Department of Statistics; Faculty of Economics and Political Science; Cairo University; Cairo, Egypt; March 25-28, 2013.

Pample, Fred C. (2000). Logistic Regression Aprimer. Sage University Paper series on Quantitative Applications in the Social Sciences. No.07-132. Beverly Halls, CA: Sage.

Pezzullo,John C.(2003). Logistic Regression. (Revised at 11/18/2003) Available at: http://members.aol.com/jonp71/logistic.html.

Poston, Duley L.(2004). Sociological Research: Quantitative Methods (Lecture notes, Lecture 7). Spring 2004.

Schmidt, Amy Elizabeth (2000). An Approximation of A Hierarchical Logistic Regression Model Used Establish the Predictive Validity of Scores on A Nursing Licensure Exam. Educational and Psychological Measurement, Vol.60, No.3, June2000, 463-478.

Soderstrom, Irina R. & Leitner, Dennis W. (1997). The Effects of Base Rate, Selection Ratio, Sample Size, and Reliability of Predictors on Predictive Efficiency Indicies Associated with Logistic Regression Models. Paper Presented at the Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL, October 15-18, 1997).

Walker,Marilyn D.(1998).Discriminant Function Analysis. Lesson8.

Westin, Lena Kallin (2005). Receiver Operating Characteristic (ROC) Analysis Evaluating Discriminance EffortsAmong Decision Support Systems. ISSN-0348-0542.

www.exeter.ac.uk