A FRAMEWORK FOR PREDICTION BANKING RISK USING MACHINE LEARNING TECHNIQUES

ASMAA SAEED EMBARK[1], RIHAM Y. HAGGAG[2], SAMIR ABOUL FOTOUH SALEH[3]

[1]Department of information system, Faculty of Commerce & Business Administration, Helwan University, Cairo, Egypt

[2]Business Information Systems department, Faculty of Commerce and Business Administration, Helwan University

[3] Department of Accounting &Information Systems, Faculty of Commerce, Mansoura University

e-mail:[1]asmaasaeed2010yahoo.com, [2]prof_samir@hotmail.com, [3]rihamhaggagg@yahoo.com

ABSTRACT

One of the main challenges facing the banks is to determine the proper bank liquidity. Risk differs widely from bank to bank, and a Careful understanding of various risk factors assists predict the likelihood of expected liquidity based on historical data, Real-world datasets often have missing values, which can cause bias in results. the most widely adopted method for dealing with missing data is to delete observations having missing values, these methods have the disadvantages represented in loss of precision and biased. The purpose of this study is to forecast banks' liquidity risk. We also present a method for dealing with missing data using powerful machine learning methods. we Used available datasets through Kaggle there are 350 cases and 19 characteristics in this dataset. SPSS and the WEKA tool were used to analyze the data. ROC and accuracy were used to assess and compare three classification models (Decision Tree, Support Vector Machine (SVM), and random forest). Results showed that the model obtained acceptably, results The 66-fold(97.47, 97.47, 97.47) respectively (DT, SVM, RF) the best accuracy among from 10-fold.

Keywords: *Liquidity Risk; Machine Learning; Decision Trees; Support Vectors; Random Forests; Missing Data*

1. INTRODUCTION

Risk management in banks has risen in prominence since the global financial crisis, with a constant focus on how risks are recognized, measured, reported, and controlled. Both academia and industry have focused on banking and risk management innovations, as well as present and emerging concerns. In parallel, machine learning is becoming more prevalent in corporate applications, with many solutions currently in place and many more being investigated.

By 2025, risk functions in banks will need to be radically different than they are now, Risk management is projected to alter due to the expansion and complexity of legislation, changing customer expectations, and the evolution of risk categories. Machine learning, which has been identified as one of the technologies with ignificant risk management implications, can help risk managers construct more accurate risk models by recognizing complicated, nonlinear patterns in vast datasets.

The predictive strength of these models can improve with each new piece of data added, resulting in improved predictive power over time. Machine learning is expected to be used in a variety of areas inside a bank's risk organization.[1]

Banks are subject to many different potential risks that range from those related to the technological and financial structure, affecting also their reputation, to those derived from the institutional and social environment. These risks are not mutually exclusive and have some intersections that make them hard to isolate and identify. liquidity risk poses a serious financial threat to banks.[2]

In this research, we will focus on the algorithms used in forecasting. the process of forecasting liquidity risks leads to what the bank must do to face these risks.

This research aims at a framework for predicting liquidity risks using machine learning techniques with a focus on the data preparation stage to compare the results before and after this stage.

2. RELATED WORK

There are many studies on liquidity risk, where researchers have used many techniques, including machine learning, deep learning, and a convolutional neural network. Below is a brief overview of research papers focusing on liquidity risk.

Andrés Alonso and José Manuel Carbó, 2020, "Machine Learning In Credit Risk: Measuring The Dilemma Between Prediction And Supervisory Cost", This paper aims to measure the costs and benefits of evaluating ML models. These algorithms were used (logistic regression, decision tree, random forest, XGBoost, deep neural network). The results obtained were observed gains in discriminatory power of up to 20% in terms of AUC-ROC when compared to more traditional quantitative methods. [4]

Ms. Usha Devi, Dr. Neera Batra, 2020, "Exploration Of Credit Risk Based On Machine Learning Tools", This paper aims to A framework with the help of tables and diagrams. These algorithms were used (SVM, MDA, RS, LR, ANN, CBR, DT, GA, KNN, DGHNL, XGBoost). The results obtained The proposed DGHNL model is capable to achieve the highest prediction accuracy. [5]

Martin Leo, Suneel Sharma, and K. Maddulety, 2019, "Machine Learning in Banking Risk Management: A Literature Review", This paper aim to This paper, seek to analyze and evaluate machine-learning techniques. These algorithms were used (Clustering analysis, Bayesian networks, Decision trees, SVM). The results reveal Most of the research appears focused on credit risk management, Market risk, and Operational risk. Liquidity risk has seen limited research. Given the implications to a bank's profitability and solvency as a consequence of a liquidity risk event materializing, liquidity risk would be a very good candidate to research extensively, predicting liquidity risk events. [1]

Noureddine Lehdili, Pascal Oswald, and Harold Gueneau, 2019, "Market Risk Assessment of a trading book using Statistical and Machine Learning", This paper aims to be interested in how machine learning algorithms can help banks. These algorithms were used (Bayesian Gaussian, Gaussian processes). The results reveal The numerical tests show that the Gaussian process regression (GPR) can drastically improve the computing time whilst ensuring an excellent level of accuracy. [6]

Saqib Aziz, Michael Dowling, 2019, "AI and machine learning for risk management", This paper aims to analyze, using current practice and empirical evidence. These algorithms were used (LASSO regression, Ridge regression, LARS regression, support vector machines, decision trees, K- means clustering). The results reveal LASSO regression zero weights independent variables with low explanatory power, while Ridge regression gives lower weights to variables in a model that are highly correlated with other variables in a model. [7]

Peter Martey Addo, Dominique Guegan , Bertrand Hassani, 2018, "Credit Risk Analysis Using Machine and Deep Learning Models", This paper aims to build binary classifiers based on machine and deep learning models on real data in predicting loan default probability. These algorithms were used (Elastic Net, Random Forest, A Gradient Boosting, Deep Learning). The results showed that observe that the tree-based models are more stable than the models based on multilayer artificial neural networks.[3]

Anastasios Petropoulos, Vasilis Siakoulis, Evaggelos Stavroulakis, and Aristotelis Klamargias, 2019, "A robust machine learning approach for credit risk analysis of large loan-level datasets using deep learning and extreme gradient boosting", This paper aims to investigate the analysis of loans corporate using machine learning techniques and deep learning neural networks and the combination of data mining algorithms. These algorithms were used (XGBoost, Deep Learning, Random Forest). The results showed selection XGBoost the methodology marginally outperforms Deep Neural Networks (MXNET) but the latter methodology provides the opportunity of increased flexibility over boosting techniques through a large combination of different structures which may optimize the bias-variance trade-off. [2]

Madjid Tavana , Amir-Reza Abtahi , Debora Di Caprio , Maryam Poortarigh, 2018, "An Artificial Neural Network and Bayesian Network model for liquidity risk assessment in banking", This paper aims to propose a model that uses Artificial Neural Networks and Bayesian Networks. These algorithms were used (An Artificial Neural Network - Bayesian Network). The numerical results the ability of the proposed two-phase ANN-BN approach to somehow “self-confirming” the results via an independent and parallel implementation of the same dataset. [8]

Chanh-Ho An, 2017, “A Study on Estimation of Financial Liquidity Risk Prediction Model Using Financial Analysis”, This paper aims to predict the financial liquidity risk. These algorithms were used (ANOVA, Apparent Error Rate). The results showed that The logit discriminant model is the most suitable model to identify the financial liquidity risk. [9]

Jordi Petchamé Sala, 2011, “Liquidity Risk Modeling Using Artificial Neural Network”, This paper aims to a theoretical introduction and a state of the art survey of the key elements needed to understand the complexity of the dealt issue. These algorithms were used (Time Series, artificial neural networks). The results showed Regarding FTDNN results, they have not been satisfactory. take into account would be liquidity risk modeling is a new issue and it has not a significant model yet. [10]

There were four previous studies [3,4,5,10] that discussed credit risk and used machine learning algorithms: logistic regression, decision tree, random forest, Random Forest, XGBoost, deep neural network and use feature selection, and don't mention missing data. one study that discussed market risks [6] and used Bayesian Gaussian, Gaussian processes did not mention the missing data and feature selection. Two studies discussed banking risk management [1,7] and used clustering analysis, Bayesian networks.

Decision trees, SVM, LASSO regression, Ridge regression ,LARS regression, K-means clustering talked about missing data and it was processed with estimation and mentioned feature selection and it was processed with PCA.

Three studies discuss liquidity risk [8,9,10] and the use of an Artificial Neural Network, Time Series study only one of them mentioned feature selection it was processed with standard deviation and contain Missing data, and it was processed with k-neighbors. By reviewing previous studies and research, the researcher found that there is a deficiency in discussing and predicting bank liquidity risks, We will make a prediction for liquidity risks using machine learning algorithms based on (Decision Tree, Support Vector Machine (SVM), random forest). As for the missing data, it was neglected by previous studies, and the studies that I mentioned used simple methods to deal with it. In our thesis, it will be treated using advanced techniques.

3. PROPOSED TABLE

In this part, the overall experimental procedures are implemented in depth. The phases of the model are shown in Figure 1 below, which illustrates two main sections of the experiment: pre-processing and modeling data. First of all, the raw data is described as follows.

3.1 Description Of Data

The objective of this paper is to forecast bank liquidity risk based on balance sheet data. It can be obtained from the Kaggle website [19] Data collected from the Bank of England. There are 350 cases and 19 variables in this dataset . Detailed descriptions of the traits are shown in Table 1.

3.2 Description Of The Dataset Framewark

Table 1: Description of the Dataset

Features	Description
Year	Year
Total Assets	Total Assets
Government debt	Denotes the Government debt
Other Government securities	Denotes the Other Government securities
Other securities	express the Other securities
Coin and bullion	It expresses the Coin and bullion
Notes in the Bank	Notes in the Bank
Notes In circulation	Notes In circulation
Notes in the Bank	Notes in the Bank
Capital	Expresses the value of a Capital
Rest	Rest
Deposits	Expresses the value of a Deposits
o/w Public deposits	Expresses the value of an o/w Public deposits
o/w Special deposits	Expresses the value of an o/w Special deposits
o/w Bankers deposits	Expresses the value of an o/w Bankers deposits
Other accounts	Expresses the value of Other accounts
7 day and other Bills	7 day and other Bills
Total Liabilities	Total liabilities
Check Assets=Liabilities	Check Assets=Liabilities

3.3 Implementation Model

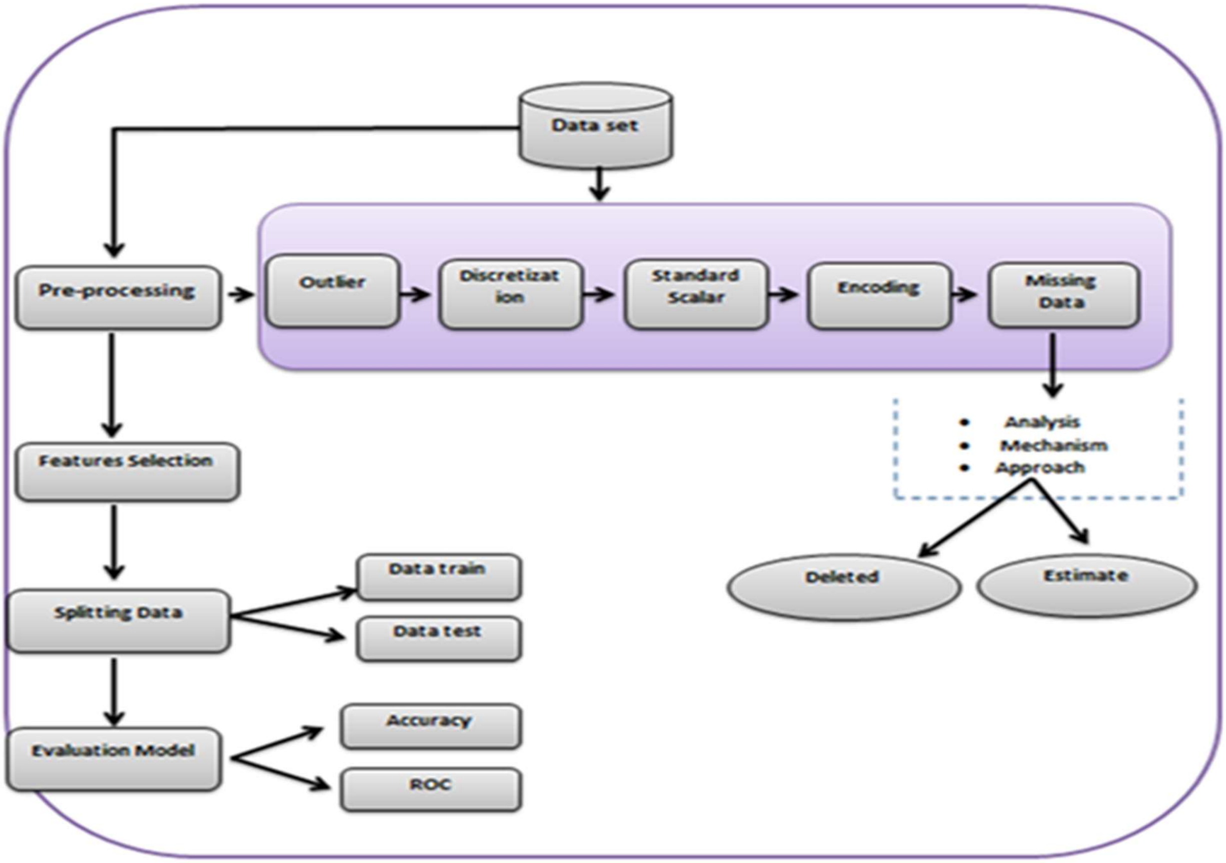


Figure. 1: Overall Structure Of The Proposed Model

In this research, we designed a model to predict Bank liquidity risk by applying Machine Learning Techniques (MLT) based on the Loans, deposits, and securities shown in figure (1). phase to prepare the proposed model consists of (1) Data collection phase, (2) Data preprocessing of the data before applying the MLT, (3) data splitting into data training and data testing, (4) Selection of classification models, (5) evaluation phase to evaluate the accuracy of the built model using a machine learning technique.

3.3.1. Data Preprocessing

To improve our proposed model's predictive effect, the raw data, which are often redundant, inconsistent, or uncertain in general machine learning, are processed and optimized in this section. Therefore, until designing a predictive model, it is important to preprocess the data. The following steps have been done to achieve enhancement. achieve enhancement.

3.3.1.1. HANDLE MISSING DATA

There are many missing values existing in the using dataset the following is an analysis of missing data using the SPSS tool in Figure 2 and Figure 3.

A) MISSING DATA ANALYSIS

The dataset has been cleaned of noise, the quantitative variables in the dataset have a large number of missing values. There are missing values in all cases. The counts and percentages of missing values are shown in figures 2 and 3.

Figure 2 illustrates that there are 9 variables out of 19 that have missing values, accounting for 47.37 percent of the total. In terms of the rows level, all cases have missing values.

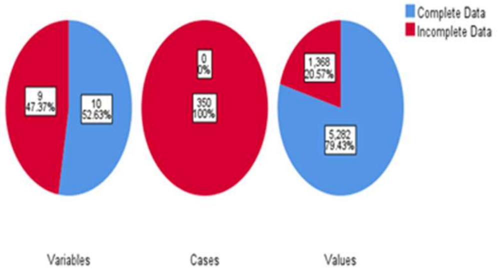


Figure 2.: Overall summary of missing data

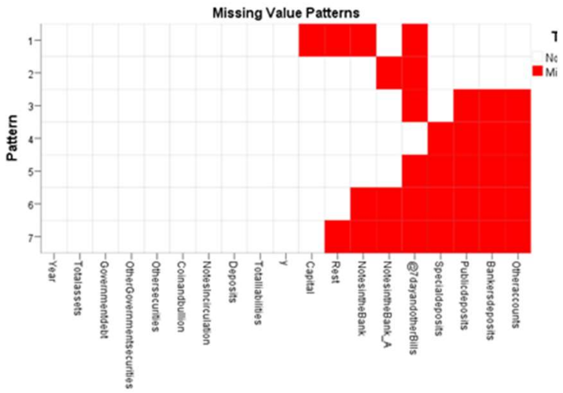


Figure 3.: Overall missing value patterns

The missing value patterns for the analysis variables are displayed in the patterns chart. Each pattern corresponds to a group of occurrences with the same incomplete and complete data pattern. All of the cases in Figure 3 suffer data loss.

3.3.1.2 MCAR TEST

Rubin and colleagues proposed a methodology for categorizing missing data issues. This study yielded three concepts known as missing data mechanisms, which show how the probability of missing data relates to the observed data. A systematic link exists between one or more measured qualities and the chance of missing values, which is referred to as missing at random (MAR). Missing completely at random (MCAR) assumes that missing data is unrelated to observed data, whereas missing not at random (MNAR) assumes that the chance of missing data on a variable Y is connected to the values of Y. In this phase, we use Little's missing totally at random test to verify if the data was fully randomly lost or not, and then we use that information to establish the mechanism by which the data was lost and the best way to handle the missing data based on that. [12]

Null hypotheses H0 = MCAR

Alternative hypothesis H1≠ MCA

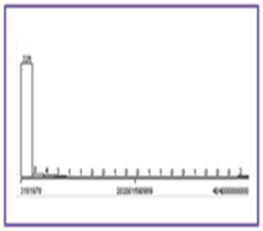
Since the sig is greater than .05 then this indicates that the missing data (= MCAR or ≠ MCAR).

B) MPUTE MISSING DATA

mechanism by which the data was lost was identified in the previous stage, and it turned out to be MAR, which indicates that the missing data has to do with Observed Data, which means that there is a bias towards certain values, requiring us to estimate these lost values rather than deleting them, and one of the most accurate of these methods is multiple imputations, where the final MI estimate is simply the average of the estimates. [13,14]

4. DISCRETIZATION

Discretization is an important data reduction technique. Its major purpose is to convert a set of continuous variables into discrete variables by dividing the scope of the variables into a finite number of disjoint intervals and then linking all intervals with denotation labels. [16]



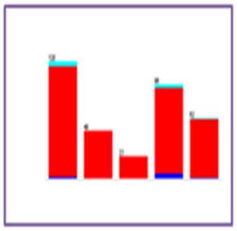


Figure 4: Total assets before discretization

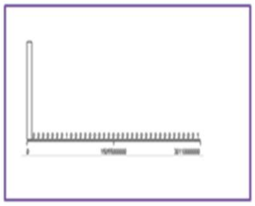
Figure 5: Total

assets after

discretization

debt before discretization

debt after discretization



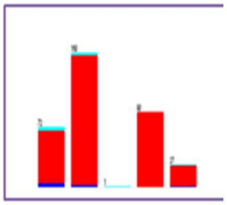


Figure6: Government *Figure7 :Government*

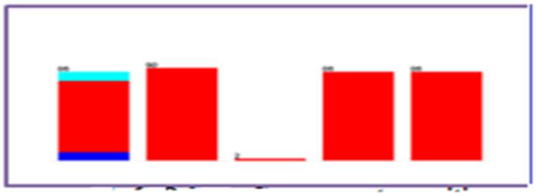


Figure 8 :Other Government securities before discretization. discretization

Figure 9 :Other Government securities after

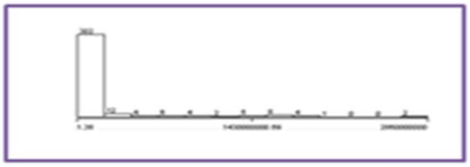


Figure 10: Other securities before discretization.

Figure 11: Other securities after discretization

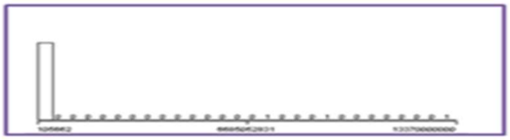
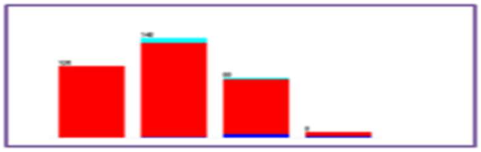


Figure 12 :Coin and bullion before discretization discretization



. Figure 13.: Coin and bullionafter after

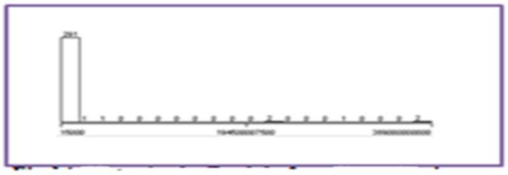


Figure 14: Notes in the Bank before discretization discretization

Figure 15: Notes in the Bank after

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

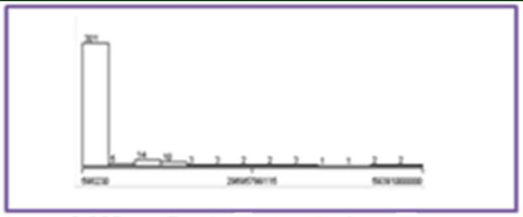
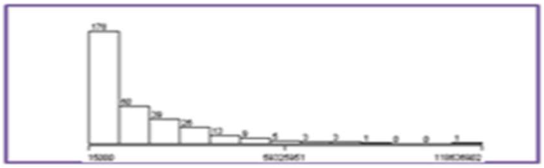


Figure 16 :Notes In circulation before discretization . Figure 17: Notes In circulation after discretization



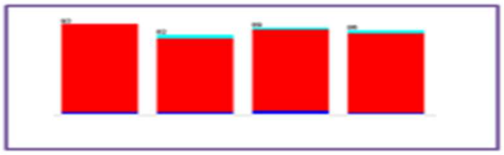
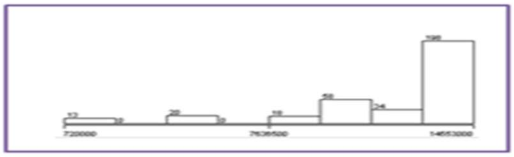


Figure 18: Notes Bank before discretization . Figure 19: Notes Bank after discretization



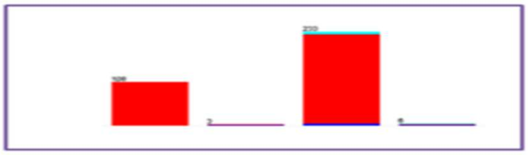
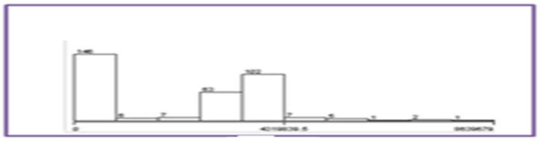


Figure 20 :Capital before discretization

Figure 21: Capital after discretization



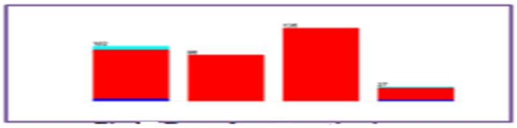


Figure 22: Rest before discretization

Figure 23 :Rest after discretization

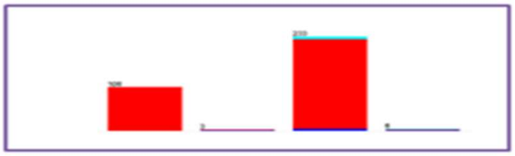


Figure 24 :Deposits before discretization . Figure 25: Deposits after discretization

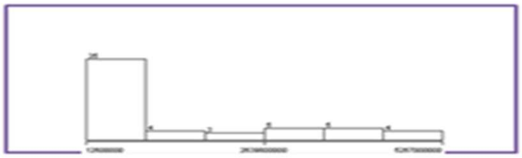


Figure 26: Public deposits before discretization

Figure 27: Public deposits after discretization

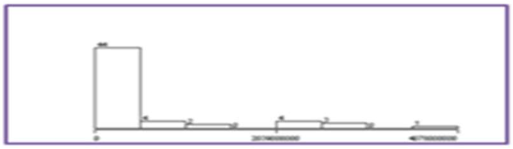


Figure 28: Special deposits before discretization

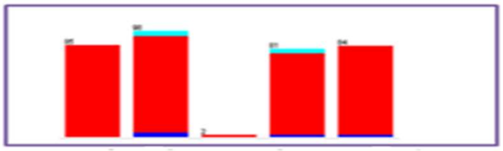


Figure 29 : Special deposits after discretization

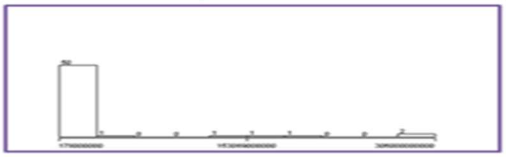


Figure 30: Bankers deposits before discretization

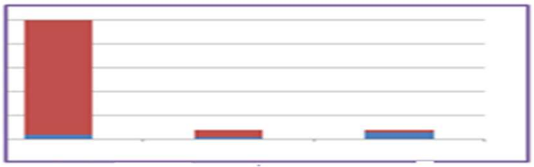


Figure 31: Bankers deposits after discretization

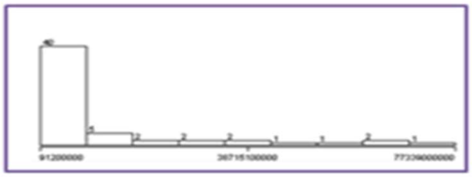


Figure 32: Other accounts before discretization

Figure 33: Other accounts after discretization

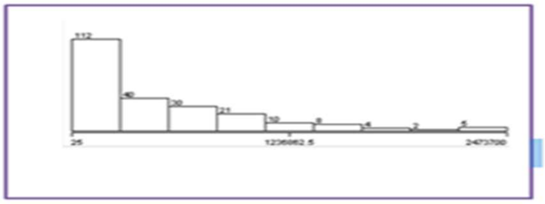


Figure 34: 7day and other Bills before discretization

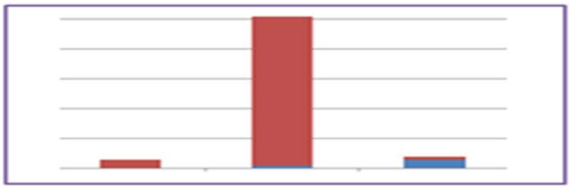


Figure 35: 7day and other Bills after discretization

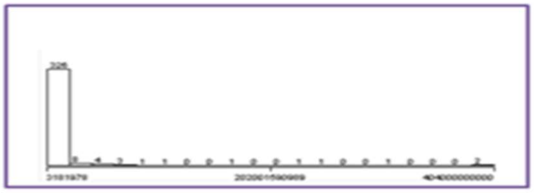


Figure 36 :Total liabilities before discretization

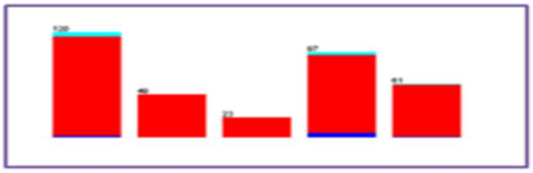


Figure 37: Total liabilities after discretization

5. STANDARDIZATION

We used standardized processing on all attributes in the data set to transform raw data into a dimensionless index, that is, each index value is at the same scale level, due to the varying qualities of the indicators.[16]

6. TRAINING MODEL

To evaluate the three machine learning models, it is necessary to first divide the dataset into two categories, training data, which represents 20% of the data, and test data, which represents 80% of the total data.

Loans, securities, and deposits are used in this classification model to anticipate the target. The training data is utilized to train a fitted and logical model with the goal of identifying probable predictors. Testing data is used to calculate the accuracy of the model prediction, which can demonstrate the model's efficiency and efficacy. Three frequently used classification models, such as DT, SVM, and RF [17] are used in this research.

7. EVALUATION MODEL AND EXPERIMENTAL RESULTS

After dividing the data into two parts at this stage, the test part (30%) was used to evaluate the model.

As we explained earlier. The next step is to enter the test data into the three proposed models to calculate their performance. We used both precisions, recall, and under the curve to calculate the accuracy of the models, the results showed that the SVM system had obtained a 95.79%, DT had obtained a 96.63% and RF had obtained a 96.65 before processing the missing data with training 10-fold, as shown in Table 3. Then we tested the three models again, but after processing the missing data, we found an improvement in the results for each of the three models: the DT, SVM, and RF, as in table 3.

The researcher can use these models to investigate the link between various data sets and anticipate the outcome. The discriminatory effects of the models stated above may be quantified by comparing the accuracies of model predictions and computing the value of the area under the ROC curve (AUC). In addition, the ROC curve is depicted at the same time. The results showed that training the model at a 66-fold is higher than training the model at a 10-Fold.

Table 2: Training Before and After preprocessing 10-fold

	Before preprocessing	After preprocessing
Techniques	10-fold	10-fold
DT	95.12	95.71
SVM	95.70	95.71
RF	95.14	95.55

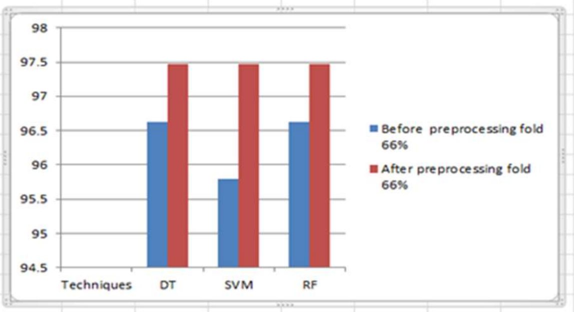


Figure 38: Training Before and After preprocessing 10-flod

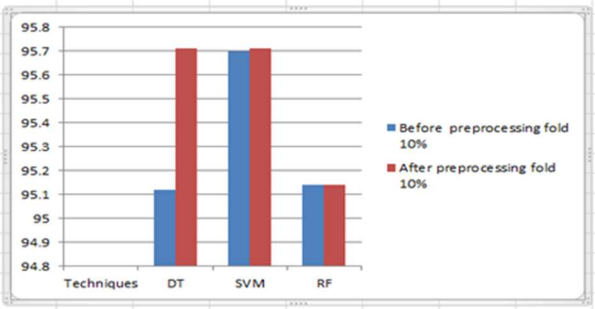


Fig 39 Training Before and After preprocessing 66-flod

This table is a comparison of three techniques for the decision tree, support vector, and random forest, where the results show that the accuracy score when training with 66-fold before preprocessing was DT 96.63, SVM 95.79, and RF 96.63 The results after preprocessing are shown DT 97.47, SVM 97.47 and RF 97.47 improved results with a higher accuracy rate in the case of training by 66-flod. As in Table (4), as in Figure (39).

Table 2 is a comparison of three techniques for the decision tree, support vector, and random forest, where the results show that the accuracy degree when training with 10% before preprocessing was DT 95.12, SVM 95.70, and RF 95.14 and the results after preprocessing show DT 95.71, SVM 95.71 and RF 95.55. Improved results accuracy degree by a very small percentage in the 10-fold training. As in Table (2), as in Figure (38).

Table 3: Training Before and After preprocessing 66-fold

	Before preprocessing	After preprocessing
Techniques	66-fold	66-fold
DT	96.63	97.47
SVM	95.79	97.47
RF	96.63	97.47

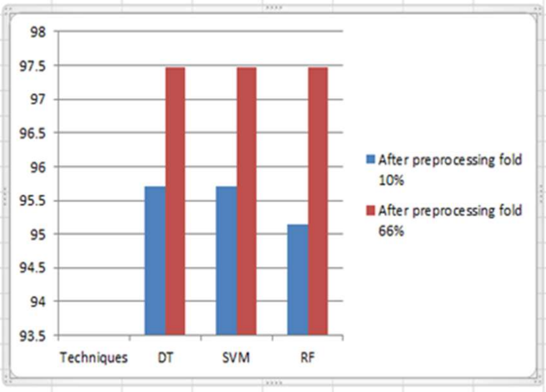


Fig 40 :Training After preprocessing 10-fold and 66-fold

The data set used consisted of 351 features after its pretreatment. The results showed that the degree of training accuracy at 66-fold is higher than the degree of training accuracy at 10-fold as in Table (3,4) as in Figure (40).

The evaluation criteria in this research was the balance sheet of the Bank of England to take advantage of the liquidity ratio in previous years, as the bank used the traditional methods of calculating liquidity, and it became clear to us through the use of machine learning algorithms that it is the best in predicting bank liquidity.

8. CONCLUSIONS

prediction of bank liquidity is one of the major challenges in banking as it represents a serious financial threat to it. prediction of liquidity risk, monitoring, reliability of the model, and effective liquidity treatment is essential for decision-making and transparency. Many banks use traditional methods to predict bank liquidity, and these methods are outdated to deal with these problems.

Real-world datasets often have missing values, Therefore, we applied advanced statistical methods and machine learning algorithms that can handle these problems.

we proposed a method that uses three machine learning techniques (decision trees - support vectors machine - random forests). The model variables are liquidity ratios and they have been selected on the basis of the available data from the balance sheet of the Bank of England.

The results showed that the accuracy of the model in training at 66-fold is higher than the accuracy of the model in training at 10-fold.

Despite the many possibilities of machine learning, it is rarely used in predicting bank liquidity risks, and therefore this study contributes to filling a very important gap that still exists.

It is obtaining acceptable liquidity for banks that do not lead to bankruptcy of the banks and does not lead to the failure to make the best use of the bank's money and profit from it as much as possible because the increase in liquidity leads to the lack of optimal exploitation of the funds and the lack of liquidity may lead to the closure and bankruptcy of the banks.

REFERENCES

[1] Leo, Martin, Suneel Sharma, and Koilakuntla Maddulety. "Machine learning in banking risk management: A literature review." Risks 7.1 (2019): 29.

[2] Petropoulos, Anastasios, et al. "A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting." IFC Bulletins chapters 49 (2019).

[3] Addo, Peter Martey, Dominique Guegan, and Bertrand Hassani. "Credit risk analysis using machine and deep learning models." Risks 6.2 (2018): 38.

[4] Alonso, Andrés, and José Manuel Carbó. "Machine learning in credit risk: measuring the dilemma between prediction and supervisory cost." Documentos de Trabajo/Banco de España, 2032 (2020).

[5] Devi, Ms Usha, and Neera Batra. "Exploration Of Credit Risk Based On Machine Learning Tools." Journal of Critical Reviews 7.19 (2020): 4698-4718.

[6] Lehdili, Noureddine, Pascal Oswald, and Harold Gueneau. "Market Risk Assessment of a trading book using Statistical and Machine Learning."2019.

[7] Aziz, Saqib, and Michael Dowling. "Machine learning and AI for risk management."

Disrupting Finance. Palgrave Pivot, Cham, 2019. 33-50.

[8] Tavana, Madjid, et al. "An Artificial Neural Network and Bayesian Network model for liquidity risk assessment in banking." *Neurocomputing* 275 (2018): 2525-2554.

[9] An, Chang-Ho. "A Study on Estimation of Financial Liquidity Risk Prediction Model Using Financial Analysis." International Journal of Applied Engineering Research 12.20 (2017): 9919-9923.

[10] Petchamé Sala, Jordi. Liquidity risk modeling using artificial neural network. MS thesis. Universitat Politècnica de Catalunya, 2011.

[11] https://www.kaggle.com/sohier/the-bank-of-englands-balance-sheet?select=balance.xlsx

[12] Enders, Craig K. Applied missing data analysis. Guilford press, 2010.

[13] Eekhout, Iris, et al. "Missing data in a multi-item instrument were best handled by multiple imputation at the item score level." Journal of clinical epidemiology 67.3 (2014): 335-342.

[14] van Ginkel, Joost R., et al. "Rebutting existing misconceptions about multiple imputation as a method for handling missing data." Journal of Personality Assessment 102.3 (2020): 297-308.

[15] Garcia, S., Luengo, J., Sáez, J. A., Lopez, V., & Herrera, F. (2012). A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering, 25(4), 734-750.

[16] Mohamad, Ismail Bin, and Dauda Usman. "Standardization and its effects on K-means clustering algorithm." Research Journal of Applied Sciences, Engineering and Technology 6.17 (2013): 3299-3303.

[17] Yadav, Sanjay, and Sanyam Shukla. "Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification." 2016 IEEE 6th International conference on advanced computing (IACC). IEEE, 2016.