مجلة التجارة والتمويل

https://caf.journals.ekb.eg/

كلية التجارة – جامعة طنطا

العدد : الرابع

ديسمبر 2022

تحليل زمن البقاء باستخدام أسلوب الشبكات العصبية الاصطناعية ونموذجي انحدار كوكس والانحدار اللوجستي

(دراسة تطبيقية)

إعداد

د. عبد الرحيم عوض عبد الخالق بسيوني

المستخلص:

استهدف البحث دراسة مقارنة بين اسلوب الشبكات العصبية الاصطناعية ونموذج انحدار كوكس ونموذج الانحدار اللوجستي وذلك بالتطبيق على عينة من 161 مريض سرطان للرئة ومجموعة من العوامل المؤثرة في زمن البقاء على قيد الحياة وتوصلت الدراسة الى ان باستخدام اسلوب الشبكات العصبية الاصطناعية اهم العوامل المؤثرة في زمن البقاء هي التدخين والمهنة ودرجة المرض ثم العمر بكفاءة تصل الى 86% وحساسية 85% ونوعية 87% ونسبة تصنيف خاطئ 14% اما نموذج انحدار كوكس توصل الى ان اهم العوامل المؤثرة في زمن البقاء وهي درجة المرض , طرق العلاج، الاصابة بكورنا والتدخين وذلك بكفاءة 80.7% وحساسية 82.5% ونوعية 79% ونسبة تصنيف خاطئ 19.8% اما باستخدام نموذج الانحدار اللوجستي فأهم العوامل المؤثرة في زمن البقاء هي المهنة، ودرجة المرض, التدخين ثم العمر بكفاءة تصل الى 67.1% وحساسية 54.3% ونوعية 76.9% ونسبة تصنيف خاطئ 32.9% وبالتالي هناك أفضلية للشبكات العصبية الاصطناعية عن نموذج انحدار كوكس ونموذج الانحدار اللوجستي في تحديد العوامل المؤثرة في زمن البقاء والتصنيف والتنبؤ بالمشاهدات الجديدة .

Abstract:

The objective of the research is to compare artificial neural networks with Cox Regression model with logistic Regression model By Appling a sample of 161 lung cancer patients and affecting of factors in survival time, so there's a better chance of artificial neural networks than the cox Regression model and the logistic Regression model of classification and prediction. Because she has higher sensitivity, higher specificity, higher efficiency and lowest rate of misclassification.

١- مقدمة:

يعتمد كثير من الباحثين في تحليل بعض الظواهر والمشكلات على أساليب احصائية احادية المتغير بالرغم من تعقد الظاهرة وتعدد المتغيرات المسببه لها مما يؤكد ضعف فاعلية التحليل الاحصائي احادي المتغير وبالتالي اصبحنا في حاجه الي أساليب احصائية متقدمة ومتعددة المتغيرات لدراسة مثل هذه الظواهر ومن الاساليب الإحصائية متعددة المتغيرات أسلوب الشبكات العصبية ونموذجي انحدار كوكس والانحدار اللوجستي وغيرها وبالتالي فان هناك وفره في الاساليب الإحصائية ولكن المشكلة هي كيفية اختيار الاسلوب الملائم للظاهرة محل الدراسة ويعد سرطان الرئة من الامراض الأكثر خطورة والمسبب الأول للوفيات من بين كل أنواع السرطان ويعتبر التوصل لنموذج احصائي يحدد العوامل المسببة لمثل هذا المرض يقلل من حدة المرض ويساعد في علاجه.

وَمن الدراسات السابقة المُتعلقة بموضوع الدراسة قامت دراسة Savitridina (1997) باستخدام الانحدار اللوجستي في مُحددات ونتائج الزواج المُبكر في إندونيسيا كما ناقشت دراسة Chau (2005) مُقارنة بين الانحدار اللوجستي ونموذج انحدار كوكس في تحليل نتائج طلبة الثانوية وتم استخدام نموذج انحدار كوكس للكشف عن الأوقات التي يكون فيها الطلاب يواجهون صعوبة أكاديمية في تجاوز شهادة الثانوية بالصين كما اعتمدت دراسة التلباني (٢٠١٢) على المقارنة بين نموذج الانحدار اللوجستي ونموذج كوكس لدراسة أهم العوامل الاقتصادية والديموغرافية المؤثرة على معرفة واتجاهات الشباب نحو قضايا الصحة الإنجابية بالجزائر كما سعت دراسة حسن، صالح (٢٠١٤) إلى استخدام نموذج انحدار كوكس لأوقات البقاء للمرضى المُصابين بسرطان الدماغ في العراق خلال الفترة من ٢٠١٠ إلى ٢٠١٢ كما تناولت دراسة صالح (٢٠١٤) الانحدار اللوجستي لدراسة زمن البقاء لمرضى سرطان الدم بالعراق واستهدفت دراسة Guglu (2015) عينة من النساء المُصابة بسرطان الثدي باستخدام انحدار كوكس

وتناولت دراسة سُليمان (٢٠١٥) المُفاضلة بين التحليل التمييزي والنموذج اللوجستي والشبكات العصبية في تصنيف المُشاهدات والتعرف على العوامل المؤثرة على كفاية دخل الأسرة في السودان كما سعت دراسة بكير (٢٠١٦) إلى المقارنة بين الانحدار اللوجستي وانحدار كوكس لتحديد العوامل المؤثرة على الحالة الزواجية للنساء في فلسطين كما اهتمت دراسة سالم (٢٠١٧) باستخدام كُلاً من الانحدار اللوجستي وانحدار كوكس لمعرفة أهم العوامل المُصاحبة للسمنة في غزة، كما استهدفت دراسة الشويلي (2019) المقارنة بين الشبكات العصبية ونموذج كوكس في التصنيف والتنبؤ لعينة من ١٨٦ مريض بمرض الثلاسيما بالعراق ،كما تناولت دراسة Yupari et al (2021) تحليل عوامل الخطر البيولوجية والاجتماعية والسريرية للوفيات من المرضى المُصابين بفيروس كورونا في بيرو واخيرًا دراسة عُثمان (٢٠٢٢) مُقارنة بين الشبكات العصبية متعددة الطبقات مع الانحدار اللوجستي في التنبؤ بمخاطر الائتمان لمنشآت الأعمال في مصر ورغم تعدد الدراسات التي تناولت أسلوب أو اثنين على الأكثر إلا إنَ في دراستنا الحالية تم تناول الأساليب الإحصائية لأول مرة في تحليل زمن البقاء لمرضى سرطان الرئة.

٢- مشكلة البحث:

يعد اختيار الاسلوب الاحصائي المناسب المستخدم في تحليل البيانات احدى المشاكل التي تواجه كثير من الباحثين حيث ان اختياره يعتمد على نوع المشكلة محل القياس والتحليل ونوع البيانات سواء كانت بيانات نوعية او كمية ولذلك فالاسلوب الملائم لدراسة احدي المشاكل قد لا يكون مناسباً لمشكلة اخري كما انه لا تخلو أي دراسة سواء كانت طبية او اقتصادية او اجتماعية من الاساليب الإحصائية المتنوعة خصوصا تلك التي تصنف بين المشاهدات مثل طرق الانحدار بانواعه المختلفة ولكن ليس بشكله التقليدى انما بشكل متطور لتلائم حاله المتغير التابع عندما يكون ثنائي القيمة او مُتَعدد وَبالتالي فإن مُشكلة الدراسة تتمثل في المفاضلة بين اسلوب الشبكات العصبية الاصطناعية ونموذج انحدار كوكس ونموذج الانحدار اللوجستي وذلك لتحديد العوامل

المؤثرة في زمن البقاء لمريض سرطان الرئة وكيفية التصنيف للمشاهدات الجديدة والتنبؤ .

٣- أهمية البحث:

يستمد البحث أهميته من الناحية الإحصائية بالمقارنة بين أسلوب الشبكات العصبية الاصطناعية ونموذج انحدار كوكس ونموذج الانحدار اللوجيستي لإيجاد أفضل نموذج من بينها ليكون نموذج كفء وفعال قادر على التصنيف والتنبؤ بصورة عالية أما من الناحية التطبيقية من خطورة سرطان الرئة حيث إنه يمثل المسبب الأول للوفيات التي تحدث نتيجة لمرض السرطان سواء بين النساء او الرجال ويحصد سرطان الرئة عددا من الضحايا سنوياً أكثر بكثير مما يسببه أي نوع اخر من السرطان والتوصل الي أهم العوامل المسببه للمرض يجعلنا نتخذ القرارات السليمة للحد من هذا المرض والمساهمه في علاجه .

٤- أهداف البحث:

يمثل الهدف الاساسي للبحث هو المقارنة والمفاضلة بين كل من اسلوب الشبكات العصبية الاصطناعية ونموذج انحدار كوكس ونموذج الانحدار اللوجستي للوصول للأفضل في التصنيف والتنبؤ ويتم ذلك من خلال الخطوات التالية:

* استخدام طريقه كابلان – مايير واختبار لوغاريتم الرتبة (Kaplan – Meier and logrank test; 1958) وذلك لتحديد أهم العوامل التي لها تأثير معنوي على زمن البقاء لمريض سرطان الرئة.
* تقدير الشبكات العصبية الاصطناعية وتحديد الاهمية النسبية للمتغيرات المؤثرة على زمن البقاء وترتيبها وتحديد نسبه التصنيف الصحيح والخاطئ والحساسية والنوعية.
* تقدير نموذج انحدار كوكس مع استخدام طريقه الحذف العكسي Back ward Elimination method وتحديد الترتيب للمتغيرات المؤثرة على زمن البقاء ونسبة التصنيف الصحيح والخاطئ والحساسية والنوعية.
* تقدير نموذج الانحدار اللوجستي مع استخدام طريقه الحذف العكسي وترتيب المتغيرات المؤشرة على زمن البقاء ونسبه التصنيف الصحيح والخاطئ والحساسية النوعية.
* المقارنة بين الاساليب الإحصائية الثلاثة بالاعتماد على المعايير الإحصائية مثل مُعامل التحديد، الكفاءة، الحساسية، النوعية، نسبة التصنيف الخاطئ، وَذلك لإظهار قوة الأساليب الثلاثة لمعرفة أيهما أكثر كفاءة في التصنيف وهل هناك شروط معينه إذا توافرت في البيانات تجبرنا استعمال واحده من الاساليب.

٥- متغيرات وبيانات البحث:

تم الحصول على البيانات من المستشفي الميري الجامعي قسم الامراض الصدرية جامعة الإسكندرية ومعهد الأورام بطنطا لعينة من 161 مريض بسرطان الرئة خلال الفترة الزمنية 2018 حتى 2022

المتغير التابع Y : (زمن البقاء المريض على قيد الحياة وهي الفترة من تاريخ تشخيص المرض وتاريخ الوفاة أو فقدان المشاهدة لاي سبب مقاسه بالشهور + الحالة 0 فقدان المشاهدة , 1 حدوث حدث الوفاة)

X_1 : التدخين
X_2: التاريخ المرضى للعائلة
X_3: المهنة
X_4 : النوع
X_5 : العمر
X_6: التعرض لعلاج اشعاعى
X_7: طرق العلاج
X_8: الحالة الاجتماعية
X_9: درجة القرابة بين الاباء
X_{10}:الإصابة بكورونا
X_{11}:الامراض المزمنة
X_{12}: محل السكن
X_{13} : درجه المرض

٦- نماذج تحليل البقاء Survival analysis Models

اهتمت البحوث الإحصائية وخاصة منها الطبية بتحليل بيانات البقاء Survival data وهذا النوع من البيانات يكون في شكل متغيرات خاصة بالمريض كَالعُمر والنوع وغيرها من المُتَغيرات التي تؤثر في زمن البقاء على قيد الحياة.

يعتبر نموذج تحليل البقاء اسلوب يدرس العلاقة بين الزمن الذي يسبق حدوث الحدث وغالبًا الوفاة كَمُتَغير تابع وَمُتَغيرات أو أكثر من مُتَغير مُستَقِل بغض النظر عن طبيعة هذه المتغيرات من حيث كونها نوعية او كمية او خليط بينهم ومن أهم أساليب تحليل البقاء نموذج انحدار كوكس (Fox ; 2002)

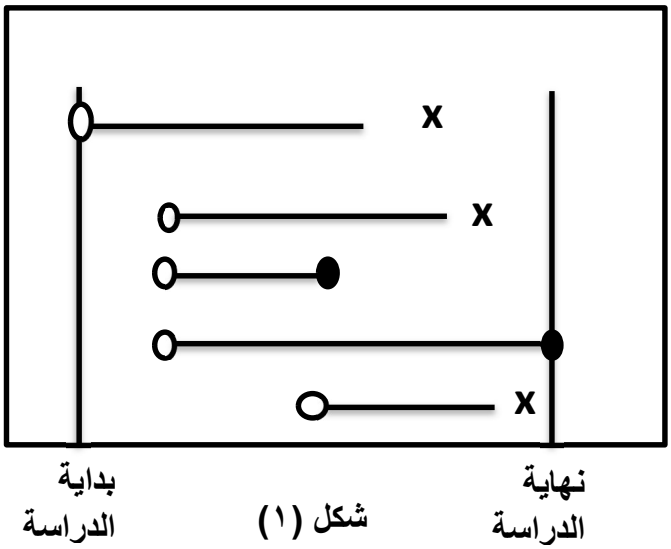
بيانات البقاء: Survival data

تم تعريف وقت البقاء على قيد الحياة بالفترة من تاريخ التشخيص او بداية العلاج حتى تاريخ وقوع الحدث (الوفاة) او تاريخ استجابة المريض او تطور حاله المريض وقد تكون الفترة بالسنوات او بالشهور او بالايام ويواجه الباحث عند حساب وقت البقاء ثلاث حالات كما هو موضع بالشكل (١):

١. وفاه المريض قبل نهاية التجربة وبالتالي يحسب وقت البقاء من تاريخ بداية العلاج حتى نهاية الوفاة وتسمى بيانات غير مبتورة (مراقبة) uncensored data.

٢. بقاء المريض على قيد الحياة حتى نهاية التجربة يحسب وقت البقاء من تاريخ بداية العلاج جتى نهاية التجربة وتسمى بيانات مبتورة (مراقبة) Censored data .

٣. فقدان المريض من المتابعه في وسط التجربة او وفاته لسبب اخر غير محل الاهتمام وتسمى بيانات مبتورة Censored data.



شكل (١)

تعريف الاختفاء (المراقبة) Definition of censoring

لتحليل بيانات البقاء لابد من توافر عدة شروط اهمها ان يكون المتغير التابع عبارة عن زمن البقاء وان يكون هناك بيانات مراقبه أو موقوفة ومتغيرات مستقله تؤثر على زمن البقاء نرغب في تحديدها.

(Rodrfguez ; 2001) والمقصود بالاختفاء وجود مفردات لا تعرف زمن حدوث الحدث لها ولا تستطيع متابعتها خلال فتره زمنية ويشترط في الاختفاء ان يكون مستقل ولا يعتمد على خطر التجربة (التلباني , 2012)

7- الأساليب الإحصائية المُستخدمة في الدراسة:

في هذا الجُزء سوف نتناول الإطار النظري للأساليب والنماذج الإحصائية المُستخدمة في الدراسة وَهيَ:

أولاً: الشبكات العصبية: Neural Networks تعرف الشبكات العصبية بانها دوال غير خطية تتمتع بالمرونه حيث انها لا تتطلب كثيراً من الافتراضات التي تحكم العلاقة بين المتغير التابع والمستقل كما انها تعالج البيانات اللامعلمية وصغيره الحجم ولا تشترط تبعية التوزيع الطبيعى وحققت نجاحاً هائلاً في مجالات الطب لإنها تقدم مستوى أفضل من النماذج الإحصائية التقليدية وتنقسم الشبكات العصبية الى نوعين هما (العباسي , 2005)

١- نماذج الشبكات العصبية الاصطناعية

Artificial Neural Networks (ANN)

٢- نماذج الشبكات العصبية الرياضية:

Neural Networks Algorithm (NNA)

وسوف يتم الاعتماد على النوع الأول.

* الشبكات العصبية الاصطناعية (ANN)

المخ البشري عبارة عن مجموع من الخلايا العصبية تسمى عصبونات وتنتشر هذه الخلايا في مجموعات تسمى شبكات وكل شبكة تتكون من الاف الخلايا العصبية المتصله والمتشابهه فيما بينها وهذه الشبكات التي تمكن المخ من اداء وظائفه المختلفة والخليه العصبية او العصبون عبارة عن وحده معالجة بها نواه في المنتصف لها بعض النهايات العصبية وهي المسئوله عن المدخلات وكذلك يوجد موصل طرفي مسؤول عن المخرجات وهذه النهايات الطرفية مندمج معها النهايات العصبية للخلية الثانية فيما يعرف بنقطة اشتباك ومن خلال هذه الخلايا العصبية يتم تخزين المعرفة عن العالم

الخارجي في العقل البشرى ولذلك يعطى المقدرة للشبكات العصبية على التعلم من خلال الخبره.

مكونات الشبكة العصبية Components of Artificial neural Networks

تتكون الشبكة العصبية من

أ- طبقه مدخلات input layer

ب- طبقه مخرجات output layer

ج- طبقه خفيه Hidden layer

د – الوصلات البينية (الاوزان) connections

* طبقه المدخلات: من خلالها يتم امداد وتغذية الشبكة بالبيانات وقد تكون المدخلات على صوره بيانات خام او مخرجات معالجة اخرى ووحدات المعالجة في طبقه المدخلات لا يتم فيها اي معالجه حسابية بل تقوم بنقل البيانات عبر الوصلات الى الطبقة الخفية او الى طبقه المخرجات إذا لم يكن هناك طبقه خفيه وأي شبكة عصبية تحتوي على طبقه واحده فقط من وحدات الادخال.

* طبقه المخرجات:

هي عبارة عن وحدات معالجة من خلالها يتم اخراج الناتج النهائي للشبكة وقد تحتوي على وحدة معالجة واحدة او أكثر من وحده وتستقبل وحدات المعالجة من طبقه المخرجات الاشارات القادمة اليها من الطبقة الخفية او من طبقه الادخال مباشره بعد اجراء المعالجة اللازمه وقد ترسل اشاره المخرجات النهائية او تقوم باعادة هذه المخرجات الى الشبكة مرة اخرى إذا لم يتم المعالجة المطلوبة لها وتحتوي الشبكة عادة على طبقه مخرجات واحدة.

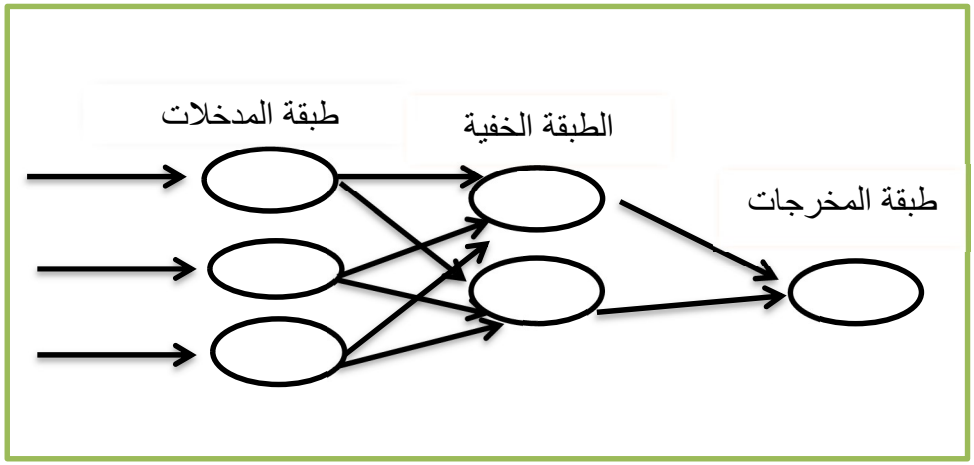
* الطبقة الخفية:

تقع بين طبقه المدخلات وطبقه المخرجات وتحتوي بعض الشبكات على طبقه خفية واحده او أكثر وقد لا تحتوي على طبقه خفية حيث انها تستقبل الاشارات من طبقه المدخلات ونقوم بالمعالجة اللازمة وارساله الى طبقه المخرجات.

* الوصلات (الاوزان)

هي التي تربط بين الطبقات او بين الوحدات التي تكون داخل كل طبقه ومسئوله عن نقل الاشارات بين الطبقات ووحدات المعالجة

وشكل (٢) يوضح المكونات الأساسية للشبكات العصبية



* وحدات المعالجة (عصبونات) processing Element

هي احدى المكونات الاساسية التي تتكون منها كل الطبقات الشبكة العصبية وتتكون كل وحده معالجة (عصبون) من الآتي.

***معاملات الاوزان** **weights coefficients** * **دالة الجمع** **summation Function**

*الة التحويل Transfer function * دالة المخرجات output function

- معاملات الاوزان:

تعتمد الشبكة العصبية الاصطناعية على الوزن الترجيحى للعنصر الذي يعبر عن الاهمية النسبية للعنصر المدخل.

- دالة الجمع: -

تعد اول عملية تقوم بها وحده المعالجة هي حساب مجموع المدخلات الموزونة القادمة الى الوحدة باستخدام دالة الجمع وتقوم هذه الدالة بحساب المتوسط للأوزان لكل المدخلات الواردة الى وحده المعالجة وذلك بضرب كل مدخله في وزنها المصاحب لحساب المجموع كالتالي:

$$S_j = \sum_{i=1}^{n} x_i w_{ij} \qquad (1)$$

حيث ان:

S_j: ناتج عمليه الجمع لكل وحده معالجه (j)

x_i: القيمة المدخله من الوحده (i) الى الوحده (j)

w_{ij}: الوزن الذي يربط الوحدة (j) بالوحده (i) من الطبقة السابقة

دالة التحويل:

وهي معروفه بدالة التحفير الخارجية وذلك لان العصبون يقوم باستقبال المدخلات من العصبونات الاخرى اومن المصادر الاخرى وهي العملية الثانية بعد الجمع ومن دوال التحويل.

أ- دالة Sigmoid

ب- دالة الخطوة step function

ج- دالة الاشارة sign function

* دالة sigmoid:

من مهام هذه الدالة تحويل المخرجات الى قيم محصورة بين الصفر والواحد وتسمى التنشيط اللوجستي الثنائي او تحويل المخرجات الى قيم محصوره بين (1+,1-) وتسمى

في هذه الحالة بدالة التنشيط اللوجستي ثنائي القطبية وهي من أكثر الدوال المستخدمة وتأخذ الشكل الآتي: -

$$F(x) = \frac{1}{1 + e^{-s}} \qquad (2)$$

* دالة الخطوة:-

تقوم هذه الدالة بتحويل دالة المخرجات الى قيم محصوره بين (0 , 1) كالنظام الثنائي وتأخذ الشكل التالي:

$$F_{(x)} = \begin{cases} \mathbf{1} & \boldsymbol{if\ s \geq 0} \\ \mathbf{0} & \boldsymbol{if\ s < 0} \end{cases} \qquad (3)$$

* الدالة الخطية: -

تعمل على توفير مخرجات تساوى المدخلات الموزونه لوحده المعالجة وعادة تستخدم وحدات الذي الدالة الخطية للتقريب الخطى وتأخذ الشكل التالي: $F(x) =$

s حيث s هي المخرجات .

* دالة الاشارة: -

تستخدم في عمليه التصنيف والتمييز وتستخدم قيمه معنية تدعي ∅ تاخد الشكل التالي:

$$F_{(x)} = \begin{cases} 1 & if\ s \geq \emptyset \\ -1 & if\ s < \emptyset \end{cases} \qquad (4)$$

* دالة المخرجات:

بعد عمليه الجمع ثم عمليه التحويل الى قيم محصوره في مدى معين تقوم دالة مخرجات بإعطاء الناتج النهائي للمشكلة.

- استخدام الشبكات العصبية في التصنيف.

أصبح من الممكن استخدام الشبكات العصبية الاصطناعية (ANN) في عمليه التصنيف وذلك من خلال تحديد دالة تنشيط خاصه بالتصنيف وهناك عده دوال للتنشيط تختلف باختلاف الناتج التي يتم الحصول عليه اضافه الى اهداف الشبكة العصبية التي يراد

تحقيقها وهنا سوف يتم استخدام دالة الخطوة لأنها تناسب عمليه التصنيف وتعطى المخرجات نتجيتين هما صفر وواحد كما يلى

$$F_{(x)} = \begin{cases} 0 \ if \ x \ \leq 0 \\ 1 \ if \ x \ > 0 \end{cases} \qquad (5)$$

وتستخدم دالة الخطوة في وحدات طبقه المخرجات بينما تستخدم الدالة اللوجستية Sigmoid في الطبقة الخفية وتأخذ الشكل الآتي: -

$$F(x) = \frac{1}{1 + \ e^{-s}} \qquad (6)$$

حيث ان s المجموع الموزون للمدخلات مضاف اليه حد التحيز bias الذى يرمز له بالرمز ∅ اي يكون المجموع:

$$S = \sum_{i=1}^{n} x_i w_j + \emptyset \qquad (7)$$

وبالتالي فإن مخرجات الشبكة اما صفر او واحد بالاعتماد على المدخلات فإذا كانت المخرجات (0) تصنف المشاهدة في المجموعة الأولى وإذا كانت المخرجات (1) تصنف المشاهدة الى المجموعة الثانية.

ثانياً: نموذج انحدار كوكس Cox – Regression

يعتبر نموذج انحدار كوكس من اهم النماذج التي يتم استخدامها في تحليل البقاء ويسمى بنموذج الاخطار التناسبية (التلبانى , 2012) ويشترط في نموذج انحدار كوكس ان يتكون المتغير الناتج من جزئين أحدهُما الزمن الذي يسبق حدوث الحدث والثاني متغير وصفي ثنائي القيمة وان يكون هناك مجموعه من المتغيرات المستقلة وقد تكون وصفيه او كميه او خليط منها. ومن اهم خصائص نموذج انحدار كوكس

(Sainani ; 2005) انه نموذج شبه معلمى ويستوعب كل القياسات المتصله والمنفصله في مرات الحدث كما يمكن دمج المتغيرات التي تعتمد على الوقت وهي

المتغيرات التي تتغير قيمتها على مدار فتره المراجعه ويهتم نموذج كوكس بتأثير المتغيرات المستقلة على معدل الخطر فإذا كانت T متغير عشوائي متصل فنموذج انحدار كوكس يصاغ كالتالي (Cox ; 1972) : -

$$H\ (t/x_i) = h_0(t)exp^{\beta x_i} \qquad (8)$$

حيث ان

$H(t/x_i)$: الخطر الشرطى لوقوع الحدث عند الزمن t بالنسبة للمفردات التي لديها متجه المتغيرات المفسرة x_i

$h_0\ (t)$: دالة الخطر الاساسية Baseline hazard function

وتعتمد على الزمن وتقابل متجه متغيرات مستقله X_i=o

β : متجه عمود $p\ x1$ معلمات نموذج انحدار كوكس

X_i: متجه صف $1x\ p$ المتغيرات المستقلة

حيث ان ثبات تأثير المتغيرات المستقلة على الخطر سواء بالزيادة او النقصان لا تتغير تبعا لتغير نقطة زمنية t

وان النسبه بين اي معدلين للخطر تكون ثابته ولا تعتمد على الزمن

$$\frac{h\ (\ t\ /\ x_1\ )}{h\ (\ t\ /\ x_2\ )} = \frac{h_0\ (t)exp^{\beta\ x_1}}{h_0\ (t)exp^{\beta\ x_2}} = exp^{\beta\ (x_1 - x_2)} \qquad (9)$$

وباخذ اللوغاريتم كتحويله لنموذج انحدار كوكس

$$h(t) = h_0\ (t)exp\ \beta_1 x_1 + \beta_2 x_2 + \cdots \ldots \ldots \ldots + \beta_p x_p \qquad (10)$$

$$\text{Log h(t)} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \sum_{i=1}^{p} \beta_i x_i \qquad (11)$$

ومن المعروف ان اللوغاريم يساهم من حل كثير من المشاكل القياسيه للموذج منها عدم ثبات

التباين والطبيعية ويحول الى شكل الانحدار الخطى عند تحليل العلاقات للبيانات ذات المتغيرات التابعه الثنائية وذلك يعنى ان أي تغير يحدث في المتغيرات المستقلة يقابله تغير في لوغاريتم معدل الخطر. وجرت العادة انه قبل تقدير نموذج الانحدار كوكس لابد من اختيار أفضل المتغيرات المؤثرة ذات دلالة احصائية على المتغير التابع وهو زمن البقاء ويتم ذلك بالاستعانة بطريقه كابلان – مايير واختبار لوغاريتم الرتبة.

Kaplan – Meier and log rank test حيث تعرف بانها احتمال البقاء للمفرده لمدة معينه من الزمن وتستعمل لتحليل بيانات البقاء بصوره فرديه لكل متغير مستقبل في متغيرات الدراسة مع المتغير التابع والذى يمثل زمن البقاء ويعد تحديد المتغيرات المؤثرة ذات الدلالة الإحصائية في التأثير على زمن البقاء يتم تقدير نموذج انحدار كوكس باستخدام طريقه الحذف العكس Back ward Elimination حيث يتم استبعاد المتغير المفسر الذي قيمته p-value له اكبر من مستوى المعنوية (α) ويتم الابقاء على المتغير صاحب P-value اقل من مستوى المعنوية.

تقدير معاملات نموذج انحدار كوكس:

تستخدم طريقه الامكان الجزئية partial likelihood المقترحة من قبل Cox لتقدير قيمه β وذلك في غياب المعرفة عن طبيعة دالة الخطر الأساسية ho (t) وتقابل هذه الطريقة طريقه المربعات الصغرى في تحليل الانحدار حيث تأخذ الدالة الجزئية الشكل التالي: -

$$L\mathbf{p}(\beta) = \prod_{i=1}^{m} \frac{e^{x_i \beta}}{\sum e^{x_j B}} \qquad (12)$$

حيث ان:

m: عدد حالات الفشل

X_i: تمثل قيم المتغير للمفردة التي وقت بقائها (t_i) ومجموعه الخطر R (t_i) أي كل المفردات المعرضة لخطر الفشل وهي

$$R(t_i) = \{j = t_j \geq t_i\}$$

وباخذ اللوغاريتم للمعادلة رقم (s)

$$Ln\ Lp(\beta) = \sum_{i=i}^{m} \{ x_i\beta - ln \sum_{j\in(ti)} e^{x_jB} \} \quad (13)$$

وبأخذ المشتقة الأولى للمعادلة (13) بالنسبةβ نحصل على تقدير β ويفسر معناها بانه مقدار التغير في لوغاريتم معدل الخطر $h(t/x_i)$ عند تغير المتغير المستقل بوحدة واحدة فإذا كان β سالبه فذلك يعنى ان زيادة المتغير المستقل لوحده واحده يقلل معدل الخطر وبالتالي فالحالة تذهب في الاتجاه الأفضل اما إذا كانت β موجبه فذلك يعنى ان زيادة المتغير المستقل بوحدة واحدة يزيد من معدل الخطر والحالة تذهب في الاتجاه السئ.

* اختيار صلاحيه نموذج انحدار كوكس Goodness of fit

لمعرفه مدي ملائمه النموذج للبيانات واختبار لقوة النموذج ومعرفه الأهمية النسبية للمتغيرات المستقلة (x_i) في تفسير سلوك المتغير التابع ($_iy$) تستخدم معامل التحديد التالي:

$$R^2_{cox} = 1 - exp^{\frac{2}{n}[log\ l_f - log\ l_0]} \quad (14)$$

حيث ان:

R^2_{cox}: معامل التحديد

l_0: لوغاريتم الترجيح الجزئي للنموذج بدون اي متغيرات مستقله

l_f: لوغاريتم الترجيح الجزئي للنموذج بكامل للمتغيرات المستقلة

n: عدد المشاهدات الكلية

* اختيار معنوية المعالم المقدرة للنموذج β:

يتم اختيار معنوية كل معلمة من معالم النموذج على حده بإحصائية wald statistic التالية:

$H_0 : \beta i = o$

$H_1 : \beta i \neq o$

إحصائية الاختبار: $W^2 = \left(\frac{\widehat{Bi}}{SE\beta_i}\right)^2$

ثالثا: نموذج الانحدار اللوجيستي Logistic Regression:

يعد الانحدار اللوجيستي اسلوب إحصائي يدرس العلاقة بين متغير تابع نوعي ومتغير أو أكثر مستقل أيًا كان نوعها ويعتبر الانحدار اللوجيستي حاله خاصه من حالات الانحدار العامة ويطلق عليه احيانًا نماذج اللوجت (عكاشه , 2002) ومن استخدامات الانحدار اللوجيستي هو التنبؤ كما يمتاز بانه أكثر مرونة من نماذج الانحدار التقليدية اذ بالإمكان افتراض علاقه معينه تربط بين المتغير التابع والمتغيرات المستقلة كما يمكن تحويله الى الشكل الخطى باستخدام تحويله Logit transformation كما انه يعاب عليه انه لا يأخذ الزمن الذي يسبق حدوث الحدث بعين الاعتبار وعدم مقدرته على التعامل مع بيانات الاختفاء وخاصة البيانات المقطعية لذلك كان هناك البديل الأفضل وهو انحدار كوكس.

* نموذج الانحدار اللوجستي:

عندما يكون المتغير التابع (Y) متغير نوعي ثنائي القيمة ولدينا متغير مستقل (x) فاننا نعطى احدى القيمتين اللتين يأخذهما المتغير التابع احدهما نجاح ويأخذ القيمة (1) والاخرى فشل وتأخذ القيمة (0) واحتمال النجاح p عندما تأخذ القيمة (1)

$$P = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \qquad (15)$$

واحتمال الفشل l-p عندما y تأخذ القيمة (o)

$$1 - P = 1 - \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{\beta_0 + \beta_1 x}} \qquad (16)$$

نسبة الترجيح (odd ratio) وهو عبارة عن نسبه بين احتمال حدوث الحدث الى احتمال عدم حدوث.

$$\text{Odd} = \frac{\text{p}}{1 - \text{p}} = \text{e}^{\beta_0 + \beta_1 \text{X}}$$

وللتغلب على مشكلة ان القيم المقدرة لا يمكن تفسيرها بوصفها احتمالات وذلك لان القيم محصورة بين (1,0) وقد تم استخدام تحويلة (Logit) وهي:

$$\text{Logit (odd)} = \log\left(\frac{\text{p}}{1 - \text{p}}\right) = \beta_0 + \beta_1 x_1 \dots \dots \dots + \beta_p x_p \qquad (17)$$

حيث ان logit تشير الي اللوغاريتم الطبيعي لنسب الترجيح.

ويتم تقدير معاملات نموذج الانحدار اللوجستي باستخدام طريقه الامكان الاعظم Maximum likelihood method وهي من اهم الطرق لتقدير معالم النماذج سواء الخطية وغير الخطية وتقابل طريقه المربعات الصغري في نماذج الانحدار العادية (Garson;2014)

وهناك أكثر من طريقه لتفسير معاملات نموذج الانحدار اللوجستي اهمها طريقه نسبه الترجيح (odd) ولقد ذكر (Edward ,2003) ان حساب نسبه الترجيح OR إذا تغير قيمه المتغير المستقل x بأكثر من وحده يمكن اجرائه كالتالي:

$$OR = e^{cb}$$

$$Log\ (OR) = C * b$$

C مقدر التغير من المتغير المستقل

b معامل الانحدار اللوجستي

* اختبار جودة نموذج الانحدار اللوجستي من خلال الآتي:

اختبار هزمر ولميشو Hosmer and lemshow test ويستخدم الاختبار هل يوجد اختلاف بين القيمة المشاهدة والقيم المتوقعة للنموذج وذلك باستخدام.

$$R_L^2 = \frac{\chi^2}{-2\ln(l_0)} \qquad (18)$$

χ^2: تمثل الفرق بين لوغاريتم الاحتمال في حاله النموذج بدون متغيرات مستقله بالحد الثابت فقط ولوغاريتم الاحتمال في حاله النموذج بكامل المتغيرات المستقلة.

$2\ln(l_0)$: لوغاريتم الاحتمال للنموذج بالحد الثابت فقط .

R_L^2 : تمثل نسبه الانخفاض في القيمة المطلقة للوغاريتم الاحتمال ومقياس لمدى تحسن سوء التوفيق نتيجة لإضافة المتغيرات المستقلة وتنحصر قيمته بين الصفر والواحد.

$R_L^2 = 0$: ليس هناك فائدة او تأثير المتغيرات المستقلة على المتغير التابع

$R_L^2 = 1$: هناك فائدة او تأثير تام للمتغيرات المستقلة على المتغير التابع

: معامل التحديد للانحدار اللوجستي R^2 يستخدم لاختبار قوة النموذج اللوجستي أي نسبة مساهمة العوامل المؤثرة التي يتضمنها النموذج المقدر على المتغير التابع (متغير الاستجابة)

ويتم ذلك من خلال:

* معامل التحديد R^2_{cs} $cox\ and\ snell$

$$R^2_{cs} = 1 - e^{\frac{2}{n}[Ln(l_f) - Ln\ (l_0)\ ]} \qquad (19)$$

* معامل التحديد R^2_N Negerkerke

$$R^2_N = \frac{R^2_{cs}}{1 - e^{\frac{2}{n}[Ln(lo)]}} \qquad (20)$$

ولاختبار معنوية معاملات نموذج الانحدار اللوجستي وبالإضافة لإحصائية Wald السابق ذكرها هناك مقياس اخر هو partial R^2 وهو أفضل من احصائية Wald حيث انه يحسب الاهمية النسبية للمتغيرات المستقلة في النموذج ويعتمد على قيمة χ^2 فإذا كانت قيمة χ^2 ذات دلالة احصائية فهذا يعنى ان المتغير المستقل له دلالة احصائية والعكس صحيح وبالتالي دخول المتغير للنموذج النهائي.

* جدول التصنيف Classification table

يوضع جدول التصنيف عدد الحالات التي تمتلك خاصيه معينه وعدد الحالات التي لا تمتلك تلك الخاصية وبين نسبه التصنيف الصحيح ونسبه التصنيف الخاطئ (Cox , DR , 1975)

التصنيف	موجب p	سالبn	مجموع
موجب p	Tp	FN	P
سالبn	FP	TN	$\grave{p}$
مجموع	Q	$\grave{Q}$	1

ويحسب منه التالي:

* الحساسية (s) Sensitivity:-

تحسب من جدول التصنيف وهي احتمال بان يكون التصنيف المتوقع موجب للحاله ويحسب:

$$S = \frac{TP}{TP + FN} = \frac{Tp}{P} \qquad (21)$$

* النوعيه Specificity (sp)

نسبه التصنيف الصحيح للحاله السالبه

$$Sp = \frac{TP}{TP + FN} = \frac{Tp}{p^1} \qquad (22)$$

* الكفاءة نسبة التصنيف الصحيح Efficiency (E_t)

$$Et = \frac{Tp + TN}{P + p^1} \qquad (23)$$

الجانب التطبيقي

يستهدف البحث دراسة مقارنة بين الاساليب الإحصائية المختلفة متمثلة في اسلوب الشبكات العصبية الاصطناعية (ANN) ونموذج انحدار كوكس – Cox Regression ونموذج الانحدار اللوجستي Logistic Regression وذلك من خلال التطبيق على عينه من 161 مريض مصاب بسرطان الرئة ويمكننا توصيف متغيرات الدراسة كالتالي: ـ

Time	زمن البقاء بالشهور + Censored (٠ : فقدان المشاهدة، 1: حدوث الحدث الوفاة)
X_1 التدخين	(1: مدخن , 2 غير مدخن)
X_2التاريخ المرضي للعائلة	(1: يوجد امراض وراثيه , 2 لا يوجد امراض وراثيه)
X_3 المهنة	(1 : موظف , 2: عامل , 3: متقاعد, 4: ربه منزل , 5: طفل, 6: طالب)
X_4النوع	(1:ذكر , ٢ :انثى)
X_5 العمر	عدد السنوات من تاريخ الميلاد حتى تاريخ تشخيص المرض
X_6 التعرض لعلاج اشعاعي	(1: تعرض لعلاج اشعاعي, 2: لم يتعرض لعلاج اشعاعي)

X_7طرق العلاج	(1: جراحه , 2: كيماوى, 3 : اشعاعى , 4 : أكثر من طريقه)
X_8 الحالة الاجتماعية	(1 متزوج , 2 : اعزب, 3 : غير ذلك)
X_9 درجه القرابة للآباء	(1: يوجد قرابه , 2 : لا يوجد قرابه)
X_{10} الإصابة بكورونا	(1 : مصاب , 2 : غير مصاب)
X_{11} الامراض المزمنة	(1: يوجد امراض مزمنه , 2: لا يوجد امراض مزمنه)
X_{12} محل السكن	(1: الحضر , 2: الريف)
X_{13} درجه المرض	(1: مرحله اولى , 2: مرحله ثانية , 3: مرحله ثالثه , 4: مرحله رابعه)

ولتحديد أكثر المتغيرات اهميه وتأثيراً في زمن البقاء على قيد الحياة اولاً يتم استخدام طريقه كابلان مايير واختيار لوغاريتم الرتبه Kaplan – Meier and logrank Method

حيث تقوم هذه الطريقة بتحليل البيانات او اتخاذ كل متغير على حِدَة مع المتغير التابع الذي يمثل وقت البقاء على قيد الحياة وذلك قبل استخدام أياً من الاساليب الإحصائية السابق ذكرها ويمثل الفرض العدمى للاختبار $H_0: h(t/x_i) = o$ وذلك يعنى عدم وجود فروق معنوية بين دوال البقاء عند مستويات المتغير (x_i) ويكون الفرض البديل للاختيار $H_1: h(t/x_i) \neq o$ وذلك يعنى وجود فروق معنوية بين دوال البقاء عن مستويات المتغير x_i وبالتالي دخول المتغير للنموذج وتم الحصول على النتائج الآتية:

جدول (١) نتائج طريقه كابلان مايير واختبار لوغاريتم الرتبة.

المتغير	χ^2	P-Value
X_1 التدخين	12.776	0.000
X_2التاريخ المرضى للعائلة	0.279	0.579
X_3 المهنة	16.137	.006
X_4النوع	0.891	.345
X_5العمر	7.56	.031
X_6 التعرض لعلاج اشعاعي	5.378	.02
X_7طرق العلاج	4.124	.038
X_8 الحالة الاجتماعية	4.958	,084
X_9 درجه القرابة للآباء	1.069	.301
X_{10} الإصابة بكورونا	3.663	.046
X_{11} الامراض المزمنة	1.145	.285
X_{12} محل السكن	0.029	.865
X_{13} درجه المرض	12.805	.012

من جدول (١) ومن نتائج طريقه كابلان مايير توصلنا الى ان كل من المتغيرات X_2 (تاريخ العائله المرضي) , X_4 (النوع) , X_8 (الحالة الاجتماعية), X_9 (درجه القرابة للآباء) , X_{11} (الامراض الزمنه) , X_{12} (محل السكن) , ليس لها تأثير معنوي على المتغير التابع وهو زمن البقاء على قيد الحياة، في حين كل من المتغيراتX_1 (التدخين) , X_3 (المهنة) , X_5 (العمر), X_6 (التعرض لعلاج اشعاعي), X_7 (طرق العلاج) , X_{10} (الإصابة بكورنا) , X_{13}(درجه المرض) متغيرات معنوية وذات دلالة احصائية في التأثير على زمن البقاء على قيد الحياة وبناء عليه سيتم الاكتفاء بالمتغيرات المعنوية فقط .

أولاً: الشبكات العصبية الاصطناعية (ANN)

باستخدام الشبكات العصبية بالاعتماد على المتغيرات المعنوية فقط الناتجة من طريقه كابلان مايير وهي X_1 (التدخين) , X_3 (المهنة) , X_5 (العمر), X_6 (التعرض لعلاج اشعاعي), X_7 (طرق العلاج) , X_{10} (الإصابة بكورنا) , X_{13} (درجه المرض) ومن خلال استخدام برنامج SPSS V23 تمت عمليه التدريب للنموذج وتم اختيار قيم عشوائية للاوزان بين طبقه المدخلات والطبقة الخفية وايضاً الاوزان بين الطبقة الخفية وطبقه المخرجات وذلك يعنى قيام الشبكة العصبية بالتعلم والتعرف على البيانات والعلاقة بين المتغيرات المختلفة تم الحصول على جدول (٢) :-

جدول (٢) التدريب والاختيار

	N	Percent
Sample Training	118	73.3%
Testing	43	26.7%
Valid	161	100%
Excluded	21	
Total	182	

ومن جدول (٢) ملاحظ انه تم استخدام البيانات بنسه 100% حيث ان عدد فترات التدريب Training هي 118 بسنه 73.3% وعدد فترات الاختبار Testing هي 43 بنسبه 26.7%

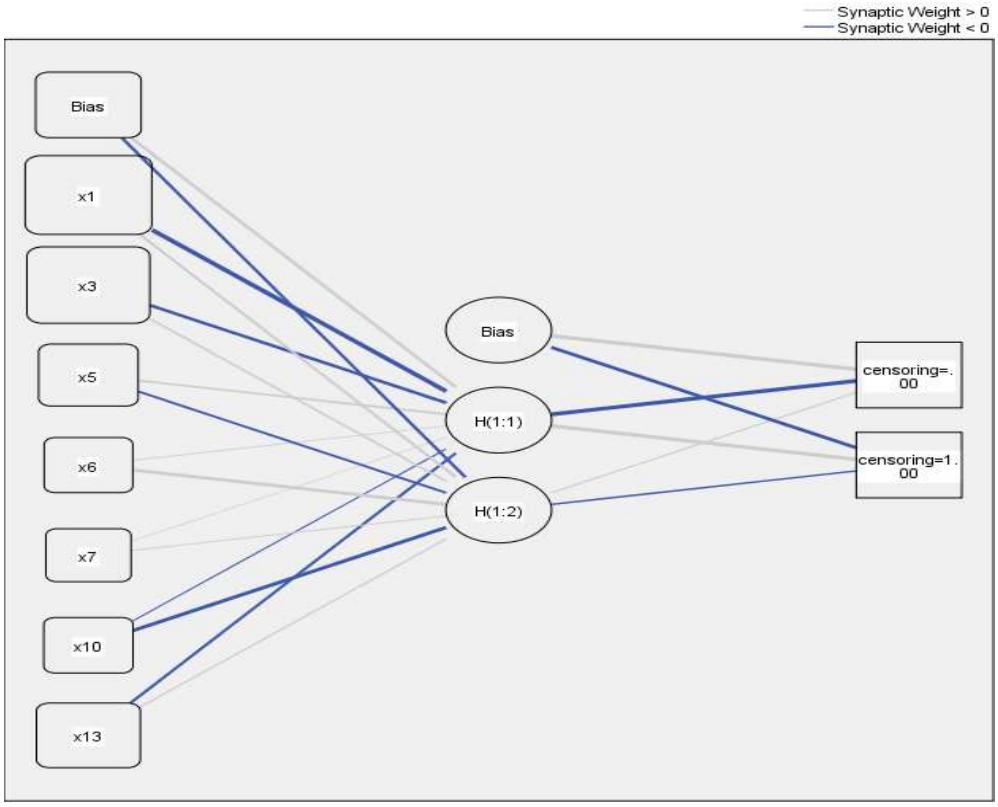
اما في جدول (٣) مكونات الشبكة العصبية :

جدول (٣) مكونات الشبكة العصبية

Input Layer	Covariates	1	x1
		2	x3
		3	x5
		4	x6
		5	x7
		6	x10
		7	x13
	Number of Units[a]		7
	Rescaling Method for Covariates		Standardized
Hidden Layer(s)	Number of Hidden Layers		1
	Number of Units in Hidden Layer 1[a]		2
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	censoring
	Number of Units		2
	Activation Function		Softmax
	Error Function		Cross-entropy

وبعض مكونات الشبكة العصبية من جدول (٣) تتكون الشبكة من الجزء الأول وهو طبقه المدخلات input layer وتتكون من سبعه متغيرات مستقله وهي X_1 التدخين , X_3 الهنه , X_5العمر, X_6 التعرض لعلاج اشعاعي, X_7طرق العلاج , X_{10} الإصابة بكورنا , X_{13} درجه المرض اي سبعة وحدات ادخال والجزء الثاني هو الطبقة الخفية وتتكون من طبقه واحده تحتوي على وحدتين معالجه (عصبون) بالإضافة إلى حد التحيز Bias ودالة التنشيط المستخدمة هي Hyperbolic والجزء الثالث هي طبقة المخرجات وتتكون من وحدتين ودالة التنشيط المستخدمة فيها هي Sofimax وهي الدالة اللوجستية وبالتالي فان رسمه الشبكة العصبية كالتالي .

شكل (٢) الشبكة العصبية



ومن شكل (٢) يتضح من الشكل البيانى الذي يظهر معماريه الشبكة المستخدمة وهي تتكون من طبقه المدخلات وتتكون من سبعه وحدات ادخال وهي المتغيرات المستقلة ويضاف اليها الثابت (bias) والطبقة الخفية وتتكون من وحدتين معالجه مضاف اليها الثابت (bias) اما طبقه المخرجات تتكون من وحدتين هما (0 فقدان المشاهدة، 1حدوث حدث الوفاة) تم الحصول على الجدول التالي:

جدول (٤) ملخص عملية التدريب والاختبار من الشبكة العصبية

Training	Cross Entropy Error	58.382
	Percent Incorrect Predictions	24.6%
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	0:00:00.08
Testing	Cross Entropy Error	15.552
	Percent Incorrect Predictions	14.0%

من جدول (٤) يتضح لنا ان نسبه التصنيف الخاطئ في عمليه التدريب بلغت 24.6% بينما نسبه التصنيف الخاطئ في عمليه الاختبار بلغت 14% وهذا يشير الى ان الشبكة تم تدريبها بشكل جيد على تصنيف المفردات الجديدة.

تم بعد ذلك تم تقدير المعلمات باستخدام الشبكة العصبية كما هو موضح في جدول (٥) التالي:ـ

جدول (٥) تقديرات المعلمات في الشبكة العصبية

Predictor		Predicted			
		Hidden Layer 1			Output Layer
		H(1:1)	H(1:2)	[censoring=.00]	[censoring=1.00]
Input Layer	(Bias)	-.239	.279		
	x1	-.515	-.233		
	x3	.122	-.832		
	x5	.436	.785		
	x6	.598	.240		
	x7	1.076	.427		
	x10	.366	-.348		
	x13	-.507	.028		
Hidden Layer 1	(Bias)			.522	-.206
	H(1:1)			-.456	.414
	H(1:2)			.255	.689

التنبؤ باستخدام الطبقة الخفية وطبقة المدخلات وطبقة المخرجات كالتالي:

$$H(1:1) = -0.239 - 0.515x_1 + 0.122x_3 + 0.436x_5 + 0.598x_6 + 1.076x_7 + 0.366x_{10} + 0.507x_{13}$$

$$H(1:2) = 0.279 - 0.233x_1 + 0.832x_3 + 0.785x_5 + 0.240x_6 + 0.427x_7 - 0.348x_{10} + 0.28x_{13}$$

$$y(0) = 0.522 - 0.456\,H(1:1) + 0.255\,H(1:2)$$

$$y(1) = 0.206 + 0.414\,H(1:1) + 0.689\;\;H(1:2)$$

وللتاكد من قدره الشبكة على التصنيف تم اعداد جدول التصنيف رقم (٦) كما يلى

جدول (٦) جدول التصنيف باستخدام الشبكات العصبية

Sample	Observed	Predicted		
		N	y	Percent Correct
Training	N	50	18	73.5%
	Y	11	39	78.0%
	Overall Percent	51.7%	48.3%	75.4%
Testing	N	20	3	87.0%
	Y	3	17	85.0%
	Overall Percent	53.5%	46.5%	86.0%

من جدول (٦) نسبه التصنيف الصحيح للحالات (0) فقدان المشاهدة 73.5% في عينه التدريب Training بينما بلغت 87% في عينه الاختبار Testing اما نسبه التصنيف الصحيح للحالات التي انتهت بحدوث الحدث (1) 78% في عينه التدريب بينما بلغت 85% في عينه الاختبار وبالتالي فان نسبه التصنيف الصحيح للشبكة هي 86% وهي نسبة جيدة جدا للتنبؤ وتصنيف المشاهدات الجديدة مما يؤكد على كفاءة الشبكة العصبية العالية

* الحساسية sensitivity (s)

$$S = \frac{17}{20} = 85\%$$

* النوعية Specificity (Sp)

ويشير نسبه التصنيف الصحيح للحالات (0)

$$\text{Sp} = \frac{20}{23} = 87\%$$

* الكفاءة Efficiency (Et)

نسبه التصنيف الصحيح

$$\text{Et} = \frac{20 + 17}{43} = 86\%$$

ولتحديد الاهمية النسبة للمتغيرات المستقلة في تأثيرها على المتغير التابع وهو زمن البقاء كما في جدول (٧) التالي:

جدول (٧) الاهمية النسبية للمتغيرات المستقلة

	Importance	Normalized Importance
x1	.337	100.0%
x3	.308	91.2%
x5	.121	36.0%
x6	.034	10.1%
x7	.005	1.6%
x10	.038	11.4%
x13	.156	46.2%

ومن جدول (٧) تبين ان اهم المتغيرات المستقلة في التأثير على زمن البقاء هو متغير التدخين (X_1) بنسبه 100% يليية متغير المهنة X_3 بسنبه 91.2% وثم درجه المرض x_{13} بنسبه 46.2% وثم العمر x_5 بنسبه 36% ثم الإصابة بكورونا X_{10} بنسبه 11.4% ثم التعرض لعلاج اشعاعي X_6 بنسبه 10.1% واخيرا x_7 متغير طرق العلاج بنسبة 1.6%

ثانياً: نموذج انحدار كوكس Cox – Regression

يُعد نموذج انحدار كوكس احد الاساليب الإحصائية المستخدمة في تحليل وقت البقاء على قيد الحياة قبل وقوع الحدث حيث من اهم شروطه ان يتكون المتغير التابع من جزئيين احدهما وقت البقاء وَالثاني متغير وصفي ثنائي وبالاعتماد على المتغيرات التي سبق اثبات معنويتها وتأثيرها في زمن البقاء بطريقه كابلان مايير وهما سبعه متغيرات هي X_1 التدخين، X_3 المهنة، X_5العمر، X_6 التعرض لعلاج اشعاعي، X_7طرق العلاج، X_{10} الإصابة بكورونا، X_{13} درجة المرض ولاختيار جوده وملائمة النموذج للبيانات توصلنا الى:

جدول (٨) اختبار جوده التوفيق لنموذج انحدار كوكس

Model	-2log likelihood	χ^2	Df	p-value
Block0	604.963	-	-	-
Block1	545	53.134	7	0.000

من جدول (٨) تبين ان Block0 هو النموذج بالحد الثابت فقط اي بدون متغيرات مستقله اما Block1 هو النموذج بعد اضافه سبعه متغيرات المستقلة والمعنوية ونلاحظ ان قيمة (p-value = 0.000) اقل من مستوى المعنوية 05. وبالتالي فان النموذج الاحصائي ذو دلالة احصائية وان المتغيرات لها تأثير معنوي على زمن البقاء وعند تقدير نموذج انحدار كوكس وباستخدام طريقه الحذف العكسي Back ward Elimination تم استبعاد المتغير التعرض لعلاج اشعاعي (X_6) وكانت النتائج كالتالي :

جدول (٩) نتائج طريقه الحذف العكسي

المتغير	βi	S.E	Wald	Sig	EXp(β)
التدخين X_1	-1.164	.339	11.809	.001	.312
المهنة X_3	-.198	.118	2.809	.049	.820
العمرX_5	.016	.006	6.442	.011	1.016
طرق العلاجX_7	.596	.151	15.629	000	1.815
الإصابة بكورنا X_{10}	1.066	.287	13.802	000	2.905
درجه المرض X_{13}	-.590	.128	21.412	000	.554
Block1	- 2 log likelihood = 528.362 χ^2 = 50.729 df =6 p-value =0.000				

ومن جدول (٩) يتضح ان هناك علاقه عكسيه بين X_1 التدخين، X_3 المهنة، X_{13} درجة المرض ووقت البقاء على قيد الحياة او الإصابة وعدم الإصابة بسرطان الرئة وان هناك علاقه طرديه بين كلا من العمر (X_5) وطرق العلاج (X_7) والإصابة بكورنا (X_{16}) واحتماليه الإصابة وعدم الإصابة بسرطان الرئة . اما عن نسبه الترجيح Exp(β) نلاحظ انها Exp(β)=.312 لمتغير X_1 التدخين وهي أقل من الواحد والعلاقة عكسيه مع X_1 التدخين والإصابة بسرطان الرئة حيث ان نقصان التدخين وحده واحده من غير مدخن (2) الى مدخن (1) تزيد من احتماليه الإصابة بسرطان الرئة وبالتالي انخفاض وقت البقاء على قيد الحياة . وتشير قيمه $Exp(\beta) = .820$ ل للمهنه (X_3) والعلاقة عكسية بين المهنة والإصابة بسرطان الرئة حيث ان نقصان المهنة بوحده واحده تودى الى زيادة الإصابة بسرطان الرئة .

وتشير في $Exp(\beta) = 1.016$ ل X_5 وهي العمر والعلاقة طردية مع الإصابة بسرطان الرئة أي كلما زاد العمر وحدة واحده تؤدى الى زيادة الإصابة بسرطان الرئة.

وتشير قيمه $Exp(\beta)$ =1.815 ل (X_7) طرق العلاج والعلاقة طردية أي كلما زادت طرق العلاج وحدة من جراحه (1) الى كيماوى (2) تؤدى لزيادة احتمال الإصابة بسرطان الرئة. وتشير قيمه $Exp(\beta) = 2.905$ ل (X_{10}) الإصابة بكورنا والعلاقة طردية تعنى ان زيادة X_{10} من (1) مصاب الى (2) غير مصاب تزيد من احتمالية عدم الإصابة بسرطان الرئة وبالتالي زيادة وقت البقاء . اما عند قيمة $Exp(\beta) = .554$ ل (X_{13}) درجة المرض والعلاقة عكسية فان زيادة درجة المرض درجة واحدة تقلل وقت البقاء وتزيد من اجمالي الإصابة بسرطان الرئة .

ويمكن صياغة نموذج انحدار كوكس رياضيًا كالتالي: -

$$h(t/x_i) = ho(t)\exp(-1.164X1 - .198X3 + .016X5 + .596X7 + 1.066X10 - .596X13)$$

ويحسب معامل التحديد كالتالي: -

$$R^2_{cox} = 1 - \exp\frac{2}{n}[\ln Lf - \ln Lo]$$

$$= 1 - exp\ \frac{2}{161}[528.362 - 604.963] \qquad = 61.4\ \%$$

وذلك يعني ان المتغيرات المستقلة تساهم بسنه 61.4% من تحديد وقت البقاء لمريض السرطان الرئة وتم تكوين جدول التصنيف بالاعتماد على دالة الخطر ودالة البقاء فكانت النتائج كالتالي:

جدول (١٠) التصنيف

	0	1	Correct percent
0	72	19	79%
1	12	58	82.5%
Over percent			80.7%

ومن جدول التصنيف نلاحظ ان هناك 72 حاله غير مصابه (0) تم تنصيفها بطريقه صحيحة بنسبة 79% و١٩ حالة غير مصابه صنفت خطا بنسبه 21% وان هناك 58

حاله مصابه (1) صنفت بطريقه صحيحة ٨٢,٥% وان هناك 12 حاله صنفت خطا بنسبة (17.5) % وان نسبة التصنيف الصحيح للنموذج هي 80.7% وهي بنسبة جيدة للتوزيع للاعتماد عليه في التصنيف والتنبؤ.

* الحساسية Sensitivity (s)

نسبة التصنيف الصحيح للحالات (1) الى التي تحمل الصفة أو الإصابة

$$S = \frac{58}{12+58} = 82.5\%$$

* النوعيه Specificity (Sp)

نسبه التصنيف الصحيح للحالات (0) أي غير المُصَابة

$$\text{Sp} = \frac{72}{19+72} = 79\%$$

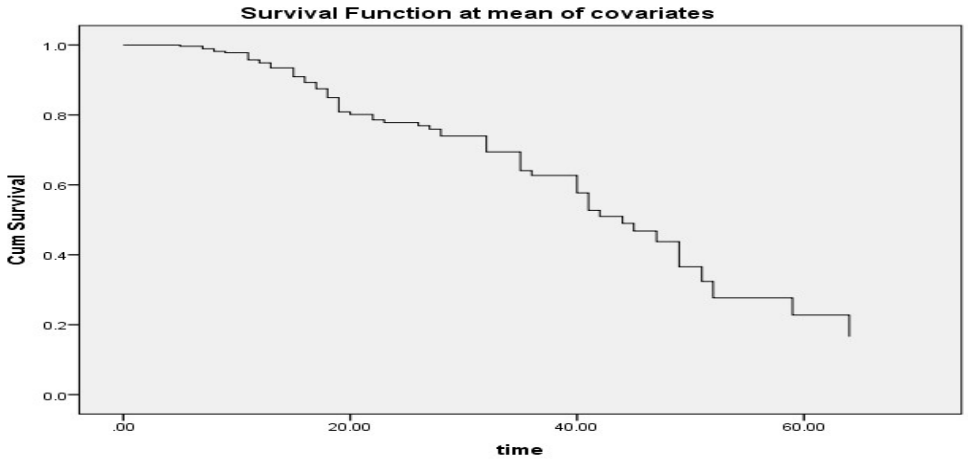
* الكفاءة Efficiency (Et)

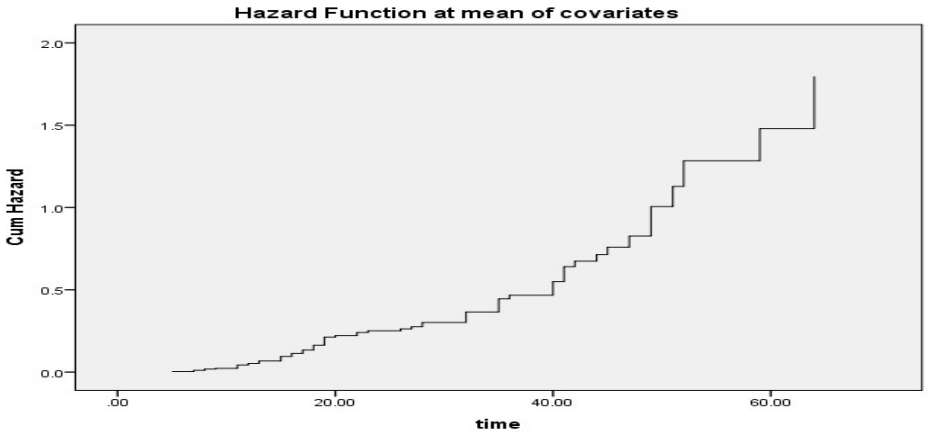
نسبه التصنيف الصحيح

$$\text{Et} = \frac{58+72}{161} = 80.7\%$$

ولمعرفه الاهمية النسبية للمتغيرات المستقلة ومساهمة كلا منها في التصنيف والتنبؤ بالإصابة بسرطان الرئة تنظر في جدول (٩) الى عمود wald كلما كانت قيمة wald كبيرة تعنى أي مساهمه المتغير في التصنيف والتنبؤ كبيره وعليه فان المتغير درجة المرض X_{13} في المرتبة الأولى بنسبه 21.4% يليه طرق العلاج X_7 بنسبة 15.629% يليه الاصابة بكورونا بنسبة 13.802% يليها التدخين X_1 بنسبة 11.809% ثم العمر X_5بنسبة 6.4%وأخيرًا المهنة X_3 بنسبة 2.809% .

اما بالنسبة لسلوك دالتي البقاء ودالة الخطر نلاحظ من الاشكال الآتية:





ومن الملاحظ ان دالة البقاء تكون في اعلى مستوى لها عند بداية الفترة الزمنية ثم تبدأ في الانخفاض تدريجيًا حتى تقل لأدنى حد لها ودالة الخطر تسلك سلوك معاكس دائمًا لدالة البقاء

ثالثا: نموذج الانحدار اللوجستي logistic Regression

يتميز نموذج الانحدار اللوجستي بانه لا يشترط تبعيه المتغيرات المستقلة والتي يتضمنها النموذج للتوزيع الطبيعى ولا يشترط خطية العلاقة بين المتغيرات المستقلة والمتغير التابع بالإضافة الى كفاءة نموذج اللوجستي في التصنيف والتنبؤ. ولاختبار مدى ملائمة نموذج الانحدار اللوجستي للبيانات وذلك بالاعتماد على المتغيرات المستقلة التي ثبت

معنويتها وتأثيرها في زمن البقاء بطريقة كابلان - مايير وَهُم سَبعة متغيرات متمثلة في التدخين (X_1)، X_3 المهنة، X_5العمر، X_6 التعرض لعلاج اشعاعي، X_7طرق العلاج، X_{10} الإصابة بكورونا، X_{13} درجة المرض، حيث تبين لنا التالي:

Chi-square	Df	p-value
43.876	7	0.000

حيث ان قيمة χ^2 والتي تمثل الفرق بين لوغاريتم دالة الترجيح للنموذج بالمتغيرات المستقلة ولوغاريتم دالة الترجيح للنموذج بالحد الثابت فقط وبدون متغيرات مستقله وحيث ان قيمة P-value = 0.000 وهي اقل من 05. وذلك دليل على جودة توفيق النموذج وملائمته للبيانات وبعد ذلك نبدأ في تقدير نموذج الانحدار اللوجستي وباستخدام طريقه الحذف العكسي Backward Elimination Method تبين معنوية كل من التدخين (X_1) X_3 المهنة،X_5 العمر، X_7طرق العلاج، X_{13} درجة المرض. وتم استبعاد كُلاً من التعرض لعلاج إشعاعي (X_6) والإصابة بكورونا (x_{10}) لعدم معنويتها إحصائيًا ولاختبار صلاحية النموذج اللوجستي الجديد في ظل وجود خمس متغيرات مستقلة وباستخدام اختبار Hosmer and lemshow تبين ان: -

Chi-sq	Df	sig
4.877	8	.771

وحيث ان sig=.771 وهي اكبر من مستوى المعنوية 05. وبالتالي قبول الفرض العدمي بمعنوية المتغيرات وملائمة النموذج للبيانات كما تبين ان المتغيرات المستقلة ساهمت بنسبة 29.5% في زمن البقاء كما هو موضح بالجدول (١١) التالي:

جدول (١١)

-2log likelihood	Cox 8 snell R square	Negelkereke R square
180.616	.219	.294

معالم نموذج الانحدار اللوجستي تم الحصول على نتائج التالية:

جدول (١٢) تقدير معالم الانحدار اللوجستي

المتغير	βi	S.E	Wald	Sig	EXp(β)
X_1	-1.288	.447	8.308	.004	.276
X_3	-.619	.187	10.980	.001	.538
X_5	.019	.009	5.012	.025	1.019
X_7	.461	.214	4.625	.032	1.585
X_{13}	-.576	.184	9.753	.002	.562
constant	2.214	1.083	4.177	.041	9.152

ومن جدول (١٢) نلاحظ ان العلاقة عكسية بين كل من التدخين X_1 والمهنة X_3 ودرجة المرض X_{13} وزمن البقاء على قيد الحياة او الإصابة بسرطان الرئة والعلاقة طردية بين كلا من العمر x5 وطرق العلاج x_7 وزمن البقاء على قيد الحياة او الإصابة بسرطان الرئة اما عن قيمة Exp (β) وهي قيمة odd Ratio والتي تستخدم للمقارنة الفئة المرجعية بالفئات الاخرى لكل متغير من المتغيرات وهي تعني احتمال حدوث حدث مقسوما على احتمال عدم حدوثه. وحيث ان قيمة Exp (β) = .276 لـ X1 التدخين وهي اقل من الواحد والعلاقة مع X_1 عكسية فان نقصان التدخين درجة واحدة من (2) غير مدخن، (1) مدخن تؤدي الى زيادة احتمال الإصابة بسرطان الرئة وانخفاض وقت البقاء.

وحيث ان قيمه Exp (β)=1.019 لـ X_5 العمر والعلاقة مع X_5 طردية فان زيادة العمر وحده واحدة تزيد من احتمالية الإصابة بسرطان الرئة.

وبالتالي نموذج الانحدار اللوجستي يتم تقديره كالتالي:

$$Log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \cdots \ldots + \beta_p x_p$$

$$= \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_5 x_5 + \beta_7 x_7 + \beta_{13} x_{13} =$$

$$2.214 - 1.288\, x_1 - .619\, x_3 + .019\, x_5 + .461\, x_7 - .576 x_{13}$$

ولاختيار جودة النموذج من التصنيف تم اعداد جدول التصنيف التالي:

جدول (١٣) جدول التصنيف بالانحدار اللوجستي

observed	Predicted		Percentage correct
Censoring	0	1.	
0	70	21	76.9%
1	32	38	54.3%
			67.1%

* الحساسية (s) :

هي نسبة التصنيف الصحيح للحالة (1)

$$S = \frac{38}{32 + 38} = 54.3\%$$

* النوعية (S_p)

نسبة التصنيف الصحيح للحالات (0)

$$Sp = \frac{70}{(70 + 21)} = 76.9\%$$

* الكفاءة (E_t)

نسبة التصنيف الصحيح

$$Et = \frac{70 + 38}{161} = 67.1\%$$

أما عن تحديد الاهمية النسبية للمتغيرات المستقلة أي ترتيبها حسب درجة التأثير وذلك حسب قيمة إحصائية wald كما موضح في جدول (١٢) تأتي المهنة في المرتبة الأولى يليها درجة المرض يليها التدخين يليها العمر وأخيرًا طرق العلاج.

* المقارنة بين الاساليب الإحصائية:

	الشبكات العصبية	نموذج انحدار كوكس	نموذج الانحدار اللوجستي
العوامل الأكثر تأثيرا في زمن البقاء	التدخين، المهنة، درجة المرض، العمر	درجة المرض، طرق العلاج. الإصابة بكورونا، التدخين	المهنة، درجة المرض، التدخين، العمر
الكفاءة	86%	80.7%	67.1%
الحساسية	85%	82.5%	54.3%
النوعية	87%	79%	76.9%
R2		61.4%	29.4%
نسبة التصنيف الخطي	14%	19.8%	32.9%

الملاحظ ان كل المؤشرات تعطي الأفضلية للشبكات العصبية يليها نموذج انحدار كوكس ثم الانحدار اللوجستي.

النتائج والتوصيات

استهدف البحث الى عمل مقارنة بين الاساليب الإحصائية متمثلة في الشبكات العصبية والاصطناعية ونموذج انحدار كوكس ونموذج الانحدار اللوجستي وذلك لتحديد اهم العوامل المؤثرة في وقت البقاء على قيد الحياة لمريض سرطان الرئة وذلك لعينه من 161مريض وباستخدام عدد من المتغيرات المستقلة وهي التدخين , تاريخ المرض للعائلة, المهنة , والنوع , والعمر , والتعرض للعلاج الاشعاعي , وطرق العلاج والحالة الاجتماعية ودرجة القرابة بين الاباء والإصابة بكورونا والامراض المزمنة , ومحل السكن ودرجة المرض وتبين أفضلية الشبكات العصبية عن نموذج انحدار كوكس ونموذج الانحدار اللوجستي حيث أنها توصلت الي ان أكثر المتغيرات تأثيرا في زمن البقاء لمريض سرطان الرئه هي التدخين، المهنه، درجة المرض واخيرا العمر وذلك بنسبة تصنيف صحيح بلغت ٨٦٪ وحساسيه بنسبة ٨٥٪ ونوعيه بنسبة ٨٧٪ ونسبة تصنيف خاطئ ١٤٪ كما ان نموذج انحدار كوكس توصل الي ان اكثر المتغيرات تأثيرا في زمن البقاء لمريض سرطان الرئه هي درجة المرض، طرق العلاج، الاصابة بكورونا، التدخين، العمر، المهنة بنسبة تصنيف صحيح ٨٠,٧٪ وحساسية بنسبة ٨٢,٥٪ ونوعية بنسبة ٧٩٪ ونسبة تصنيف خاطئ ١٩,٨٪ كما توصل نموذج الانحدار اللوجستي الي ان أكثر المتغيرات تأثيرا في زمن البقاء لمريض سرطان الرئه هي المهنه، درجة المرض، التدخين، العمر، طرق العلاج وذلك بنسبة تصنيف صحيح ٦٧,١٪ وحساسية بنسبة ٥٤,٣٪ ونوعية بنسبة ٧٦,٩٪ ونسبة تصنيف خاطئ ٣٢,٩٪.

ولذلك يوصي الباحث بـــ:ـــ

١) التوسع في استخدام الشبكات العصبية ونموذج انحدار كوكس واللوجستي في الدراسات الاقتصادية بدلا من استخدامها في المجالات الطبيه والاجتماعية.

٢) اجراء دراسات اخرى على سرطان الرئة مع اضافة متغيرات جديده للدراسة مثل الحالة التعليمية وغيرها.

٣) ضروره وجود الاحصائي في المستشفيات حتى يتم اثبات البيانات بطريقه صحيحة مما يسهل الدراسات والبحوث الإحصائية.

٤) ضرورة الكشف المبكر يقلل من خطورة الامراض المزمنة.

٥) يوصي الباحث في مجال التصنيف والتنبؤ بالمشاهدات الجديدة استخدام الشبكات العصبية بدلا من الانحدار اللوجستي وانحدار كوكس إذا كانت المتغيرات المستقلة لا تتبع التوزيع الطبيعي.

٦) يُرجع الباحث أفضلية نموذج كوكس عن نموذج اللوجستي لأسباب أهمها قصر فتره المتابعة عن خمس سنوات وعدم تعامل الانحدار اللوجستي مع بيانات الاختفاء وزمن البقاء لذلك يوصي الباحث بإعادة الدراسة مع زيادة فتره المتابعة عن عشر سنوات.

المراجع العربية

١-التلباني شادي إسماعيل (2012) " دراسة مقارنة بين نموذج الانحدار اللوجستي ونموذج انحدار كوكس لتحديد أهم العوامل الاقتصادية والديموغرافية المؤثرة على معرفة واتجاهات الشباب نحو قضايا الصحة الإنجابية" جامعة أبو بكر بلقيد، تلمسان، كلية علوم الاقتصاد وعلوم التفسير والعلوم التجارية، رسالة دكتوراه، الجزائر.

٢-الشويلى، مرتضى فرحان (2019) " استعمال الشبكات العصبية الاصطناعية ونموذج انحدار كوكس للتنبؤات المستقبلية"، جامعة كربلاء، كلية الإدارة والاقتصاد، قسم الإحصاء، رسالة ماجيستر، العراق.

٣- العباسي، عبد الحميد محمد (2005) "المقارنة بين اسلوب الشبكات العصبية الاصطناعية والاساليب الإحصائية التقليدية للتنبؤ بإعداد الوفيات الناتجة عن حوادث المرور بالكويت" المجلة المصرية للسكان وتنظيم الاسرة، معهد الدراسات والبحوث الإحصائية، جامعة القاهرة، مصر.

٤- بكير، عز الدين عيسى رمضان (2016) "استخدام الطرق الإحصائية لدراسة محددات الحالة الزواجية للنساء في فلسطين" جامعة الازهر، كلية الاقتصاد والعلوم الإدارية، قسم الإحصاء التطبيقي، غزة.

٥- حسن، رعد فاضل – صالح، عائدة هادي (2014) "استخدام نموذج انحدار كوكس Cox – regression لأوقات بقاء المرض المصابين بسرطان الدماغ في العراق"، الجامعة المستنصرية، كلية الإدارة والاقتصاد، قسم الإحصاء، مجلس كلية الرافدين الجامعة للعلوم، العدد 34.

٦- سالم، إسماعيل جميل (2017) "استخدام الانحدار اللوجستي وانحدار كوكس في تحديد العوامل المؤثرة على السمنة لدى السيدات المتزوجات في قطاع غزة" جامعة الأزهر، كلية الاقتصاد والعلوم الإدارية، قسم الإحصاء، رسالة ماجيستير، غزة.

٧-سليمان، علي أبشر (2015) "المقارنة بين التحليل التميزي والنموذج اللوجستي الثنائي والشبكات العصبية في تصنيف المشاهدات" رسالة دكتوراه غير منشورة، جامعة السودان للعلوم والتكنولوجيا، كلية الدراسات العليا.

٨-صالح، عائد هادى (2014) "تحليل الانحدار اللوجستي لدراسة زمن البقاء لمرضي سرطان الدم" جامعه كربلاء، مجلس الإدارة والاقتصاد، المجلد الثالث، العدد التاسع، العراق.

٩-عكاشة، محمود (2002) "استخدام نظام Spss وتحليل البيانات الإحصائية" الطبعة الأولى، جامعة الازهر، غزة، فلسطين.

١٠-عثمان، حسام محمد (2022) "استخدام الشبكات العصبية متعددة الطبقات في التنبؤ بمخاطر الائتمان لمنشآت الاعمال: دراسة تطبيقية" جامعة القاهرة، كلية التجارة، مجلة الدراسات المالية والتجارية، العدد الأول.

المراجع الأجنبية:

1. Chau- K., G. (2005)" Analyzing Student learning out comes Usefulness of logistic and Cox Regression Models", VOL. 5, Association Institutional Research.
2. Cox, D.R, (1975), "partial Likelihood" Biometrika,vol-62,No(3),p 269-276.
3. Cox, D.R. (1972). Regression models and life tables Journal of The. 7. Royal Statistical Society Series B.34, 187.220.
4. Edwards, C. (2003). "A Assessing Association: Logistic Regression and Logit Analysis" Biometry, FRWS 6500, Fall 2003.

5. Fox, Y. (2002), Cox proportional hazard regression for Survival data.
6. Garson, G. D. (2014). Logistic regression: Binary and multinomial. Asheboro, NC.
7. Guglu, I., (2015) Cox proportional Hazards Model in Social science The Journal of Academic Social Science studies.
8. Hosmer, D. & lemeshow, S& May, R. (2007)"Applied Survival Analysis: Regression Modeling of Time to Event Data", 2nd edition wiley, New York, USA.
9. Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282), 457-481.
10. Rodrfguez, G., (2001)," Generalized Linear models" princeton university USA.
11. Sainani, K. (2005), introduction to Cox regression" Department of Health. Research and policy, stand ford university USA.
12. Savitridina, R(1997). "Determinants and Consequences of Early Marriage in Java", Indonesia, Asia pacific population Journal, vol.12, June 1997.
13. Yupari, A.I., Bardales, A.L., Rodriguez, A.J., Barros, S.J. and Rodriguez, D. A, (2021), COVID-19 mortality risk factors in hospitalized patients; A Logistic regression model,Rev.Fac.Med.Hum.Enero,21(1),19 – 27 .