Advances in Economics, Business and Management Research, volume 203
Proceedings of the 2021 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021)

The Application of Business Analytics in the Era of Big Data

Tianhuiqi Chen[1, *, a,†], Bowen Gu[2,*,b, †], Zhenxin Jin[3,*,c, †],

[1] *School of St. George, University of Toronto, Toronto, Ontario, M5S, Canada.*
[2] *School of Resources and Safety Engineering, Central South University, Changsha, Hunan 300000, China*
[3] *College of Biomedical Engineering & Instrument Science, Zhejiang University, Hangzhou, Zhejiang 310000, China*
[*] *Corresponding author. Email:[a]thq.chen@mail.utoronto.ca, [b]8210190809@csu.edu.com, [c]3180101350@zju.edu.cn,*
† These authors contributed equally.

ABSTRACT
With the advent of the big data era, data-based business analytics is more and more widely used in all industries, in which banking with the loan business is one of the most important businesses. In order to conduct more intelligent risk control, banks often build prediction models based on loan records to assess whether future clients will default on loans, and the factors generally considered include income level, loan amount, interest rate, etc. This ultimately helps banks to optimize the loan business, avoid credit risks, and reduce losses. This paper focuses on the solutions of choosing appropriate prediction models, classifying the clients, and predicting their defaults. It clarifies the four-step framework of business analytics in the first part. Then this paper introduces three typical statistical analysis models, including the logistic regression model, the decision tree and random forest model, and the K-mean cluster model. The bank loan risk control dataset includes 20,000 borrowers and the details of personal information and loan information. Based on the dataset, the best prediction results are obtained using a random forest model that area under the curve is 0.741 and the clients are divided into four clusters. The logistic regression indicates a negative coefficient between the annual income and the default, while the coefficients are positive between other factors, like the loan amount and the interest rate and the default. In practical application, these models can be combined to give full play to their advantages and make a better prediction.

Keywords: *Business Analytics, Big Data, Banking Industry.*

1. INTRODUCTION

Business analytics is about using data to discover the information that is useful at present and in the future. Business analytics is around leveraging esteem from information. Rather than being alluded to as the 'sludge of the data age,' information has as of late been considered 'the unused oil.' Whereas information can be utilized for the purposes such as recognizing modern openings, recognizing advertise specialties, as well as creating modern products and administrations, it is additionally famously nebulous and difficult to extricate esteem from [1]. Big data may be a potential investigation zone that gets considerable attention from the scholarly community and IT communities. Within the advanced world, the sums of information created and put away have extended inside a brief period of time. Subsequently, this quick developing rate of information has made numerous challenges [2]. This challenge happened to the banks as well. The amount of data stored by banks is rapidly increasing and provides the opportunity for banks to conduct predictive analytics and enhance their businesses. As a result, data scientists face large challenges to handle the massive amount of data efficiently and generate insights with real business value. Thus, the banking customer behaviors can be analyzed from banking big data through analytical modeling methodologies and techniques designed for a key business scenario.

Business intelligence (BI), decision support, and analytics are core to making business decisions in many organizations. As of late, conventional approaches to utilizing organizational information have been addressed as companies grasp voluminous, high-velocity information in an assortment of designs (i.e., multi-structured) that are for the most part surrounded as "big data". Expanded competitiveness and efficiency within the industry have formed the basis for enormous

information analytics and its advances. Intrigued in huge information inquiry, it is developing exponentially as evidenced by the increment of the number of papers, tracks, and scaled down tracks centered on analytics, and enormous information in driving information system (IS) conferences. [3]

2. BACKGROUND INTRODUCTION

The dataset regarding risk control of loans has 32 categories of information about 20,000 borrowers to build a model based on the dataset. It includes details of the person's identification (ID), employment title and length, the purpose of the loan, and just to name a few. The model is for determining the borrower's ability to repay, which demonstrates whether the person is able to pay back the specific amount of money the person will borrow. Three essential variables to build the model are annual income, loan amount, and interest rate, respectively. Base on the dataset, we calculate the maximum, minimum, mean, and variance of each category. Maximum and minimum tell a range of each variable; the median gives the value in the middle, and the variance explains the spread of each person from the mean.

Table 1. Maximum, minimum, mean, and variance of 20,000 borrowers' annual income, loan amount, and interest rate.

	Maximum	Minimum	Median	Variance
Annual Income	10999200	0	65000	4753759642
Loan Amount	40000	500	12000	75970158
Interest Rate	30.99%	5.31%	12.74%	22.71244%

2.1. Annual Income

The distribution shown as figure 1, the annual income has a minimum of 0 Yuan and a maximum of 10,999,200 Yuan. The median is 65000 Yuan which means half of the borrowers' yearly income is lower than 65000 Yuan. The variance is 4753759642, and it explains there is a large spread from each person to the mean, which the difference in each person's annual income is enormous.

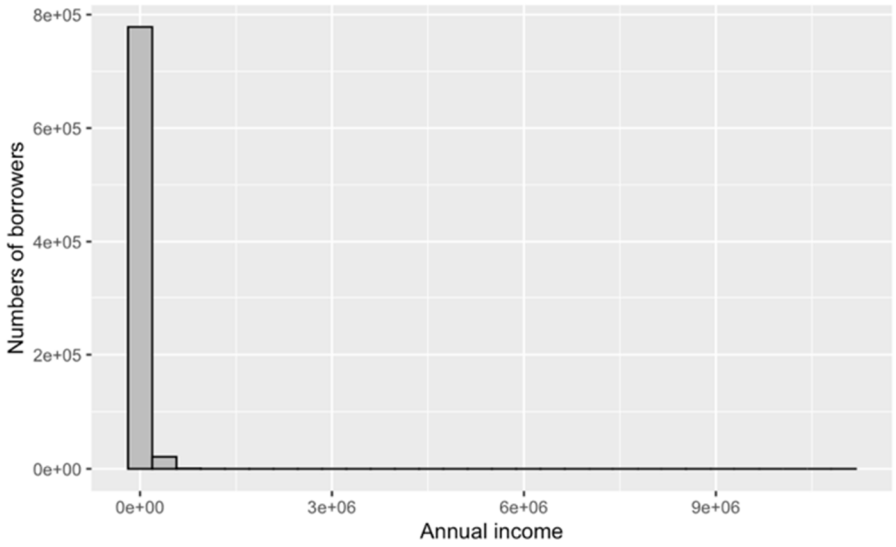

Figure 1. Histogram of annual income

2.2. Loan Amount

The loan amount borrowers apply has the lowest amount of 500 Yuan, while the largest amount is 40,000 Yuan. It has a middle cut-off amount of 12,000 Yuan, which medium amount of the borrower’s request is 12,000 Yuan. The variance is 75970158 shows the significant difference in total borrowers' demands. The distribution shown as figure 2.

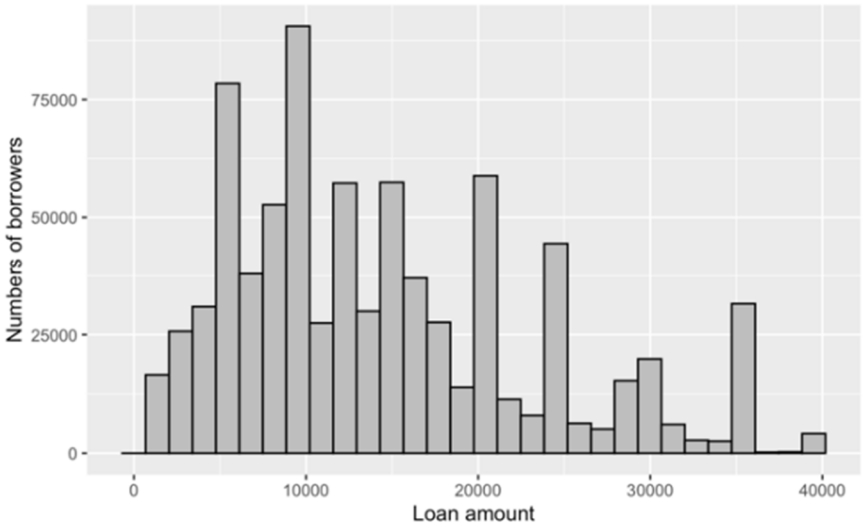

Figure 2. Histogram of loan amount

2.3. Interest Rate

The lowest interest rate is 5.31%, and the largest interest rate is 30.99%. It has a median of 13.24%, of which 50% of borrowers' interest rate is higher or lower than this figure; the variance of 22.71 demonstrates the spread in each person's interest rate is relatively compressed.

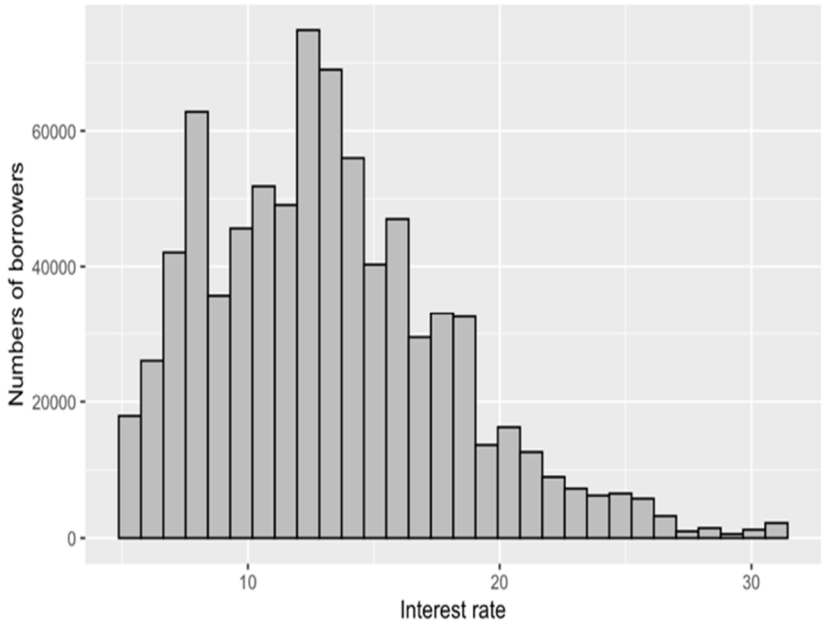

Figure 3. Histogram of interest rate

3. THE FRAMEWORK OF BUSINESS ANALYTICS

3.1. First Step: Demand Analysis and Goal Setting

The first step in business analytics is demand analysis and goal setting. To begin with, we should identify client demand and business goals clearly, which would basically guide the further analysis. Then we tie analytics to different elements in the case, especially the business drivers and returns. Defining performance indicators (KPI) is also necessary to clarify the goals and assess the analysis. Key performance indicators are tools for change, performance management, and sustainable business improvement. The KPIs measure business performance and drive change as well as monitor and sustain ongoing business performance [4]. KPIs should follow the SMART criteria and are firmly based on the business drivers and returns. Then we can set comprehensive plans and strategies for analytics.

3.2. Second Step: Data Mining and Modeling

Since data is the foundation of business analytics, the second step is data mining and modeling. We can collect and lean data from internal and external sources and do exploratory data analysis. Besides, feature engineering is supportive for decreasing information dimensionality, diminishing forecast show complexity, and handling the issue of undermined and loud data [5]. It incorporates a handle of conceptualizing or testing highlights, choosing what highlights to make, making highlights, checking how the highlights work with the demonstrate, progressing the highlights if required, and going back to brainstorming/creating more highlights until the work is done. According to the ultimate model, we can deliver business insights and advisory.

3.3. Third Step: Operational Integration

The third step is operational integration, which integrates the market demand and operation resources of enterprises. It links solutions to business operations by creating corresponding strategies for operations. To optimize those solutions, we should engage users to obtain feedback. Therefore, we can ensure the continued effectiveness of the solutions through iterations and updates.

3.4. Last Step: Technology and Management Solution Delivery

The last step is technology and management solution delivery. The technology solution delivery part involves database and data warehouse interface, where database mainly serves for business while data warehouse for analytics. It may also involve cloud computing, automatic process creation, etc. The management solutions related to communications, human resources management, and performance management are delivered as well. In the process of business analytics, this step provides a certain basis for the follow-up project implementation.

4. METHODS

4.1. LOGISTIC REGRESSION

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable [6]. Moreover, it is a branch of classification belonging to the machine learning method, which is often used for showing the probability of certain things happening in dichotomy problem, like whether the mail is the spam mail, the probability of someone clicking the advertisements (which is yes or no). The advantage of logistic regression is that it is easy to achieve and understand due to the low circulation cost, making it widely applicable to industrial situations. While classification, thanks to the less amount of circulation, is a faster and low-storage-resource-costing way as well. What's more, you can observe probability scores more conveniently. For logistic regression, there are no more difficulties in multiple linear problems with the help of the combination with L2 regularization. The disadvantages of logistic regression are that it has worse performance when the feature space is too big, usually gets into the underfitting problem, and has lower accuracy. It cannot handle many multiclass features or variables but is only used to solve linear separable two-class problems (the SoftMax develop from this can used for multi-classification). For the nonlinear feature, you have to transform.

In brief, logistic regression is generally quick compared to other directed classification methods such as kernel SVM or ensemble methods but endures to a few degrees in its precision. It too has the same issues as linear regression as both procedures are distant as well short-sighted for complex connections between factors. At long last, logistic regression tends to underperform when the choice boundary is nonlinear" [7].

In this paper, we take business analytics in banking as an example, focusing on the classification problem in the issue of financial risk management to predict the probability of loan defaults. We apply logistic regression to this real situation, and we can quantify the consequence of paying off the loan, marking the payoff as one while default as 0. Then we can consider this problem as a two-class problem. As for the independent variable, we can collect both constant or discrete variables as a feature, like the annual income, interest rate, and loan amount. Through logistic regression, we can see the weight of all these dependent variables.

4.2. Decision Tree

In machine learning, supervised learning is a training method, including two main tasks of regression and classification. Decision trees are the simplest modeling techniques and are most appropriate for modeling interventions in which the relevant events occur over a short time period [8]. It uses a tree structure to reach the ultimate classification through reasoning step by step. The tree consists of 3 elements: root node, which contains the complete set of samples; internal node, which refers to a test to specific characteristics; and leaf node, which represents the result of the decision. The process of decision starts from the root node, and some judgments at different internal nodes lead it to the leaf node, which symbolizes the classification has been completed. This algorithm is based on "if-then-else" rules, and the results are obtained by training instead of manual work.

4.2.1. Three Steps of decision tree algorithm

The first step when performing decision tree learning is feature selection. The most relevant features are filtered out, which help improve the classification ability. The second step, decision tree generation, is about establishing internal nodes based on the different values of the characters in the former node. The last step is pruning, with the purpose of reducing the risk of overfitting in decision tree learning.

4.2.2. Three typical types of decision tree algorithm

There are three types of decision tree algorithms, which are ID3 algorithm (ID3), C4.5 algorithm, and Classification and Regression Tree (CART), respectively. ID3 is the earliest decision tree algorithm proposed. It uses information gain to select features. C4.5 algorithm, instead of directly using information gain, includes the "information gain ratio" indicator as the basis for feature selection. CART can be used for both classification and regression problems, using the Gini coefficient instead of the information entropy model.

4.2.3. Advantages and disadvantages of decision tree

The advantages of a decision tree are various. Firstly, it can be visualized and analyzed, easy to understand, explain, and extract rules. Secondly, it can process nominal and numerical data at the same time. Thirdly, it is suitable for processing samples with missing features; besides, it is able to deal with irrelevant features. Fourthly, it has a fast-running speed when testing the data set. Lastly, it is feasible and effective results for large data sources in a relatively short time. However, it accompanies by other disadvantages. It is prone to overfitting and easy to ignore the correlation of features in the data set—furthermore, different feature selection tendencies in different judgment criteria. Plus, the bias towards the feature with more numerical values in the ID3 algorithm.

4.3. Random Forest

Random Forest (RF) algorithm is one of the best algorithms for classification. RF is able for classifying large data with accuracy. It is a learning method in which the number of decision trees are constructed at the time of training and outputs of the modal predicted by the individual trees [9]. When new input samples enter in classification tasks, each decision tree in the forest is judged separately, classified separately, and gets its own classification result. The final result of random forest is the decision tree result which appears the most often.

4.3.1. Four Steps of Random Forest Algorithm

The first step is to collect a sample with a size of N is drawn N times with replacement, one sample at a time, and finally, N samples are formed. The selected N samples are used to train a decision tree as root nodes of the decision tree. Next, each sample has M attributes. When each node of the decision tree needs to be split, randomly select m ($m << M$) attributes from these M attributes. Then choose a certain strategy (such as information gain) to select one attribute from these m attributes as the split attribute of the node. After that, in the process of decision tree formation, each node must be split following step 2 until it can no longer be divided. Finally, follow steps 1 to 3 to build a large number of decision trees, which constitute a random forest.

4.3.2. Advantages and Disadvantages of Random Forest

The random forest has several advantages. It produces high-dimensional data without dimensionality reduction or feature selection, and it is able to judge the importance of features and the mutual influence between different characteristics. Moreover, it is not easy to overfit, relatively fast training speed, and simple to implement. It is able to balance errors or unbalanced data sets. It maintains accuracy when a large part of the features is missing. In contrast, the disadvantages are overfitted in some classification or regression problems with large noise, and the attribute weights produced by the random forest on data with many different attribute values are unreliable.

4.4. K-mean Cluster

A cluster refers to a collection of data points aggregated together because of certain similarities. K-mean clustering groups the data points based on their similarity or closeness to each other, in simple terms, the algorithm needs to find the data points whose values are similar to each other, and therefore these points would then belong to the same cluster [10]. K-means clustering partitions a data space into k clusters, each with a mean value. Each individual in the cluster is placed in the cluster closest to the cluster's mean value [11]. Due to the number of transactions in the banking sector is rapidly growing and huge data volumes are available, representing the customers' behavior, and the risks around loan are increased [12]. Thus, to control loan risk, it is crucial to classify clients to understand the common behavior of the groups. As a result, the K-mean clustering logarithm is able to sort data into vectors. The use of K-mean clustering is simple to implement, easy to interpret the clustering results [13], and efficient in terms of computational cost [14]. The dataset of clients in the field of loan risk often contains tens of thousands of information; besides, K-mean clustering can deal with a

huge dataset in a short time [15]. However, using K-mean clustering often produces clusters with a relatively uniform size even if the input data have different cluster sizes [16]; plus, the model is only suitable for numerical data.

5. RESULTS AND DISCUSSION

We apply k-means clustering to the data and obtain the figure. The whole dataset of transactions in the banking sector tends to be aggregated to 4 clusters. The outlier can be ignored. Then we can use four clusters with different features instead of the whole tanglesome dataset, the detailed data shown in the table 2, which helps us a lot to classification.

Table 2. AUC value of three models

Model	Logistic Regression	Random Forest	Neutral Networks
AUC	0.541832674	0.741015462591921	0.565059077952 08

Through analysis by logistic regression, we can obtain the coefficient of the independent variables, which can help us see the weight of different features that may affect the consequence of paying off the loan (which is 1 or 0 in this model). Based on the K-mean clustering graph displayed as figure 4 we can see the coefficient of annual income is negative, which means the higher the annual income someone has, the lower probability of loan default is. So are the other features with a negative coefficient, like the term, postcode and n0, etc., the higher value features have, the lower probability of loan default is. The feature of n8 has a bigger absolute value, which can help us to know that it can impact the consequence a lot. So maybe the risk manager should highlight this feature in risk evaluation. The rest of the features with positive coefficient shows that: the bigger they are, the higher probability of loan default is, like the bigger loan amount someone has, the higher probability of his lack of ability to afford the loan is.

Figure 4. Scatter diagram of K-means clustering

Through observing the ROC curves [17] and circulating the AUC [18]value of the three models, we can see that the model of the random forest has a larger area of ROC curve and a bigger AUC value [19], which means the model of random forest achieves better performance in classification [20].

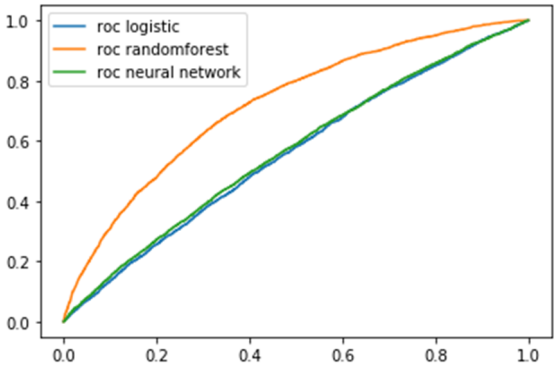

Figure 5. ROC curve of three models

Table 3. Coefficient of variables in logistic regression model

Feature	Coefficient
Loan Amount	2.13E-05
Term	-0.001975952
Interest Rate	0.002179915
Annual Income	-1.19E-05
Post Code	-0.001063319
n0	-8.09E-05
n1	-0.002007863
n2	-0.001885662
n3	-0.001885662
n4	-0.003949215
n5	-0.008256758
n6	-0.003861128
n7	-0.006264627
n8	-0.014418813
n9	-0.001931452
n10	-0.008730307
n11	1.17E-06
n12	7.62E-06
n13	-5.12E-05
n14	0.001474194

6. CONCLUSION

In conclusion, we mainly apply the logistic regression model, random forest model, and neural network model to complete the classification of clients. According to the results, the random forest model shows the best performance in client classification with an AUC value of 0.741, which indicates this model has a relatively accurate prediction on client loan default. But the results of logistic regression and neural network are quite close with AUC values of 0.542 and 0.565, respectively. Also, K-mean clustering helps divide the clients into four clusters.

In logistic regression, the model built shows the principles of typical factors, like annual income, loan amount, and interest rate, influencing the loan default and predicts default based on those attribute variables on clients. Higher annual income, lower loan amount, and lower interest rate can all lead to a smaller probability of default. Now that the interaction relationship between these factors and the default has been proved. And banks should pay more attention to this basic information to avoid risks.

In practical application, the models mentioned above can be combined with each other to give full play to their advantages. With the increase of the number of clients and information categories, banks should also flexibly select the appropriate model. This is necessary for banks to make a more accurate prediction, take corresponding intervention measures in time, and avoid potential credit risks.

REFERENCES

[1] Acito, F., & Khatri, V. (2014). Business analytics: Why now and what next? Business Horizons, 57(5), 565-570. doi: 10.1016/j.bushor.2014.06.001

[2] Yaqoob, Ibrar & Hashem, Ibrahim & Gani, Abdullah & Mokhtar, Salimah & Ahmed, Ejaz & Anuar, Nor & Vasilakos, Athanasios. (2016). Big Data: From Beginning to Future. International Journal of Information Management. 36. 10.1016/j.ijinfomgt.2016.07.009.

[3] Phillips-Wren, G., Iyer, L.S. Kulkarni, U., & Ariyachandra, T. (2015). "Business Analytics in the Context of Big Data: A Roadmap for Research," Communications of the AIS, Vol. 37, #23.

[4] Greeff, G., & Ghoshal, R. (2004). *Practical E-manufacturing and supply chain management*. Oxford: Newnes.

[5] Fan, C., Sun, Y., Zhao, Y., Song, M., & Wang, J. (2019). Deep learning-based feature engineering methods for improved building energy prediction. *Applied Energy*, *240*, 35-45. doi: 10.1016/j.apenergy.2019.02.052

[6] Edgar, T., & Manz, D. (2017). *Research Methods for Cyber Security*. [Place of publication not identified]: Elsevier Science and Technology Books, Inc.

[7] Chang, A. *Intelligence-based medicine*.

[8] Culyer, A. (2014). *Encyclopedia of health economics*. Amsterdam: Elsevier.

[9] Paul, S., & Bhatia, D. *Smart Healthcare for Disease Diagnosis and Prevention*.

[10] B. Meyer, Applying "Design by Contract", Computer 25(10) (1992) 40–51. DOI: https://doi.org/10.1109/2.161279

[11] "K Means Clustering with Simple Explanation for Beginners." Analytics Vidhya, 1 Mar. 2021, www.analyticsvidhya.com/blog/2021/02/simple-explanation-to-understand-k-means-clustering/.

[12] Aboobyda Jafar Hamid, and Tarig Mohammed Ahmed. "Developing Prediction Model of Loan Risk in Banks Using Data Mining." Machine Learning and Applications: An International Journal, vol. 3, no. 1, 2016, pp. 1–9., doi:10.5121/mlaij.2016.3101.

[13] Steinley, Douglas. "K-means clustering: a half-century synthesis." British Journal of Mathematical and Statistical Psychology 59.1 (2006): 1-34.

[14] Alsabti, Khaled, Sanjay Ranka, and Vineet Singh. "An efficient k-means clustering algorithm." (1997).

[15] Velmurugan, T., and T. Santhanam. "Computational complexity between K-means and K-medoids clustering algorithms for normal and uniform distributions of data points." Journal of computer science 6.3 (2010): 363.

[16] Kodinariya, Trupti M., and Prashant R. Makwana. "Review on determining number of Cluster in K-Means Clustering." International Journal 1.6 (2013): 90-95.

[17] Bradley, Andrew P. "The use of the area under the ROC curve in the evaluation of machine learning algorithms." Pattern recognition 30.7 (1997): 1145-1159.

[18] Cortes, Corinna, and Mehryar Mohri. "AUC optimization vs. error rate minimization." Advances in neural information processing systems 16 (2003): 313-320.

[19] Lobo, Jorge M., Alberto Jiménez-Valverde, and Raimundo Real. "AUC: a misleading measure of the performance of predictive distribution models." Global ecology and Biogeography 17.2 (2008): 145-151.

[20] Tharwat, Alaa. "Classification assessment methods." Applied Computing and Informatics (2020).