RESEARCH PAPER ON LOAN APPROVAL PREDICTION USING MACHING LEARNING

PRAKASH GIRI (DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING) GALGOTIAS UNIVERSITY,GREATER NOIDA.

prakash.20SCSE1010601@galgotia suniversity.edu.in

SWAPNIL (DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING)

GALGOTIAS UNIVERSITY,GREATER NOIDA.

swapnil7684@gmail.com

Dr. ARSHAD HUSAIN

(Assistant Professor)

(DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING)

GALGOTIAS UNIVERSITY,GREATER NOIDA

# ABSTRACT

Now a days, a lot of people are applying for bank loans for education, house construction, buying cars, and many more expenditures. Some of them are not able to return back the money because of some reasons.

In such scenarios, bank faces a loss. Also, bank haslimited amount of assets which they want to spend only on those customers which can repay the loan amount. Thus, finding the right customers for the banks is one of the most challenging task.

So, we try to reduce the losses faced by banks by selecting the right customers who will repay the loan. We have collected some old data of customers from thebanks andusing these data and some machine learning algorithms, we will trainthe machine and then will test on some test cases.

We will be focusing on the classification problem of “whether the person should be provided the loan or not?”. We have divided the paper in some under-named parts:-

* Collection and distribution of Data into train and test part
* Different results obtained from already available models
* Designing more accurate model
* Results of our model

# INTRODUCTION

The main business for the banks are the loan. They earn a large amount of their money from the interests on the loan they provide. There is a 90% system, according to which, a bank can give loan amount upto the 90% of the capital they have. Thus, they have to choose the right candidate for providing loan so that they does not face loss.

The banks have a very high level of validation checking procedures even though there is no guarantee that whether the chosen person is theright among the remaining ones.

Loan prediction is very important for the banks as well the customers. It will efficientlypredict whether the should

provide the loan to a particular candidate or not. It will consider some parameter and on the weightage of those parameters, it will output the result. The parameters which will affect the loan are selected by consulting banks and reading some articles on it. Some of the parameters which will be considered are:

* **Self Employed:** Whether the person is employed or not.
* **Educational Qualification:** Whether the person is educated or not.
* **Income:** How much the candidate earns.
* **Dependents:** Number of family members which are dependent on thecandidate.
* **Loan Amount:** How much amount the candidate demands.
* **LoanAmountPeriod:** Loan tenure in months
* **Credit History:** Credit history of the applicant
* **Property Area:** Area of the total property the applicant holds

# BASIC WORKING

To evaluate the effectiveness of the models, a comparison is made between the two techniques on five different sector companies namely, JP Morgan, Nike, Johnson and Johnson, Goldman Sachs and Pfizer using both ANN and RF models.

Loan prediction analysis involves assessing credit risk and predicting the likelihood of a borrower's repayment. In this field, credit scores play a crucial role as numerical indicators of creditworthiness, derived from an individual's credit history and financial behavior. Machine learning techniques are commonly employed, necessitating careful feature engineering to enhance model performance. The evaluation of these models involves metrics such as accuracy, precision, recall, and F1 score, with overfitting and underfitting

# BASIC DESIGN/WORKING

The historical data for the five companies has been collected from Yahoo Finance . The dataset includes 10 year data from 4/5/2009 to 4/5/2019 of Nike, Goldman Sachs, Johnson and Johnson, Pfizer and JP Morgan Chase and Co. The data contains information about the stock such as High, Low, Open, Close, Adjacent close and Volume. Only the day-wise closing price of the stock has been extracted.

# BASIC CONCEPTS

We have done the project using google collab which is a free open system software designed by google to execute tasks using cloud processor.

Colab is one of the most used tools by data scientists for executing machine learning and deep learning projects. It uses Jupyter Notebook on the back-end torun ML projects. Some of the libraries used in the projects are:

* Pandas
* Numpy
* Mathplotlib
* Seaborn
* Sklearn

We have used Pandas for importing and getting some statistical values like mean and deviation of the data. Pre-processing has been done using “pandas” and “numpy”.

We have used the visuals at most of the place in our project to understand the data better and most of these graphs like count of employed participant, loan amount-density, loan amount term are made using the powerful “sea born” library. Some of the plots were also made by using “Mathplotlib”. The “skylearn” library contains all the machine learning algorithms like ”gradient Boosting Classifier”, “Random forest classifier”, “Decision tree classifier” and K Neighbours Classifier”. We have used these algorithms in our project to get the desired precision.

# PROBLEM STATEMENT

Account firms and banks need to automatize the credit qualification activity (continuously) essentially dependent on data given by customers when rounding out an online structure. Sex, Marital Status, Education, Number of Dependents, Salary, Loan Amount, Credit History, and different subtleties are incorporated. To digitize this interaction, they made an issue to group the client sections that can apply for a credit sum, permitting them to focus on these clients explicitly. They have introduced a fractional informational collection for this situation. “Approval of Loan is a very common real-life problem that every company faces in their lending operations. If the loan approval process is automated, it can save a lot of man hours and improve the speed of service to the customers.

The increase in customer satisfaction and savings in operational costs are significant”. “However, the rewards can only be realised if the bank has a sturdy model in place to accurately forecast which client's loans it should accept and which it should reject, in order to reduce potential risk.

# LITERATURE SURVEY

Vaidya had suggested a method for approving loan forecasts using logistic regression”“Logistic Regression is one of the most popular and very useful classification based algorithm” The purpose or the importance of using Logistic Regression was that it uses the concept of predictive analysiswhich was suitable enough for describing the data “M. Bayraktar et al. proposed a method for credit risk analysis using machine learning. Boltzman machine was used to make the analysis for risk calculation of loan “Y. Shi and P. Song proposed a method for evaluating project loans using risk analysis.

The method evaluated the risk involved in loans of commercial banks”“V. C. T. Chan et al. proposed a credit approval system using web-services. For clients loan as approved by the system. The consumer provides extra relevant information with the credit application. This information’s are processed by Credit Approval System which finally give credit score to the applicant. The paper developed a web services based solution of this problem After going through this, it is found that loan approval prediction task is very crucial for banking system. Machine learning algorithm are very helpful in predicting outcomes even when data is huge in size.

# MACHINE LEARNING ALGORITHM

Random Forest :

Random Forest is a prominent learning method in Supervised Machine Learning, and it is effective for Regression & Classification tasks in ML

1. It generates random forests and then uses these random forests to seek thesolutions.
2. It is an ensemble learning in which significant number of classifiers are used to solve a complex problem.
3. Random forest analyzes each tree for prediction rather than just one to avoidoverfitting issues.
4. The greater the number of trees, the greater the accuracyin problem solving.

### SUPPORT VECTOR MACHINE (SVM) SVM is a

well known Supervised Machine Learning algorithm. Currently, the SVM classifier is the most common classifier 1.SVM has demonstrated a wide range of exceptional abilities, particularly in classification problems.

**DECISION TREE** is a non-parametricsupervised machine learning algorithm. 1. It can also be usedfor both classification and regression, although it is more widely used for classification technique. 2. It works with categorical and continuous variables.

Adaboost( ADB):

Adaptive Boosting is also known as adaboost. It is delicate to noisy information data and outliers. It&#39;s some what different from neural network systems and SVM because Adaboost preparing methodology chooses just those tricks known to enhance the foreseeing power of the model, dwindling dimensionality and possibly improving running time as potential features do not need to be reused.

Pandas:

Pandas is a package, present in python, furnishing presto, flexible, and suggestive data structures designed to make working with structured (irregular, multidimensional, potentially miscellaneous) and time series data both easy and intuitive. It aims to be the abecedarian high- position structure block for doing practical, real world data analysis in Python. also, it has the broader thing of getting the most important and

flexible open source data analysis/ manipulation toolavailable in any language. It&#39;s formerly well on its way toward this

Thing.

The initial analysis of our data shows a maximum number of null values in the data . A deep analysis of the records having null entries proved that all missing values were

not actually random . Thus , it was compulsory to fill them up with appropriate values as removing them from the data might lead to wrong predictions . All the missing entries were filled up using KNN or K-Nearest Neighbours algorithm .

KNN is an algorithm which estimates the value of a variable using its nearest or closest neighbours. The number of nearest neighbours which are to be taken into consideration is denoted by value K .

Linear Models( LM):

The Linear Model is numerically different to a colorful retrogression analysis yet burdens its felicity for both different qualitative and multitudinous quantitative variables.

Neural Network( Nnet):

Neural networks are non-linear statistical data modelling tools. They are generally needed to model complex connections between inputs and labors, to find patterns in these data, or to capture the statistical structure in an unknown common probability distribution between observed variables.

# Implementation

# It was chosen from the original dataset using K Means clustering algorithm of ML which divides given data into various clusters or groups . This ensures all the groups of the majority class remain in our reduced dataset thus making it a true representative of the original dataset. The exploratory data analysis (EDA) of the new dataset showed the skewness that is, measure of asymmetry of distributed data, of some of the features like “loan amount “ , “interest rate” etc . Also , boxplot analysis oF ML gave us the amount of outliers present which could potentially ruin the model . This issue was corrected using feature scaling techniques in which every technique was analysed using probability distribution , QQ plot and box plot .

# The aim was to find a suitable transformation of the feature which will help us to make the probability distribution close to normal , points of the QQ plot closer to the straight line and have least outliers in the box plot . After that , the following ensemble models were used : Random Forest - creates several instances of the training set and trains a decision tree on each

# instance . Finally , the predictions of each and every tree are ensembled while working on the test set or data . Gradient Boosting , further , helps to ensure that for each new tree the incorrect predictions of the previous trees are given higher weights. This helps in reducing the incorrect predictions of the model . XG- Boost or Extreme Gradient Boosting has further enhancements like pruning of each trees which regularises the model thus preventing overfitting.

# The receiver operating characteristics curve, also called as roc curve, for Random Forest , Gradient Boosting and our final model was plotted . This helped visualize the performance of the best performing models in this data.

# LOAN REJECTION OVERVIEW

A personal loan is a versatile unsecured loan and it can be rejected due to several reasons. Multiple personal loan rejections could increase your hard enquiries, ultimately affecting your credit score negatively. Hard enquiries are a sign that you are hungry for more credit and the lenders might assume you are a risky borrower. If your recent personal loan application has been rejected, you may want to find out the reasons for the denial. Read on to know more why your personal loan application could be rejected.

## REASON FOR PERSONAL LOAN REJECTION:

The reasons for a personal loan rejection can be determined broadly under six reasons, however these vary from one lender to another, each lender have their own internal policy to determine the creditworthiness and risk rating.

* Scant credit score
* Unstable employment
* Insufficient income
* Working in a delisted private company
* Not falling into the age limit
* Living in a location where defaults are high

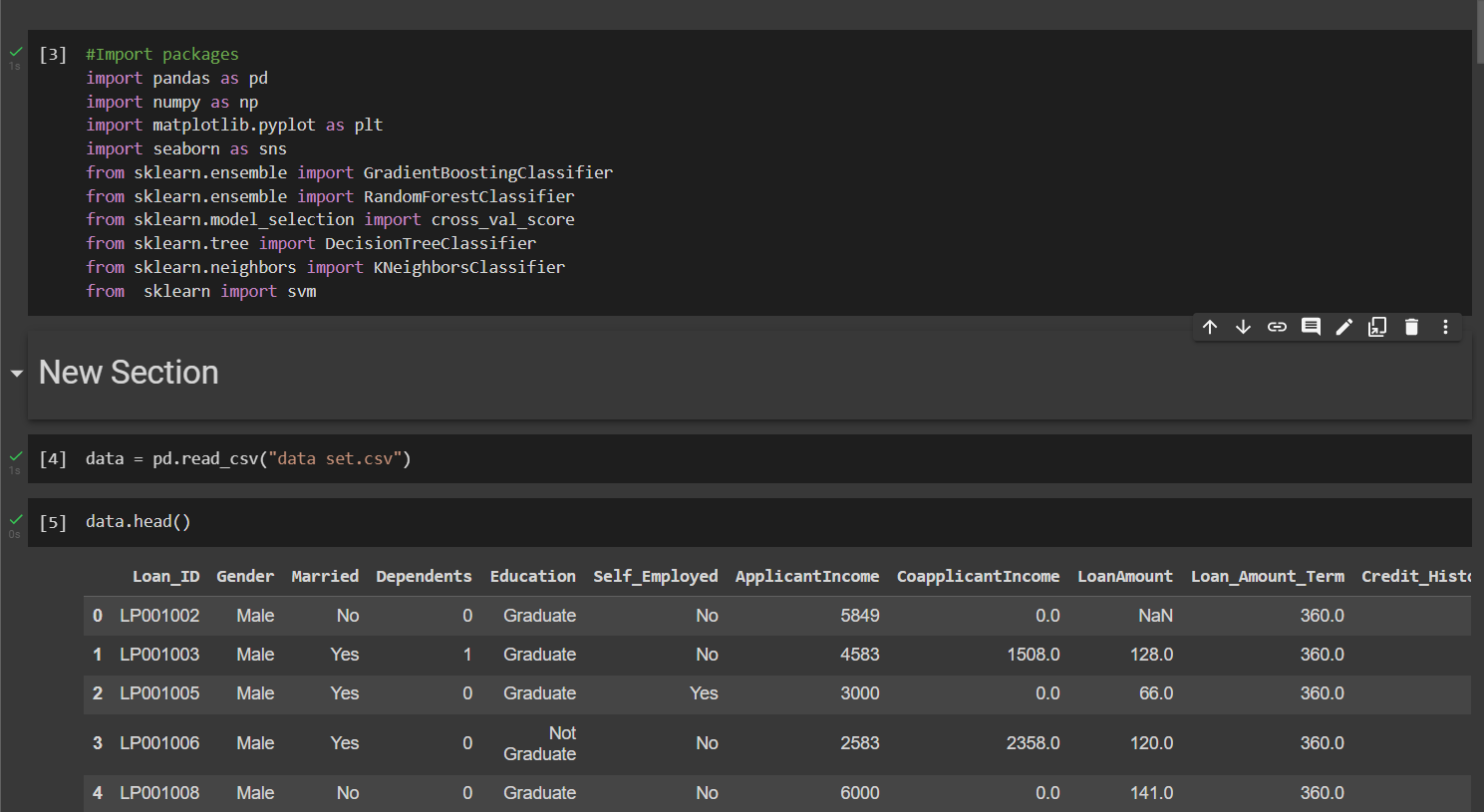
## PERSONAL LOAN REJECTION CRITERIA:

Every bank has its own eligibility criteria for approval of personal loans. Likewise, the reasons for personal loan rejection could vary with the rejection criteria of the other banks. The following could be the main reasons why your personal loan application could be rejected.

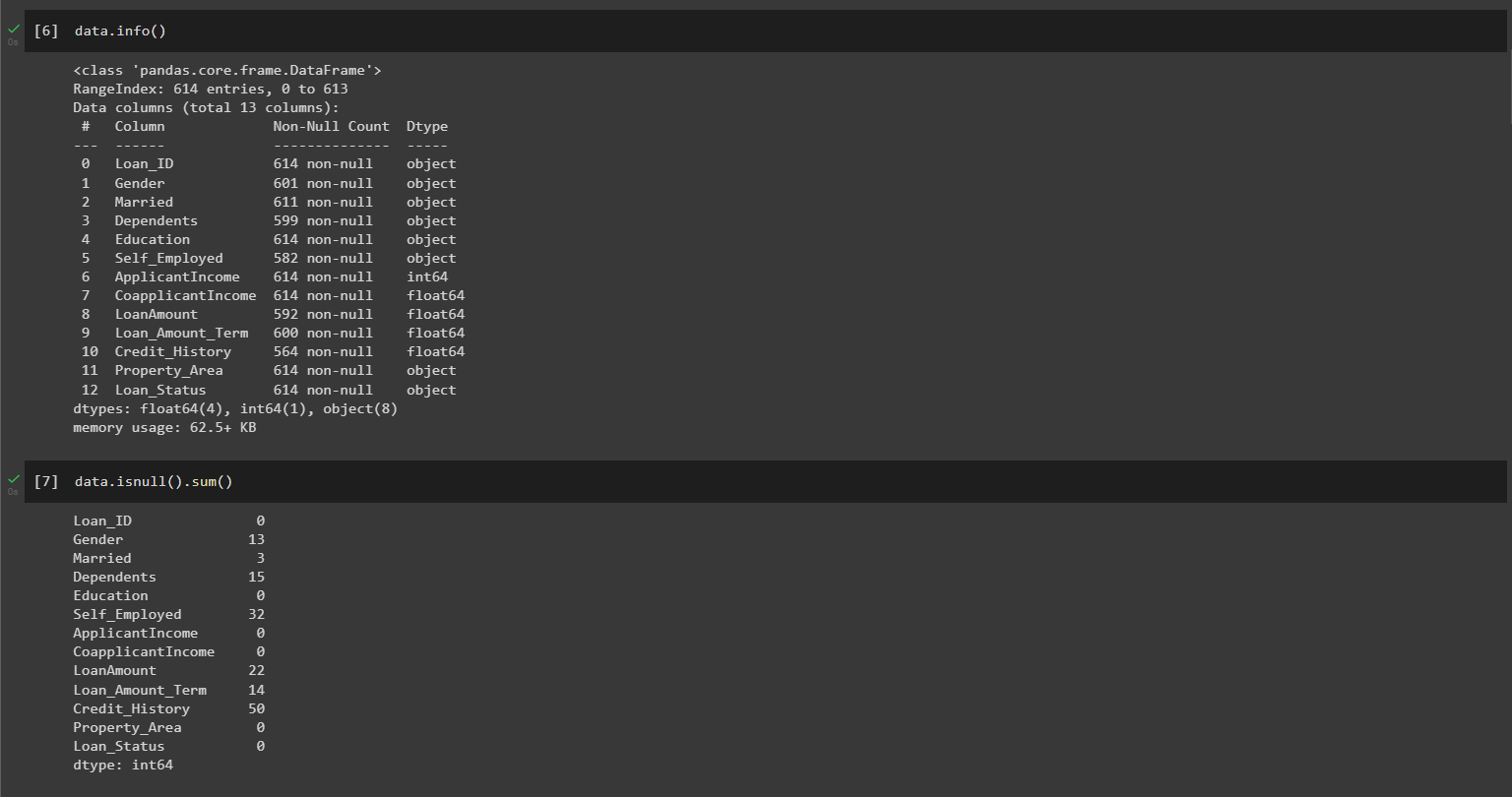
## LOAN REJECTION IMPACTS:-

The answer to this question depends on the level of your personal loan application being processed. The borrower generally applies for a personal loan after checking the eligibility criteria and proceed with document submission. If there are difficulties in furnishing the relevant documents, the personal loan might get rejected.

If all your documents are verified, the lenders will check your credit score to assess your creditworthiness. It is called hard enquiry. In worst case scenario, if your credit score is low and the personal loan gets rejected, the hard enquiry will have an impact on your credit score further in a small way. Hence, if you are faced with multiple hard enquiries, your credit score is sure to take a hit with it.



**Figure 1: Loan Train Data set**



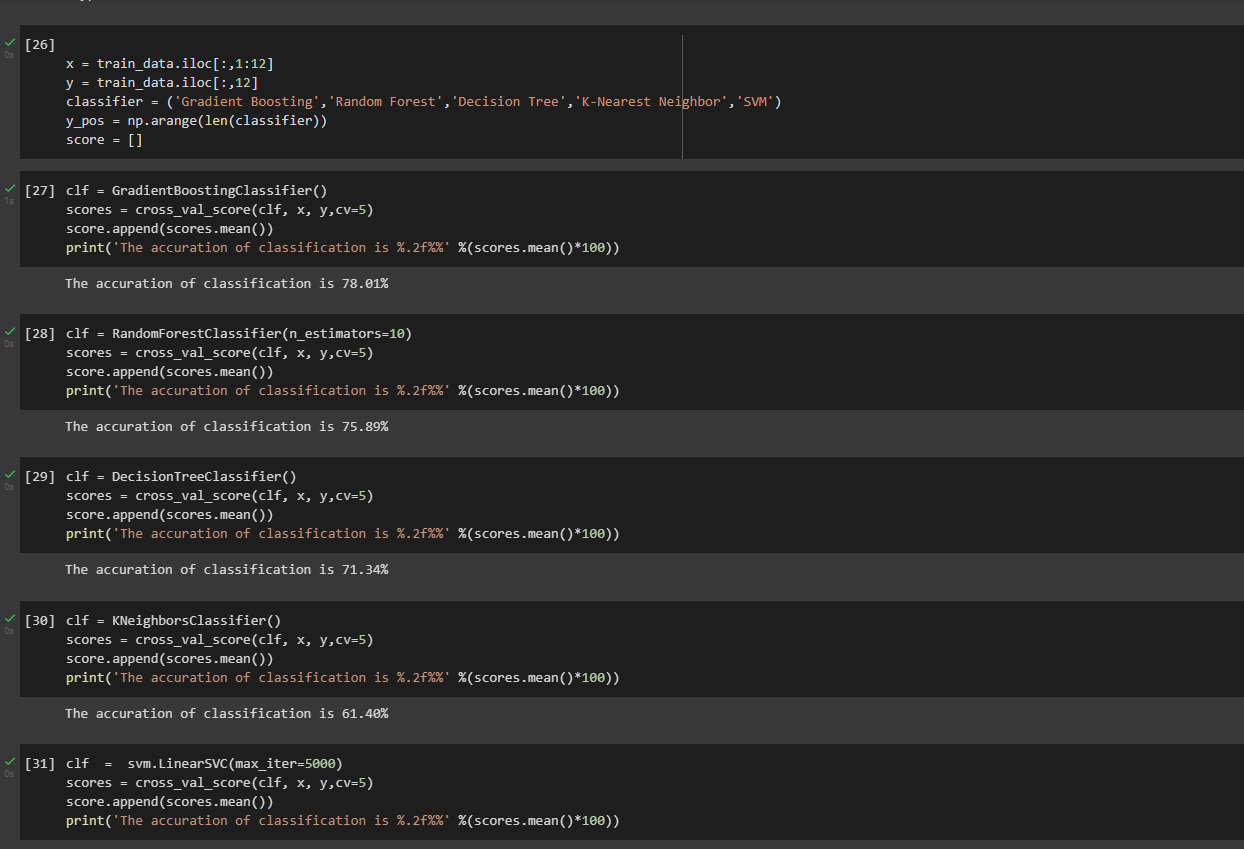
**Figure 2: Loan Train Data on Dependents**

****

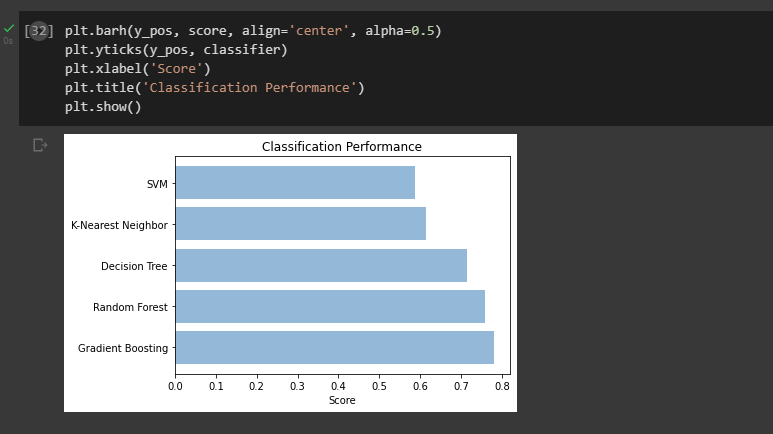
**Figure 3: Loan Train Dataset with Bar Chats**

****

**Figure 4: Loan Train Dataset on dependent with Bar Chat**

****

**Figure 5: Loan Train Data set accuracy precision and recall**

****

**Figure 6: Loan Train Data set accuracy precision & bar graph**

# CONCLUSION

In this paper, machine learning was used to predict loan acceptance. The prediction method begins with data pre- processing, filling the missing values, experimental data analysis. After evaluating model on test dataset, each of these algorithms obtained a precision rate between 70% and 80%. Although here it can be concluded with certainty that the Support Vector Machine model is very efficient and produces superior results than other models.

The comparative analysis based on RMSE, MAPE and MBE values clearly indicate that ANN gives better prediction of stock.

For future work, deep learning models could be developed which consider financial news articles along with financial parameters such as a closing price, traded volume, profit and loss statements etc., for possibly better results.

**Acknowledgement**

This paper would not have been possible without the support of Mr. Tarun Kumar Sir. They inspired us to take this project. I wish to extend special thanks to our reviewers who provided us good recommendations with his deep insights and understandings of the topic.

# REFERENCES

1. Tejas Yadav, J., T. Manmohan Kavita, P. Satripathi, and Venkata Rao Maddumala. "Loan Approval Prediction Based On Machine Learning Hard computing." Journal of Engineering Science
2. Arjun Kumar, GargIshani, and Kaur Sameeti. "Loan Approval Predictions based on Machine Learning Techniques."
3. Vaidiya, Ashleshal. "Predictive and probabilistic approach using logistic regression: Application to predict the loan approval." In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-6. IEEE, 2017.
4. Sheikh, Mohammad Ahmad, Amit Kumar Goel, and Tapas Kumar. "An Approach for Prediction of Loan Approval using Machine Learning Algorithm." In 2020 International Conference on Electronics and Communication of Sustainable Systems (ICESC), IEEE, 2020.
5. Rathi, Golik Gujrat Debasish Das, and Biswas Ranjan Acharya. "Modern Approach for Loan Sanctioning in Banks Using Machine Learning." In Advances in Machine Learning and Computational Intelligence. Sprint, Denmark, 2020.
6. Kartik, R., N. Ambika, and K. E. Kannammal. "A Review on Machine Learning Classification Technique for Bank Loan Approval." In 2019 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-6. IEEE, 2019.
7. Sriram Kumar, Abhi. "Predicting Loan Defaults using Machine Learning Techniques." PhD diss., CalifornUniversity, USA ,2019.
8. Boukerche A, Wang J. Machine learning-based traffic prediction models for intelligent transportation systems. Comput Netw. 2020;181
9. Breiman L. Bagging predictors. Mach Learn. 1996;24(2):123–40. doi: 10.1007/BF00058655.

[CrossRef] [Google Scholar]

1. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. doi: 10.1023/A:1010933404324.

[CrossRef] [Google Scholar]