Prediction of Loan Approval using Machine Learning Techniques

Shrutinamra Mahapatra, Nihar Ranjan Palleyi, Siblu Das, Nikita Rojalin Sahoo, Dr. Niranjan Panda

Dept. of Computer Science and Engineering, Institute of Technical Education and Research, Bhubaneswar, India

ABSTRACT

Lending loan and earning profit from it is how banks do their business. Thousands of applicants fill the application for a particular loan. Thus, it becomes very difficult for the banks and the financial institutions to choose the most genuine applicant who will repay the loan properly within a particular time frame. Earlier banks used to choose out the applicants manually based on certain criteria, but it was very time consuming and there was a chance of a lot's manual errors. With advancement in the field of machine learning, to predict the most suitable applicant, banks have started using machine learning algorithms. But this process is at its initial stage and is not very efficient. In this paper few machine learning algorithms i.e., Logistic Regression, Decision tress, Random Forest, SVM, KNN are being used to predict the most suitable applicant for a particular loan.Confusion Matrix, accuracy, precision, recall and support are the metrics that are being used to evaluate the models and then the results are being compared. Finally, we get that *Random Forest* model has the highest accuracy of *83.59%.*

Keywords: Loan, Machine Learning, Logistic Regression, random Forest, SVM, Decision Tree, KNN.

1. INTRODUCTION

Loan lending is the core business model of all financial institutions. People and financial organizations take loan to overcome financial limitations to achieve their personal goals. So, lending of loan is a win-win situation for both the financial institutions and the borrower. But it also comes with its risks. Thousands of applicants apply for a particular loan, so it becomes very difficult for the financial institutions to filter out the best applicant who is in need of the loan and also, if he has the capacity to pay the loan back along with the interest or not. Earlier this process of filtering out applicants was done manually. Although nowadays machine learning models are being adopted by financial institutions to make the process easier, faster and risk free, but still, it is at a nascent stage. In this project we aim to predict the loan approval using various machine learning techniques.

2. LITERATURE SURVEY

Pidikiti Supriya et al. in paper [1] presented a study in which attributes like property area, education, loan amount and credit history were being considered to train the machine learning models. Four algorithms were used to predict the loan approval namely KNN, Decision tree, gradient boosting and SVM. The result showed that decision tree achieved the highest accuracy of 81.1 percent. Here in paper [1], applicants with high income applying for loans with low amount is more likely to get selected, others with low income in need of the loan will have a less chance of getting it. Also, some basic characteristics like marital status and gender are not taken into consideration.

J. Tejaswini et. al. in [2] proposed a system in which they have used three machine learning techniques such as Random Forest, Decision Tree and Logistic Regression to predict the approval of loan for a particular customer. The dataset they

used contained 13 attributes. They finally concluded that the accuracy of Decision Tree is better as compared to other machine learning approaches namely Logistic Regression and Random Forest. They had used an old dataset to train the model so in future for new datasets it may not predict the results as accurately as it was doing here.

Kshitiz Gautam et. al. in [3] proposed a system for predicting approval of loan. They took two machine learning approaches i.e., Random Forest and Decision tree. They used a dataset with 12 attributes such as gender, marital status, qualification, income, etc. They concluded that 85.75% was the best result and it was given by Random Forest technique. They used only two algorithms, but they could have taken some more algorithms into consideration which could have helped them to explore more options and find better results.

Mehul Madan et al. in paper [4] proposed a model where they used two machine learning algorithms i.e., decision tree and random forest. They first started with data pre-processing, then they moved on to exploratory data analysis and then finally did model building and evaluation. They eventually got the best accuracy which was 80% through Random Forest Classifier. Some of the attributes such as gender, marital status etc. were not being taken into consideration because of which the algorithms put some of the non-defaulters such as borrowers who do not own property and are applying for a small business ventures or marriage loan into default class.

Pratik Dutta in [5] took into consideration three machine learning approaches namely Logistic Regression, Random Forest and Decision Tree for the enhancement of loan prediction. In his study he found that logistic regression gives the best results with an evaluation of 89.7%.

Vishal Singh et. Al. in [6] proposed a system which would predict whether a customer is eligible for a particular loan or not by using machine learning techniques trained. For this purpose, they used three machine learning algorithms such as XGBoost, Random Forest and Decision Tree. They used a dataset with 13 attributes. They finally concluded that XGBoost is giving the best result of 77.7 %.

3. METHODOLOGY

Below is the diagram which represents all the steps included during building our project

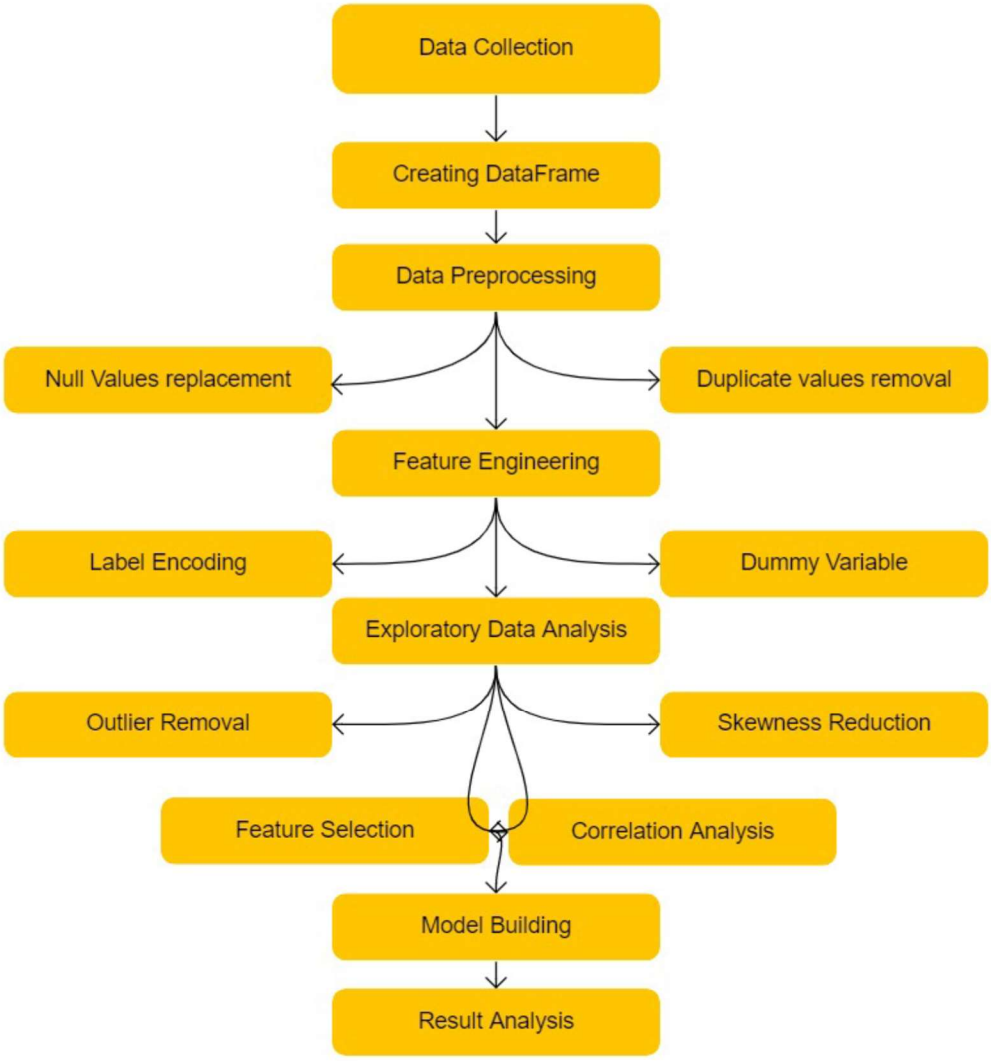


Fig 1. Flow chart which gives the overview of various steps involved in the paper

4. DESIGN STEPS

In this section we have described the design steps we have followed in our project.

4.1 Data Collection

The dataset is collected from Kaggle. It includes 13 attributes of 614 customers.

The 13 columns that the dataset contains are:

Loan_Id, Gender, Marital_Status, Dependents, Education, Employment_status, Applicant_income, Coaaplicant_income, Loan_amount, Loan_amount_term, Credit_history, Property_area, Loan_status.

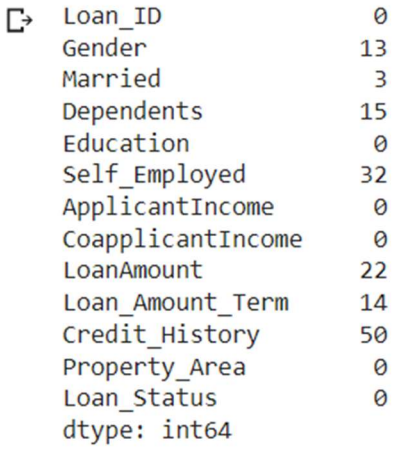
The column Loan_status represents the feature that we are going to predict, the status value or result. If the Loan_status is Y means the customer is applicable for loan or the customer is approved for taking loan and if the loan_status is N means the customer is not approved for his loan.

4.2 Data Frame Creation

Data Frame creation is the process where data is represented in a 2-Dimensional tabular form having columns and rows. Here, Panda's read_csv () function is used to create a Data frame from csv file. The Data frame consists of 614 rows and 13 number of columns as follows.

4.3 Data Pre-processing

Data pre-processing is a technique in machine learning where data which is in raw format is transformed to a format which is more relevant and efficient. Here, firstly, all the duplicate rows are dropped using pandas drop_duplicates () function and duplicate columns are dropped by first transposing the data frame and then using drop_duplicates () function. Then, Null and Zero values are replaced with Mean and Mode values of the respective columns using fillna () and replace () functions.



```
Loan_ID              0
Gender              13
Married              3
Dependents          15
Education            0
Self_Employed       32
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount          22
Loan_Amount_Term    14
Credit_History      50
Property_Area        0
Loan_Status          0
dtype: int64
```

Fig 2. Total Number of columns in the Dataset

4.4 Feature Engineering

The process in which raw data is transformed and manipulated into relevant features to be used in models of machine learning is called feature engineering. In this step, Label Encoding is used to transform data from labels to numeric forms or machine-readable form. Label Encoding can be done using LabelEncoder () function or manually replacing labels to numeric forms. Dummy variables are created for property area and dependents columns are created using panda's get_dummies () function to create more new features.

4.5 Exploratory Data Analysis

The process which is used to analyze data by graphical representation and other means is called exploratory data analysis. Removal of outliers is done Manually with quantile () function by setting the threshold limit to 97%. Then Distribution of data is analyzed using seaborn's distplot () function and histograms and skew () function is used to reduce skewness of the data.

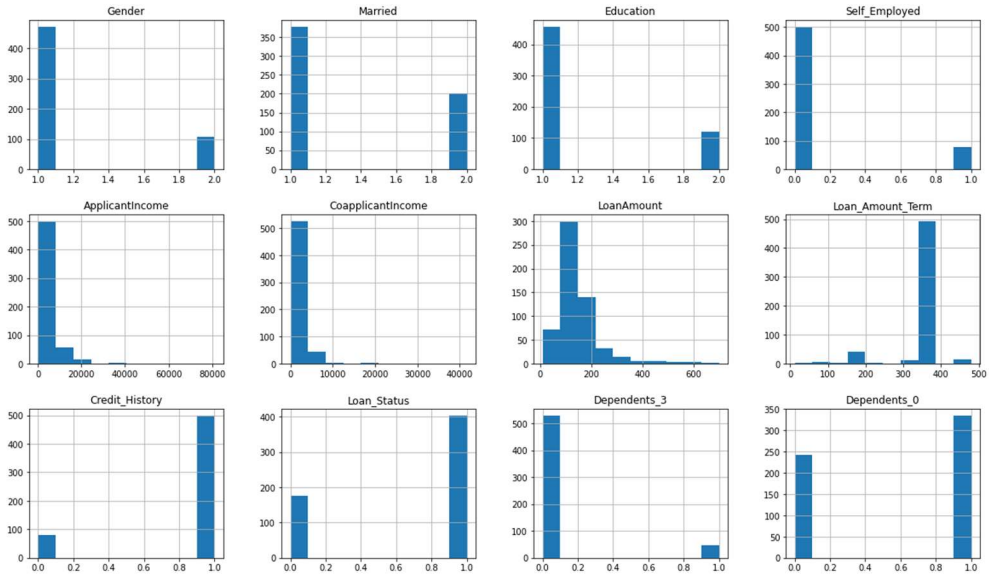


Fig. 3.Histogram of each attribute of the dataset

	Skew
Dependents_3	3.305486
Self_Employed	2.322912
ApplicantIncome	2.098468
Dependents_1	1.872356
TotalIncome	1.772322
CoapplicantIncome	1.765195
Dependents_2	1.738318
Gender	1.599822
Education	1.317622
Property_Area_Rural	0.885695
Property_Area_Urban	0.785681
Married	0.617274
LoanAmount	0.612981
Property_Area_Semiurban	0.477056
Dependents_0	-0.392768
Loan_Status	-0.875449
Loan_Amount_Term	-2.028681
Credit_History	-2.067084

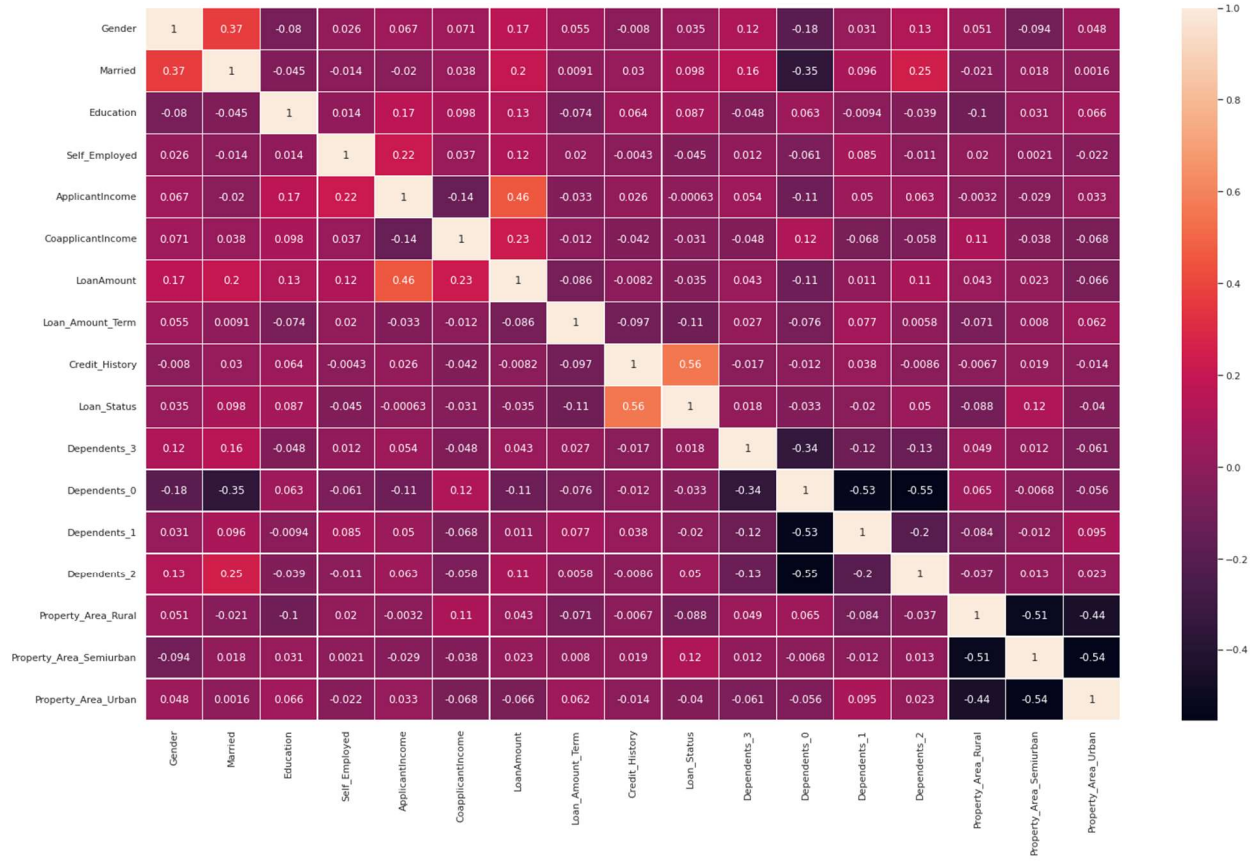
fig 4. Skewness of Features

4.6 Feature Selection

The step in which only relevant features are selected to give the machine learning models as input is called feature selection. Then in this step, Seaborn's Heatmap is plotted to analyze the correlation of the Predictor variables with the Target variables. From sklearn ensemble Random Forest Classifier's feature_importances_ function is taken to select the best predictor variables to train

the models and then *Credit_History, ApplicantIncome, Loan_Amount_Term, CoapplicantIncome, Married, Property_Area_Semiurban, Dependents_2, LoanAmount, TotalIncome* are finalized to be the best features.

Fig.5. HeatMap of the dataset



	Features	Importances
9	Dependents_3	0.009134
12	Dependents_2	0.010582
15	Property_Area_Urban	0.014589
11	Dependents_1	0.016545
3	Self_Employed	0.016732
0	Gender	0.016788
10	Dependents_0	0.017276
13	Property_Area_Rural	0.017619
14	Property_Area_Semiurban	0.018868
2	Education	0.021016
1	Married	0.023206
7	Loan_Amount_Term	0.030901
5	CoapplicantIncome	0.085911
6	LoanAmount	0.139477
4	ApplicantIncome	0.146938
16	TotalIncome	0.153201
8	Credit_History	0.261217

Fig 6. Feature Importance

4.7 Machine Learning Algorithms

In this section we have briefly describe the algorithms that are used for model prediction in our project.

4.7.1 K-Nearest Neighbors

K-Nearest Neighbor is a type of machine learning algorithm that puts the data in the category with which it founds more similarity. It implies that when a new data appears it can be placed into a well-suited category easily. We can use it for both classification and regression.

4.7.2 Logistic Regression

Logistic Regression is one of the most common supervised machine learning approaches which can be used to predict variables which are categorical dependent using a set of variables which are independent.

4.7.3 Decision Tree

Decision tree falls under supervised machine learning algorithms category. It can be used for both classification and regression. It contains a tree structure where the features of a dataset are represented by internal nodes, decision rules are represented by branches and outcomes are represented by leaf nodes.

4.7.4 Random Forest

Random forest is a popular supervised machine learning algorithm. It takes into consideration the prediction from several decision trees instead of one and based on the majority gives the output.

4.7.5 Support Vector Machine

It is a well-known supervised machine learning algorithm which can be used for both classification and regression but is mostly used for classification. In the SVM algorithm, n-dimensional space is segregated into classes by best line or decision boundary so that it would be easier in future to put new data points in the correct place. This decision boundary is also known as hyper plane.

4.8 Metrics used for result evaluation

Below are the different metrices for evaluating performance.

4.8.1 Accuracy

The ratio we get by dividing predictions made correctly by total predictions made is called accuracy.

Accuracy = (TP + TN) / (TP+TN+FP+FN) ----------(1)

4.8.2 Precision

The ratio we get by dividing true positive values by all values predicted to be positive is called precision.

Precision = (TP) / (TP + FP) ----------(2)

4.8.3 Recall

When we divide the total number of positives correctly predicted by total number of positives, we get recall.

Recall = (TP) / (TP+FN) ----------(3)

4.8.4 Specificity

The ratio we get when we divide the total number of negatives predicted correctly by total number of negatives, we get specificity.

Specificity = (TN) / (TN+FP) ----------(4)

4.8.5 F1-Score

We get the value of F1-Score by taking into consideration both recall and precision.

Sometimes, with uneven distribution of class, accuracy may not be the best measure of performance, in that case we use F1-Score

F1-Score = (2 X Precision X Recall) / (Precision + Recall) ----------(5)

4.8.6 ROC Curve

A ROC curve also called receiver operating characteristic curve is a graph which helps us to determine the model's capability in distinguishing the classes. Here, when AUC (which means area under the curve) is more, better is the model.

4.8.7 Confusion Matrix

A table that contains incorrect and correct predictions of the classification algorithm is called confusion matrix.

We use skLearn confusion_matrix() function to create confusion matrix, which takes values which are used (y_test) and values which are predicted(y_predict).

In confusion matrix rows represents the real classes while predicted classes are represented by columns.

The confusion matrix returns values which are divided into four categories as follows.

i. **True Positive: -**

The real value is positive, and the model predicted result is positive.

ii. **True Negative: -**

The real value is negative, and the model predicted result is negative.

iii. **False Positive: -**

The real value is negative, but model predicted value is positive.

iv. **.False Negative: -**

The real value is positive, but model predicted value is negative.

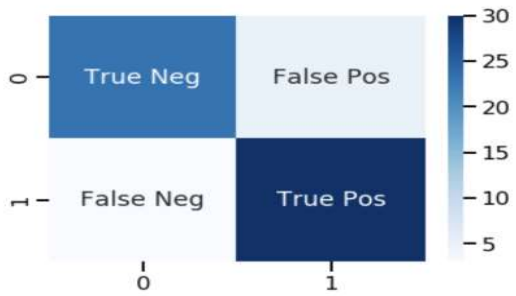


Fig.7. Structure of a confusion Matrix

The observations correctly predicted are there in the diagonal from the top left to the bottom right.

4.9 Result Analysis

The performance various using metrics such Classification Accuracy, Confusion Matrix, F1 Score, Precision, Recall is evaluated in the final step.

Metrics	Logistic Regression	Random Forest Classifier	KNN Classifier	Decision Tree Classifier	SVM Classifier
Accuracy	0.8281	0.8359	0.8203	0.8281	0.7812
Precision	0.8190	0.8148	0.8000	0.8131	0.7748
Recall	0.9663	0.9888	0.9888	0.9775	0.9663
F1-score	0.8866	0.8934	0.8844	0.8878	0.8600
Confusion Matrix	[[20 19] [3 86]]	[[19 20] [1 88]]	[[17 22] [1 88]]	[[19 20] [2 87]]	[[14 25] [3 86]]

Fig.8. Table for representing performance metrices of all algorithms used

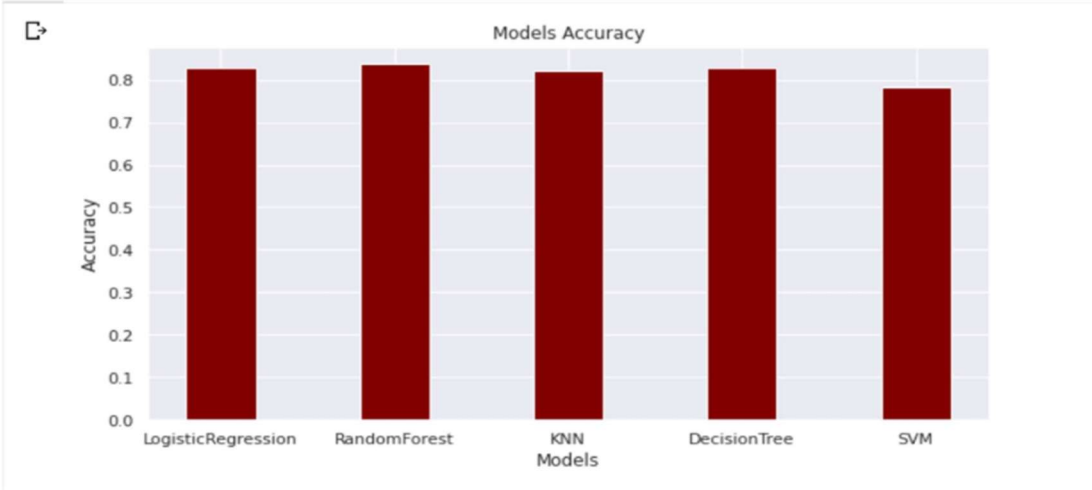


Fig 9. Bar graph of accuracy of all algorithms

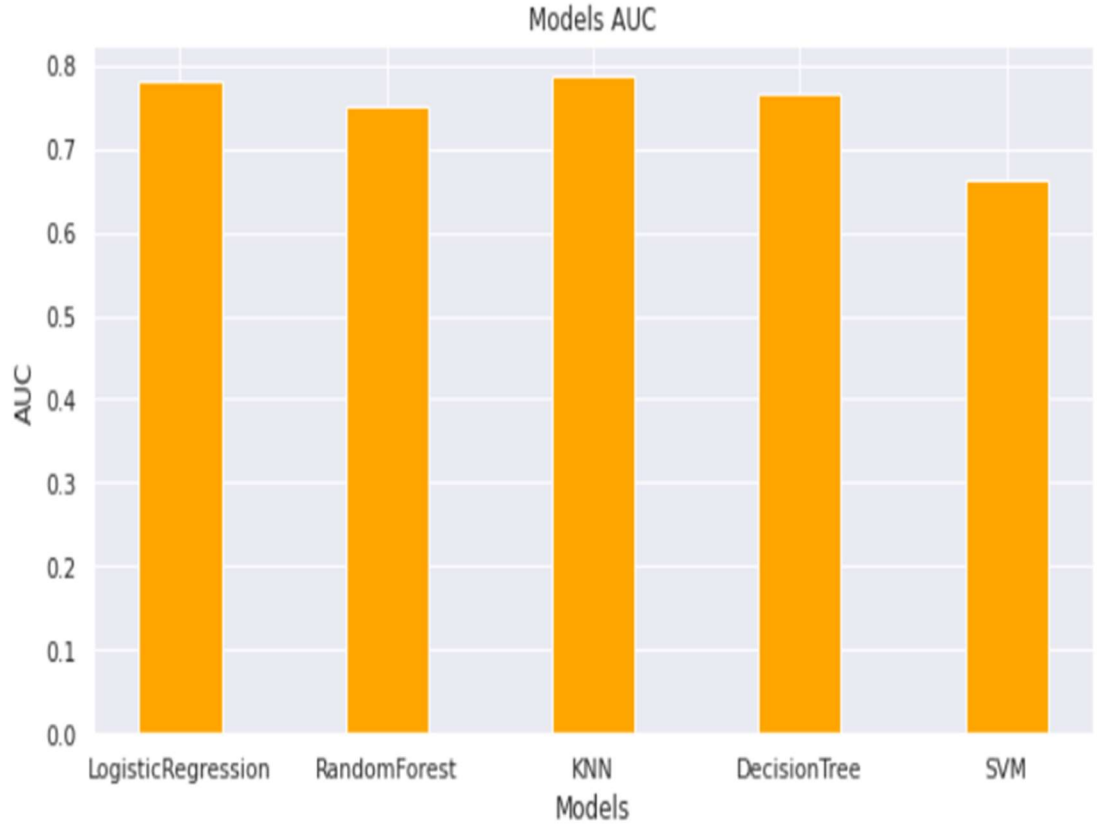


Fig.10. Bar Chart displaying AUC(Area Under The Curve) for each algorithm

5. CONCLUSION

The objective of this paper is to create a system using machine learning to predict the best choice for loan approval by taking into consideration certain attributes of the applicant. Five machine learning algorithms such as SVM, decision tree, random forest, KNN, logistic Regression are being used here. Evaluation of performance of these models are done using metrics such as accuracy, confusion matrix, f1 score, precision, recall and support. Finally, we found that amongst all the algorithms, *Random Forest* is giving the best results with an accuracy of *83.59% and F1-Score of 89.34%.* This system may not fit well when a new dataset with a greater number of attributes comes, so there may be a chance of over fitting. So, this paper could be modified in future by taking part in new testing to be made such as to pass new test cases. This system can assist companies in making the best judgment on whether to approve or deny a customer's loan request to automate the entire system. It can also assist the banking industry in establishing more effective distribution routes. It can be integrated with the automated processing system module in the near future.

6. REFERENCES

[1] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma Namburi Vimala Kumari, K Vikas (2019),” Loan Prediction by using Machine Learning Models” International Journal of Engineering and Techniques - Volume 5 Issue 2

[2] J. Tejaswini, T. Mohana Kavya, R. Devi Naga Ramya, P. Sai Triveni Venkata Rao Maddumala (2020) “ACCURATE LOAN APPROVAL PREDICTION BASED ON MACHINE LEARNING APPROACH”, Vol 11, Issue 4, ISSN NO: 0377-9254

[3] Kshitiz Gautam, Arun Pratap Singh, Keshav Tyagi, Mr. Suresh Kumar (2020)” Loan Prediction using Decision Tree and Random Forest” International Research Journal of Engineering and Technology (IRJET)

[4] Mehul Madaan Aniket Kumar, Chirag Keshri, Rachna Jain and Preeti Nagrath (2021)” Loan default prediction using decision trees and random forest: A comparative study” *IOP Conf. Ser.: Mater. Sci. Eng.* 1022 012042

[5] Prateek Dutta (2021) “A STUDY ON MACHINE LEARNING ALGORITHM FOR ENHANCEMENT OF LOAN PREDICTION” International Research Journal of Modernization in Engineering Technology and Science

[6] Vishal Singh, Ayushman Yadav, Rajat Awasthi (2021),” Prediction of Modernized Loan Approval System Based on Machine Learning Approach”, 2021 International Conference on Intelligent Technologies (CONIT)

[7]. P. Maheswari and C. V. Narayana, "Predictions of Loan Defaulter - A Data Science Perspective," *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, 2020, pp. 1-4, doi: 10.1109/ICCCS49678.2020.9277458.

[8]. Xiaojun Ma, Jinglan Sha, Dehua Wang, Yuanbo Yu, Qian Yang, Xueqi Niu, Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning,
Electronic Commerce Research and Applications, Volume 31,2018, Pages 24-39, ISSN15674223, https://doi.org/10.1016/j.elerap.2018.08.002.
[9]. Odegua, R.: Predicting bank loan default with extreme gradient boosting. arXiv preprint arXiv:2002.02011 (2020)

[10] Turkson, R.E., Baagyere, E.Y., Wenya, G.E.: A machine learning approach for predicting bank credit worthiness. In: 2016 Third International Conference on Ar-ti_cial Intelligence and Pattern Recognition (AIPR). pp. 1{7. IEEE (2016)

[11] Meer, K.: Machine learning models for mortgage default prediction in pakistan. In:
2021 International Conference on Artifcial Intelligence (ICAI). pp. 164{169. IEEE (2021)