تطبيق تقنية تعلم المكائن للتنبوء بمخاطر التمويل البنكي – دراسة حالة بنك العمال الوطني

وضـاح عبدالله حسـن[1] ، سلاّم عثمان فقيري[2]
جامعة الزعيم الإزهري، كلية علوم الحاسوب وتقانة المعلومات[1،2]
waddah142@gmail.com

Received on: 16-6-2018
Accepted on: 17-7-2018

المستخلص– *التطور الهائل فى مجال تكنولوجيا المعلومات والإتصال أدى إلي توفر كمية كبيرة من البيانات يمكن أن تساهم بصورة مباشرة فى دعم إتخاذ القرار إذا ما تم تحليلها بالصورة المطلوبة، تكمن المشكلة فى أن البنك قيد الدراسة يقوم بعملية تمويل لعملائه ولكنه يواجه بعض المخاطر فى كيفية وفاء العملاء بإسترداد التمويل حسب القيد الزمني المحدد للإسترداد، كذلك لا يستفيد البنك من الكمية الهائلة للبيانات المتوفرة والتي يساعد تحليلها في الحد من مخاطر هذا النوع من التمويل. تهتم هذه الورقة بتطبيق تقنيات تعلم المكائن فى التنبؤ عن مخاطر الإئتمانية في التمويل البنكي. تم إستخدام خوارزمية شجرة القرارات (Tree Decission) و خوارزمية المستقبل متعدد الطبقات (MultiLayer Perceptron) ، أظهرت النتائج أنه ومن خلال إستخدام النظم الذكية يمكن للبنك التنبوء بالمخاطر والإئتمانية مما يمكن البنك من وضع الإحتياطات اللازمة لحل هذه المشكلات. لقياس دقة إكتشاف المخاطر تم إستخدام مصفوفة الارتباك والتي تحتوي معيار الصواب والخطأ و علي صنف الموجب والسالب، تم استخدام شجرة القرارات في عملية التصنيف واستخدمت لإكتساب المعلومات وتقسيم الشجرة للتنبؤ بحدوث الخطر كما أن خوارزمية شجرة القرارات (Tree Decission) حققت نسبة دقة 100% فى إكتشاف المخاطر مقارنة بخوارزمية المستقبل متعدد الطبقات(MultiLayer Perceptron) والتي كانت نسبة دقة نتائجها 97.7064 %, من خلال بيانات رقمية وحرفية بحجم 3847 عميل وتم تنظيف البيانات وذلك بملء البيانات المفقودة في مجموعة البيانات Dataset والتخلص من الحقول الغير ضرورية.واستخدم ايضا خوارزمية التجميع لتجميع عدة بيانات اعتماداً على خصائصها إلى K تجمع، وتتم عمليه التجميع من خلال تقليل المسافات بين البيانات ومركز التجمع.*

الكلمات المفتاحية : تعلم المكائن، تنقيب البيانات، التصنيف

ABSTRACT- The tremendous development in the field of Information and communication technology has led to the availability of a large amount of data that can directly contribute to decision support if analyzed as required, also the bank does not benefit from the vast amount of available data for analysis to help reducing the risk of this type of finance. The problem is that the bank under the investigation faces some financial risks on how the customers recover their funds according to the time limit scheduled for the recovery. This paper is concerned with the application of machine learning techniques in predicting credit risk. The Decission Tree algorithm and the MultiLayer Perceptron algorithm were used. The results showed that, by using intelligent systems, the Bank can predict the risks may occur during the credit given to the customers, enabling the Bank to establish the necessary precautions to overcome these problems. To evaluate the auccuracy of the risk detection the confusion matrix are used, which contains true and false postive and negative measure. Decission Tree algorithm has achieved 100% accuracy in risk detection compared to MultiLayer Perceptron with an accuracy of 97.7064%. Using numeric and character of a professional dataset of 3847 customers. The data was preprocessed by filling in the missing data in the Dataset, also eliminating the unnecessary fields. The assembly algorithm is also used to collect several data depending on their attributes. The assembly process is done by reducing the distances between the data and the assembly center.

المقدمـــة :

من أجل كسب المزيد من العملاء والحفاظ علي العملاء الموجودين في ظل التنافس الكبير بين المؤسسات البنكية، تبحث البنوك عن وسائل وخدمات إضافية تعمل علي تحفيذ العملاء، وبالتالي المحافظة عليهم من الانتقال والتعامل مع مؤسسات أخري منافسة. من ضمن هذه الخدمات تقديم التمويل البنكي للعملاء حسب شروط وقيود معينة يفرضها البنك علي عملائه. ولكن المشكلة فى أن البنك قيد الدراسة يقوم بعملية تمويل لعملائه ولكنه يواجه بعض المخاطر فى كيفية وفاء العملاء بإسترداد التمويل حسب القيد الزمني المحدد للإسترداد، كذلك لا يستفيد البنك من الكمية الهائلة للبيانات المتوفرة والتي يساعد تحليلها في الحد من مخاطر هذا النوع من التمويل.

أدى الانتشار الواسع لتقنية المعلومات وسهولة إتاحتها إلى تضخم حجم المعلومات بصورة استباقية لم يشهدها التاريخ من قبل، مما جعل من قضية البيانات الضخمة على الإنترنت مثاراً للجدل، من حيث جدوى وجودها بهذه الصورة العشوائية. وعندما نتحدث عن البيانات الضخمة، فإننا نتحدث عن كميات لا يمكن تخيلها من البيانات متعددة الأنواع والمصادر بحجم يصل إلى المئات من التيرابايت أو حتى البيتابايت (البيتابايت هو الرقم واحد متبوعا بـ 15 صفر).

من هنا ظهر ما يسمى باستخراج البيانات Data Mining كتقنية تهدف إلى استنتاج المعرفة من كميات هائلة من البيانات، تعتمد على الخوارزميات الرياضية والتي تعتبر أساس التنقيب عن البيانات وهي مستمدة من العديد من العلوم مثل علم الإحصاء والرياضيات والمنطق وعلم التعلم، والذكاء الاصطناعي والنظم الخبيرة، وعلم التعرف على الأنماط ،وعلم الآلة.

وغيرها من العلوم والتي تعتبر من العلوم الذكية وغير التقليدية.ظهر التنقيب في البيانات (Data mining) في أواخر الثمانيات وأثبت وجوده كأحد الحلول الناجحة لتحليل كميات ضخمة من البيانات، وذلك بتحويلها من مجرد معلومات متراكمة وغير مفهومة (بيانات) إلى معلومات قيِّمة يمكن استغلالها و الاستفادة منها بعد ذلك.وقد اجتذبت مرحلة التنقيب في البيانات الكثير من الاهتمام في الأوساط البحثية على مدي العقد الماضي، في محاولة لتطوير خوارزميات قابلة للتوسع والتكيف مع كميات متزايدة من البيانات في البحث عن أنماط معرفية ذات معنى. .[1]

وقد نمت حزم من الخوارزميات والبرمجيات و بشكل كبير خلال العقد الماضي، إلى حد أن التوسع قد جعل من الصعب على العاملين في هذا الحقل تتبع التقنيات المتاحة لحل مهمة معينة

الدراسات السابقة:

سنتناول في هذا الجزء بعض الدراسات التي تم إجراؤها في مجال إدارة المخاطر، و نتطرق إالي الاستثمار العقاري والأزمة المالية العالمية والاقتراحات بمنع حدوثهاعام 2010م: [2] بدءاً من العام 2007م أدى الإنخفاض فى أسعار العقارات وعدم سداد مقرضيتها والاختلاس وحالات الرهن العقارى والغش فيها إلى حدوث مشكلات كثيرة فيما يتعلق بالأوراق المالية المتعلقة بهذه الضمانات والرهن العقارى وبصفة خاصة الأدوات المعقدة التى تعتمد على الرهن العقارى. تهدف هذه الدراسة بصفة أساسية لاستقراء احدث الأدبيات المحاسبية التي تناولت مشاكل القياس والإفصاح المحاسبي عن الاستثمارات العقارية ودور نموذج القيمة العادلة في الأزمة المالية العالمية . تنبع أهمية الدراسة كذلك من أن التقييم السليم للاستثمارات العقارية يعد حجر الزاوية في اتخاذ القرار الاستثماري الملائم .

تقنيات التنقيب عن البيانات واهميتها في ادارة العمليات المصرفية في البنوك الاردنية عام2012 : [3] ويمكن تلخيص مشكلة الدارسة من خلال طرح التساؤلات التالية:

1 – هل أن وجود كم هائل من البيانات والمعلومات المحاسبية و غير المحاسببية المخزنة أو المسترجعة أو المعاد استخدامها في البنوك الاردنية، يتطلب استخدام تقنيات التنقيب عن البيانات لتحقيق ادارة العمليات المحاسبية والمصرفية في البنوك الأردنية؟

2 – هل أن استخدام تقنيبت التنقيب عن البيانات في البنوك الاردنية تؤدي إلى كفاءة ادارة العمليات المحاسبية والمصرفية في البنوك الأردنية؟

تهدف الدارسة إلى التعرف على مدى اهتمام البنوك الأردنية بتكنولوجيا التنقيب عن البيانات ومدى اهميتها في ادارة العمليات المصرفية والمحاسبية في هذه البنوك، وكانت النتائج هي أن ترتيب المجالات التي قد ينظر إليها عند البحث في تطبيق مفاهيم التنقيب عن البيانات والتي تناولتها هذه الدارسة، قد كانت 5 بحسب اهميتها ومستوى الاهتمام بها لدى أفارد مجتمع الدارسة، كالتالي: (نظم البحث كأداة تبادل للمعرفة مع المحيط الخارجي، ثم نظم البحث والتنقيب كأداة

داعمة لإدارة المعرفة، ثم بيئة العمل المعرفي مع تكنولوجيا المعلومات، ثم نظم البحث كأداة تبادل للمعرفة مع المحيط الداخلي، ثم فرص تعزيز نظم المعرفة مع تطوير بيئة نظم البحث، نظم البحث والتنقيب كأداة داعمة للقرار).

دراسة قياسية لتوقع خطر القروض البنكية، 2009م: [4] تتمثل مشكلة الدراسة في إمكانية بناء نموذج قياسي يسمح لنا بتوقع خطر قروض بنك تجاري؟ هدفت الدراسة الي تطوير أساليب الوقاية من خلال السيطرة على خطر القرض في البنك التجاري من خلال عدة أساليب منها الأساليب الوقائية و تسمى بالتسيير الوقائي.

أن منح قرض بنكي يمثل مشكلة كبيرة بالنسبة لأي بنك تجاري، فأي خطأ في منحه قد يجر إلى مشاكل تؤدي حتى إلى إفلاسه لذلك لابد من دراسة معمقة لحالة العميل ماضيا و مستقبلا باستعمال مختلف التقنيات الحديثة كأسلوب وقائي قبل المنح قبل الدخول في الأساليب العلاجية أو التسيير العلاجي و الذي قد يتعب البنك و يوصله إلى حالات ميئوس منها.

تقنيات استخراج البيانات وتطبيقاتها في القطاع المصرفي: [5] أصبح استخراج البيانات مجالا استراتيجيا هاما للعديد من منظمات الأعمال بما في ذلك القطاع المصرفي. وهي عملية تحليل البيانات من وجهات نظر مختلفة وتلخيصها في معلومات قيمة. يساعد استخراج البيانات البنوك على البحث عن نمط مخفي في مجموعة واكتشاف علاقة غير معروفة في البيانات. كانت تقنيات تحليل البيانات المبكرة موجهة نحو استخراج الخصائص الكمية والإحصائية للبيانات. وتسهل هذه التقنيات تفسيرات البيانات المفيدة للقطاع المصرفي لتجنب استنزاف العملاء. الحفاظ على العملاء هو العامل الأكثر أهمية لتحليلها في بيئة الأعمال التنافسية اليوم. كما أن الاحتيال يمثل مشكلة كبيرة في القطاع المصرفي. ومن الصعب اكتشاف الإحتيال ومنعه، تحلل هذه الورقة تقنيات استخراج البيانات وتطبيقاتها في القطاع المصرفي مثل منع الاحتيال والكشف، واحتفاظ العملاء، والتسويق وإدارة المخاطر.

قاضي بشير أحمد (قسم علوم الحاسوب، كلية علوم الحاسوب وتكنولوجيا المعلومات، لاتور، مس)، الهند (وزارة التجارة وتكنولوجيا المعلومات، كلية السيد ، أورانجاباد، (مس)، الهند.[6] في عولمة اليوم وقطاع المنافسة في البنوك هنالك مكافحة من أجل الحصول على ميزة تنافسية. وبصرف النظر عن تنفيذ العمليات التجارية، وإنشاء قاعدة المعرفة واستخدامها لصالح البنك أصبحت أداة استراتيجية للمنافسة. في السنوات الأخيرة زادت القدرة على توليد التقاط وتخزين البيانات بشكل كبير. وقد شهدت الصناعة المصرفية في جميع أنحاء العالم تغييرا هائلا في الطريقة التي يتم بها العمل.

وقد بدأت الصناعة المصرفية تدرك الحاجة إلى تقنيات مثل استخراج البيانات التي يمكن أن تساعدهم على المنافسة في السوق. تستخدم البنوك الرائدة أدوات تعدين البيانات لتجزئة العملاء والربحية وتسجيل الائتمان والموافقة عليها والتنبؤ بالتخلف عن السداد والتسويق واكتشاف المعاملات الاحتيالية وما إلى ذلك.

تعريف المشكلة:

تُعد عمليات الرهن من أهم الاخطار الائتمانية الواجب البحث فيها لضمان ديمومة وكفاءة العمليات المصرفية المنفذة والمعمول بها لدى البنوك ومعرفة أهميةالحد من خطر الرهونات الائتمانية التي لها دور في تأخير المعلومات اللازمة لتلبية احتياجات المستخدم الداخلي والخارجي للبنوك، والتي قد يلزم الإبقاء على تخزينها لفترا ت طويلة لاسترجاعها وفقاً لأنواعها أو مصادر الحصول عليها أو صفات استخدامها أو بأي شكل من الإشكال التي قد تتطلب البحث في كيفية الاستخدام الفعال لها لغايات التعرف على مختلف المتغيرات فيما بينها وكيفية تأثيرها على عمل البنوك وعملياتها الائتمانية.

وبرزت أهمية دراسة تقنيات التنقيب عن البيانات في إدارة العمليات المصرفية ، اعتقادا من الباحث في أن أهمية التنقيب عن البيانات بمثل أهمية البيانات والمعلومات اللازمة لتنفيذ العمليات المصرفية والائتمانية في البنوك.

الهدف من الدراسة:

تهدف الدراسة إلي إستخدام و تطبيق خوارزميات تعلم المكائن وبناء وترشيح النموذج المناسب للتنبؤ بإكتشاف حدوث خطر يهدد البنك بالتعرف علي الخصائص التي تشير الي الخطر القادم علي البنك،, و تتثمل المخاطر في عدم إيفاء العملاء بسداد ما عليهم من التزامات والذي بدوره يحدث إختلالا في أداء البنك بسبب تعثر الوفاء بالالتزامات من قبل المدينين. وبالتالي فإن إستخدام تقنيات التنقيب عن البيانات و بالإستدلال بهذه النماذج من خلال البيانات المتاحة، تمكن متخذي القرار من التعرف علي الخطر في فترة زمنية معقولة قبل حدوث الخطر والسبب الرئيسي من هذه الدارسة هو الحفاظ علي البنك من الخطر المستقبلي.وتم استخدام شجرة القرارات و المستقبل متعدد الطبقات كنماذج للتنبؤ من خلال تصنيف بيانات العملاء

الموجودة كنوع التمويل، قيمة التمويل، الفترة المحددة للسداد، والتي يؤدي تحليلها لمساعدة البنك في التنبؤ واكتشاف الخطر.

المنهجية:

لدي البنوك قواعد بيانات ضخمة تقوم بتوليد كمية ضخمة من البيانات ويتم توليد البيانات بواسطة مجموعة من العمليات والإجراءات المختلفة لكل (عميل) وبحسب سياسات البنك المتبعة.

المرحلة (1) تعريف المشكلة problem identification:

المشكلة هي كيفية إستخدام تقنيات التنقيب عن البيانات في البنك لتحسين العمل الوقائي وإكتشاف الخطر بإستخدام مجموعة بيانات (Dataset) تتعلق بالعملاء ثم تطبيقها علي كل الأقسام بعد نجاح النموذج، ولابد أن نشير إلي أن القيد الزمني للتعرف علي ما يشكله العميل من خطر علي البنك يختلف من عميل إلي آخر لان عملية التمويل في حد زاتها تختلف من عميل إلي آخر وحسب المعطيات وحسب القيد الزمني المسموح به لكل عميل علي، ولكن يمكن تطبيق عمليات التنبؤ الاولية من خلال البيانات الاساسية للعميل كمرحلة أولي وكذلك عند إنقضاء نصف مدة التمويل.

المرحلة (2) جمع البيانات Data collection:

إستخدم البحث بيانات علي مدي عام واحد مع تحديثها بإستمرار ومقارنتها مع عدة بنوك وتم التركيز علي بنك العمال، يعتمد البحث علي الأوصاف النصية للكشف عن الخطر من كتب المخاطر المتعلقة بالبنوك من مهارات الموظفين وموظف المراجعة والتفتيش [8-7] .

المرحلة (3) إختيار البيانات Data selection:

خوازرميات التنقيب عن البيانات تتطلب مجموعة بيانات Dataset مدخلات والتي تحتوي علي متجه Vector لقيم الخصائص في هذه المرحلة يتم إختيار البيانات وفقاً لتوجهيات موظف المراجعة والتفتيش، جدول (2) يوضح نموذج لبعض البيانات المختارة في عملية التنبؤ

الإختيار يتم من البيانات الخام (الأصلية Raw Data) التي تم جمعها من البنك ويتم إختيار العمليات التي قد تتسبب بصورة متكررة بالخطر القادم من العميل.

المرحلة (4)المعالجة المسبقة للبيانات Data Preprocessing:

في هذه الخطوة يتم تنظيف البيانات وذلك بملء البيانات المفقودة في مجموعة البيانات Dataset والتخلص من الحقول الغير ضرورية، التعامل مع البيانات المزعجة تم إستخدام عدد من الأدوات في هذا البحث للقيام بعمل تنظيف البيانات.

تسمية البيانات Data labeling:

العملية التنبؤية لكشف الخطر يمكن أن تعتبر كمهمة تصنيف Classification Task تحتوي علي صنفين:

- إكتشاف خطر تأخذ القيمة NO .
- عدم وجود خطر تأخذ القيمة YES.

نموذج التصنيف يقوم كل إجراء جديد للعميل في واحد من الصنفين السابقيين.

المرحلة (4) تحديد المهام والتقنيات:

1/ خوارزمية (J48) Decision Tree :

عملية التصنيف باستخدام شجرة القرارات تستخدم إكتساب المعلومات (Information Gain) لتقسيم الشجرة.

الخطوه الأولى هى إكتساب المعلومات لكل خاصية (Attribute)الخاصية ذات أكبر كمية معلومات مكتسبة ستكون عقدة الجذر (Root Node) لشجرة القرارات. [9]

وتهدف تقنية شجرة القرار إلى تقسيم قاعدة البيانات بهدف معين سبق وأن تم تحديده، ويصبح وجود عنصر معين في إحدى المجموعات، وهي ممثلة هنا بالفروع، هو نتيجة لأنه حقق سلسلة الشروط الموضوعة وصولاً إلى هذا الفرع وليس فقط لأنه يشبه بقية عناصره، بالرغم من أنه لم يتم تعريف التشابه في هذه الحالة.

إن شجرة القرار والخوارزميات التي تستخدم لإنتاجها يمكن أن تكون معقدة ولكن النتائج التي تؤدي لها يمكن إظهارها بشكل مبسط وسهل الفهم وبفائدة عالية المستوى. [9]

2/ خوارزمية البيرسبترون Perceptron: [11]

شبكه الـPerceptron تعد من أقدم وأسهل أنواع الشبكات العصبيه ، وهي نوع مبسط من ال Feed-Forward Neural Network حيث هناك نوع منها يحتوي على طبقه واحده Single Layer والأخر يحتوي على أكثر من طبقه Multi-Perceptron Layer واختصارا MPL. وبشكل عام مهمه هذا النموذج هي في التصنيف Classification.

وشبكه الSingle Layer Perceptron تسمى عاده linear classifier بمعنى أن هذه الشبكه تحل المشاكل التي يمكن فصلها بشكل خطي linearly separable.

3/ خوارزمية K-means Clustering K-Means:

تستخدم هذه الخوارزمية لتجميع عدة بيانات (أمثلة) اعتماداً على خصائصها إلى K تجمع، وتتم عمليه التجميع من

خلال تقليل المسافات بين البيانات ومركز التجمع (clustercentroid). [12] ,واستخدمت خوارزمية التجميع للوصول الى (Attribute) من Datasetالتي تم استخدامها في شجرة القرارات للتنبؤ بموقف العميل إذا كان يشكل خطر على البنك او لا.

المرحلة (5) التطبيق(Implementation)

فى هذه المرحلة يتم إستخدام مجموعة البيانات التى تم الحصول عليها فى مرحلة تسمية البيانات (Data Labeling) لتعليم النموذج المرغوب ، هنالك العديد من تقنيات تعدين البيانات متوفرة لإستنتاج (Infer) النموذج الذى نريده هذه التقنيات تشمل شجرة القرارات إعتماداً على نوع التقنية، قد تحتاج إلى خطوة ما قبل المعالجة (Preprocessing) متقدمة و تكون على النحو التالى:

أ. إختيار مجموعة البيانات الفرعية المناسبة للخصائص (Attributes).

ب. تطبيع .(Normalization) الخصائص الأولية.

ج. إنشاء خصائص جديدة من البيانات الإبتدائية أو تجزئة (تقطيع) (Discretize) الخصائص المستمرة.

تم إقتراح مناهج (Approaches) لتنفيذ هذه المهام فى مجالات مختلفة مثل تعلم الآلة (Machine Learning) ، الأحصاء (Statistics) التعرف على الأنماط (Pattern Recognition)

وتعدين البيانات. لتحسين وقياس أداء النماذج المستخدمة في التنبوء نجد أن أهم مقياس للأداء هو دقة التنبؤ الذي يتم الحصول عليه بعد إختبار النماذج من خلال مجموعة البيانات المستخدمة فى الدراسة.

الدقة تعرف باستخدام خطأ التنبؤ (Forecast Error) وهو الفرق بين القيم الحقيقية والقيم التى تم التنبؤ عنها هنالك مقياس آخر هو الحساسية (Sensitivity) و النوعية (Specificity).

المرحلة (6) التقييم(Evaluation)

يتم تعريف مقاييس التقييم فى مسائل التصنيف من مصفوفة تحتوى على أعداد من الأمثلة تم تصنيفها بطريقة صحيحة وغير صحيحة لكل صنف تسمى هذه المصفوفة بمصفوفة الإرتباك (Confusion Matrix) ، تحتوي مصفوفة الإرتباك على صنفين فقط هما سالب و موجب (Positive and Negative) كما موضح في الجدول رقم (1).

جدول (1) مصفوفة الإرتباك (Confusion Matrix)

	Predicted Class	
True class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

من الجدول(1) يمكننا التعرف على المصطلحات التالية [11]:

أ. False Negative: هي الأمثلة التى تم التنبؤ عنها على أنها موجبة وهى ضمن الصنف السالب.

ب. False Positive: هى الأمثلة التى تم التنبؤ عنها على أنها سالبة وصنفها الحقيقي هو موجب.

ج. True Positive:هى الأمثلة التى تم التنبؤ عنها على أنها موجبة وتنتمي لصنف

د. موجب .True Negative:هى الأمثلة التى تم التنبؤ عنها على أنها سالبة وتنتمي لصنف سالب.

مقياس الإستخدام الأكثر إستخداماً هو معدل الدقة (ACC) يقوم بتقييم كفاءة (Effectiveness) المصنف (Classifier) بواسطة النسبة المئوية للتنبؤات الصحيحة ، المعادلة (1) توضح كيف يتم حساب معدل الدقة (ACC) .

$$Acc = \frac{|TN| + |TP|}{|FN| + |FP| + |TN| + |TP|} \quad (1)$$

Equation (1): Accuracy

هنالك مقياس آخر مكمل لمعدل الدقة هو معدل الخطأ (Error Rate) كما موضح في المعادلة (2) والذى يقوم بتقييم المصنف عن طريق النسبة المئوية للتنبؤات غير الصحيحة، معدل الدقة ومعدل الخطأ هى مقاييس يمكن أن تستخدم في مسائل التصنيفات متعددة الأصناف (Multi Classes Classification Problems).

$$Err = \frac{|FN| + |FP|}{|FN| + |FP| + |TN| + |TP|} \quad (2)$$

Equation (2): Error Rate

يقوم مقياس الـ Recall ومقياس النوعية (Specificity) بتقييم فعالية المصنف لكل صنف في المسائل الثنائية .يسمى (Recall) أيضاً بالحساسية (Sensitivity) أو المعدل الموجب الصحيح True Positive Rate هو الجزء من الأمثلة التى تنتمى إلى المصنف الموجب والتى يتم

التنبؤعنها بصورة صحيحة على أنها موجبة. النوعية (Specificity) هى النسبة المئوية التى يتم التنبؤ عنها بصورة صحيحة على أنها سالبة.

المعادلتان (3) و (4) توضحان طريقة حساب (Recall) و(Specificity).

$$R = \frac{|TP|}{|TP| + |FN|} (3)$$

Equation (3): Recall

$$Spe = \frac{|TN|}{|FP| + |TN|} (4)$$

Equation (4): specificity

6- مجموعة بيانات (دراسة الحالة): Data set (case study):

تم إستخدام مجموعة بيانات التنبؤ للمخاطر الإئتمانية والتى تم الحصول عليها من بنك العمال، تم جمع بيانات العام 2017، والتي تحتوي على كل بيانات العميل بالإضافة إلى البيانات المسجلة أثناء الإجراءات التي تخص كل عميل في كل مراحلها. هذه البيانات يتم الحصول عليها بعد الرجوع الي ذوي الخبرة في المجال المعين لتحديد الحقول التي تعتمد عليها عملية التعدين

عملية تنظيف البيانات Preprocessing:

سوف يتم إستبعاد عدد من الحقول و تتم إضافة حقل آخر وهو حقل التعدين يوجد خطر/ لا يوجد خطر.

7- التنفيذ والتطبيق العملي:

1. إعداد البيانات:

وتم تنظيف البيانات من أجل تحسين جودة التجمعات، يتم تحويل القيم الرقمية إلى نطاقات تم تحويل القاعدة الى ملف بامتداد csv .

2. معالجة البيانات :

تم حذف الحقول الغير ضرورية ذات صلة مباشرة بالتحليل. هناك العديد من تقنيات معالجة البيانات. تنظيف البيانات المطبقة لإزالة الضوضاء والتناقضات الصحيحة في البيانات. وهذا يمكن أن يحسن دقة وكفاءة خوارزميات التعدين.

3. خوارزمية Clustering K-Means:

تستخدم هذه الخوارزمية لتجميع عدة بيانات (أمثلة) اعتماداً على خصائصها إلى K تجمع، وتتم عمليه التجميع من خلال تقليل المسافات بين البيانات ومركز التجمع (clustercentroid).

يوضح الجدول رقم(2) نتيجة حساب المتوسط للخاصية lon_class لكل تجمع من البيانات للكشف عن الارتباطات داخل كل تجمع التي ادت الى التعرف الى اسباب الارتباطات مثلا داخل cluster 0 اكثر التجمعات داخل الخاصية (sub_sec) =4 وهي شفرة تدل الى المهن الصناعية والحرفية , واكثر التجمعات داخل الخاصية (type –invest)=2 وهي شفرة تدل قرض استثماري , واكثر التجمعات داخل الخاصية (area–creadit)=1 وهي شفرة تدل الى المناطق الخالية من الحروب وهكذا , مما أدى ذلك التعرف الى حالة العميل من خلال متوسط اكثر التجمعات داخل كل خاصية وكانت النتيجة هي(clint_type) =yes.

يوضح الجدول رقم(3) نتيجة حساب المتوسط للخاصية lon–type لكل تجمع من البيانات للكشف عن الارتباطات داخل كل تجمع و التعرف على اسباب الارتباطات مثلا داخل cluster 0 اكثر التجمعات داخل الخاصية (sub_sec) =4 وهي شفرة تشير الى المهن الصناعية والحرفية , واكثر التجمعات داخل الخاصية (type–invest)=2 وهي شفرة تشير قرض استثماري , واكثر التجمعات داخل الخاصية (area–creadit)=1 وهي شفرة تشير الى المناطق الخالية من الحروب وهكذا , مما أدى ذلك التعرف الى حالة العميل من من خلال متوسط اكثر التجمعات داخل كل خاصية وكانت النتيجة هي(clint_type) =no.

يوضح الجدول رقم (4) نتيجة حساب المتوسط للخاصية creadit_area لكل تجمع من البيانات للكشف عن الارتباطات داخل كل تجمع التي ادت الى التعرف الى اسباب الارتباطات مثلا داخل cluster 0 اكثر التجمعات داخل الخاصية (sub_sec) =1 وهي الى خدمات النقل , واكثر التجمعات داخل الخاصية (type _ invest)=2 وهي الى التمويل الاصغر, واكثر التجمعات داخل الخاصية (lon_type)=1 وهي الى المناطق الحضارية او الخالية من الحروب وهكذا , مما أدى ذلك التعرف الى حالة العميل من

خلال متوسط اكثر التجمعات داخل كل خاصية وكانت النتيجة هي(clint_type) =no.

يوضح الجدول رقم(5) نتيجة حساب المتوسط للخاصية clint_type لكل تجمع من البيانات للكشف عن الارتباطات داخل كل تجمع التي ادت الى التعرف الى اسباب الارتباطات مثلا داخل 0 cluster اكثر التجمعات داخل الخاصية (sub_sec) =1 وهي خدمات النقل, واكثر التجمعات داخل الخاصية (invest – type)=2 وهي طريقة الدفع بالتقسيط , واكثر التجمعات داخل الخاصية (creadit – area)=1 وهي المناطق الحضارية او الخالية من الحروب وهكذا .

كذلك تم إستخدام خوارمية K–Mean في عملية التحليل جدول رقم (6) و بمقارنة التجميع بين التكرارت ، نلاحظ أن التجمعات من خلال المتوسط. هذا يعني أن عملية الحسابات في خوارزمية k–means clustering وصلت إلى حالة الاستقرار، وهذا يعني أن هذه الخوارزمية لم تعد بحاجة إلى المزيد من التكرار، وبالتالي حصلنا على النتيجة النهائية للتجميع.

6- النتائج :

للحصول علي النتائج المناسبة والتي تمكن البنك من تدارك المخاطر تم تنفيذ النماذج بإستخدام خوارزمتين مختلفتين

1- تنفيذ النموذج بإستخدام خوارزمية المستقبل متعدد الطبقات :Multilayerperceptron

ندخل مجموعة البيانات (Training andTesting) على هيئة ARFF ، وذلك من خلال التعامل مع خوارزمية المستقبل متعدد الطبقات (Multilayerperceptron) لتقوم بتصنيف مجموعة البيانات إلى قسمين هما :الأول هل هنالك خطر من العميل والثانى هو لا يوجد خطر من العميل.

في الجداول رقم (7) و (8) علي التوالي تم بناء النماذج في تجربة بإستخدام مجموعة بيانات تدريب مختلفة ومجموعة بيانات إختبار مختلفة في التجربة لتقدير أفضل تصنيف لحساب الخطأ لبيانات التدريب ودقة المصنف من مجموعة البيانات لإختبار ومن ثم قياس هذه الخوارزمية خلال التجربة بواسطة الدقة (Accuracy) ، الحساسية (Sensitivity) و النوعية (Specificity) ، وتم حسابها عن طريق مصفوفة الإرتباك (Confusion Matrix) وذلك للخوارزميتين معا.

جدول (7) خوارزمية المستقبل متعدد الطبقات Multilyer Perceptron

2- تنفيذ النموذج باستخدام خوارزمية شجرة القرارات ((DecisionTree(J48):

فى هذا النموذج تم إستخدام خوارزمية شجرة القرارات ((Decision Tree(J48) ، وذات مجموعة البيانات ، نفس خطوة المعالجة المسبقة التى استخدمت في الخوارزمية السابقة بالإضافة إلى نفس مقاييس الأداء.

3- مقارنة خوارزمية المستقبل متعدد الطبقات Multilyer Perceptron مع خوارزمية شجرة القرارات ((DecisionTree(J48):

فى جدول (9) يوجد فرق بين الخوارزميتين لصالح شجرة القرارات (J48) حيث أن متوسط الدقة هو الأعلى وهو(100) و متوسط perseptron هو (97.7064) و الفرق هو 2.2936لصالح شجرة القرارات (J48).

فـي جـدول (10) أعـلاه يمكـن شـرح تفاصـيل التحليـل كمـا يلي :-

- **Kappa statistic**

هو مقياس يقارن بين الدقة الملحوظة والدقة المتوقعة

- **Mean absolute error**

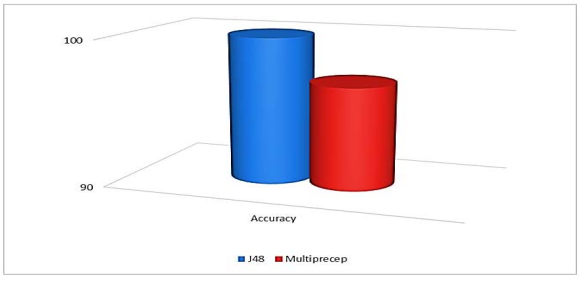
متوسـط الخطـأ المطلـق بـين القيمـة المرصـودة (العينـة) والمتوقعة

- **Rote mean squared error**

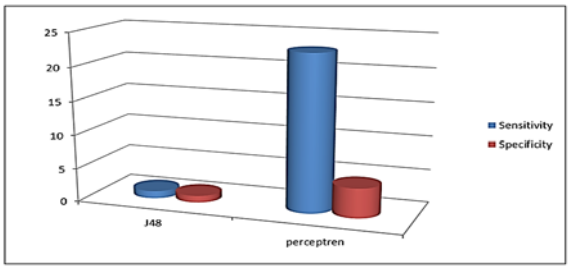
قياس الفروق بين القيمة (العينة) التي يتنبأ بها النموذج والقيم التي لوحظت بالفعل.

- **Relative absolute error**

شكل رقم (1) يوضح المقارنة الدقة بين خوارزمية (J48) مع خوارزمية Perceptron

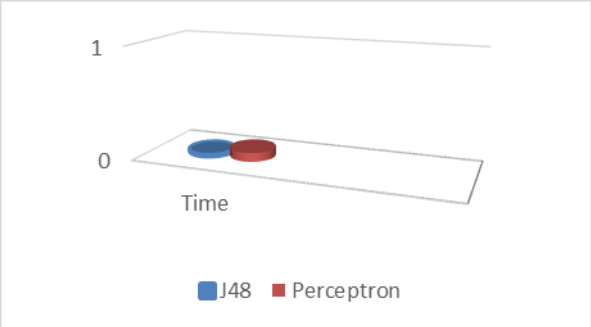


في الشكل رقم (1) يوجد فرق بين الخوارزميتين لصالح شجرة القرارات (J48) حيث أن معدل الاخطاء هو الاقل مقارنة مع معدل اخطاء perseptron .



شكل رقم (2) يوضح مقارنة الحساسية والنوعية بين خوارزمية (J48) مع خوارزمية Perceptron

في الشكل رقم (2) يوجد فرق بين الخوارزميتين لصالح شجرة القرارات (J48) حيث أن الحساسية والنوعية هو الاقل مقارنة مع الحساسية والنوعية perseptron .



شكل رقم (3) يوضح الحساب الزمني للتنبوء بين خوارزمية (J48) مع خوارزمية Perceptron

يوجد فرق بين الخوارزميتين لصالح شجرة القرارات (J48) حيث أن معدل الزمن الذي يستهلك بواسطة المعالج هو الاقل مقارنة مع معدل الزمن المستغل بواسطة المستقبل متعدد الطبقات Multi layer perseptron.

7- النتائج :

من التجربتين التى استخدمت فى الفصل السابق ، تم إستخدام تقنيات مختلفة في عملية التصنيف هى شجرة القرارات (J48) Decision Tree ، والمستقبل متعدد الطبقات perceptron ، مقاييس الأداء التى استخدمت هى الدقة وهى النسبة المئوية لمجموعة صفوف الإختبار التى تم تصنيفها بصورة صحيحة ، الحساسية وهى الجزء من الصفوف الموجبه التى تم تصنيفها بصورة صحيحة والنوعية Specificity وهى الجزء من الصفوف السالبة والتي يتم تصنيفها بصورة صحيحة . الجدول رقم (7) و (8) علي التوالي يوضحان نتائج التجربتين و تقنيات التصنيف حيث يوضح نتائج مختلفة تعتمد على طبيعة وحجم الخصائص . خوارزمية التصنيف التى تشير الى أعلى يتم أختبارها على أنها أفضل تقنية تصنيف لمجموعة البيانات , نجد أن شجرة القرارات (j48) هى الأفضل لتحقيق مهمة التصنيف وذلك لأن الدقة accuracy 100 %، والحساسية (Sensitivit) 1.089% النوعية (Specificity).1.000%

و يمكن تلخيص النتائج على النحو التالى :

1. يمكن التنبؤ عن حالة العميل بدرجة دقة عالية و قبل فترة كافية .
2. تقنية التصنيف الأفضل للتنبؤ في حالة العميل هي شجرة القرارات (j48) مع عدم وضع زمن التنفيذ فى الأعتبار (Efficiency) .
3. بعد إجراء المقارنة بين تقنية شجرة القرارات (j48) و perceptron ، نجد أن شجرة القرارات هى الأفضل ويرجع إلى أن الدقة العالية و الحساسية العالية أفضل من النوعية .
4. إذا أخذنا مقياس آخر يعتمد على زمن التنفيذ أى يعتمد على الكفاءه (Efficiency) . يمكن أن تجد تقنية أخرى هى الأفضل من شجرة القرارات ستكون سريعة ولكن بدقة أقل قد تفضل فى بعض الأحيان السرعة على الدقة لتعطيها تنبؤ سريع حالة العميل وبأقل دقة وعليه يمكن أن نتخذ القرار المناسب قبل حدوث خطر.
5. نلاحظ فى الجدول (10) اختلاف كبير فى قيم التحليل الاحصائي لمقاييس الأداء فى كل التقنيات هذا يعزى إلى أن بعض التقنيات كانت تعانى من البيانات المزعجة (NoisyData) أثناء خطوه ماقبل المعالجة (Preprocessing).[13] .

خلاصة المقارنات

تمت المقارنة بين الخوارزميات و بين مقاييس الأداء و بعدة طرق في الفقرات السابقة و من خلال جميع المقارنات.

لا نجد فروقات ذات دلالة إحصائية بين خوارزمية شجرة القرارات (J48) و خوارزمية perceptron.

عليه من كل هذه المقارنات التي تمت و بطرقها المختلفة نجد أن الخوارزمية الأفضل للتنبؤ بمعرفة حالة العميل (خطر,عدم وجود خطر) هي الخوارزميتان خوارزمية شجرة القرارات (J48) و خوارزمية perceptron وذلك لأنهما قدمتا نتائج أفضل .

لكن عند المقارنة بين خوارزمية شجرة القرارات (J48) و خوارزمية perceptron لإختيار الأفضل نجد أن الخوارزمية الأفضل للتنبؤ بمعرفة حالة العميل (خطر,عدم وجود خطر) هي خوارزمية شجرة القرارات (J48) و ذلك للأسباب التالية :

أ. خوارزمية شجرة القرارات بسيطة الفهم والتفسير .

ب. في خوارزمية شجرة القرارات يمكن تحديد أسوأ وأفضل التوقعات لقيم السيناريوهات المختلفة و هذا لا يتم في خوارزمية perceptron.

جـ. خوارزمية perceptron تحتاج لذاكرة كبيرة و زمن أكبر للقيام بعملية المطابقة بين الحالات بينما خوارزمية شجرة القرارات لا تحتاج لذلك.

د. خوارزمية شجرة القرارات تتعامل مع البيانات المزعجة Noisy Data بصورة أفضل من خوارزمية perceptron و هذا يزيد من الدقة في خوارزمية شجرة القرارات شجرة .

هـ. خوارزمية شجرة القرارات تعمل بطريقة فرق تسد (Divide and Conquer) التي تقوم على مبدأ تقسيم المشكلة إلى مشاكل فرعية ثم تجميع الحلول (التجزئة و الضم) و هي من أفضل طرق تصميم الخوارزميات. أما خوارزمية perceptronتعمل بطريقة التعليم المتكاسل (Lazy Learnig).

و. بالعكس من خوارزمية perceptron نجد أن خوارزمية شجرة القرارات لها قدرة عالية في عملية إختيار الخصائص أو المتغيرات (Feature Selection or Variable Screening).

ز. في خوارزمية شجرة القرارات يبذل المستخدم مجهود أقل من الجهد الذي يبذل في خوارزمية perceptron في تجهيز البيانات(Data Preparation).

ح. خوارزمية شجرة القرارات لا تتأثر بالعلاقات غير الخطية بين المتغيرات عكس خوارزمية (Nonlinear RelationshipsBetween Variables perceptron) و هذا يعني أنه يمكن إستخدام خوارزمية شجرة القرارات مع العلم أن المتغيرات غير مرتبطة ببعضها خطياً (the parameters are nonlinearly related

الخاتمة

ضمن الخدمات التي تقدمها البنوك لعملائها، تمويل العملاء ولكنها تواجه بعض المخاطر فى كيفية وفاء العملاء بإسترداد التمويل حسب القيد الزمني المحدد للإسترداد، تهتم هذه الورقة بتطبيق تقنيات تعلم المكائن فى التنبؤ عن مخاطر الإئتمانية. تم إستخدام خوارزمية شجرة القرارات (Tree Decission) و خوارزمية المستقبل متعدد الطبقات (MultiLayer Perceptron) للتنبوء الائتمانية للبنوك وتم تطبيق الدراسة علي عملاء بنك العمال الوطني، أظهرت النتائج أنه ومن خلال إستخدام النظم الذكية يمكن للبنك التنبوء بالمخاطر الإئتمانية مما يمكن البنك من وضع الإحتياطات اللازمة لحل هذه المشكلات، مع العلم أن مجموعة البيانات (Dataset) تحتوي علي كم هائل من البيانات المرتبطة ببعضها البعض والتي تحلل إحصائيا عبر خوارزميات مخصصة تعمل علي تدريب الخوارزميات و من ثم تجعلها قادرة علي التصنيف والتنبوء بأي بيانات مشابهة غير مصنفة، أثبتت الدراسة أن خوارزمية شجرة القرارات (Tree Decission) حققت نسبة دقة 100% فى إكتشاف المخاطر مقارنة بخوارزمية المستقبل متعدد الطبقات(MultiLayer Perceptron) والتي كانت نسبة دقة نتائجها 97.7064 %.

المراجع:

[1] الكاشف، محمود يوسف (2000) ، "مدخل مقترح لتطوير دور المعلومات المحاسبية في إطار المفهوم المتكامل للجودة الشاملة"، مجلة الإدارة العامة، المجلد 94 ، العدد 3، اكتوبر.

[2] محمد محسن عوض مقلد، دراسة باسم الاستثمار العقاري والأزمة المالية العالمية والاقتراحات بمنع حدوثها عام 2010م.

[3] عبدالرازق الشحادة، المؤتمر العلمي السنوي جامعة الزيتونة الأردنية، دراسة باسم تقنيات التنقيب عن البيانات واهميتها في ادارة العمليات المصرفية في البنوك الاردنية عام2012م.

[4] شيخي محمد – جامعة ورقلة ، و بن قانة اسماعيل – جامعة ورقلة، دراسة قياسية لتوقع خطر القروض البنكية، 2009م.

[5] عبدالرزاق الشحادة، تأثير تطبيق تقنيات التنقيب عن البيانات في إدارة العمليات المصرفية (دراسة ميدانية في البنوك التجارية الأردنية) مجلة جامعة دمشق للعلوم الاقتصادية والقانونية – المجلد 29 –العدد الثاني–2013

[6] د. قاضي بسير أحمد (قسم علوم الحاسوب، كلية علوم الحاسوب وتكنولوجيا المعلومات، لاتور، مس)، الهند (وزارة التجارة وتكنولوجيا المعلومات، كلية السيد ، أورانجاباد، (مس)، الهند.

[7] هيا مروان ابراهيم لظن. مدى فاعلية دور التدقيق الداخلي في تقويم إدارة المخاطر وفق إطار COSO الجامعة االسالمية – غزة. رسالة ماجستير – 2016

[8] سلطة النقد الفلسطينية. دليل القواعد والممارسات الفضلى لحوكمة المصارف في فلسطين. 2017

[9] (دراسة تطبيقية على القطاعات الحكومية في قطاع غزة)

[10] التحليل المتقدم وتنقيب البيانات، البحث العلمي العصري"، د. مصطفى عبيد، (2017)، شركة أمازون العالمية-

[11] Avuçlu, Emre, and Fatih Baçiftçi. "New Approaches to determine Age and Gender in Image Processing Techniques using Multilayer perceptron network." AppliedsoftComputing (2018)

[12] Tan, p.n (2006).introduction to data mining.Pearson education India Chapter8.http//www.users.cs.umn.edu/Kumar/dmbook/index.php

[13] Singh, Poornima, Sanjay Singh, and Gayatri S Pandi-Jain. "Effective Heart Disease Prediction System Using Data Mining Techniques." International Journal of Nanomedicine 13 (2018): 121-124. PMC. Web. 31 Aug. 2018.

[14] Suthar, Trilok, Digvijaysinh Mahida, and Pinkal Shah. "Review of Data Pre-processing Techniques for Classification." 2nd International Conference on Current Research Trends in Engineering and Technology Volume 4 Issue 5 (2018).

جدول رقم (2) يوضح نتيجة حساب المتوسط للخاصية Lon class

Attribute	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
lon_reqamt	2458740.1111	63119.9041	84329.5758	493550.9243	183974.8565
Profit_rate	469516.4106	12548.4553	20585.5854	90920.2673	33065.5831
sub_sec	4	4	1	1	4
lon_type	3	1	2	3	1
secur_val	2488316.9013	67081.9436	84565.3875	499223.2472	172146.4822
overdue_ amount	61955.725	11787.0838	361.4855	12233.7535	177.3707
invest_type	2	2	2	2	2
credit_area	1	1	1	1	1
npa_amount	36998.7182	9932.6597	239.3673	10608.0305	85.0902
branch	1	1	11	3	17
client_type	Yes	No	Yes	yes	Yes

الجدول (3) نتيجة حساب المتوسط للخاصية (lon_ type))

Attribute	Cluster 0	Cluster 1	Cluster 2
lon_reqamt	1548417.8952	140095.3456	249556.7362
profit_rate	291214.0868	26861.0047	48800.4857
sub_sec	4	1	1
Lon_class	2	5	2
secur_val	1575929.2196	141073.5871	246931.5409
overdue_amount	61955.725	11787.0838	361.4855
invest_type	2	2	2
credit_area	1	1	1
npa_amount	43186.3789	16077.2119	28.6755
branch	1	7	11
client_type	No	No	Yes

الجدول (4) نتيجة حساب المتوسط للخاصية (creadit_area)

Attribute	Cluster 0	Cluster 1
lon_reqamt	217903.0803	399446.9004
profit_rate	40982.6011	76975.2814
sub_sec	1	1
Lon_class	5	2
lon_type	1	1
secur_val	221572.1307	399657.8058
overdue_amount	47814.9744	124.749
invest_type	2	2
npa_amount	33914.9049	78.7199
branch	1	11
client_type	No	Yes

الجدول رقم (5) نتيجة حساب المتوسط للخاصية Client–type

Attribute	Cluster 0	Cluster 1
lon_reqamt	1121317.5102	107520.2482
profit_rate	212059.1126	21837.6612
sub_sec	1	1
Lon_class	2	2
lon_type	3	1
secur_val	1136728.5231	103419.6242
overdue_amount	32119.9138	1236.5906
invest_type	2	2
Credit_area	1	1
npa_amount	22547.2116	946.7035
branch	1	11

جدول رقم (6) التحليل بواسطة K-Mean

Classes to clusters evaluation attribute	No. of clusters	Cluster Instances	No. of Iterations	Within clusters sum of squared errors	Time taken to build model	Correctly clustered instance	Incorrectly clustered instances
Loan-class	5	311 (8%) 636 (17%) 1249 (32%) 618 (16%) 1032 (27%)	8	5373.909920439123	0.28 seconds	36.5575	63.4425
Loan-type	3	385 (10%) 501 (13%) 2960 (77%)	6	5929.436290908403	0.02seconds	43.0837	56.9163
Credit-area	2	723 (19%) 3123 (81%)	5	8271.998137145973	0 seconds	77.1212	20.8788
Client-type	2	978 (25%) 2868 (75%)	6	7539.080450666685	0.02 seconds	65.0806	34.9194

	NO. of Instances	Percentage
Correctly Classified Instances	1278	97.7064 %
Incorrectly Classified Instances	30	2.2936 %

Accuracy=97.7064 % Sensitivity= 23% Specificity= 1.215%

جدول (8) خوارزمية شجرة القرارات J48

	NO. of Instances	Percentage
Correctly Classified Instances	1308	100 %
Incorrectly Classified Instances	0	0 %

Accuracy= 100% Sensitivity= 1.089% Specificity= 1.000%

جدول (9) مقارنة خوارزمية المستقبل متعدد الطبقات Multilyer Perceptron مع خوارزمية شجرة القرارات (DecisionTree(J48)):

Method	Accuracy (%)	precision	Recall	f-easure	Computation all time
J48	100 %	1.000	1.000	1.000	0.02
perceptron	97.7064 %	0.961	0.900	0.930	0.08

جدول (10) يوضح بعض التحاليل الإحصائية

Method	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Coverage of cases (0.95 level)	Mean rel. region size (0.95 level)	Total Number of Instances
J 48	1	0	0	0%	0%	100%	50%	1308
perceptron	0.9159	0.0347	0.1343	11.5413%	35.7771 %	99.9235 %	55.4281 %	1308