LOAN ELIGIBILITY PREDICTION
A LITERATURE REVIEW

Project submitted to the
SRM University – AP, Andhra Pradesh
for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology
In
Computer Science and Engineering
School of Engineering and Sciences

Submitted by
Navaneetha Yandrapragada (AP20110010411)
Meghana Amara (AP20110010577)
Sai Monika Padarthi (AP20110010637)
Sumanapriya Kavuri (AP20110010390)

Under the Guidance of
Dr Krishna Prasad

SRM University–AP
Neerukonda, Mangalagiri, Guntur
Andhra Pradesh – 522 240

[December,2022]

Certificate

Date: 12-Dec-22

This is to certify that the work present in this Project entitled "**LOAN ELIGIBILITY PREDICTION : A LITERATURE REVIEW**" has been carried out by **Navaneetha Yandrapragada, Meghana Amara, Sai Monika Padarthi and Sumana Priya Kavuri** under my supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology in School of Engineering and Sciences.

Supervisor

(Signature)
Dr. Krishna Prasad
Associate Professor,
Department of Computer Science and Engineering,
School of Engineering and Applied Science.

Acknowledgements

We express our sincere gratitude to **Dr Krishna Prasad sir**, our mentor in charge, for his guidance and support in finishing the Undergraduate Research Opportunities Project (UROP). His patience and enormous knowledge helped us to overcome many obstacles which occurred at each and every phase of this project. Our overall experience while doing this project was just indescribable. Under his supervision we came to know about many things which we were always curious about. We could not have imagined a better supervisor.

Navaneetha Yandrapragada
Meghana Amara
Sai Monika Padarthi
Sumana Priya Kavuri

Table of Contents

Abstract

Every individual wishes to have their own properties. People, who need financial support, are depending on banks in terms of loans. Major financial asset of the bank depends on the repayment of loans and its interest. Predicting accurately whether the person is eligible or not can decrease the risk. This became the first step for this study. Machine learning helps in predictive analysis. Loan eligibility prediction is a yes or no problem. It is a supervised classification problem. Many researchers did their research on this problem through ML models. According to that, accuracy varies depending on the dataset, features selected and performance measures. Every ML model has its own advantages and limitations. So, by using any one of the ML models we may not get the optimal trained model. We have analyzed various ML algorithms, which were tested on various datasets in the literature, based on their accuracies.

Keywords - SVM,KNN,Adaboost,XGBoost,Random Forest,Logistic Regression.

Abbreviations

ML	Machine Learning
XG-Boost	Extreme Gradient Boosting
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
AdaBoost	Adaptive Boosting

List of Tables

List of Figures

1. Introduction

World is progressing at a rapid pace towards automation, Artificial Intelligence is one of such fields regarding the development of automation. Field of artificial intelligence deals with machines and computers to replicate human intelligence through programming. A machine that can work at a human level intelligence can perform mundane tasks repeatedly using fewer resources. AI is being used in a wide range of fields. One of the fields is business analytics.

One of the applications in the Banking sector is Loan eligibility prediction. Predicting the solution to a problem can be done using machine learning. We can create analytical and predictive models using machine learning and its algorithms use data to find patterns and make inferences. Machine Learning is broadly classified into two types. They are supervised machine learning and unsupervised machine learning. Further supervised machine learning can be divided into two categories: classification and regression. Classification problem is a yes or no type of problem.

Predicting whether a customer is eligible for a loan or not is a difficult task and has a great need in the banking sector. And it is also a yes or no type of problem. So, loan eligibility prediction is a classification problem.

There are several algorithms to solve a classification type of problem. Every algorithm has its own advantages and limitations. So, by using any one of the algorithms we may not get the optimal trained model. But once an optimal trained model is found, then approving a loan for a candidate becomes an effortless task.

1.1 Types of Loans

Loans are classified into two types based on the purpose. They are

Secured loans

Secured loans are the ones where you have to pledge an asset as security. In case if you cannot repay the loan, the lender still has some means to get back their money.Home loan, gold loan,car loan refer to secured loans.

Unsecured loans

Unsecured loans are loans which do not require collateral. Personal loans, educational loans refer to unsecured loans.

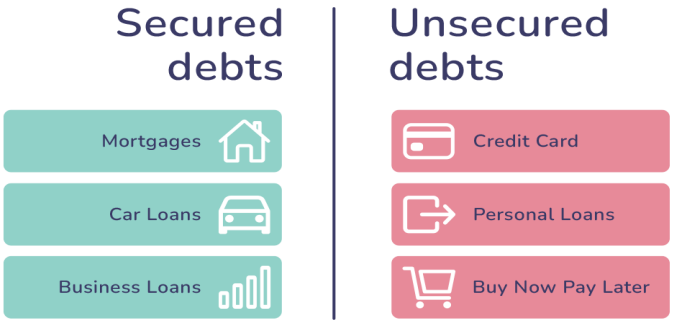

Figure 1. Secured and Unsecured loans

1.2 Factors considered for loan approval

Choosing a candidate to approve a loan depends on many independent variables such as,

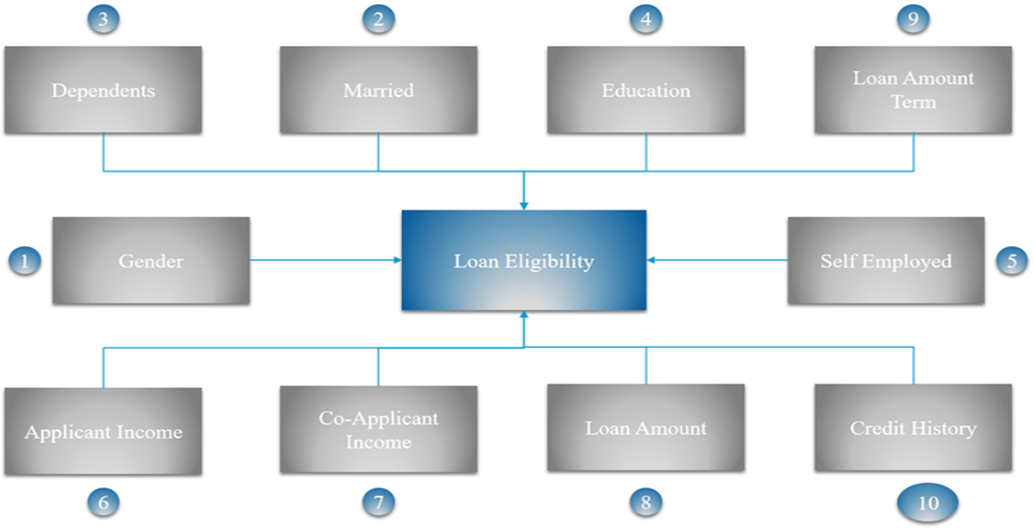

Figure 2. Factors considered for loan approval

Some variables that should be taken note are,

- **Credit scoring**

A key element in determining a borrower's eligibility for a loan is their credit score. A statistical approach of determining a customer's credit risk, whether they are current or prospective, is credit scoring. A borrower's credit score is a quantitative indicator of how creditworthy they are.

Human lenders are biased, which can result in poor judgment. So,Credit scoring can be automated using machine learning for unbiased judgment. Credit risk is the likelihood of experiencing a loss as a result of a borrower's failure to make loan payments. The majority of credit scoring models are built on machine learning algorithms, which examine the many features mentioned before in order to forecast their future behavior. Making wise financing selections depends on the accuracy of these forecasts.

- **Income verification**

It is a crucial component of loan eligibility. It is a procedure used by lenders to verify that a borrower has the resources to repay a loan. In order to determine the borrower's capacity to repay the loan, lenders must confirm the borrower's income. Additionally, proving your income can be difficult, especially if you're self-employed or have many sources of income. In this situation AI and ML can aid.

The process of verifying an individual's income can be automated using machine learning. Machine learning models can learn to recognise patterns that are predictive of loan default by leveraging data from prior loan applications, tax returns, and bank statements. The income of new loan applicants can then be automatically verified using these patterns. Classification models can be used to categorize whether the applicant's income was validated or moved to an exceptional workflow for human processing. Compared to manual income verification, this is not only more effective but also results in more reliable choices.

- **Loan amount**

Determining how much money to lend and under what conditions presents another difficulty relating to loan eligibility. The borrower's income and credit score are often taken into consideration when making this choice. To evaluate loan terms and amounts, there are various considerations that might be made. For instance, the loan size may be determined by the loan's purpose (such as purchasing a car versus launching a business). The borrower's employment history might also be considered when determining the loan's terms.

Automating the selection of a loan's terms and amount is possible with machine learning. Machine learning algorithms can learn to recognise patterns that are indicative of loan default by analyzing data from prior loan applications.The loan amount and terms for new loan applicants can then be automatically determined using these patterns.

2. Literature Review

Authors [1] compared Random Forest, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Decision Tree and Gradient Boost.Dataset used is kaggle's historical dataset "Loan Eligible Dataset".Authors used the best techniques to pre-process and analyze the data. One of the techniques they used for treating imbalance classification problems is Synthetic Minority Oversampling Technique (SMOTE).Authors one of the observations is that there are more male applicants, more married and more applicants with good credit scores.They concluded Random Forest has high accuracy of 95.56%.

Authors [2], implemented Decision tree, random forest, support vector machine, k-nearest neighbor and an ensemble model in order to find the best model which predicts the loan eligibility. In ensembled model, 2 classifiers, Decision tree and AdaBoost technique were mixed in order to improve the performance. Dataset from Kaggle which is based on loan and which is publicly available was used. The machine learning package which was used to train the model is PySpark. After the implementation Decision Tree with AdaBoost got 0.84 as the test accuracy which was the highest accuracy among the implemented models.

Authors [3], compared and calculated the accuracies of ML algorithms Random forest,Logistic regression,Support Vector Machine(SVM),K-Nearest Neighbour(KNN).The authors proposed a logistic regression method for loan approval prediction with a sample set for loan approval applications.
To predict loan approvals,three machine learning algorithms are applied to the test-data.After comparing all the accuracies of the algorithms,found that the random forest algorithm(81%)has the highest accuracy.

Dagar, Akash [5]**,** used Logistic Regression,Random Forest Classifier, XGBoost,SVM. Author objective is to find a plane that has maximum margin between data points of both classes.For training and testing the models, K-fold Cross Validation is used. For comparison, the Confusion Matrix is used. In this paper, Logistic Regression has the highest accuracy.

Sarkar, A. [6] , used Logistic Regression, Random Forest and Decision Tree. Main Observation is that there are more male applicants and most of them are married.Concluded that Logistic regression had highest accuracy.

Dutta, Prateek. [10] proposed a model which includes Logistic Regression, Decision Tree, Random Forest. Dataset used is kaggle's dataset. In this paper, It is concluded that Logistic Regression is the best model with highest accuracy.
Candidates with low credit risk and high income will be chosen for the loan.

In paper [13], they selected Logistic Regression, Decision Tree, Random Forest as their models. They fit the models and then improved the models to decrease the risk.They calculated accuracy score and cross validation scores for three models. Even though accuracy of Decision tree is higher than other models, they showed that Random Forest shows best response in generalization.

[14]Models used for the analysis are XGBoost, Random Forest, Decision Tree. Authors chose two data sets where one data set contains Loan_ID, Gender, Married and another dataset contains Dependents, Education, Self_Imployed, Applicant_Income, Co_Applicant_Income, Loan_Amount, Loan_Amount_Term, Credit_History, Property_Area and Loan_Status.Their model gave the best Results for both dataset. They also mentioned that their model will not give the best result when the applicant faces a financial problem suddenly.Authors felt that zip code and credit code plays an important role.

[15] Whether an applicant is approved or not is predicted using SVM. This paper analyzed four kernels of SVM: Linear, Poly, Sigmoid, Rbf. Poly SVC has highest accuracy (97.2), Rbf(96.7), Linear(95.1), Sigmoid(83.3)

2.1 COMPARISON TABLE

AUTHORS	DATASET USED	MODELS	ACCURACY (In %)
Orji, Ugochukwu E., Chikodili H. Ugwuishiwu, Joseph CN Nguemaleu, and Peace N. Ugwuanyi.[1]	Kaggle's historical dataset 'Loan Eligible Dataset	Random Forest	95.56
		Gradient Boost	93.33
		K-Nearest Neighbor	93.33
		Decision Tree	91.11
		Support Vector Machine	84.44
		Logistic Regression	80.00
Kumar, Ch Naveen, D. Keerthana, M. Kavitha, and M. Kalyani. [2]	Public Repository which contains 614 records with 13 attributes	Decision Tree with AdaBoost	84.00
		Random Forest	72.00
		Support Vector machine	70.00
		Decision Tree	69.00
		K-Nearest Neighbor	59.00
Tumuluru, Praveen, Lakshmi Ramani Burra, M. Loukya, S. Bhavana, H. M. H. CSaiBaba, and N. Sunanda.[3]		Random Forest	81.00
		Logistic Regression	77.00
		Support Vector Machine	73.20
		K-Nearest Neighbor	68.00
Dagar, Akash[5]	Kaggle dataset	Logistic Regression	80.94
		XGBoost	79.15
		Random Forest	78.50
		Support Vector Machine	69.70
Sarkar, A[6]		Logistic Regression	80.78
		Random Forest	79.79
		Decision Tree	70.51
Dutta, Prateek[10]	Kaggle dataset	Logistic Regression	89.70

		Decision Tree	85.40
		Random Forest	77.45
Dosalwar, S., Kinkar, K., Sannat, R., & Pise, N.[11]		Logistic Regression	78.5
		Naive Bayes	77.9
		Random Forest	77.3
		XGBoost Classifier	77.3
		Decision Tree	66.2
		Support Vector Machine	65.0
		K Neighbors Classifier	61.9
KHAN, AFRAH, EAKANSH BHADOLA, ABHISHEK KUMAR, and NIDHI SINGH[13]		Decision Tree	93.648
		Random Forest	83.388
		Logistic Regression	80.945
Singh, Vishal, Ayushman Yadav, Rajat Awasthi, and Guide N. Partheeban.[14]	Data collected directly from customers	XGBoost	77.77
		Random Forest	76.38
		Decision Tree	64.5

Table 1. Summary of literature reviews

3. Methodology

The process or methodology for prediction analysis in ML is as follows,

Figure 3. Process of predictive analysis in ML

- **Dataset:**

Collect the dataset with a clear cut view of all the attributes. The decision variable (loan_status) is our target or dependent variable. Remaining all are independent variables.

- **Data preprocessing:**

The data collected is raw data. That should be cleaned. Here, in this step we will handle the missing or incomplete values by replacing them with appropriate substitution of values.

- **Exploratory Data Analysis:**

EDA is the process where data is analyzed visually. Clean Data is visualized through graphs, plots etc... and also each feature weight on the target variable is also examined.

- **Splitting the data:**

Now, we select the features and then split the data into train sets and test sets. The ratio can be 6:4 or 7:3 or 8:2. Train set is used to train the model and the test set is used to test the model.

- **Model fitting:**

All the required libraries should be imported. And then the train set should be fit into various models. Trained model is tested on the test set.

- **Performance measure:**

Different performance metrics or evaluation metrics, like precision, confusion matrix, F-score etc.…, are used to evaluate the performance or quality of the model.

3.1 ML Models

Logistic Regression

It is a classification model which estimates by using logit function.Logit function is a log of odds in favor of events. If the response variable is binary but the explanatory variables are continuous, Logistic regression performance is more efficient.But, if number of features are greater than the number of observations then Logistic regression may lead to overfitting.

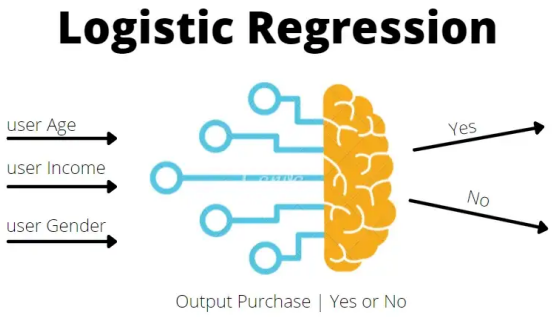

Figure 4. Logistic Regression

Support Vector Machine

It is a supervised ML algorithm. It finds a hyperplane which classifies the data points clearly. The dimension of the hyperplane depends on the number of features.It is used mostly when data is not regularly distributed. It, mostly, will not suffer from overfitting problems and its generalization is better than other models.

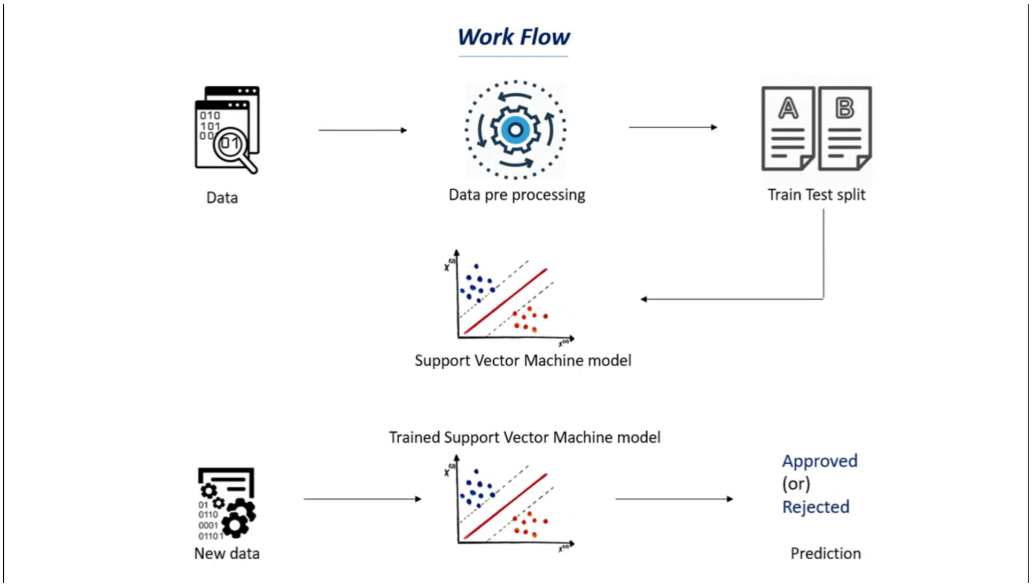

Figure 5. Support Vector Machine

Decision Tree

It is used to perform the tasks of classification and regression. It comprises several branches, leaf nodes, and root nodes. This algorithm generates a structure like a tree by classifying the instances and utilizing a Recursive Portioning Algorithm(RPA). A class label represented by a leaf node and branches represent test results. These tests are represented by internal nodes for an attribute.

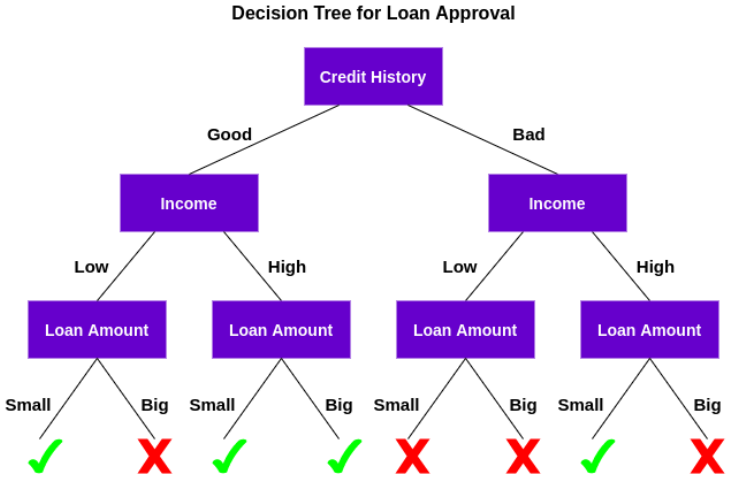

Figure 6. Decision Tree

Random Forest

It is supervised machine learning used for classification and regression. A predictor ensemble is built with several decision trees that expand in randomly selected data subspaces. Immunity to overfitting, accurate classification or regression and more efficient results on large databases are the advantages of random forest over other machine learning models. It combines the output of multiple Decision Trees to generate the final output.

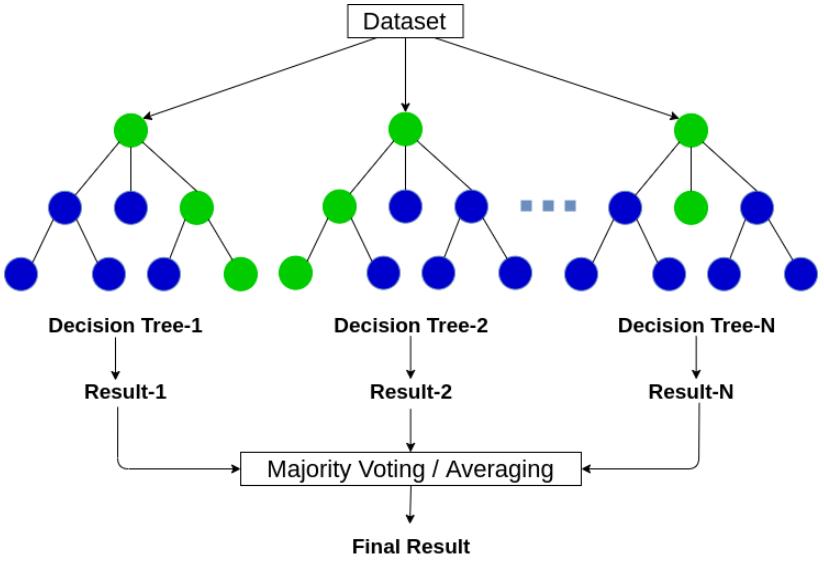

Figure 7. Random Forest

K-Nearest Neighbor

The K-NN algorithm detects similarities between new instances/data and existing cases and assigns each new instance to the category that most closely resembles the current category. No matter where the data originates from, the KNN approach can quickly sort it into a relevant portion.

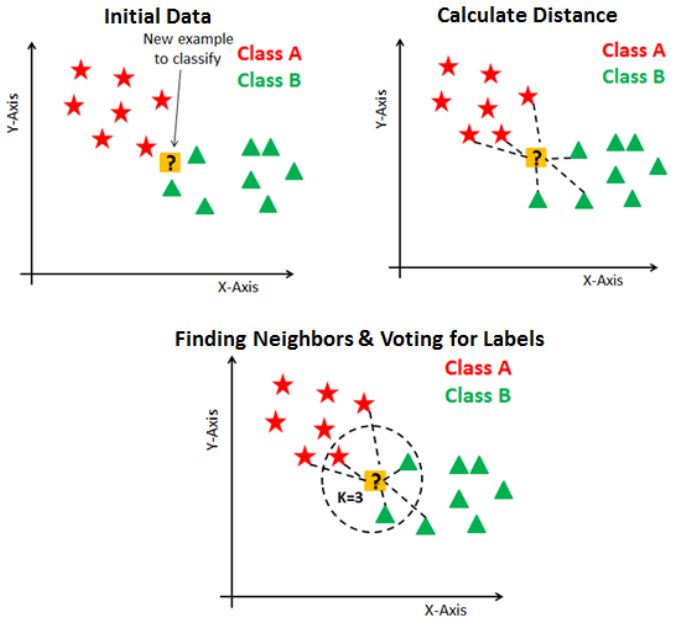

Figure 8. K-Nearest Neighbor

XG Boost

It is simply the implementation of gradient boosted trees algorithm. It is a supervised machine learning algorithm used for both regression and classification on large data sets. XG Boost accurately predicts the target variable by simply combining the estimates from simpler and weaker models.

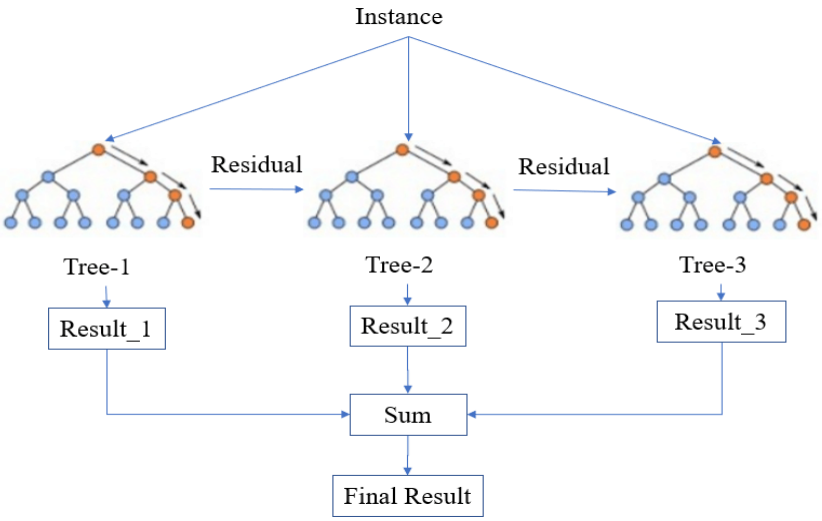

Figure 9. XG Boost

MODEL	ADVANTAGES	LIMITATIONS
Random forest	• It can handle missing values automatically. • No feature scaling is required. • It is very stable.	• It takes more time to train since it creates a large number of trees and makes decisions based on the majority of votes. • Requires more computational power as it creates more trees to compute the output.
Support vector machine	• It can handle nonlinear data efficiently. • Even though the data is not enough,it gives good results.	• Requires feature scaling • It takes more time to train long datasets.
Logistic Regression	• Easy to implement. • Non-linear effects are handled. • Discrete variables, either true or false, are provided as output.	• Independent variables are required for estimation. • For parameter estimation a large sample is required. • Continuous outputs cannot be provided by the logistic model.
K-Nearest Neighbor	• Simple to implement. • No training required - It constantly evolves with new data. • Can learn non-linear decision boundaries.	• It is sensitive to outliers. • Requires high memory • High prediction complexity is required for large datasets.
Decision Tree	• Simple to understand and Interpret. • Requires little data preparation. • It takes the consideration of all possible outcomes of a decision and traces each path to a conclusion.	• It creates over- complex trees. • It can be unstable because of small variations in the data. • Predictions here are neither smooth nor continuous.

XGBoost	• Effective with large data sets. • Works well even if data is non-linear or with segregated clusters.	• Can overfit the data. • It is sensitive to outliers • Doesn't perform well on unstructured data.

Table 2. Advantages and Disadvantages of various models

Discussion

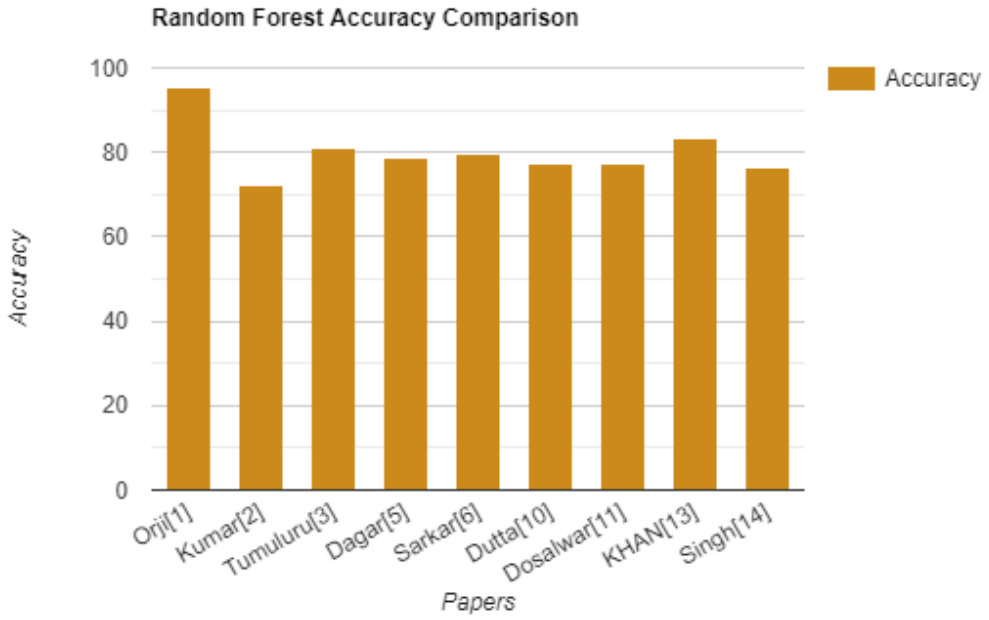

Figure 10. Accuracy comparison of Random Forest

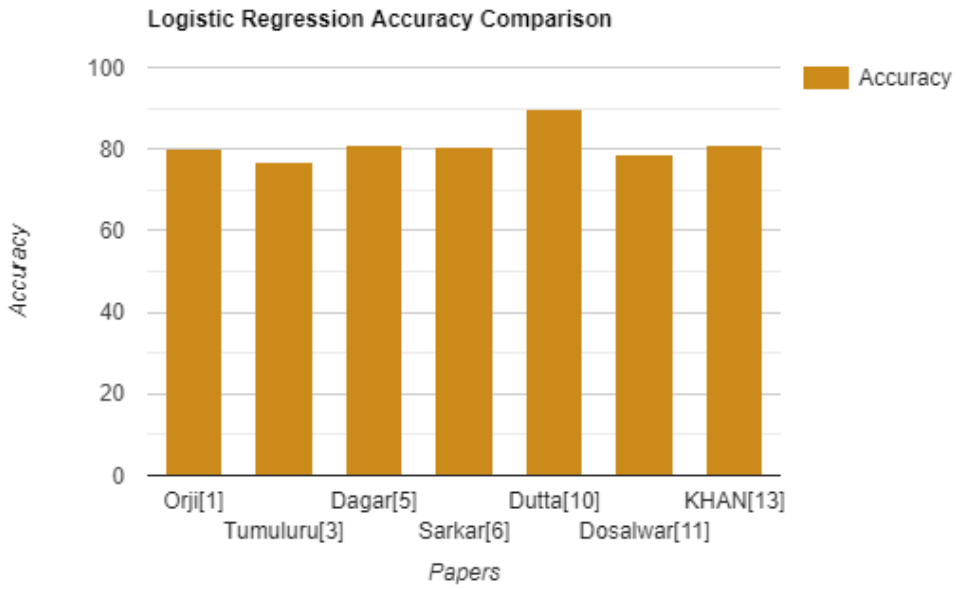

Figure 11. Accuracy comparison of Logistic Regression

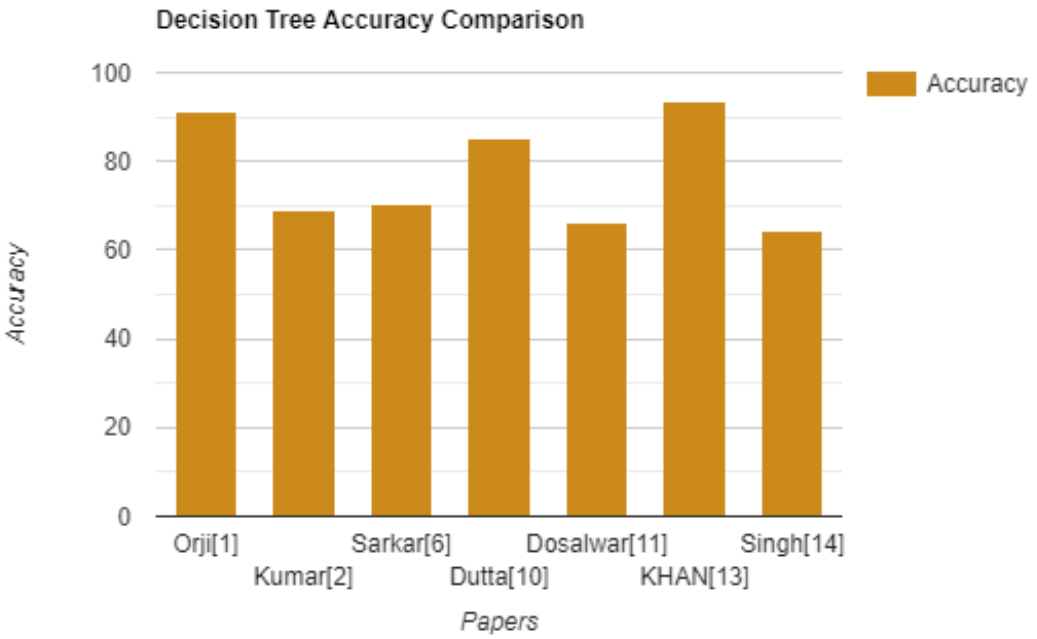

Figure 12. Accuracy comparison of Decision Tree

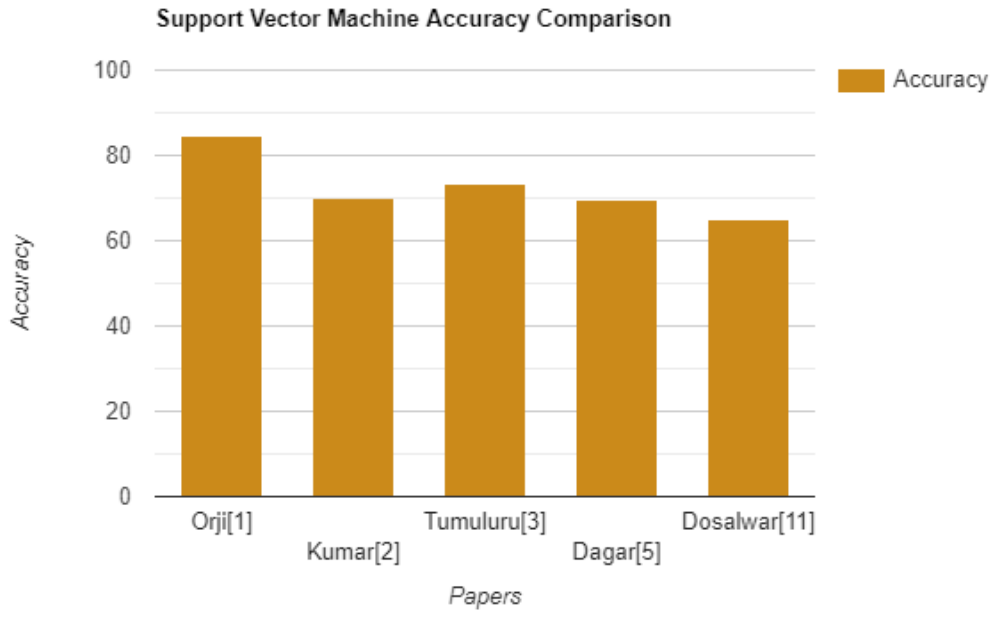

Figure 13. Accuracy comparison of Support Vector Machine

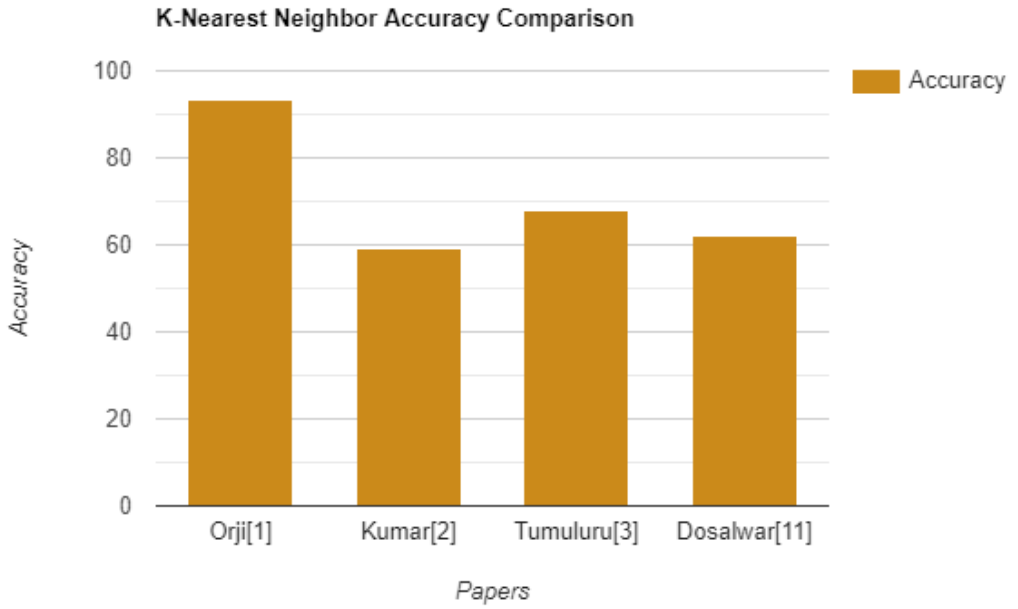

Figure 14. Accuracy comparison of K-Nearest Neighbour

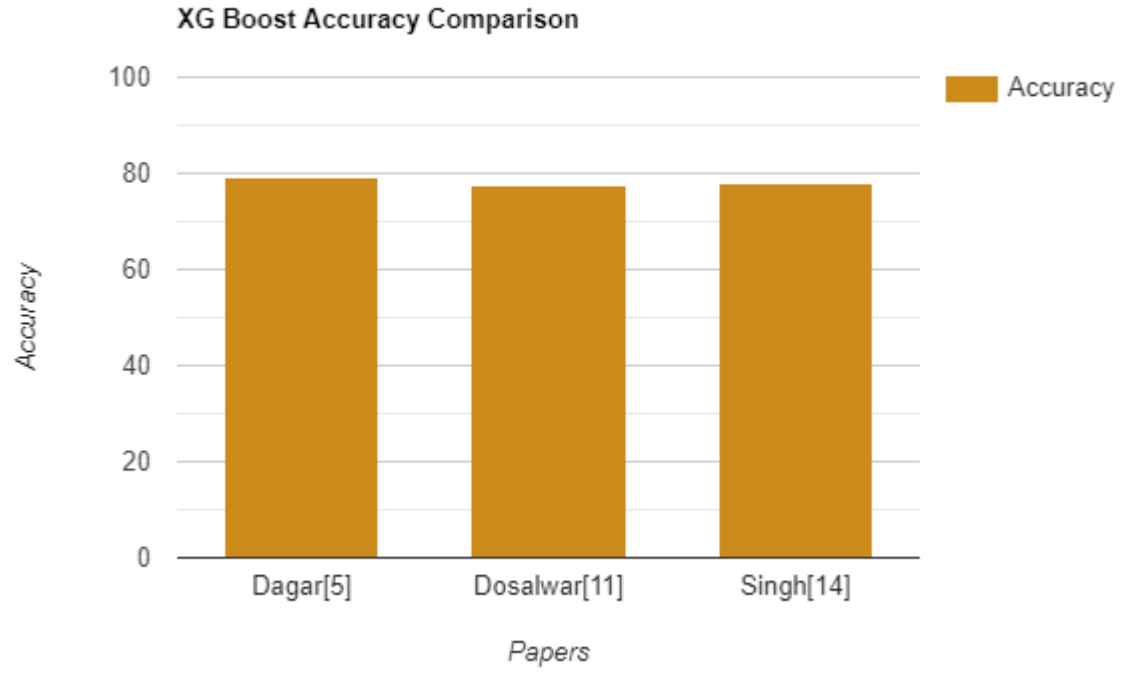

Figure 15. Accuracy comparison of XG Boost

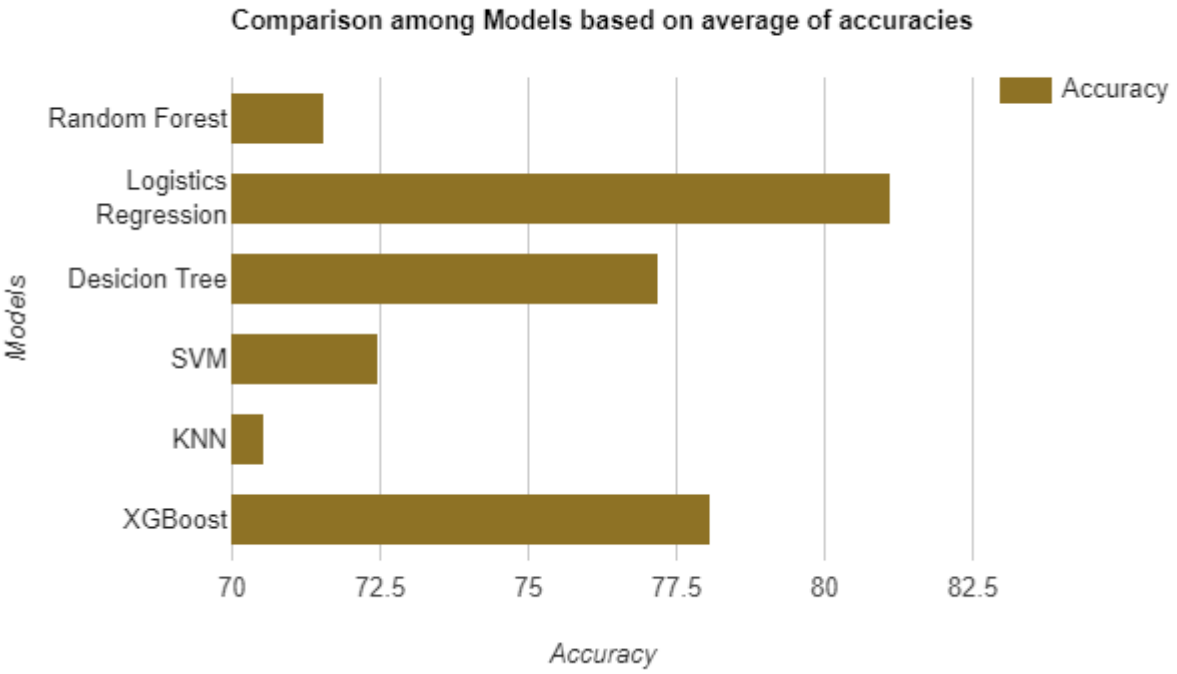

Figure 16. Comparison of Random Forest, Logistic Regression, Decision Tree, SVM, KNN and XG Boost based on accuracy average

Concluding Remarks

This study proposes the usage of machine learning algorithms to forecast loan acceptance. After comparing all the algorithms we found that Random Forest has the highest accuracy and KNN has the least accuracy. The highest accuracy algorithm is used to train the model to produce best results. The accurate results are used to predict whether a borrower should be lent money or not.

After comparing the algorithms we can conclude the following points,

Random forest model provides the highest accuracy of all the categorization techniques that are currently available. And it can also handle large amounts of data with hundreds of different variables.

Logistic regression analysis is valuable for predicting the likelihood of an event. In the case of binary output, it helps us make a final decision.

We prefer SVM in cases when data is not regularly distributed. It also lowers the likelihood of data errors. It does not experience overfitting.

Decision trees take less work to prepare the data during pre-processing than other methods do. It determines the best models to evaluate credit risk.

KNN algorithms can compete with the most accurate models by producing extremely accurate predictions. It deals with risk associated with each applicant in repaying the loans.

Compared to other algorithms, the XGBoost model has the best combination of processing time and prediction performance.

References

[1] Orji, Ugochukwu E., Chikodili H. Ugwuishiwu, Joseph CN Nguemaleu, and Peace N. Ugwuanyi. "Machine Learning Models for Predicting Bank Loan Eligibility." In 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON), pp. 1-5. IEEE, 2022.

[2] Kumar, Ch Naveen, D. Keerthana, M. Kavitha, and M. Kalyani. "Customer Loan Eligibility Prediction using Machine Learning Algorithms in Banking Sector." In 2022 7th International Conference on Communication and Electronics Systems (ICCES), pp. 1007-1012. IEEE, 2022.

[3] Tumuluru, Praveen, Lakshmi Ramani Burra, M. Loukya, S. Bhavana, H. M. H. CSaiBaba, and N. Sunanda. "Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms." In 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), pp. 349-353. IEEE, 2022.

[4] Tejaswini, J., T. Mohana Kavya, R. Devi Naga Ramya, P. Sai Triveni, and Venkata Rao Maddumala. "Accurate loan approval prediction based on machine learning approach." Journal of Engineering Science 11, no. 4 (2020): 523-532.

[5] Dagar, Akash. "A Comparative Study on Loan Eligibility." Int. J. Sci. Res. Eng. Trends 7, no. 3 (2021): 1646-1649.

[6] Sarkar, A. "Machine learning techniques for recognizing the loan eligibility." International Research Journal of Modernization in Engineering Technology and Science 3, no. 12 (2021)

[7] Naik, Sanskruti, and Ganesh Manerkar. "Education Loan Prediction Analysis."

[8] 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 Sustainable Communication Systems (ICESC), pp. 490-494. IEEE, 2020.

[9] Reddy, Chenchireddygari Sudharshan, Adoni Salauddin Siddiq, and N. Jayapandian. "Machine Learning based Loan Eligibility Prediction using Random Forest Model." In 2022 7th International Conference on Communication and Electronics Systems (ICCES), pp. 1073-1079. IEEE, 2022.

[10] Dutta, Prateek. "A Study On Machine Learning Algorithm For Enhancement Of Loan Prediction." International Research Journal of Modernization in Engineering Technology and Science 3 (2021).

[11] Dosalwar, S., Kinkar, K., Sannat, R., & Pise, N. (2021). Analysis of Loan Availability using Machine Learning Techniques.

[12] Vaidya, A. (2017). Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval. 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT).

[13] KHAN, AFRAH, EAKANSH BHADOLA, ABHISHEK KUMAR, and NIDHI SINGH. "LOAN APPROVAL PREDICTION MODEL A COMPARATIVE ANALYSIS." (2021).

[14] Singh, Vishal, Ayushman Yadav, Rajat Awasthi, and Guide N. Partheeban. "Prediction of modernized loan approval system based on machine learning approach." In 2021 International Conference on Intelligent Technologies (CONIT), pp. 1-4. IEEE, 2021.

[15] Akça, Mehmet Furkan, and Onur Sevli. "Predicting acceptance of the bank loan offers by using support vector machines." International Advanced Researches and Engineering Journal 6, no. 2 (2022): 142-147.