

Predicting Bank Loan Risks Using Machine Learning Algorithms

Maan Y. Alsaleem

maanyounis1983@gmail.com

Directorate of Education in Nineveh
Tenahî, Duhok, Ramy Land B9, Iraq

Safwan O. Hasoon

dr.safwan1971@uomosul.edu.iq

College of Computer Science and Mathematics
University of Mosul, Mosul, Iraq

Received on: 10/03/2020

Accepted on: 25/03/2020

ABSTRACT

Bank loans play a crucial role in the development of banks investment business. Nowadays, there are many risk-related issues associated with bank loans. With the advent of computerization systems, banks have become able to register borrowers' data according their criteria. In fact, there is a tremendous amount of borrowers' data, which makes the process of load management a challenging task. Many studies have utilized data mining algorithms for the purpose of loans classification in terms of repayment or when the loans are not based on customers' financial history. This kind of algorithms can help banks in making grant decisions for their customers. In this paper, the performance of machine learning algorithms has been compared for the purpose of classifying bank loan risks using the standard criteria and then choosing (*Multilayer Perceptron*) as it has given best accuracy compared to *RandomForest*, *BayesNet*, *NaiveBayes* and *DTJ48* algorithms.

Keywords: Bank loans, machine learning algorithms, Multilayer Perceptron, classification, accuracy, ROC.

التنبؤ بمخاطر القروض المصرفية باستخدام خوارزميات تعلم الآلة

صفوان عمر حسون

كلية علوم الحاسوب والرياضيات
جامعة الموصل، الموصل، العراق

معن يونس عناد

مديرية تربية نينوى، الموصل، العراق

تاريخ قبول البحث: 2020/03/25

تاريخ استلام البحث: 2020/03/10

المخلص

تلعب القروض المصرفية دوراً حاسماً في تطوير الأعمال الاستثمارية للبنوك. في الوقت الحاضر ، هناك العديد من القضايا المرتبطة بمخاطر القروض المصرفية. مع ظهور أنظمة الحوسبة ، أصبحت البنوك قادرة على تسجيل بيانات المقترضين وفقاً لمعاييرها. في الواقع ، هناك كمية هائلة من بيانات المقترضين ، مما يجعل عملية اتخاذ القرار مهمة صعبة. استخدمت العديد من الدراسات خوارزميات تنقيب البيانات لغرض تصنيف القروض من حيث الايفاء أو عدم الايفاء بسداد القرض بالاعتماد على بيانات المقترضين السابقة . يمكن لهذا النوع من الخوارزميات مساعدة البنوك في اتخاذ قرارات المنح لعملائها. في هذه الورقة ، تمت مقارنة أداء خوارزميات التعلم الآلي لغرض تصنيف مخاطر القروض المصرفية باستخدام المعايير القياسية ثم اختيار الشبكات العصبية متعددة الطبقات (*Multilayer*

Perceptron) حيث أنها أعطت أفضل دقة مقارنة بخوارزميات (NaiveBayes , BayesNet , RandomForest) (DTJ48& .)

الكلمات المفتاحية: القروض المصرفية، خورزميات تعلم الآلة، شبكات العصبية متعددة الطبقات، التصنيف، الدقة .ROC,

I Introduction:

Granting loans is an essential part of the work of any bank. Most of the banks' profits come from the benefits that are taken on these loans and most of the capital is in them. These days many banks agree on the loan after verification and validation, but there is still no whether the applicant is the appropriate applicant [13, 14].

Machine learning algorithms have been used in many areas of business, business administration, human resource management and medical purposes and have shown good success in data mining and decision support systems [3].

In the proposed paper, the use of neuronal networks as one of the machine learning algorithms for the purpose of classifying bank loans in terms of risks depending on the data in banks for previous loans and training algorithms in classifying loans using the characteristics of the borrower and comparing the performance of the neuronal network algorithm with decision trees j48, random forests and statistical methods (NaiveBayes , BayesNet).

II Methods:

In this section, we summarize the machine learning algorithms that were used for the purpose of classifying and forecasting loan risks.

A. *DT J48*, is one of algorithms used for making a decision tree developed by Ross Quinlan [1, 8]. The tree is built in the same way as building Iterative Dichotomiser 3 (ID3), where the contract is chosen based on the concept of gain, where the attribute with the highest classification ability (highest gain) is considered as the root of the tree that is branched into leaves. These leaves also choose (in the same way) the attribute with the ability to rank higher than the remaining attribute at the next level. This separation continues until the entire tree is built [5, 11]. The attribute selection for each node is based on the following three measurements:

$$Entropy(t) = - \sum_{i=0}^{c-1} p(i|t) \log_2 p(i|t) \quad (1)$$

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2 \quad (2)$$

$$Classification\ error(t) = 1 - [p(i|t)] \quad (3)$$

Where c is the number of class and $p(i|t)$ indicates the probability of records belonging to that class.

B. *Random Forest* the RF approach, it is based on the creation of many taxonomy trees based on different subsets of data using random subsections of available variables. The overall result of this approach creating and refining a set of correct theories and assumptions represented by trees, and combine trees in a "forest of classifiers" that its final decision depends on the results of the different decision trees. An additional powerful advantage of this approach, it is based on decentralized group behavior without any central

or hierarchical learning structure [11]. Each tree is built similar to J48 and the final result depends on the average output of those trees, as in the following equation:

$$RF_i = \frac{\sum T_j}{N} \quad (4)$$

RF sub (*i*): the end result of feature *i* from all trees in the RF model

T sub (*j*): the output of tree for *i* in tree *j*

C. Bayes's theorem

It is one of statistical probability theories used to predict the occurrence of a particular event based on the attribute of that event. It can be performed by calculating the probability of each attribute and its impact on the occurrence of that event [10]. Bayes's theorem is mathematically represented by the following formula:

$$p(x) = \frac{p(y)p(y)}{p(x)} \quad (5)$$

Where is:

$p(x)$: The probability that *y* will occur if event *x* occurs.

$p(y)$: The overall probability of a result of that class.

$p(x)$: The probability that event will occur for all events within a particular attribute.

In this paper, two types of Bayes' theory have been used

1. *BayesNet*, it is a probabilistic graphical model that uses Bayesian reasoning to calculate probabilities. The Bayesian network relies on conditional dependence, causation, and inferring from random variables by calculating these probabilities and according to the influence of each factor.
2. *NaiveBayes*, is one of probabilistic classifiers family that based on Bayesian theorem. This model is based on the principle of a maximum a posteriori decision rule and takes. The probability of each attribute independently without considering the relationships between those attributes.

D. Multilayer Perceptron, it is a mathematical model that derives its principle from the way neurons work in the human brain. The network consists of a group of artificial cells linked by connections. The work of Neuronal networks is based on the principle of parallelism that enables the network to analyze many problems with multiple variables [8, 11]. The multi-layered neural network is composed of an input layer that in turn receives the input values for the network and a number of hidden layers depending on the network structure. In this research we use a network with one hidden layer and this layer is called hidden because it is considered as a black box for the user as its inputs are the outputs of the input layer and its results are the inputs of the last layer in the sense that its inputs and outputs are not visible to the user and finally the output layer which consists of one [11]. The mathematical formula of a neuron is the following equation:

$$Y_k = f(\sum W_{jk}X_j + BK) \quad (6)$$

Where is:

$\{X_1, X_2, \dots, X_j\}$: Input signal.

$\{W_1, W_2, \dots, W_j\}$: Weights for the neuron *k*.

BK : It represents bias that can be counted as one of the weights.

f : Activation function.

III Dataset

The dataset used for classification purpose entitled "German Credit data" collected from UCI repository that contains 1000 Instances, 11 attributes as shown in Table (1).

Table 1: German Credit dataset

| No | Age | Sex | Job | Housing | Saving accounts | Checking account | Credit amount | Duration | Purpose | Risk |
|----|-----|--------|-----|---------|-----------------|------------------|---------------|----------|---------------------|------|
| 1 | 67 | male | 2 | own | NA | little | 1169 | 6 | radio/TV | good |
| 2 | 22 | female | 2 | own | little | moderate | 5951 | 48 | radio/TV | bad |
| 3 | 49 | male | 1 | own | little | NA | 2096 | 12 | Education | good |
| 4 | 45 | male | 2 | free | little | little | 7882 | 42 | furniture/equipment | good |
| 5 | 53 | male | 2 | free | little | little | 4870 | 24 | Car | bad |
| 6 | 35 | male | 1 | free | NA | NA | 9055 | 36 | Education | good |
| 7 | 53 | male | 2 | own | quite rich | NA | 2835 | 24 | furniture/equipment | good |
| 8 | 35 | male | 3 | rent | little | moderate | 6948 | 36 | Car | good |
| 9 | 61 | male | 1 | own | rich | NA | 3059 | 12 | radio/TV | good |
| 10 | 28 | male | 3 | own | little | moderate | 5234 | 30 | Car | bad |

After the preprocessing step, dataset become 24 numerical attributes and 1 binary classifier because when converting columns with categorical data to numeric, they will become more than one column [4]. The dataset is divided into two subsets, 80% of the data for training, and then 20% of these data was used for testing. The chosen dataset contains two formats of data (original data, numerical data). The numerical dataset was used to compare it with various machine learning algorithms.

VI Implementation

The performance comparisons were applied to 1,000 cases, including 700 loan repayments and 300 payment default loans. The *weka* Version 3.8.4 environment was used for this research for the purpose of model building and testing [6]. The proposed algorithms were trained using a dataset consisting of 800 instances of loans through supervised learning and targeted data (YES, NO) YES in the case of loan repayment and NO in case of payment defaulted after training the algorithms, a trainer model for each algorithm was obtained. The algorithms were tested using a test set of 200 Instances then we obtained the results of each algorithm and analyzed those results as shown in the figure 2.

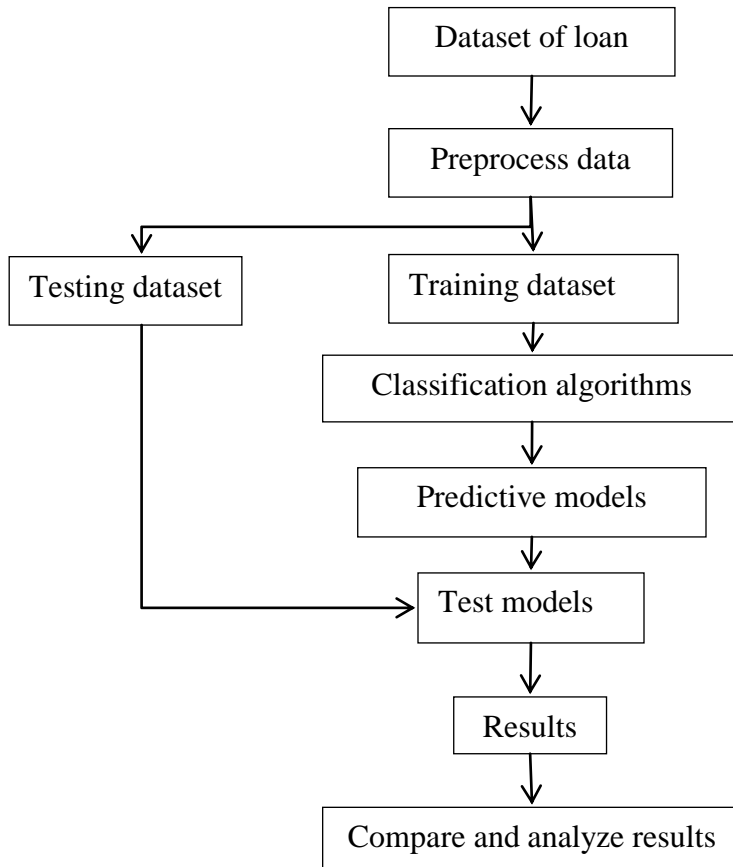


Figure (2): Proposed frame work

The results were analyzed by comparing the performance of each algorithm using several measurements as shown in Table 2 where it appeared that neuronal networks possess the highest accuracy compared to the other of the algorithms.

Table (2): Compare the performance of the algorithms

| Classification algorithm | Root relative squared error | Relative absolute error | Root mean squared error | Mean absolute error | Kappa statistic | Incorrectly Classified Instances | Correctly Classified Instances |
|--------------------------|-----------------------------|-------------------------|-------------------------|---------------------|-----------------|----------------------------------|--------------------------------|
| DT J48 | 105.3969% | 76.8996% | 0.4762 | 0.3198 | 0.3599 | 26.5 % | 73.5 % |
| BayesNet | 89.0378% | 78.5019% | 0.4023 | 0.3264 | 0.33 | 25 % | 75 % |
| NaiveBayes | 87.3059% | 65.147 % | 0.3945 | 0.2709 | 0.4268 | 22.5 % | 77.5 % |
| RandomForest | 85.048% | 74.587 % | 0.3843 | 0.3101 | 0.4271 | 21.5 % | 78.5 % |
| Multilayer Perceptron | 85.3731% | 58.26% | 0.3823 | 0.2411 | 0.4631 | 20 % | 80 % |

In machine learning algorithms, there are standard measurements used to explain the performance of each algorithm with respect to the targeted data. In this paper, there are two targeted values YES in the case of payment and NO in the case of default. The performance of each algorithm was as shown in Table 3.

Table (3): Performance measures of machine learning standard

| | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC Area | PRC Area | Class |
|-----------------------|---------|---------|-----------|--------|-----------|-------|----------|----------|-------|
| DT J48 | 0.804 | 0.439 | 0.821 | 0.804 | 0.813 | 0.360 | 0.699 | 0.820 | yes |
| | 0.561 | 0.196 | 0.533 | 0.561 | 0.547 | 0.360 | 0.699 | 0.467 | no |
| Weighted Avg | 0.735 | 0.369 | 0.739 | 0.735 | 0.737 | 0.360 | 0.699 | 0.719 | |
| BayesNet | 0.881 | 0.579 | 0.792 | 0.881 | 0.834 | 0.338 | 0.788 | 0.899 | yes |
| | 0.421 | 0.119 | 0.585 | 0.421 | 0.490 | 0.338 | 0.788 | 0.550 | no |
| Weighted Avg | 0.750 | 0.448 | 0.733 | 0.750 | 0.736 | 0.338 | 0.788 | 0.799 | |
| NaiveBayes | 0.867 | 0.456 | 0.827 | 0.867 | 0.846 | 0.428 | 0.824 | 0.922 | yes |
| | 0.544 | 0.133 | 0.620 | 0.544 | 0.579 | 0.428 | 0.824 | 0.629 | no |
| Weighted Avg | 0.775 | 0.364 | 0.768 | 0.775 | 0.770 | 0.428 | 0.824 | 0.839 | |
| RandomForest | 0.902 | 0.509 | 0.816 | 0.902 | 0.857 | 0.436 | 0.819 | 0.915 | Yes |
| | 0.491 | 0.098 | 0.667 | 0.491 | 0.566 | 0.436 | 0.819 | 0.678 | No |
| Weighted Avg | 0.785 | 0.392 | 0.774 | 0.785 | 0.774 | 0.436 | 0.819 | 0.848 | |
| Multilayer Perceptron | 0.902 | 0.470 | 0.835 | 0.902 | 0.867 | 0.469 | 0.831 | 0.924 | Yes |
| | 0.530 | 0.098 | 0.673 | 0.530 | 0.593 | 0.469 | 0.831 | 0.659 | No |
| Weighted Avg | 0.800 | 0.367 | 0.791 | 0.800 | 0.792 | 0.469 | 0.831 | 0.851 | |

One important metric for measuring the performance of binary classification algorithms is Receiver Operating Characteristic (ROC), which showed the separability of each algorithm depending on (true positive rate) and (false positive rate). Multi-layer neuronal networks have the potential for higher separation compared to the rest of the algorithms on this type of dataset as shown in Figure (3-7).

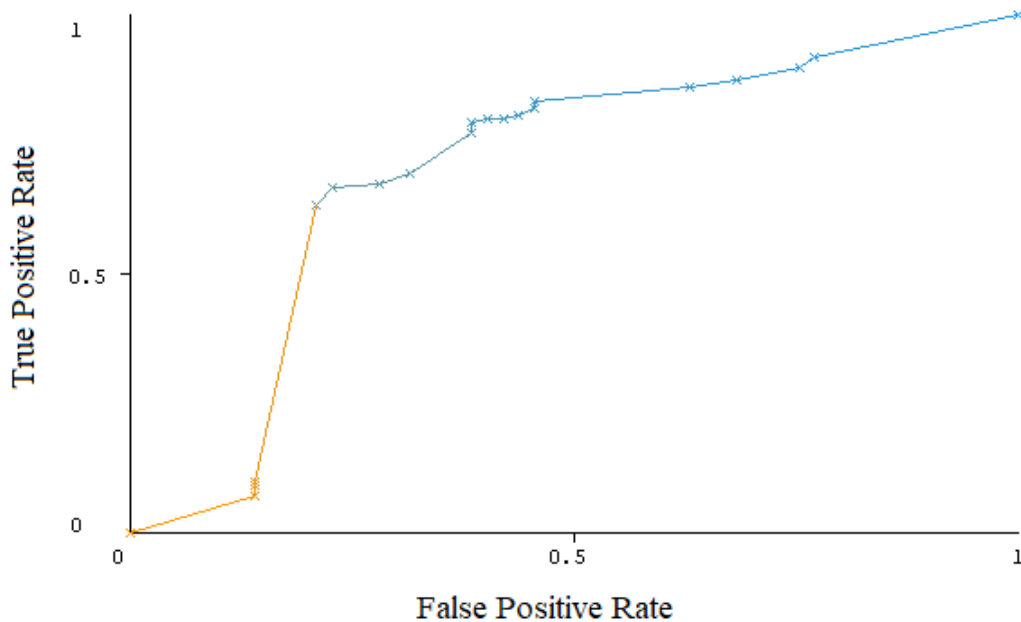


Figure (3): ROC curve for the model DT j48

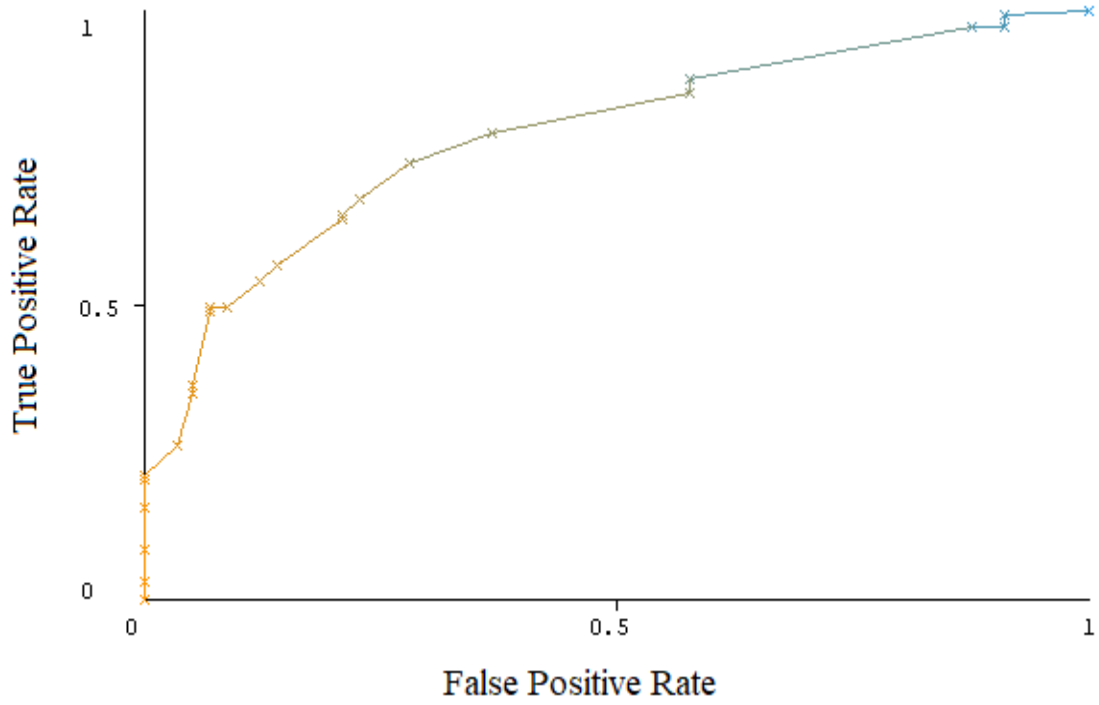


Figure (4): ROC curve for the model BayesNet

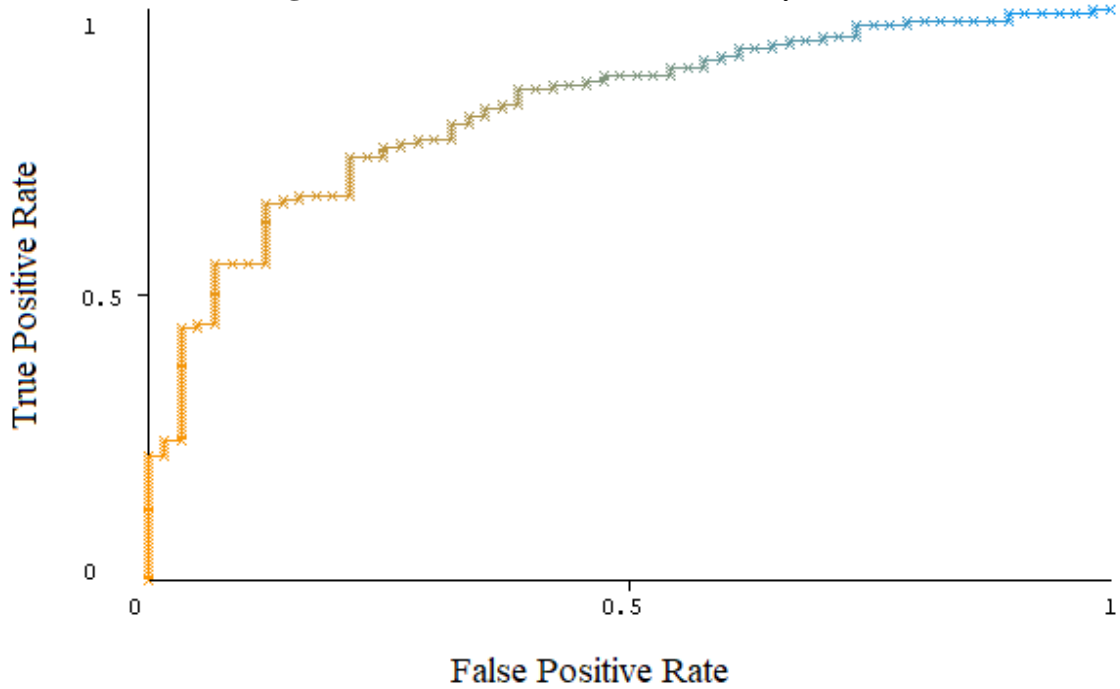


Figure (5): ROC curve for the model NaiveBayes

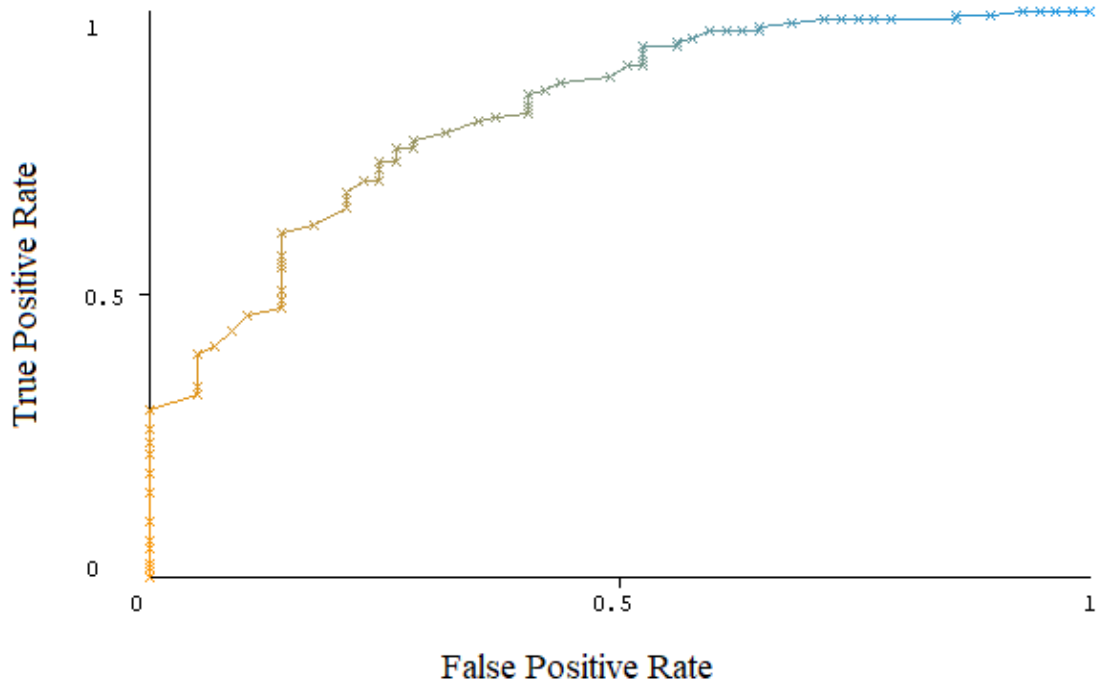


Figure (6): ROC curve for the model RandomForest

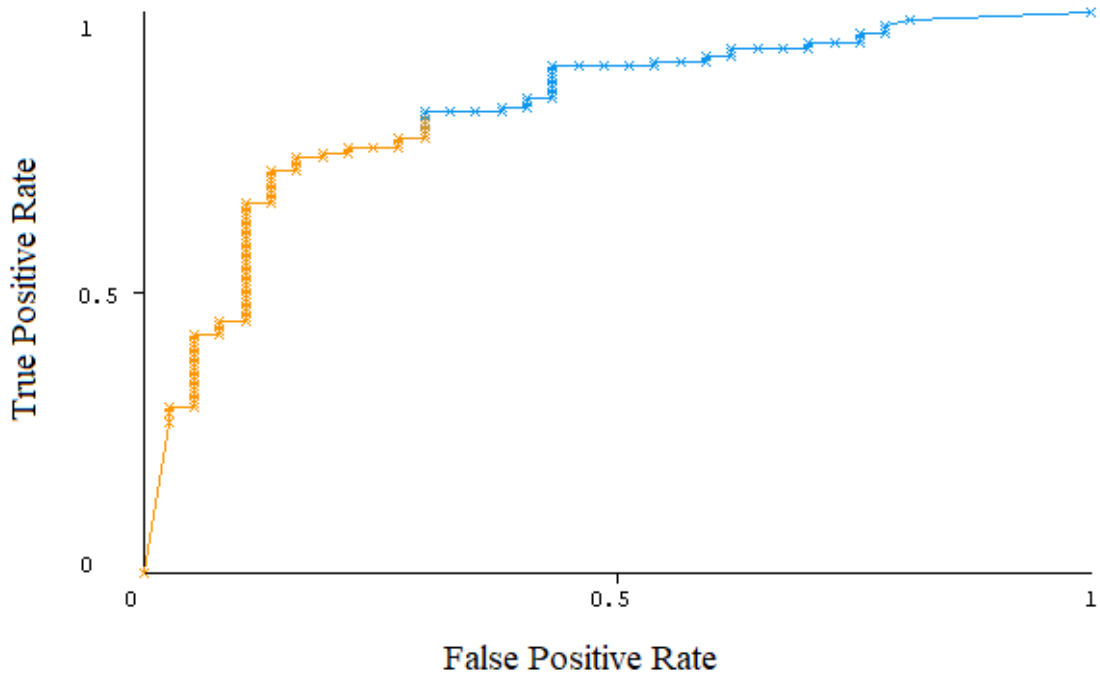


Figure (7): ROC curve for the model Multilayer Perceptron

Conclusion:

Machine learning algorithms play a significant role in predicting the risks of bank loans and decision support systems. The choice of the algorithm used to make the decision (whether the borrower will default), which is the key to addressing decision management when issuing a loan. In this paper, the performance of machine learning algorithms has been tested and their performance compared to standard measurements used on a dataset that includes 1000 loans and their repayment status. Finally, the results showed the possibility of using the proposed algorithms for this purpose with acceptable accuracy rates and superiority of the neural networks for this purpose.

REFERENCE

- [1] Alsaleem, Maan & Hasoon, Safwan. (2020). COMPARISON OF DT& GBDT ALGORITHMS FOR PREDICTIVE MODELING OF CURRENCY EXCHANGE RATES. *EUREKA: Physics and Engineering*. 1. 56-61. 10.21303/2461-4262.2020.001132.
- [2] Belcastro, L. & Marozzo, Fabrizio & Talia, Domenico & Trunfio, Paolo. (2016). Using Scalable Data Mining for Predicting Flight Delays. *ACM Transactions on Intelligent Systems and Technology*. 8. 10.1145/2888402.
- [3] Chen, M. C., & Huang, S. H. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications*, 24, 433–441.
- [4] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.
- [5] Eletter, Shorouq & Yaseen, Saad. (2010). Applying Neural Networks for Loan Decisions in the Jordanian Commercial Banking System. 10.
- [6] Hall et al.. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1). 2009.
- [7] Huang Wei , Lai K. K. , Nakamori Y. and Wang Shouyang, (2004). Forecasting Foreign Exchange Rates with Artificial Neural Networks: A Review, *International Journal of Information Technology & Decision Making* Vol. 3, No.1.
- [8] J. R. Quinlan. Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research*, 4:77-90, 1996.
- [9] Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics and Data Analysis*, 50, 111
- [10] Mattei, Pierre-Alexandre. (2019). A Parsimonious Tour of Bayesian Model Uncertainty.
- [11] Salzberg, Steven. (1996). Book Review: "C4.5: Programs for Machine Learning" by J. Ross Quinlan.
- [12] Stahl, F., & Jordanov, I. (2012). An overview of the use of neural networks for data mining tasks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(3), 193–208. doi: 10.1002/widm.1052
- [13] Supriya, P., Pavani, M., Saisushma, N., Vimala, N. and Vikas, K. (2019). Loan Prediction by using Machine Learning Models. *International Journal of Engineering and Techniques*, 5(22), pp.144-148.
- [14] Wang, Di & Wu, Qi & Zhang, Wen. (2019). Neural Learning of Online Consumer Credit Risk.