IRAQI Academic Scientific Journals

University of Kirkuk Journal For Administrative and Economic Science (2024) 14 (3): 261-273

UKJAES

University of Kirkuk Journal For Administrative and Economic Science

ISSN 2222-2995
University of Kirkuk Journal For Administrative and Economic Science

Anad Fahad Hussein. Comparison between logistic regression model and Vector machine to classify observations of COVID-19 patients at Al-Hussein General Hospital. ***University of Kirkuk Journal For Administrative and Economic Science*** **(2024) 14 (3):261-273.**

Comparison between logistic regression model and Vector machine to classify observations of COVID-19 patients at Al-Hussein General Hospital

Fahad Hussein Anad [1]

[1] *University of Thi Qar / Department of Studies and Planning, Thi Qar, Iraq*

fahadh@utq.edu.iq[1]

Abstract: This research studied the logistic regression model, which is one of the important methods of estimation and one of the most common models in processing binary data, as well as the support Vector machine method, which is one of the methods used to classify medical and other data and compare them with each other. The research aims to determine the best method for classifying data by comparing them using statistical classification methods (accuracy, specificity, sensitivity, error positive rate) and the correct classification rate criterion, and after applying statistical methods to Covid-19 patients, which is one of the biggest health problems that the world has faced. A random sample size of (80) people was used from Al-Hussein General Hospital in Dhi Qar, where the results showed that the Support Vector Machine (SVM) algorithm had achieved very effective performance in classifying real data, and this shows that classifying patients' cases It has relative importance to the factors studied using the support wave machine method.

Keywords: Logistic Regression, Support Vector Machine, Correction Ratio, Corona Epidemic.

مقارنة بين انموذج الانحدار اللوجستي والة الموجة المساند لتصنيف مشاهدات مرضى كوفيد-١٩ في مستشفى الحسين العام

م.م. فهد حسين عناد[١]

[١] *جامعة ذي قار/ قسم الدراسات والتخطيط*

المستخلص: تناول هذا البحث دراسة انموذج الانحدار اللوجستي الذي يعد من الطرائق المهمة في التقدير وهو من اكثر النماذج شيوعا في معالجة البيانات الثنائية، فضلا عن أسلوب الة الموجة المساند وهي من الأساليب التي تستعمل لتصنيف البيانات الطبية وغيرها ومقارنتهما مع بعضهما. اذ يهدف البحث الى تحديد افضل أسلوب للتصنيف البيانات من خلال المقارنة بينهما باستعمال طرائق التصنيف الإحصائية (الدقة ، النوعية ، الحساسية، نسبة الخطأ الإيجابي) ومعيار نسبة التصنيف الصحيح وبعد تطبيق الأساليب الإحصائية على مرضى كوفيد-١٩ الذي يعد من اكبر المشاكل الصحية التي واجها العالم اذ تم استعمال عينة عشوائية بحجم (٨٠) شخصا من مستشفى الحسين العام في ذي قار ، حيث اظهرت النتائج ان خوارزمية الة الموجة المساند (Support Vector (SVM

(Machine قد حققت اداء فعال جدا في تصنيف البيانات الحقيقية وهذا ما يبين ان تصنيف حالات المرضى له أهمية نسبية للعوامل المدروسة باستعمال أسلوب الة الموجة المساند.

الكلمات المفتاحية: الانحدار اللوجستي، الة الموجة المساند، نسبة التصحيح، وباء كورونا.

Corresponding Author: E-mail: fahadh@utq.edu.iq

المقدمة (introduction):-

يعد تصنيف البيانات من المواضيع المهمة في مختلف العلوم الاقتصادية والاجتماعية والطبية ومع التقدم التكنولوجي ادى الى امكانية الاستفادة من التصنيف الاحصائي للبيانات لذا تم استعمال الانحدار اللوجستي الذي اخذت في الاونة الاخيرة اهتمام الكثير من الباحثين لتطبيق البحوث الطبية ومعرفة العوامل الذي تؤثر في المرضى لذلك تم استعمال انموذج الانحدار اللوجستي وأسلوب الة الموجة المساند لتصنيف البيانات الخاصة بالمرضى اذ يصنف متغير الاستجابة الى صنفين (مصاب وغير مصاب) ،وان الهدف الاساسي للبحث هو عرض اسلوب الانحدار اللوجستي وأسلوب الة الموجة المساند لمرضى كوفيد ١٩ وبيان دقته بالتصنيف الصحيح ومن خلال استعمال معيار نسبة التصنيف الصحيح للمشاهدات حيث تكمن اهمية البحث في التصنيف الثنائي للبيانات وكان الاعتماد على العوامل المؤثرة في تحديد (كوفيد-١٩) ، اذ قام الباحث (H. Yusuff:2012.) بدراسة حول تحليل مرض السرطان باستعمال الانحدار اللوجستي كان يهدف البحث الى استعمال الانحدار اللوجستي كأداة لتصوير الاشعة السينية من اجل تشخيص العوامل المؤثرة في سرطان الثدي وقام الباحث (عبدالرزاق :٢٠١٦) بدراسة تطبيق انموذج الانحدار اللوجستي لمعرفة تأثير الضغط والتوترات النفسية على الاصابة بالضغط وتم استعمال طريقة الامكان الاعظم لتقدير المعلمات وتوصل من خلال استعمال اختباري (T,Wald) ان هناك تأثير على الاصابة بضغط الدم وقام الباحثان (Alrahamneh&Hawamdeh:2017) باستعمال اسلوب الانحدار اللوجستي لتصنيف المرضى المصابين والعوامل المؤثرة عليهم وتم تطبيق طريقة الامكان الاعظم لتقدير المعلمات وتوصلوا من خلال النتائج ان اسلوب الانحدار اللوجستي هو فعال ويعطي نتائج اكثر دقة في تصنيف البيانات من خلال عدة اختبارات.

المبحث الأول
منهجية البحث

اولاً: مشكلة البحث

بسبب صعوبة تصنيف المشاهدات لمرضى كوفيد-١٩ (مصابين ام غير مصابين) تكمن مشكلة البحث في ايجاد طريقة فعالة وناجحة لتصنيف المشاهدات.

ثانياً: هدف البحث

يهدف البحث الى عرض أسلوب الانحدار اللوجستي وأسلوب الة الموجة المساند لتحديد دقة التصنيف للبيانات ومن هنا تبرز أهمية الاسلوبين للحصول على تصنيف أكثر دقة وفق المعايير الإحصائية المستعملة.

ثالثاً: اهمية البحث

تكمن اهمية البحث في ايجاد طرائق جديدة اكثر دقة في تصنيف البيانات لمواكبة التطور الحاصل في تكنولوجيا المعلومات اذ تم استعمال اسلوب الة الموجة المساند بسبب دقتها العالية في تصنيف البيانات مقارنة بالطرائق الاخرى.

رابعاً: منهجية البحث

لتحقيق هدف البحث تم الاعتماد على تعريف مبسط عن وباء كورونا وذكر بعض اعراضه الرئيسية وتوضيح الانحدار اللوجستي وخصائصه وأسلوب الة الموجة المساند.

خامساً: مرض كورونا واعراضه

فيروس كورونا يعتبر فصيلة منتشرة بشكل واسع ومعروفة بأنها تسبب أمراضاً متنوعة تتراوح من نزلات البرد الشائعة إلى الاعتلالات الأشد وطأة مثل متلازمة الشرق الأوسط التنفسية (MERS)، ومتلازمة الالتهاب الرئوي الحاد الوخيم (السارس) ومرض كوفيد-١٩ هو المرض الناجم عن فيروس كورونا المُستجد المُسمى فيروس كورونا (سارس) لقد اكتشفت منظمة الصحة العالمية هذا الفيروس المُستجد لأول مرة في (٣١) كانون الأول/ ديسمبر (٢٠١٩)، بعد الحصول على مجموعة من حالات الالتهاب الرئوي الفيروسي في (يوهان بجمهورية الصين الشعبية) ، وسبب هذا الفايروس عدة اعراض (الحمى والسعال الجاف والاجهاد) وتوجد اعراض اخرى اقل شيوعا التي تصيب المرضى (فقدان الذوق ، احتقان الأنف، احمرار العينين، ألم في الحلق، صداع في الرأس، آلام في المفاصل، الطفح الجلدي، ، الإسهال، الرعشة في الجسم) وهنالك بعض الأعراض خفيفة، يصاب بها بعض الأشخاص بالعدوى ولكن لا تظهر عليهم إلا أعراض خفيفة أو لا تظهر عليهم. هنالك عدة إشارات تدل على ان الشخص مصاب بمرض كورونا مثل الاختناق وفقدان الشهية ودرجة حرارة الشخص تصل الى اكثر من (٣٩) ويجب على جميع الأشخاص الذين يصابون بحمى او فقدان الشهية او الاختناق يتوجهون للفحص على الفور لتجنب المضاعفات الناتجة عن المرض. (World Health Organization) (WHO).

المبحث الثاني
الطرائق المستعملة للتصنيف

اولاً: اسلوب الانحدار اللوجستي Method Logistic Regression

يعد أسلوب الانحدار اللوجستي من الطرائق الاحصائية المهمة في تصنيف البيانات ويعتبر حالة خاصة من أساليب نماذج الانحدار الخطية، ويمكن تسميته بانموذج لوجت (logit model)، وهو بتحليل العلاقات بين المتغيرات المستقلة (التوضيحية) والمتغيرات التابعة (الاستجابة) ويمكن استعماله في عدة مجالات منها الطبية والعلوم الاجتماعية والإنسانية وغيرها من العلوم.

١- الافتراضات الأساسية للانحدار اللوجستي [2] Basic Assumptions of logistic regression model

- لا يفترض الانحدار اللوجستي علاقة خطية بين المتغيرات المستقلة والمتغيرات التابعة.
- لا يتبع الخطأ توزيعاً طبيعياً ويكون التباين غير متجانس.
- يفترض ان تكون المتغيرات المستقلة والمتغيرات التابعة ثنائية مثلا تصنيف الاشخاص مصابين او غير مصابين بمرض معين.
- لا يفترض الانحدار اللوجستي مساواة التباين في كل فئة ولا يتطلب ان تكون المتغيرات المستقلة والتابعة من النوع المستمر ولا تتوزع توزيعا طبيعيا وهذا ما يميز الانحدار اللوجستي عن بقية نماذج الانحدار.
- يتطلب ان تكون كل مفردة تنتمي الى فئة واحدة ومحددة من فئات الدراسة.
- الانحدار اللوجستي يتطلب حجم عينة كبير نسبيا وتكون أكبر من انموذج الانحدار الخطي لان يتم تقدير معلمات الانحدار اللوجستي باستعمال طريقة الامكان الاعظم وهي من الطرائق تحتاج الى حجم عينة كبير نسبياً

٢- انموذج الانحدار اللوجستي الثنائي

يعرف انموذج الانحدار اللوجستي على انه احد نماذج الانحدار اللاخطية الذي تكون فيه العلاقة بين المتغير التابع y (متغير الاستجابة) والمتغيرات التوضيحية $(x_1,x_2,x_3,\ldots x_k)$ غير خطية ، وان الانحدار اللوجستي يبنى على فرض اساسي وهو ان المتغير التابع (y) ثنائي الاستجابة يأخذ أحدى القيمتين (0,1) اما نجاح او فشل اي حدوث الاستجابة (Success) باحتمال p_i او فشل عدم حدوث الاستجابة $Failure$ باحتمال $(1 - p_i)$ لذلك فأن المتغيرالتابع Y_i يتوزع توزيع برنولي [7],[20].

اي ان :-

$$Y_i \sim Ber p_i) \quad i = 1,2,\ldots..n$$

وان دالة الكثافة الاحتمالية تكون بالصيغة الاتية: -

$$P_r(Y=Y_i) = p_i^{Yi}\,(1-p_i)^{1-Yi} \qquad (1)$$

$$Y_i = 0,1$$

اذ ان: -

Y_i : متغير تابع ثنائي الاستجابة

p_i : احتمال حدوث الاستجابة عندما $Y_i = 1$

$1 - p_i$: احتمال عدم حدوث الاستجابة عندما $Y_i = 0$

لذلك فأن توقع المتغير Y_i يشير الى احتمالية حدوث الاستجابة p_i

$$E(Y) = P_r(Y = 1) = p_i \qquad (2)$$

اما بالنسبة لتباين المتغير Y_i حسب توزيع برنولي

$$V(Y_i) = p_i(1 - p_i) \qquad (3)$$

لتكن $(x_1.x_2.x_3.\ldots\ldots x_k)$ تمثل مجموعة من المتغيرات المستقلة (التوضيحية) و(n) تمثل مشاهدات المتغيرات والتي تكون عبارة عن المصفوفة:-

$$X = (x_{ij}) * n * k \qquad (4)$$

حيث ان:

X: مصفوفة المتغيرات المستقلة

n : عدد المشاهدات (حجم العينة) $\quad i = 1,2,3,4,\ldots\ldots,n$

k : عدد المتغيرات المستقلة $\quad j = 1,2,3,4,\ldots\ldots,k$

$y_i = [y_1.y_2.y_3 \ldots\ldots.y_n]$ عبارة عن عينة عشوائية من المتغير ثنائي الاستجابة وان $y_i \in \{0.1\}$.

وذلك من الممكن التعبير عن انموذج الانحدار اللوجستي بالمعادلة الاتية:

$$y_i = p + \varepsilon i \qquad (5)$$

لذا يمكن تمثيل دالة الانحدار اللوجستي μ بالشكل الاتي :

$$\mu_i = p(y = 1) = \frac{e^{xi'\beta}}{1+e^{xi'\beta}} \qquad (6)$$

β : متجه المعلمات ابعاده (P× 1).
$x_i = \{x_{i1}. x_{i2} \dots x_{ip}\}$متجه صفي من المتغيرات المستقله ابعاده (1×p)
εi:تمثل الخطأ العشوائي
P: عدد المعالم
$= y_i - p\varepsilon_i$
وان حد الخطأ له متوسط يساوي صفر وتباين مساوياً الى تباين المتغير التابع وكالاتي: -

$$E(\varepsilon_i) = E(y_i) - E(p) = p_i - p_i = 0 \quad (7)$$
$$V(\varepsilon_i) = V(y_i) = p_i(1 - p_i) \quad (8)$$

يتم تحويل هذا الانموذج الى الشكل الخطي يتمثل بعلاقة خطية من خلال الموجه الصفي $(\acute{X}_\iota)$ من المتغيرات التوضيحية مع لوجت الاحتمال $[logit\ p(Xi)]$ وحسب الصيغة الرياضية التالية: -

$$Z = logit\ p(\mathrm{X_i}) = ln\left[\frac{p(x_i)}{1-p(x_i)}\right] = \beta_O + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} \quad (9)$$

$$= [1\ X_{i1} \dots X_{ip}] \begin{bmatrix} \beta_O \\ \beta_1 \\ 0 \\ 0 \\ 0 \\ \beta_p \end{bmatrix} = X_\iota' \beta$$

$$\therefore Zi = X_\iota' \beta + \varepsilon i$$

لتقدير المعلمات $\beta_0. \beta_1. \beta_2. \dots . \beta_p$ نستعمل احدى طرائق التقدير طريقة الامكان الاعظم (Maximum Likelihood Estimation) [11]
عندما تكون لدينا عينة من مجموعة مشاهدات مستقلة بحجم (n) لكل زوج ($\mathrm{X_i}$,$\mathrm{Y_i}$) (i=1,2,……,n)
اذ ان
Yi: تمثل رتبة متغير الاستجابة الثنائي للمفردة i
Xi: تمثل قيمة المتغير المستقل للمفردة i
اذ ان

$(Yi = 1 \backslash X) = f(X)^{Yi}$ P عندما تكون Yi=1
$(Yi = 0 \backslash X) = [1 - f(X)]^{1-Yi}$ P عندما تكون Yi=0

ويمكن ان نعبر عن دالة الامكان الاعظم بالصيغة الاتية

$$L(B) = \prod_{i=1}^{n} f(X)^{Yi} [1 - f(X)]^{1-Yi} \quad (10)$$

بعد اخذ اللوغارتيم للطرفين تصبح المعادلة بالشكل الاتي:

$$L(B) = \ln L(B) = \sum_{i=1}^{n} \{Yi \ln[f(X)] + (1 - Yi)\ \ln[1 - f(X)]\} \quad (11)$$

حيث يتم اشتقاق المعادلة اعلاه للمعلمات المراد تقديرها (Bi) ونجعلها تساوي صفر بذلك نحصل على عدد من المعادلات التي لا يمكن حلها الا من خلال خوارزمية تكرارية يطلق عليها خوارزمية المربعات الصغرى الموزونة التكرارية.

ثانياً: الة الموجة المساند (Support Vector Machine) (SVM)

تعد هذه الطريقة احدى خوارزميات تعليم الالة لطريقة الموجه (Supervised) تستخدم للتصنيف تم اقتراحها على اساس نظرية التعليم الاحصائي (Statistical Learning theory)[12] وهي الفكرة التي قدمها لأول مرة الباحث (Vladimir N.Vapnik) في عام (١٩٦٣) وفي عام(1995) تم تطويرها، وتقنية آلة الموجه المساند (SVM) ابتكرت اصلا لحل المسائل المتعلقة بتمييز الانماط بواسطة تحديد المستوى الفاصل او الحل الامثل (Hyper plane) للبيانات المطلوب فصلها ويعرف المستوى الفاصل (Hyper plane) بانه عبارة عن السطح الذي يفصل العناصر أو البيانات الى صنفين حسب نوع بيانات المسافة بين (Hyper plane) واقرب عنصر من اي صنف يسمى الهامش (Margin) وقد طبقت في الكثير من البحوث في مجالات مختلفة واصبحت من الطرائق المهمة في التصنيف وتعرف بآلة الموجه المساند للتصنيف (Support Vector Machine For Classification) وكذلك تم استعمالها لتقدير معالم انموذج الانحدار والتنبؤ وسميت بآلة الموجة المساند للانحدار (Support Vector Machine For Regression) والهدف الاساسي لمشكلة التصنيف هو العثور على مصنف يمكنه

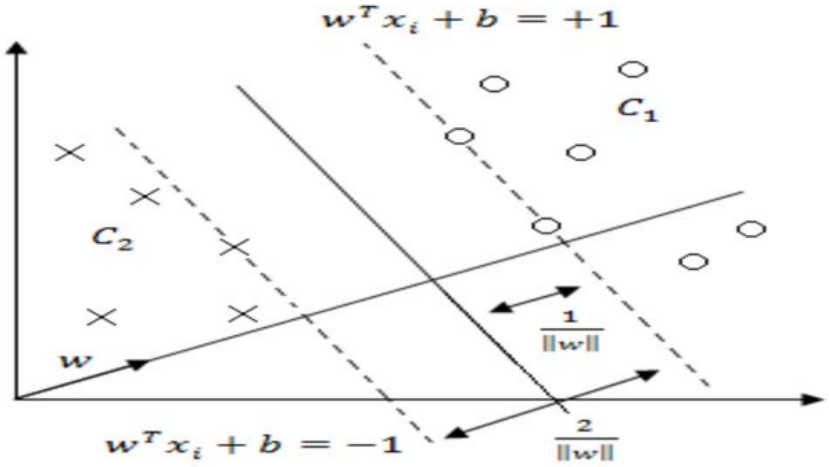
ذلك من تسمية بيانات جديدة غير مرئية او غير واضحة بشكل صحيح وهذا يتحقق عند وجود بيانات محددة وتعمل آلة الموجة المساند (Support Vector Machine) (SVM) على جعل المستوى الفاصل (Hyper plane) ابعد ما يمكن عن نقاط الصنفين.

إن الفكرة الاساسية لطريقة (SVM) تقوم بتصنيف البيانات الى ابعاد، اي اكثر من فاصل خطي واحد لأنها مصممه في الاساس لفصل صنفين عن بعضهما البعض ولكن يمكن تعميمها على مشاكل متعدد المتغيرات .وهم ما يميز طريقة (SVM) السرعة في تحليل النتائج والدقة بالتصنيف لذلك أعطت الأفضلية في كثير من المشاكل مثل مشاكل تشخيص امراض سرطان الثدي وغيرها من المشاكل الطبية والاجتماعية التي يمكن دراستها وكذلك لها القدرة على معالجة مشاكل البيانات الكبيرة ومشاكل التنبؤ وابسط مشاكل التنبؤ هو التصنيف الثنائي أي هو محاولة تميز الأشياء.

١ـ تطبيق آلة الموجة المساند Support Vector Machine (SVM)

تستعمل خوارزمية (SVM) في التصنيف الثنائي والبحث عن الحل الاولي الأمثل ويتم ذلك باستعمال الصيغ الرياضية الخاصة بخوارزمية (SVM) وهذه الخوارزمية تعمل مع جميع البيانات سوى كانت كبير او صغيرة الحجم الا انه البيانات الكبير تحتاج الى عمليات حسابية معقدة. ان الهدف الاساسي هو ايجاد أفضل مستوي يفصل اعلى مقدار من الهوامش (Hyper plane) Maximum Margin بين مجموعة نقاط. اذا تم افتراض لدينا n من النقاط وان لككل قيمة x_i يمكن ادخال لها العديد من الابعاد (d-Dimensionality) اما قيم (y_i) تكون تساوي (موجب واحد او سالب واحد) والبيانات المدربة تكون وفق الاتي:

$z_i = +1$ عن مجموعة $z_i = -1$ اي ان(y_i) هو تصنيف المصنف [+1,-1] بحيث تكون مجموعة النقاط الاقرب الى الحل الامثل او المستوى الفاصل (Hyper plane) تكون قيم النقاط اما (+١) يمثل المجموعة الاولى او (-١) يمثل المجموعة الثانية وكما مبين في الشكل (١):



شكل (١): يوضح عملية الخط الفاصل بين المجموعتين

ويمكن توضيح السطح الفاصل بالصيغ الاتية:

$$w.x_i + b = 1 \quad (12)$$

$$w.x_i + b = -1 \quad (١٣)$$

حيث ان:

$w.x_i + b = 1$: تعني المستوى الثانوي الاول والذي تقع عليه (SVM).

$w.x_i + b = -1$: تعني المستوى الثانوي الثاني والذي تقع عليه (SVM).

$\frac{2}{\|w\|}$: يمثل المسافة العمودية من (Hyper plane).

w: تمثل متجه الاوزان ويكون القياسي (Normalization) للحل الامثل.

b: حد التحيز (Bias).

x: تمثل مصفوفة المتغيرات التوضيحية ذات الابعاد $n * d$.

تحقق خوارزمية SVM توقع افضل عن طريق التعظيم واذا تم افتراض ان البيانات يمكن فصلها خطيا ويكون المستوى الفاصل (Hyper plane) بين السطحين هو($\frac{2}{\|w\|}$) ومن المفترض ان لا تكون النقاط داخل الهامش (Marginal) لذا تم اضافة الشرطين الآتيين لتصنيف البيانات [22].

$$w.x_i + b \geq 1 \quad for \ z_i = +1 \quad (١٤)$$

$$w.x_i + b \leq 1 \quad for \ z_i = -1 \quad (١٥)$$

حيث ان: $i = 1,2, \ldots \ldots . L$

بعد عملية دمج المعادلتين اعلاه نحصل على الشرط الاتي:

$$z_i(w.x_i + b) \geq 1 \quad \text{for } i = 1,2,\dots\dots L \qquad (١٦)$$
$$z_i(w.x_i + b) - 1 \geq 0 \quad \text{for } i = 1,2,\dots\dots L \qquad (١٧)$$

وبناءً على ذلك يمكن تعريف المسافة بين (SVM) و (Hyper plane) بانها هي هامش الة الموجة المساند (SVM Margin) التي تجعل المستوى الفاصل (Hyper plane) يبتعد عن الة الموجة المساند. لذلك نحتاج الى تعظيم الهامش (Margin) وتمثل المسافة بين المستويين الثانويين في المعادلة (١٦) و(١٧) وتكون مساوية الى $\frac{1}{2}\|w\|^2$ وبما ان الغاية من المستوى الفاصل هو ايجاد قيم w, b يجب تصغير الاوزان وفق المعادلة الاتية:

$$\min \frac{1}{2}\|w\|^2 \qquad (١٨)$$
S.t
$$z_i(w, x_i + b) \geq 1 \quad \text{for } i = 1\ 2\ \dots,,\ L \qquad (١٩)$$

ولكي نتمكن من تحقيق قيود هذه الدالة يمكن استعمال مضاعف لاكرانج (Lagrange Multipliers α) وبذلك يمكن كتابة معادلة الخط الفاصل (Hyper plane) وفق المعادلة الاتية [9],[17]

$$L(W, b, \propto) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{L} \alpha_i[z_i(w.x_i + b) - 1] \qquad (٢٠)$$
$$L(w, b, \propto) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{L} \alpha_i z_i w x_i + b\sum_{i=1}^{L} \alpha_i z_i + \sum_{i=1}^{L} \alpha_i \qquad (٢١)$$

نحصل على قيمة w , b من خلال مشتقة الصيغة $L(w, b, \propto)$ وذلك بجعل المشتقة الأولى تساوي صفر ثم نحصل على قيم المتغيرات المطلوبة [5].

$$\frac{dL(w,\propto,b)}{dw} = w - \sum_{i=1}^{L} \alpha_i . z_i . x_i \qquad (٢٢)$$
$$\frac{dL(w,\propto,b)}{dw} = 0$$
$$w - \sum_{i=1}^{L} \alpha_i . z_i . x_i = 0$$
$$w = \sum_{i=1}^{L} \alpha_i . z_i . x_i \qquad (٢٣)$$

إذ ان:

α_i: مضاعف الاكرانج (Lagrange Multipliers): اذ يتم افتراضه بين القيمتين (٠ و c) حيث ان (c) يتم افتراضها بالخوارزمية.

$$0 \leq \alpha_i \leq C$$

C: هو عبارة عن الموازنة بين مقدار التباعد وخطأ التقدير عن المحور الرئيسي.

نجد المشتقة بالنسبة b

$$\frac{dL(w,\propto,b)}{d(b)} = \sum_{i=1}^{L} \alpha_i . y_i \qquad (٢٤)$$

نجعل المشتقة بالنسبة الى b تساوي صفر

$$\frac{dL(w,b)}{d(b)} = 0$$
$$\sum_{i=1}^{L} \alpha_i . y_i = 0 \qquad (٢٥)$$

اي نقطة تحقق المعادلة (٢٥) سوف تكون الة الموجة المساند ونرمز لها بالرمز x وسوف تكتب بالشكل الاتي

$$z(w.x + b) = 1 \qquad (٢٦)$$

وبالتعويض عن صيغة w من خلال الصيغة (٢٢) نحصل على

$$z\sum_{i=1}(\propto_i . z_i . x_i . x + b) = 1 \qquad (27)$$

نضرب الطرفين في zتصبح المعادلة بعد التبسيط بالشكل الاتي

$$z^2\sum_{i=1}(\propto_i . z_i . x_i . x + b) = y$$

اذ ان z^2 تساوي واحد لان z تمثل فقط قيمة المساند والذي قيمتهُ تكون 1, 1- وعند تربيع قيمة zتساوي واحد اذ ان

$$(w.x_i - b) = z$$
$$b = z - \sum_{i=1}(\propto_i . z_i . x_i . x_i) . x_i \qquad (٢٨)$$

وبالقسمة على العدد الكلي للمتجهات نحصل على قيمة b [9]

$$b = \frac{1}{N}\sum_{i=1} z - (\propto_i . y_i . x_i . x) . x_i \qquad (29)$$

وبهذا يمكن تصنيف اي نقطة لأي مجموعة من مجموعات التصنيف وفق الصيغة الاتية:

$$z_i = (w.x_i + b) \qquad (30)$$

ثالثاً: معيار جدول التصنيف Classification Table Criterion (CTC)

يبين جدول التالي تصنيف عدد الحالات التي تم تصنيفها بالشكل الصحيح وعدد الحالات التي تم تصنيفها بالشكل الخطأ كما موضح في الجدول (1) بالشكل الاتي:

جدول (١): يمثل التصنيف

Classification	المتوقع			SUM
	P Pos.	N Neg.		
المشاهدة	TP	FN	P	TP+FN
	FP	TN	N	TN+FN
SUM	TP+FP	FN+TN		FN+TN+TP+FP

يمكن حساب المعايير بالاعتماد على المؤشرات التالية:
١- True Positive (TP) تشخيص العينة غير مصابة بصورة صحيحة.
٢-True Negative (TN) تشخيص العينة مصابة بصورة صحيحة.
٣- False Positive (FP) تشخيص العينة غير مصابة بصورة غير صحيحة.
٤- False Negative (FN) تشخيص العينة مصابة بصورة غير صحيحة.
بالاعتماد على المؤشرات اعلاه يمكن حساب مجموعة معايير منها:

١- الدقة (Accuracz): هي عبارة عن نسبة بين النتائج االتي يتم تشخيصها بصورة صحيحة (Positive, Negative) الى النتائج الكلية ز

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} * 100\% \quad (31)$$

٢- النوعية (Specificitz): تمثل النسبة بين العينات التي تم تشخيصها مصابة بصورة صحيحة (True Negative) الى العينات الكلية التي تم تشخيصها مصابة بصورة صحيحة مع العينات التي تم تشخيصها بصورة خاطئة.

$$\text{SPE} = \frac{\text{TP}}{\text{TN+FP}} * 100\% \quad (32)$$

٣- الحساسية (Sensitivitz): تمثل النسبة بين العينات التي تم تشخيصها غير مصابة بصورة صحيحة الى كل العينات المشخصة غير مصابة (بصورة صحيحة وخاطئة)

$$\text{SEN} = \frac{\text{TP}}{\text{TP+FN}} * 100\% \quad (33)$$

٤- نسبة الخطأ الايجابي (False Positive rate): تمثل النسبة بين العينات التي تم تشخيصها غير مصابة بصورة خاطئة الى جميع العينات التي تم تشخيصه بصورة سليمة (بصورة صحيحة (TN) او بصورة خاطئة (FP))

$$\text{FPR} = \frac{\text{FP}}{\text{TN+FP}} * 100\% \quad (34)$$

رابعاً: قانون نسبة التصنيف الصحيح

يمكن حسابة وفق الصيغة الاتية

$$PC = \frac{TP+TN}{TP+TN+FN+FP} * 100 \quad (35)$$

المبحث الثالث
الجانب التطبيقي

يتضمن الجانب التطبيقي تصنيف البيانات باستعمال الانحدار اللوجستي الثنائي والة الموجة المساند وتم استعمال معيار نسبة التصنيف الصحيح لبيان كفاءة ودقة الطريقة المستعملة واستند هذا البحث على بيانات تخص وباء كوفيد-١٩. تم جمع البيانات من مستشفى الحسين العام في ذي قار لكونها من الجوانب المهمة في الجانب الصحي حيث شمل البحث على عينة حجمها (٨٠) مريضا وتوزعوا المرضى الى نوعين مصاب وغير مصاب وكان عدد المصابين (٣٣) اما غير المصابين (٤٧) وتم استعمال ست متغيرات حيث ان يمكن تعريفها بالشكل الاتي:

X_1: الجنس، X_2: التحصيل الدراسي، X_3: المهنة ، X_4: العمر، X_5: الوزن وX_6: التدخين) والمتغير التابع يأخذ (مصاب وغير مصاب) .

و تم تطبيق هذه الطرائق بعد كتابة البرنامج بلغة بايثون حيث اجريت التجربة على عينات كوفيد-١٩ وكانت على نوعين: النوع الاول يمثل العينات المصابة والنوع الثاني يمثل العينات (غير المصابة) اذ تم استعمال انموذج الانحدار اللوجستي وخوارزمية (SVM) بتقسيم العينات الى طورين: طور التدريب باستعمال عدد مختلط من النوعين (مصاب وغير مصاب) لغرض التدريب،اما الطور الثاني فقد تم استعماله لغرض اختبار دقة التصنيف وتم حساب معايير دقة المصنف وفق معايير اعلاه.

حيث تم تطبيق المؤشرات وكانت نتائج الانحدار اللوجستي كما موضح في الجدول (٢).

جدول (٢): نتائج الانحدار اللوجستي

Classification	المتوقع		SUM
	P Pos.	N Neg.	
المشاهدة P	29 (TP)	4 (FN)	TP+FN
N	17 (FP)	30 (TN)	TN+FN
SUM	TP+FP	FN+TN	FN+TN+TP+FP

يوضح الجدول اعلاه المستخلص للتصنيف وفق انموذج الانحدار الثنائي للمصابين وغير المصابين وتم تصنيف (29) مريض من مجموع (33) من المصابين اي ان نسبة التمييز الصحيح (88%) وتكون نسبة الاشخاص الذي تم تصنيفهم بالشكل الصحيح هو (88%) من الاشخاص الذي تم تصنيفهم بالشكل الخاطئ هو (12%) وهي نسبة مرتفعة قليلاً للمصابين فعلاً بأن تم تصنيفهم بأنهم غير مصابين.

اما بالنسبة لغير المصابين فالعدد الفعلي هو (47) فقد تم تصنيف (30) بشكل صحيح اي ان نسبة التصنيف الصحيح (64%) وبخصوص ال (17) فقد تم تصنيفهم مصابين فتكون نسبتهم (36%) اما نسبة التصنيف الكلية لكلا النوعين وفق اسلوب الانحدار اللوجستي هي (74%).

١- نسبة التصنيف الصحيح للمصابين

$$\frac{29}{33} * 100\% = 88\%$$

٢- نسبة التصنيف غير المصابين

$$\frac{30}{47} * 100\% = 64\%$$

٣- نسبة التصنيف الكلية

$$\frac{29 + 30}{80} * 100\% = 74\%$$

وتم حساب مواصفات الاداء للمصنف للانحدار اللوجستي حسب المعايير التي تم ذكرها اعلاه وكما موضح في الجدول (٣)

جدول (٣): معايير انموذج الانحدار اللوجستي

Acc	SPE	SEN	FPR
74%	61%	88%	36%

واعتمادا على هذه المعايير والمواصفات تبين ان طريقة (SVM) قامت بعزل المشاهدات المصابين وغير المصابين وهذا يؤشر تصنيف المشاهدات باستعمال الانحدار اللوجستي يعطي دقة اقل.

ويمكن حسب النسب للمصابين وغير المصابين وفق القوانين الاتية:

1-True Positive Rate

$$TPR = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100\% = 88\%$$

2- True Negative Rate

$$TNR = \frac{\text{TN}}{\text{TN} + \text{FP}} * 100\% = 64\%$$

3- False Positive Rate

$$FPR = \frac{\text{FP}}{\text{TN} + \text{FP}} * 100\% = 36\%$$

4- False Negative Rate

$$FPR = \frac{\text{FN}}{\text{TP} + \text{FN}} * 100\% = 12\%$$

واما جدول (٤) يوضح احصاءات $R^2_{cox\&snel}$ و $R^2_{Nagelkerke}$

جدول (٤): Model Summary

Step	-2 Log likelihood	$R^2_{cox\&snel}$	$R^2_{Nagelkerke}$
1	373.563	.367	.593

يبين جدول (٤) احصاءة ($R^2_{cox\&snel}$)و($R^2_{Nagelkerke}$)اذا كانت قيمة ($R^2_{cox\&snel}$) تساوي (0.367) هذا يبين ان (36.7%) التغيرات في المتغير التابع يتم تفسيره من خلال الانموذج اما ($R^2_{Nagelkerke}$) قيمته تساوي (0.593) اي ان (59.3%) من التغيرات في المتغير المعتمد يتم تفسيره من خلال اسلوب الانحدار اللوجستي.

وجدول (٥) يبين كفاءة الانموذج

جدول (٥): Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	181.377	6	.000
	Block	181.377	6	.000
	Model	181.377	6	.000

يوضح جدول (٥) دقة وكفاءة الانموذج باستخدام نسبة الامكان الاعظم التي يتبع توزيع χ^2 وبدرجة حرية تساوي عدد المتغيرات المستقلة (٦) والتي يمكن حسابها من خلال الصيغة الاتية.

$$\chi^2 = -2LogL_0 - (-2LogL_1)$$

اي ان

$$\chi^2 = 543.939 - 362.562 = 181.377$$

اذ تم استخدام مستوى معنوية (α=0.001) لان كانت (*sig*= 0.00) وهذا يؤكد كفاءة ومعنوية الانموذج.

وجدول (٦) يبين اختبار H&L

جدول (٦) Hosmer and Lemeshow Test

Step	Chi-square	Df	Sig.
1	١٦.١20	8	.020

يوضح جدول (٦) اختبار H&L اذ توضح احصاءة H&L التي تتبع توزيع χ^2 تساوي (16.120) وهي معنوية لان (sig=0.020) وهي اقل من 0.05 لذلك نرفض فرضية العدم والتي تنص على عدم وجود فرق معنوي بين القيم المشاهدة والقيم المتوقعة.

اما جدول (٧) يبن القيم التقديرية للمعلمات النموذج

جدول (٧): Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	x1	.258	.487	.289	1	.581	1.327
	x2	.129	.158	.510	1	.469	1.136
	x3	-.063	.047	.920	1	.325	.928
	x4	-.014	.019	2.343	1	.137	.957
	x5	.095	.020	93.121	1	.010	1.114
	x6	.466	.323	1.969	1	.150	1.537
	Constant	-7.243	1.050	48.018	1	.030	.002

يوضح الجدول (٧) تقدير معلمات الانحدار اللوجستي باستعمال طريقة الامكان الاعظم ويكون النموذج وفق الاتي

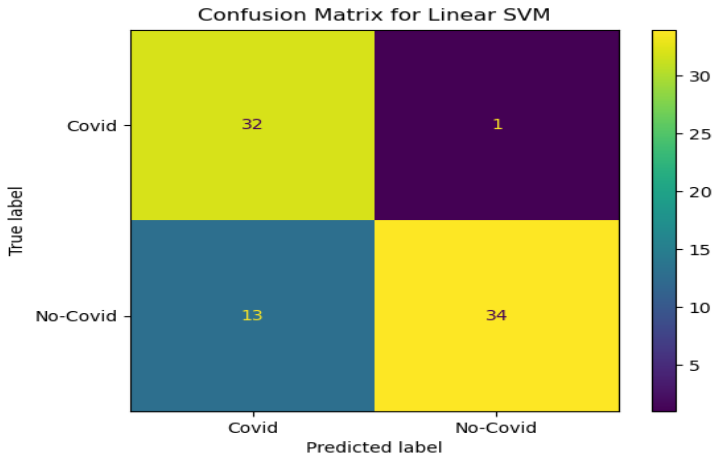
$$\hat{Y} = -7.243 + 0.258X1 + 0.129X2 - 0.063X3 - 0.014X4 + 0.095X5 + 0.466X6$$

ويبين الجدول اعلاه قيم الأخطاء المعيارية لكل معلمات الانموذج وكذلك يوضح قيم اختبار (Wald) وقيم المعنوية لكل معلمة من معلمات الانموذج ويبين ايضا ((Exp(B) الذي يمثل نسبة الارجحية حيث ان Exp(B) تشير الى مقدار التغيرات التي تحصل في نسبة الارجحية في حالة اذا كان شخص المصاب عند حدوث تغيرات في قيم المتغيرات المستقلة المرتبطة بالمعلمة B واذا كانت Exp(B) اكبر من الواحد تزداد نسبة الارجحية في حالة حدوث شخص مصاب اما اذا كانت اقل الزيادة في المتغير المستقل يؤدي ذلك الى انخفاض تلك النسبة.

بعد تطبيق أسلوب الانحدار اللوجستي يمكن تطبيق خوارزمية الة الموجة المساند (SVM)الذي تقوم بتقسيم عينات او بيانات الدراسة الى طورين: الاول هو طور التدريب باستعمال عدد مختلط من النوعين من البيانات (مصاب وغير مصاب) لغرض التدريب،اما الطور الثاني يتم استعماله لغرض اختبار ومعرفة دقة التصنيف ويتم ذلك حساب معايير دقة التصنيف المذكورة في اعلاه.والجدول (8) يوضح تصنيف البيانات وفق أسلوب الة الموجة المساند.

جدول (٨): مؤشرات آلة الموجة المساند (SVM)

SUM	**المتوقع**			**Classification**
	N Neg.	**P Pos.**		
TP+FN	1 (FN)	32 (TP)	P	المشاهدة
TN+FN	٣٤ (TN)	13 (FP)	N	
FN+TN+TP+FP	FN+TN	TP+FP	٨٠	SUM



شكل (٢): يوضح مؤشرات آلة الموجة المساند (SVM)

يوضح الجدول والشكل اعلاه المستخلص للتصنيف وفق اسلوب آلة الموجة المساند Support Vector Machine (SVM) للمصابين بعد اجراء عملية التدريب على المشاهدات وتم الاختبار وفق النسبة المحددة وهي (20%) للمصابين وغير المصابين وتم تصنيف (32) مريض من مجموع (33) من المصابين اي ان نسبة التمييز الصحيح (96%) اما الواحد المتبقي لإكمال العدد ال (33) للمصابين فعلاً فقد تم تصنيفه بانه غير مصاب.

اما بالنسبة لغير المصابين فالعدد الفعلي هو (47) فقد تم تصنيف (34) بشكل صحيح وبخصوص ال (13) فقد تم تصنيفهم مصابين اي ان نسبة التصنيف الصحيح (72%) فتكون نسبة التصنيف الكلية لكلا النوعين وفق اسلوب آلة الموجة المساند.

٤- نسبة التصنيف الصحيح للمصابين

$$\frac{32}{33} * 100\% = 96\%$$

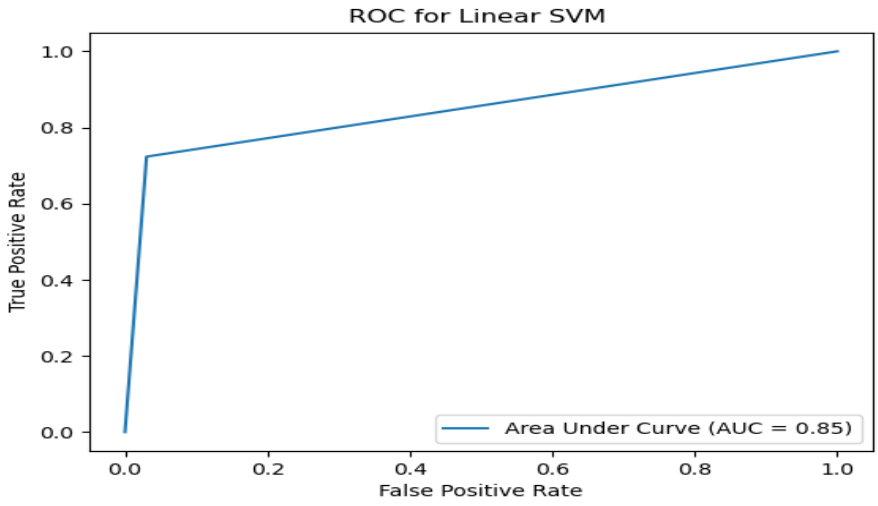
٥- نسبة التصنيف غير المصابين

$$\frac{34}{47} * 100\% = 72\%$$

٦- نسبة التصنيف الكلية

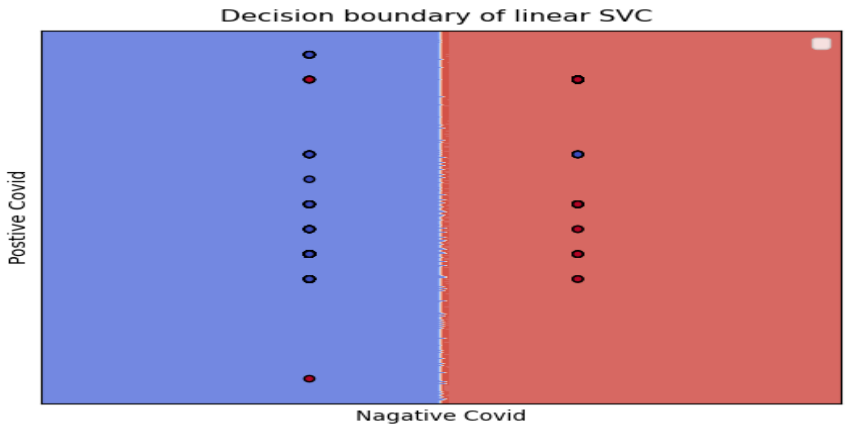
$$\frac{32+34}{80} * 100\% = 83\%$$

يمكن توضيح (ROC) و Area under curve (AUC) من خلال الشكل (٣)



شكل (٣): يوضح ROC

حيث يبين الشكل اعلاه نسبة التصنيف الممتازه لخوارزمية (SVM) وفق معيار (ROC) بحيث كلما تكون المساحة كبيرة تحت (Curve) تزداد الدقة اكثر بحيث كانت قيمة (AUC) تساوي(٨٥٪) وهذه القيمة تعبر عن دقة التشخيص العالية للتصنيف المشاهدات بحيث تكون النسب المحددة ROC تكون محصورة وفق النسب الاتية (اذا كانت النسبة من ٥٠٪ الى ٧٠٪ هذا يعني لايوجد تشخيص دقيق اما في حالة من ٧٠٪ الى ٨٠٪ هذا يدل عن وجود تشخيص مقبول وفي حال النسبة من ٨٠٪ الى ٩٠٪ هذا يدل على وجود تشخيص ممتاز ومقبول).اما الشكل (٤) يوضح كيفية فصل المشاهدات للمصابين وغير المصابين بخط فاصل .



شكل (٤): يوضح عملية انتشار البيانات

إذا نلاحظ من خلال الشكل اعلاه كيفية التصنيف الصحيح للمشاهدات بملاحظة نقطة واحدة للمصابين في الجانب الايمن ونقطتين في الجانب الاخر لغير المصابين باللون الاحمر وهذا يوضح لنا دقة التصنيف الصحيح للمشاهدات.
وتم حساب مواصفات الاداء للمصنف (SVM) حسب المعايير التي تم ذكرها اعلاه وكما موضح في الجدول (٩)

جدول (٩): تعايير آلة الموجة المساند (SVM)

Acc	SPE	SEN	FPR
83%	69%	97%	3%

واعتمادا على هذه المعايير والمواصفات تبين ان طريقة (SVM) قامت بعزل المشاهدات المصابين وغير المصابين بشكل جيد وهذا يؤشر بان طريقة (SVM) ذات كفاءة عالية في تصنيف المشاهدات.
ويمكن حسب النسب للمصابين وغير المصابين وفق القوانين الاتية

1-True Positive Rate

$$TPR = \frac{\text{TP}}{\text{TP} + \text{FN}} * 100\% = 97\%$$

2- True Negative Rate

$$TNR = \frac{\text{TN}}{\text{TN} + \text{FP}} * 100\% = 73\%$$

3- False Positive Rate

$$FPR = \frac{\text{FP}}{\text{TN} + \text{FP}} * 100\% = 3\%$$

4- False Negative Rate

$$FPR = \frac{\text{FN}}{\text{TP} + \text{FN}} * 100\% = 0.3\%$$

والجدول (١٠) ادناه يوضح المقارنة بين التصنيف بالانحدار اللوجستي وأسلوب الة الموجة المساند.

جدول (١٠): مقارنة تصنيف الانحدار اللوجستي وتصنيف أسلوب الة الموجة المساند

الطرائق	Acc	SPE	SEN	FPR
SVM	83%	69%	97%	3%
الانحدار اللوجستي	74%	61%	88%	36%

حيث تبين النتائج في الجدول (١٠) عن دقة وكفاءة الة الموجة المساند في التصنيف للمشاهدات.

الاستنتاجات والتوصيات

من خلال النتائج اظهرت آلة الموجة المساند تفوق واضح في تصنيف المشاهدات للمصابين وغير المصابين بنسبة تعرف صحيحة بلغت ٨٣٪ وهذا يبين تفوق الطريقة في التصنيف البيانات وفضلا عن تم تقدير معلمات النموذج باستعمال الانحدار اللوجستي واستعمال بعض الاختبارات ليكون ممثلا جيداً للبيانات .ونوصي بالاستفادة من خوارزمية آلة الموجة المساند واستعمالها في دراسات مختلفة (الطبية والصناعية وغيرها) واستعمالها في مجالات التصنيف والانحدار للمختصين في علوم الاحصاء .ونوصي استعمال طرائق اخرى ومقارنتها مع اسلوب آلة الموجة المساند مثل طريقة RNN(Recurrent Neural Network) في دراسات مستقبلية ونوصي باجراء دراسات متقدمة لتقدير الاصابات بمرض كورونا على مستوى العراق ووضع الحلول والاستراتيجيات المناسبة للحد من هذا المرض ونوصي باستخدام خوارزمية الة الموجة المساند في الانحدار للتنبؤ.

المصادر

اولاً: المصادر العربية

١- عبد الرازق، محمد صادق و زعلان، ريسان عبد الامام، (٢٠١٦)،" استخدام اسلوب الانحدار اللوجستي لتحليل اثر الضغط النفسي على الاصابة بضغط الدم" ,البصرة , مجلة الاقتصاد الخليجي, الصفحات ٤٨ - ٦٦.

ثانياً: المصادر العربية المترجمة

1-Abdul Razzaq, Muhammad Sadiq and Zalan, Risan Abdul Imam, (2016), “Using the logistic regression method to analyze the effect of psychological stress on the incidence of blood pressure”, Basra, Gulf Economics Journal, pp. 48-66.

ثالثاً: المصادر االاجنبية

1. Abdulqader, Q. M., (2015),"Comparison Of Discriminant Analysis and Logistic Regression Analysis: An Application or Caesarean Births and Natural Births Data”, Zakho Tecnical Institute Duhok, Polytechnic University Duhok, Duhok, Iraq, INAS, pp.34-46.
2. Ahmad. A., Khalid. M. and Zusof R.(2002)."Machine Learning Using Support Vector Machines", Universiti Teknologi Malazsia,pp.1-8, abdrahim@uniten.edu.mz.
3. Alrahamneh, A. & Hawamdeh, O., (2017) , "The Factors Affecting Eye Patients (Cataract) In Jordan by Using the Logistic Regression Model", MAS, pp. 38-42.
4. Camci, F., Chinnam, R. B., and Ellis, R. D.,(2008)."A robust kernel-distance multivariate control chart using support vector principles", International Journal of Production Research, Volume 46, Issue 18, pp 5075-5095
5. Cramer, J. S. (2002). "The origins of logistic regression".pp.3-7.
6. Duan. J.-C. , Härdle, W.K. , Gentle .J.(2012)."Handbook of computational finance", Series: Springer Handbooks of Computational Statistics, Publisher: Springer-Verlag Berlin Heidelberg, ISBN: 978-3-642-17253-3,978-3-642-17254-0.
7. Efron, B. (1979)."Bootstrap Methods", Another look at the jackknife. Ann. Statistics. pp.(1–7) from, http://data.princeton.edu/wws509/notes/c3.pdf
8. Fletcher, T. (2009)."Support vector machines explained", Universitz College London, London, pp.1-18
9. Gani ,W., Taleb, H.,andLimam, M.,(2008)."Statistical Process Control using Support Vector Machines: A Case Studz", Taken from The International Journal for Qualitz and Standards, ISG,Universitz of Tunis,pp.1-17
10. Gayou, O., & et al., (2008)," A genetic algorithm for variable selection in logistic regression analysis of radiotherapy treatment outcomes", American Association of Physicists in Medicine, pp. 5426 – 5433.
11. Ge S., Gao Z. and Wang R., (2007)., ACM,"Least Significant Bit Steganographz Detection with Machine Learning Techniques ", National Laboratorz for Novel Software Technologz Nanjing Universitz 210093 Nanjing, Jiangsu, China,
12. Gelman, A. & et al., (2014)," Bayesian Data Analysis", Texts in Statistical Science , CRC , LC , Chapman and Hall Book.
13. H E Fahad, Gh J Asmaa. (2020)." Using Support Vector Machine To Determine the Limits of Multivariate Control Charts",ASSOCIATION In cooperation with FACULTZ OF SCIENCE / TANTA UNIVERSITZ / EGZPT 22-23 November2020, Statistics Scope
14. Hosmer, D., Lemeshow, S. & Sturdivant , R. ,(2013)," Applied Logistic"
15. Ivanciue, O., (2007)."Applications of support Vector Machines in Cjemistrz",Wilez-VCH, John Wilez & Sons, Volume 23, pp. 291-400.
16. Lee, M.C. & To, C.,(2010)."Comparison of Support Vector Machine and Back Propagation Neural Network in Evaluating the Enterprise Financial Distress", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.3,pp31-43
17. Moradi, M., Sardasht, M. S., andEbrahimpour, M.(2012)."An Application of Support Vector Machines in Bankruptcz Prediction; Evidence from Iran", World Applied Sciences Journal,Vol. 17,NO. 6,pp. 710-717, ISSN 1818-4952
18. Müller, Marlene, (2004) , " Generalized Linear Models " Fraunhofer Institute for Industrial Mathematics (ITWM) , (Germany) .
19. Rodriguez ,G.,(2007), "Logit Models for Binary Data" ,Chapter(3) ,Retrieved
20. TAX ,D AVID. M.J., DUIN, ROBERT. P.W.,(2004)."Support Vector Data Description", Machine Learning,Vol. 54, PP.45–66
21. VAPNIK, V., ed., (1995)."The nature of statistical learning theorz", New Zork, Springer, pp1-201
22. Wuensch, K. L. (2014). "Binary logistic regression with SPSS. Retrieved March", 18, 2015.pp 2.