ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

Prediction Of Loan Status In Commercial Bank Using Machine Learning Classifier

1. E.Rupa ,Assistant Professor,CSE,Sri Indu Institute of Engineering&Technology(SIIET), Sheriguda, Ibrahimpatnam,Hydarabad

2. A.Shiva Shankar,Assistant Professor,CSE,SIIET,Sheriguda,Ibrahimpatnam,Hydarabad
3.M.Ramya Student,CSE,SIIET,Sheriguda,Ibrahimpatnam,Hydarabad

3. 4.N.Ditwar Student,CSE,SIIET,Sheriguda,Ibrahimpatnam,Hydarabad

4. 5.T.Bharath Student,CSE,SIIET,Sheriguda,Ibrahimpatnam,Hydarabad
6.N.Sunilyadav Student,CSE,SIIET,Sheriguda,Ibrahimpatnam,Hydarabad

Abstract:

The revenue from the loans directly accounts for the majority of the bank's profit. Additionally, one of the key characteristics of the banking industry is the credit danger. The vaticination of credit defaulters is one of the delicate tasks for any bank, but by vaticinating the loan defaulters, the banks surely may reduce their loss by reducing its non-profit means so that recovery of approved loans can take place without any loss and it can play as the contributing parameter of the bank statement. This makes the study of this loan eligibility vaticination important. Machine learning ways are veritably pivotal and useful in the vaticination of these types of data. In order to save a lot of money and bank sweats, we attempt to lessen the threat element that drives people to choose the secure person in this paper. This is accomplished by mining the Big Data belonging to the previously loan issued individuals, and based on this data, machine learning models were taught to produce the most accurate results. This paper's main goal is to determine which machine algorithm performs best at predicting whether or not a person is qualified for a loan.

Keywords: Prediction, Training, Testing, Loan, Machine Learning.

I. Introduction:

A loan is a third-party gift of money, property, or other tangible items in exchange for the subsequent early repayment of the loan amount including interest [4]. Until the loan is returned, the grantee incurs a debt for which they are often liable to pay interest. Numerous people are seeking for bank loans as a result of the banking industry's progress, but because banks only have limited resources to distribute among their customers, it is usually best for them to take a chance on someone they know. Credit threat evaluation is, as we all know, critically important. Threat position computation can be done in a number of different methods [12]. Even though the bank authorises the loan following a lengthy verification and validation process, there is still no guarantee that the grantee chosen is +secure. When done manually, this operation takes a long time. Every business that engages in lending encounters the difficulty of conformation of loans on a regular basis. However, it can reduce the number of man-hours required and speed up client service. The improvement in customer satisfaction and cost savings by automating the loan conformation procedure are substantial [13]. However, in order to remove the implicit threat, the bank must have a reliable model in place that allows it to directly read which customer loans it

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

should accept and which it should reject. Only then will the benefits be realised. The credit system controlled by banks is one of the most crucial elements that determine our nation's thriftiness and financial situation [14]. We can prognosticate whether or not a specific seeker is secure, and the entire validation process is done automatically using machine learning techniques. However, the appropriate characteristics include things like gender, education, dependents, income, loan amount, conjugal status, history of credit and others [9]. If the company want to partially automate the loan qualifying procedure based on information furnished by the customer during the online operation form submission. The customer as well as a bank's retainer both benefit greatly from loan prediction. Even if many people are seeking for loans, it can be challenging to choose the sincere candidate who would pay back the loan [11]. Many misconceptions may arise when choosing candidates manually, and since accurate predictions are crucial for maximising returns, it's crucial to understand the differences between the various approaches and compare them [5]. As a result, we are creating a method for automatically determining if a loan seeker is qualified by reviewing various machine learning models. Both clients and bank personnel will benefit from this. There will be a significant reduction in the loan approval process's timeframe.

The organization of this article is in following manner i.e. Section-II describes the research background, where we reviewed and analysed about all the literatures, Section-III denotes about the existing regime, Proposed methodology explained in Section-IV, results and discussions are demonstrated in Section-V and finally conclusions are noted in Section-VI.

II. Research Background:

Observation 1:

In their research paper titled "Prediction of Loan Risk using NB and Support Vector Machine," S. Vimala and K. C. Sharmili suggested a loan valuation model that utilised both Support Vector Machine and Naïve Bayes methods [1]. A self-reliant presumption method called Naive Bayes includes probability propositions related to data stratification. SVM, on the other hand, stratifies predictions using a statistical learning model. In order to estimate the suggested system, a data set from the UCI depository containing 21 attributes was employed. The results of the trials showed that the merging of the Support vector Classifier and Naïve Bayes resulted in an efficient stratification of loan eligibility rather than the independent efficiencies of the classifiers.

Observation 2:

Ranpreet Kaur and Anchal Goyal's research work, "Loan prediction using ensemble technique" offered a useful vaticination method that aids bankers in predicting the credit threat for consumers who have filed for loans [2]. The paper describes a prototype that organisations can use to decide correctly whether to accept or reject the applicants' loan request. In addition to three separate models, the paper employs the ensemble model, which combines the three models and analyses the credit threat for the best results (Support Vector Machine Model, Tree Model for inheritable Algorithm and Random Forest Network).

Observation 3:

The study "Overdue Prediction of Bank Loans Based on LSTM-SVM" by authors Xianzhong Long, Xin Li, Guozi Sun, Huakang Li and Geng Yang elaborates the current successful background, traditional threat soothsaying system and substantially introduces the main operation of the Long Short Term Memory-Support Vector Machine model in customer loan threat vaticination [6,15]. On this foundation, the LSTM framework and SVM system-based vaticination

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

methodology is suggested, the vaticination outcomes are compared to those of the conventional approach, and the model's viability is validated. However, the LSTM-SVM system suggested in this paper has a number of shortcomings and has to be improved in subsequent research.

Observation 4:

The study paper by Sandip Pandit, Mrunal Surve, Priya Shinde, Swati Sonawane and Pooja Thitme, titled "Data mining techniques to analyse risk giving loan (bank)" primarily focuses on identifying and analysing the threat of providing a loan to commercial banks [7]. They have employed data mining methodology to determine risk when disbursing loans. It entails analysing and recycling data from various agencies in order to recapitulate it as priceless data. For forecasting the threat chance for a person to grant loans, they employed the C4.5 stratification algorithm.

Observation 5:

The authors of the study "A machine learning strategy for forecasting bank credit worthiness" Regina Esi, Turkson, Edward Yeallakuor Baagyere, and Gideon Evans Wenya have delved into the plethora of machine learning models that may be used to predict a loan applicant's eligibility [8]. In order to discover which algorithms are the best fits for analysing bank credit data sets, they applied 15 different algorithms to the data set. Among the algorithms utilised are Linear Regression, Logistic Retrogression, Discriminant Analysis, Naive Bayes, K-Nearest Neighbor, Neural Networks, Decision Trees and Ensemble expert systems [15,16]. Excluding the Gaussian Naive Bayes and Nearest Centroid, the trial showed that the remaining algorithms perform cogently well in accordance of accuracy and sundry performance indicators.

III. Existing Regime:

The existing regime, CIBIL, is a credit scoring system that gathers and keeps track of information about individual and marketable reality payments related to loans and credit cards. Every month, banks and other lenders send these records to TransUnion, a credit reporting firm. With the use of this data, a CIBIL Score and profile for each individual is created. A three-number numerical representation of your credit history is known as a CIBIL Score. The CIBIL Report's credit history is used to calculate the score. A CIR is a record of a person's credit payment history over a period of time across several loan kinds and credit agencies. Your funds, financing, or fixed income securities are not mentioned in a CIR [3]. This allows lenders to organise and authorise loan operations. While the CIBIL Score is vital in the loan application process, this doesn't always provide a full view. The CIBIL Score is the issuer's first conviction; the higher the score, the more likely the loan will be evaluated and sanctioned. The lender alone must decide whether to advance, and they must also take into account a number of other variables. The loan or credit card should not be sanctioned, according to CIBIL. Based on the information supplied by the customer, such as their gender, conjugal status, earnings, quantity of dependents, loan amount, degree, history of credit, etc., the lender must manually analyse each application and determine whether it is creditworthy.

IV. Proposed Methodology:

Here, under this approach, we determine a client's eligibility for a loan from a bank and contrast various models to determine which one is the best fit for this use. In order to do this, we created an automatic loan predication platform using machine learning approach. We trained the machine using the prior dataset so that it could analyse and comprehend the process. It will also determine whether the loan applicant is eligible. For the purpose of determining the best functional model for determining loan eligibility, five learning algorithms are assessed. Thirteen aspects make up the dataset we're using: Loan ID,

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

Gender, Married, Dependents, Education, Self-Employed, Applicant Income, Co-applicant Income, Loan Amount, Loan Amount Term, Credit History, Property Area, and Loan Status.

The following inflow illustration demonstrates how the proposed system contains various stages. In order to facilitate communication between the user and the system, a website must be developed.

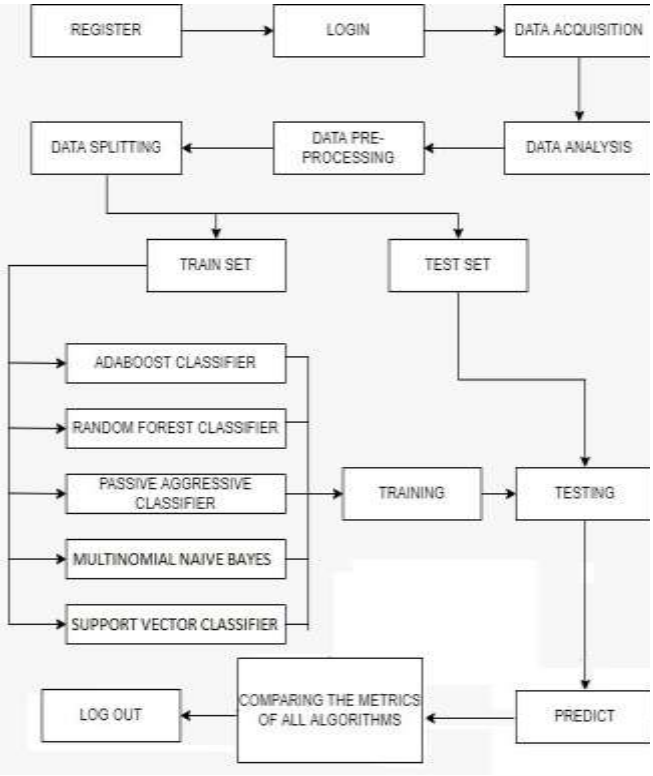

Figure 4.1. Flow diagram for the proposed methodology

The structure of the process described here is seen in figure 4.1 above.

1. **Register:**

The user registers by providing information such their first name, password, email, and other details that are recorded in the SQLite database.

2. **Login:**

The user may also log in using their username and password, which, if they are accurate, will take them to the website's home page.

3. **Data collection:**

Next, the dataset required for training and test sets is acquired from the internet.

4. **Data Analysis:**

Using descriptive statistics and visualizations, we carry out original examinations on data to look for trends, identify deviations, put theories to the test, and confirm suppositions

5. **Data Pre-processing:**

The raw data which has been acquired during the data accession process is converted into a form that can be understood by the machine.

6. **Data Splitting:**

The dataset is partitioned into two non-overlapping sets, one with a stratification column and one without it.

7. **Test Set:**

A reasonable sampling of the dataset's data that is used to give an unprejudiced evaluation of a final model fit on the training dataset.

8. **Train Set:**

It's a portion of a data set used to fit a model for the vaticination or stratification of values that are familiar in the training set, but unknown in other (hereafter) data.

a) Instantiating the algorithms:

Starting the training process by feeding the train data to the various algorithms, each of which can receive input and produce results. The various algorithms this system employs include:

i. Adaboost Classifier:

During the data training phase, the Adaboost Classifier creates a specific amount of decision trees known as stumps. The incorrectly classified record in the first model is given precedence as the first decision tree / model is constructed. For the alternative model, just these records are sent

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

as input. Until we define the quantity of base learners we would like to produce, the procedure continues. The following formula is used to update sample weights:

Current Weight of Sample = Weight of Sample * e^ (Performance)------- Equation- 1

ii. Passive Aggressive Classifier:

Belonging to the order of online learning algorithms in machine learning, responds passively in cases of proper stratification and aggressively in cases of misinterpretation.

iii. Random Forest Classifier:

In order to increase the predictive accuracy of the supplied dataset, this classifier employs several decision trees across numerous sections of the data. The following information gain calculation serves as the foundation for the data splitting:

Gain (V, Y) = Entropy(V)-Entropy (V, Y) ----------

Equation- 2

Where,

- V is the target variable
- Y is a point to be resolved on
- The entropy determined when the data is resolved on point X is called entropy (V, Y).

iv. Multinomial Naïve Bayes:

Based on the Bayes theorem, this method forecasts the label of a word that resembles a dispatch or review composition. It determines the likelihood of each label for a particular sample and outputs the label with the highest likelihood. The chance of the label appearing in the word can be determined using the formula below. P(E|X) = P(E)*P(X|E)/ P(X). ------ Equation- 3

Where,

- P(X) is predictor X's is prior probability
- P(E) is class E's prior probability
- P(X| E) is the eventuality of predictor X given class E probability

v. Support Vector classifier:

A Linear SVC (Support Vector Classifier) should fit the data you provide and deliver a "pre-eminent fit" hyperplane that partitions or classifies your data. After obtaining the hyperplane, one could additionally provide their classifier with a few features to determine what the "prognosticated" class is.

9.Training:

Algorithms for machine learning get knowledge from data. From the training data learners are provided, they establish connections, gain understanding, form views, and gauge their level of confidence.

10.Testing:

After the algorithm has been trained, we test it using data from the training set. These results allow us to assess the performance of each approach.

11.Predict:

The models that have been trained using the dataset also prognosticate whether the client could acquire a loan or not.

12.Comparing the metrics of all algorithms:

Recall, Precision and Accuracy are the three primary criteria used to estimate a stratification model. In this case, we use accuracy, which is determined by the formula below.

Accuracy = correct number of predictions/total number of predictions ------- Equation- 4

13. Logout:

If we click on the logout link it'll come out from the front page of the website.

V. Results Analysis:

The models trained on the dataset performed distinctly as given by the accuracies below:

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

Random Forest Classifier:

Figure 5.1. The Random Forest Classifier's Accuracy on the Dataset

Passive Aggressive classifier:

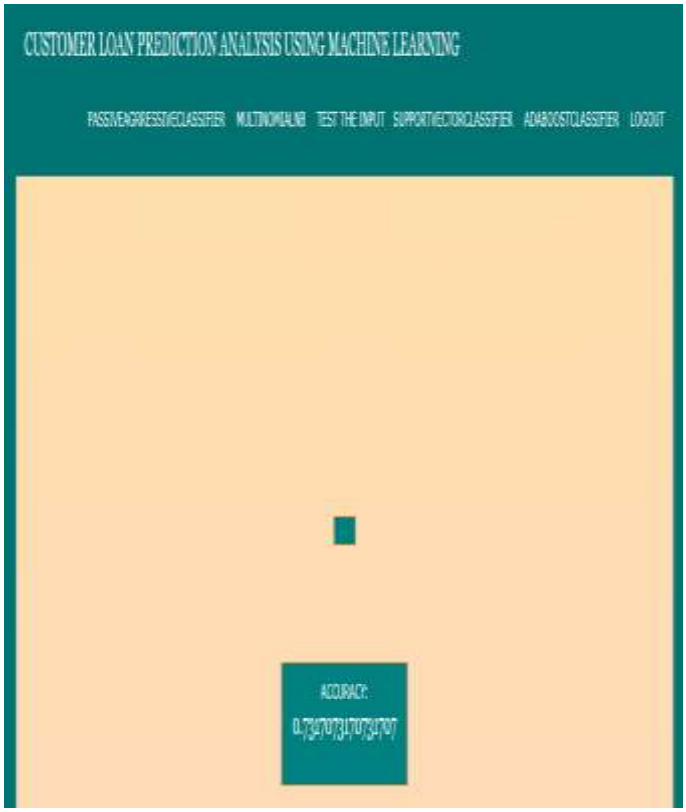

Figure 5.2. The Passive Aggressive Classifier's Accuracy on the Dataset

Multinomial Naïve Bayes:

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

Figure 5.3. The Multinomial Naïve Bayes's Accuracy on the Dataset

Support Vector Classifier:

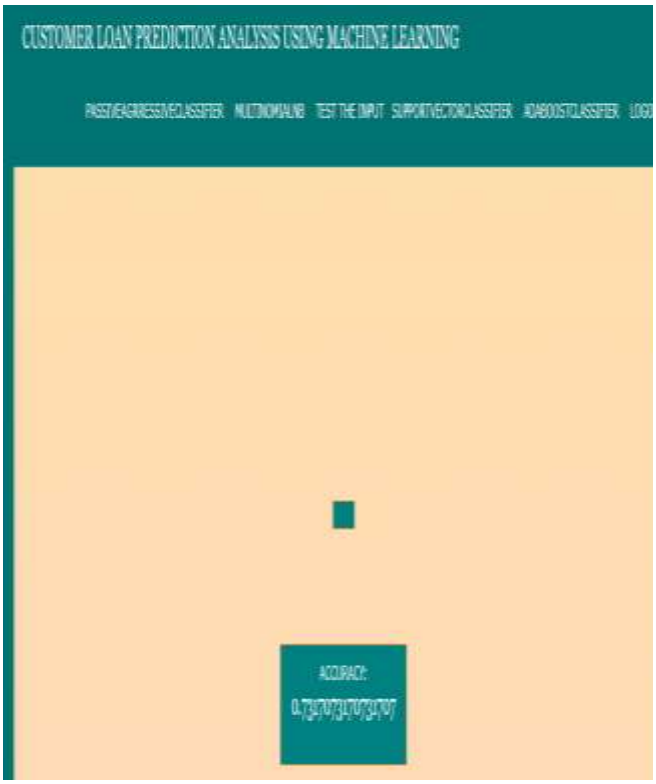

Figure 5.4. The Support Vector Classifier's Accuracy on the Dataset

Adaboost Classifier:

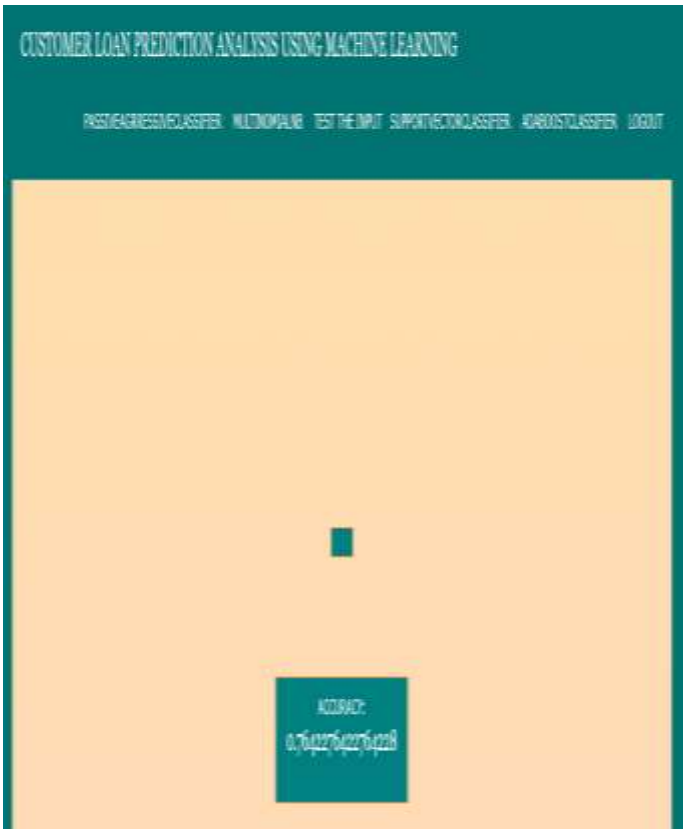

Figure 5.5. The Adaboost Classifier's Accuracy on the Dataset

Table 1: Accuracies of the five algorithms on the given dataset

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

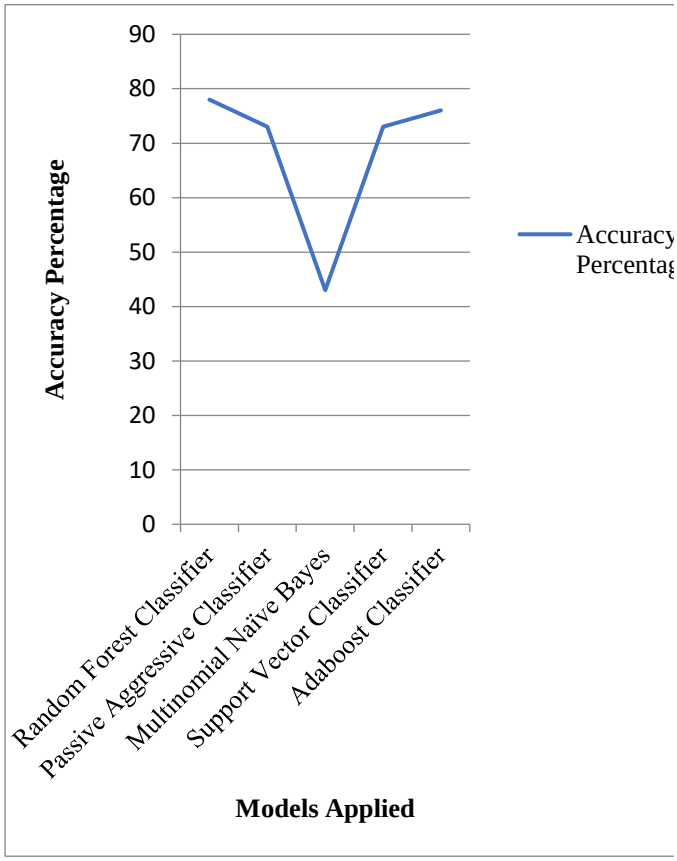

Figure 5.6. The accuracy percentages of the five algorithms depicted in a line chart.

S. No	Algorithm used	Accuracy
1	Random Forest Classifier	0.7820773930753564
2	Passive Aggressive Classifier	0.7317073170731707
3	Multinomial Naïve Bayes	0.43089430894308944
4	Support Vector classifier	0.7317073170731707
5	Adaboost Classifier	0.7642276422764228

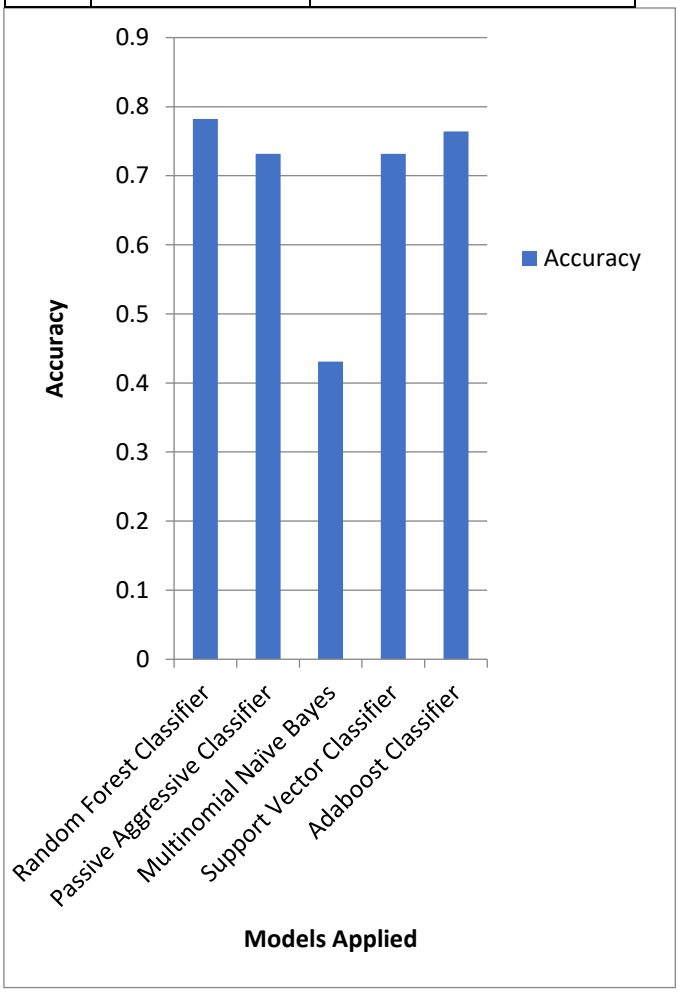

Figure 5.7. The accuracies of the five algorithms depicted in a Bar Graph.

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

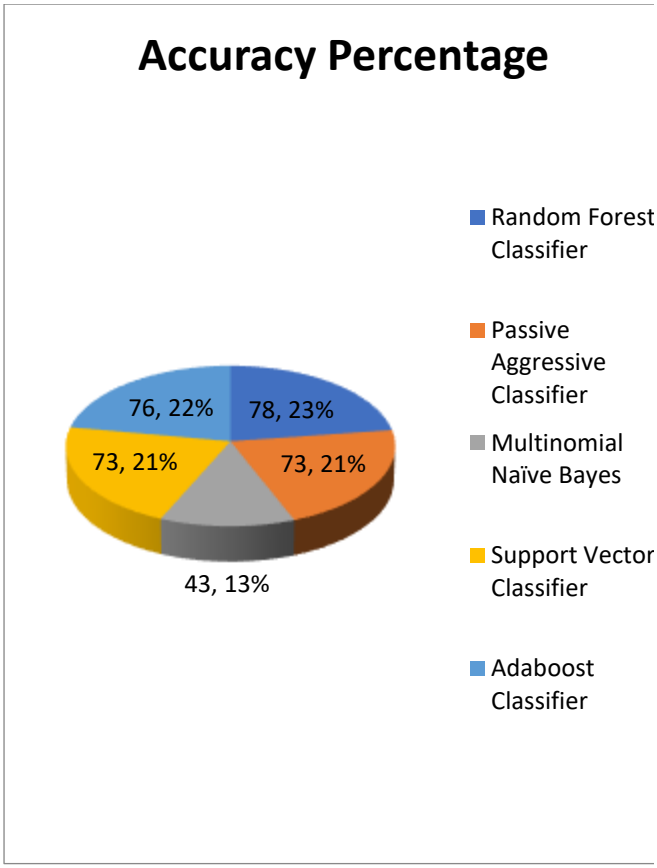

Figure 5.8. The accuracy percentages of the five algorithms depicted in a Pie chart.

VI. Conclusion:

To identify a dominant model, we have provided various machine learning models utilising various machine learning techniques. The model built using the Random Forest Algorithm surpassed all other models in correctly classifying the data with a rate of 78 percent, according to the examination of the results. It is safe to say that the solution is a generally effective element after a proper review of the element's advantages and limitations. This service is operating duly and in compliance with banking standards. This construct is simple to plug into a plethora of other systems. These results, in our opinion, will help researchers learn more about the topic of developing a vaticination study that can predict a client's loan eligibility.

Acknowledgement:

We are grateful to St.Peter's Engineering College, department of CSE for helping us with the laboratory and for continuing support to prepare this paper in a brighter manner.

References:

[1] S. Vimala, K.C. Sharmili, "Prediction of Loan Risk using NB and Support Vector Machine", International Conference on Advancements in Computing Technologies (ICACT 2018), vol. 4, no. 2, pp. 110-113, 2018.

[2] Anchal Goyal, Ranpreet Kaur, "Loan Prediction Using Ensemble Technique", International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, no. 3, March 2016.

[3] "Understand Your Credit Score and Report" [online] Available: https://www.cibil.com/faq/understand-your-credit-score-and-report.

[4] Julia Kagan, "loan", april 2021, [online] Available: https://www.investopedia.com/terms/l/loan.asp.

[5] Ashwini S. Kadam, Shraddha R. Nikam, Ankita A. Aher, Gayatri V. Shelke, Amar S. Chandgude, "Prediction for Loan Approval using Machine Learning Algorithm", International Research Journal of Engineering and Technology (IRJET), Volume: 08, Issue: 04, Apr 2021.

[6] Xin Li, Xianzhong Long, Guozi Sun, Geng Yang, and Huakang Li "Overdue Prediction of Bank Loans Based on LSTM-SVM" 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 2018, pp. 1859-1863, doi: 10.1109/SmartWorld.2018.00312.

[7] Mrunal Surve, Pooja Thitme, Priya Shinde, Swati Sonawane, and Sandip Pandit, "Data mining techniques to analyze risk giving loan (bank)", Internation Journal of Advance Research and Innovative Ideas in Education Volume 2, Issue 1, 2016.

[8] Turkson, Regina Esi, Edward Yeallakuor Baagyere, and Gideon Evans Wenya, "A machine learning approach for predicting bank credit worthiness.", 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR). IEEE, 2016.

[9] L. Udaya Bhanu, Dr. S. Narayana, "Customer Loan Prediction Using Supervised Learning Technique", International Journal of Scientific and Research Publications, Volume 11, Issue 6, June 2021.

ISSN: 2278-4632
Vol-10 Issue-12 No.03 December 2020

[10] Mohammad Ahmad Sheikh, Amit Kumar Goel,Tapas Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm", International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020.

[11] Dr. C K Gomathy, Ms.Charulatha,Mr.AAkash ,Ms.Sowjanya, "THE LOAN PREDICTION USING MACHINE LEARNING", International Research Journal of Engineering and Technology (IRJET), Volume: 08, Issue: 10, Oct 2021.

[12] J. Tejaswini1, T. Mohana Kavya, R. Devi Naga Ramya, P. Sai Triveni Venkata Rao Maddumala, "Accurate loan approval prediction based on machine learning approach", www. jespublication.com, page 523 Issue 4, April/ 2020 ISSN NO: 0377-9254.

[13] V. C. T. Chanetal., "Designing a Credit Approval System Using Web Services, BPEL, and AJAX", 2009 IEEE International Conference on e-Business Engineering, Macau, 2009, pp. 287-294.doi: 10.1109/ICEBE.2009.46.

[14] Nitesh Pandey, Ramanand Gupta, Sagar Uniyal, Vishal Kumar, "Loan Approval Prediction using Machine Learning Algorithms Approach", IJIRT, Volume 8, Issue 1,2021.

[15] Amjan Shaik, et al, "Sentiment Extraction and analysis using Machine Learning Tools: Survey", IOP Conference series: Material Science & Engineering, SCOPUS, December 2018.

[16] Amjan Shaik, et al, "Analysis of effective medical record storage formats and demonstration of time efficient secure storage framework", in European Journal of Molecular & Clinical Medicine (EJMCM), Volume: 7, Issue 6, Pages: 2744-2763, ISSN: 2515-8260, December 2020.