Academic Task -I

Course Instructor: Dr. Rajan Kakkar

Section and Course: Q2155 and INTM574

LOAN APPROVAL PREDICTION USING MACHINE LEARNING

Bikram Roy: 12102780

Table of Content

1. INTRODUCTION

A loan is a bank's main source of revenue. The profits earned through loans account for most of the bank's profits. Even though the bank accepts the loan following a lengthy verification and testimony process, there is no guarantee that the chosen candidate is the right one. When done manually, this operation takes a long time. We can predict whether a given hopeful is safe or not, and the entire testimonial process is automated using machine literacy. Loan Prognostic is beneficial to both bank retainers and hopefuls.

Considering the two most important banking issues amongst others:

1) What is the borrower's risk level?
2) Given the danger, should we lend to the borrower?

The borrower's interest rate is determined by the answer to the first question. The interest rate, together with other factors (such as the time value of money), assesses the borrower's riskiness, the higher the interest rate, the riskier the borrower. Based on the interest rate, we will determine whether the applicant is eligible for the loan. Lenders (investors) provide loans to creditors in exchange for interest-bearing repayment guarantees. The lender only gets paid (interest) if the borrower pays back the loan. Whether he or she repays the loan or not, the lender loses money. Customers are given loans by banks in exchange for a guarantee of payback. Some people would fail on their loans because they were unable to repay them for various reasons. In the event of a default, the bank keeps insurance to reduce the risk of collapse. The insured amount might be used to cover the entire loan or just a portion of it. Banking operations rely on manual procedures to evaluate whether or not a borrower is qualified for a loan. When there were a significant number of loan applications, manual techniques were usually effective, but they were insufficient. Making a decision at the time would take a long time. As a result, the machine learning model for loan prediction can be used to assess a customer's loan condition and develop plans. This model extracts and introduces the key characteristics of a borrower that determine the loan status of the consumer. Finally, it produces the desired result (loan status). These reports make the job of a bank management easier and faster.

2. LITERATURE REVIEW

We begin our review with more general systematic reviews that focus on the application of machine learning in the field of banking risk management. Since the global financial crisis, risk management in banks has become increasingly important in guiding bank decision-making.

1. **Paper Name:** Loan Prediction by using Machine Learning Models

   Authors: Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma

   The modules discussed in this study were data collecting and pre-processing, implementing machine learning models, training, and testing the data, identification of outliers and imputation removal processing as well as removal during the pre-processing stage, this is done to estimate the chances SVM, of the present status of the loan approval process, In this study, DT, KNN, and gradient boosting models were applied divide the dataset into training and testing groups. The 80:20 guideline was applied to the procedures. Experimentation of the Decision Tree was found to have much higher loan approval rates. The other models have a lower forecast accuracy.

   Results: Accuracy achieved: 0.811

   Model used: Decision Tree

2. **Paper Name:** Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval

   Authors: Vaidya and Ashlesha

   In another study, the author, Ashlesha Vaidya, uses logistic regression as a machine learning tool to demonstrate how predictive techniques can be applied to real-world loan approval challenges. His study employs a statistical model (Logistic Regression) to forecast whether a loan should be authorised or denied based on a collection of applicant records. Even power terms and nonlinear effects can be used in logistic regression. This approach has some drawbacks, such as the need for independent variables for estimate and a high sample size for parameter estimation.

   Model Used: Logistic Regression

   Results: Accuracy achieved: 0.791

3. **Paper Name:** An Exploratory Data Analysis for Loan Prediction Based on Nature of the

Clients

Authors: X. Francis Jency, V.P.Sumathi, Janani Shiva Sri

Exploratory Data Analysis (EDA) was introduced in this research as a tool for forecasting loan amounts based on the nature of the client and their needs. The major factors concentrated during the data analysis were annual income versus loan purpose, customer trust, loan tenure versus delinquent months, loan tenure versus credit category, loan tenure versus credit category, loan tenure versus credit category, loan tenure versus the number of years in current job, and chances for loan repayment versus homeownership. Finally, the goal of this research was to deduce the constraints that a client experiences while asking for a loan and produce a payback prediction. Customers were also more interested in short-term loans than long-term loans, according to the findings.

Model Used: Exploratory Data Analysis (EDA)

3. NEED AND OBJECTIVES OF STUDY

3.1 Need of the Study

Loan Acceptance is a crucial step for banking institutions. The system approves or rejects the loan applications. In a bank's financial accounts, loan recovery is a substantial contributor. It is incredibly difficult to predict if a customer will repay the loan or not. So, if a company wants to automate the loan eligibility procedure (in real time) using client information submitted into an online application form. Gender, marital status, education, dependents, income, loan amount, credit history, and other information are all included. They established a problem to identify the client segments that are eligible for a loan amount, allowing them to target these customers individually. We predict loan acceptance using machine learning and SPSS software.

3.2 Objectives of the Study

Loan Prediction is extremely beneficial to both bank employees and applicants. The purpose of this paper is to provide a quick, straightforward, and efficient method of selecting qualified applicants. Due to fierce competition, banks are struggling to get an advantage over one another to improve overall business. Banks have realized that retaining customers and preventing fraud must be the strategic tool for healthy competition.

The objectives of this paper are to achieve two goals. They are as follows:

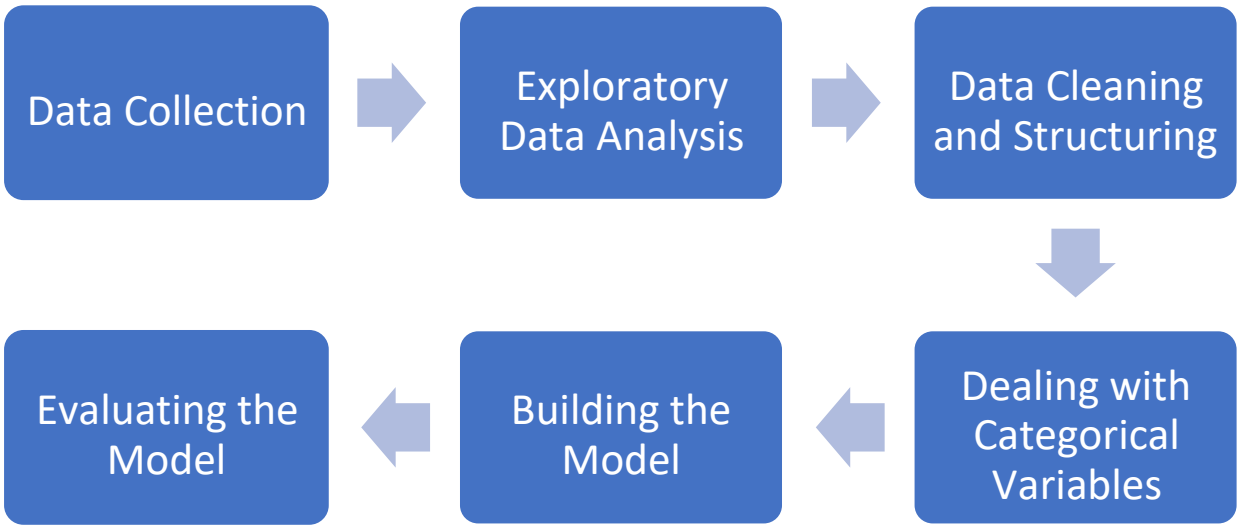
1. Identification of significant characteristics that indicate a borrower's ability to repay the loan.
2. Identifying the optimal model(s) for assessing credit risk and approving the loan application.

The goal of the problem is to determine which customers will be able to pay their bills and which customers will most likely be unable to do so. Clearly, we'll need to develop a classification system. Algorithms like logistic regression, decision trees, and random forests must be used. We need to develop a model that is accurate and has a low error percentage. The major goal of this study is to determine whether or not designating a loan to a certain person is safe. In this study, we use machine learning algorithms such as classification, logic regression, Decision Tree, and gradient boosting to predict loan data.

4. RESEARCH METHODOLOGY

4.1. Research Design

The proposed research design begins with data gathering, in which we leveraged a public dataset from Kaggle for loan prediction. Then, in the data exploration phase, we used various sorts of tables and figures to try to comprehend the data and relationships between characteristics, such as box plots, bar plots, histograms, heatmaps, and so on. The third phase entails using the examined data to train three or four classification algorithms based on the best-found factors that may contribute to the prediction process.



4.2. Population of Study

The population of study is the group we wish to make inferences about. Here, in this study were interested to find out all those people who have applied for any bank loan and want to know either their application was accepted or rejected. This would be impossible task to get the data of all these people single headedly. So, sampling is done to get the subset of the population which can represent the population. The population's mean and standard deviation should be roughly equal to the sample's mean and standard deviation then the sample can represent the population.

4.3. Sample Size and Data Description

I used a public dataset for loan prediction from Kaggle to benchmark the results with other models in order to validate the proposed model, as mentioned earlier. The dataset utilized consists of 13 features divided into two sets: a training set with 615 rows and a testing set with 368 rows. The features of the dataset are depicted in the table below, along with their proper descriptions.

Variable	Datatype	Description
Loan_ID	Categorical(non-numeric)	Unique loan ID
Gender	Categorical(non-numeric)	Male/female
Married	Categorical(numeric)	Yes/no
Dependents	Categorical(numeric)	No. of dependents
Education	Categorical(non-numeric)	Graduate/not-graduate
Self_employed	Categorical(non-numeric)	Yes/no
Applicant Income	Numeric feature	Applicant income
Loan_ID	Categorical(non-numeric)	Unique loan ID
Gender	Categorical(non-numeric)	Male/female
Married	Categorical(numeric)	Yes/no
CoapplicantInc	Numeric feature	Co-applicant income
LoanAmount	Numeric feature	Loan amount in thousands
Loan_Amount_Term	Numeric feature	Term of the loan in months
Credit_History	Categorical(numeric)	0/1
Property_Area	Categorical(non-numeric)	Urban/semi-urban/rural
Loan_Status	Categorical(non-numeric)	Yes/no

First 5 rows of dataset:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban	Y
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	Y
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	Y
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	Y

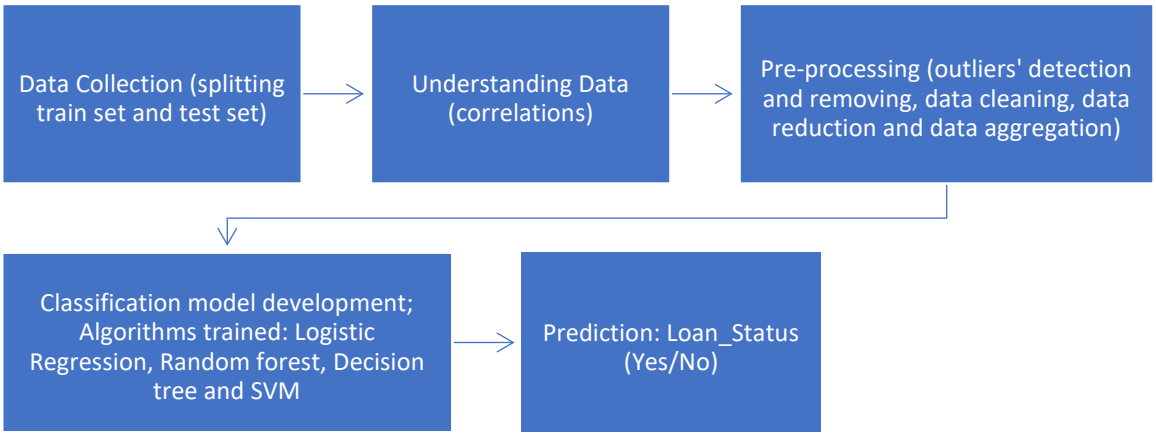
4.4. Sampling Technique

We used Stratified random sampling for our proposed sampling technique. Stratified random sampling is a sampling technique that divides a population into smaller sub-groups called strata. The strata are implemented based on members' shared traits or characteristics such as income or educational level in stratified random sampling, or stratification.

Implementing Stratified random sampling makes our training split represent the proportion of each value in the prediction variable. For example, in our dataset, if 25% of folk didn't get the loan approved and 75% got loan approved, setting Stratified random sampling will ensure that the random split has 25% of folk didn't get the loan approved and 75% who got loan approved.

4.5. Tools for analysis

Python, Jupyter Notebooks, MS Excel and knowledge of various Machine Learning Models are among the tools proposed for the analysis. The model is trained using classification methods such as Logistic Regression, Decision Tree, Random Forest Classifier, and Support Vector Machine because predicting the acceptance of a loan application is a classification problem. The steps for constructing the model are detailed in following figure.

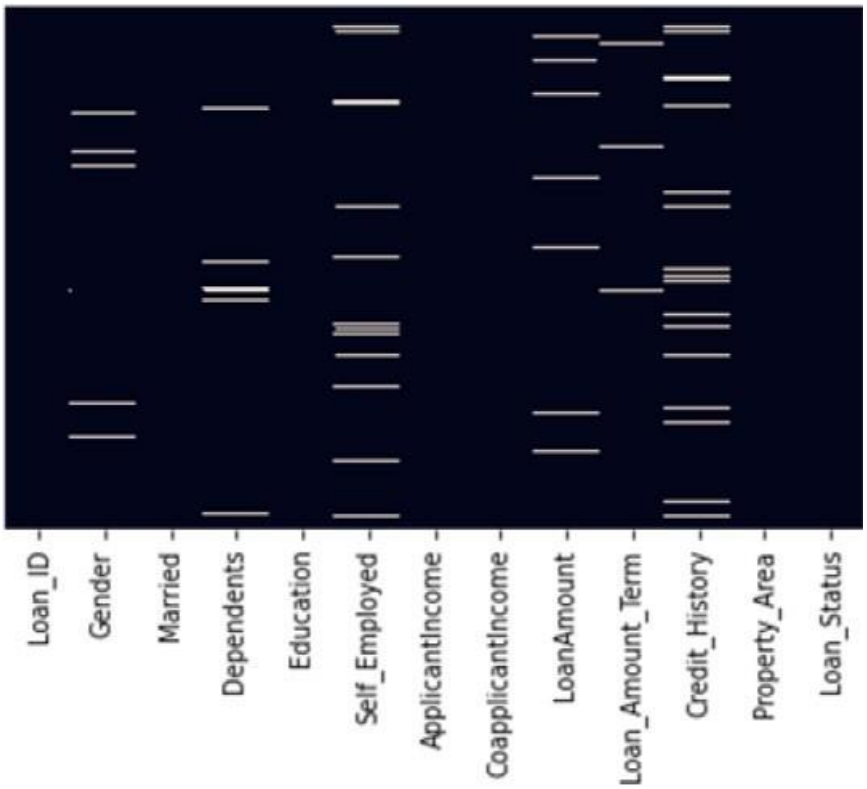


5. DATA PREPROCESSING

5. 1. Missing values Treatment:

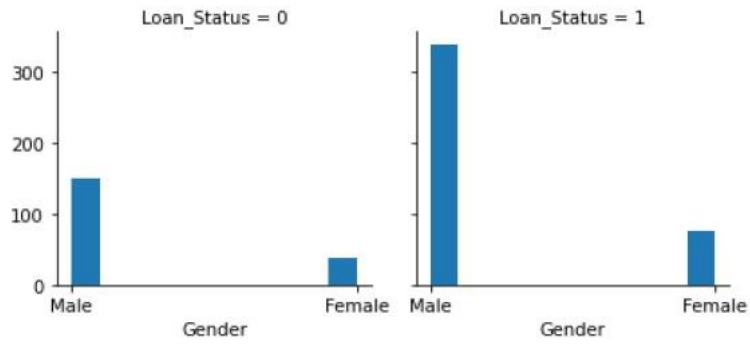
The data preprocessing begins by seeing whether any missing values are present in dataset or not. So, heat map was used to about the missing values and it also shows no duplicate missing values in any features.

As a result, there is no need to remove any features except Loan_ ID because it has no effect on the target variable. It is worth noting that a feature is eliminated if it has a high number of missing values in comparison to its length, as it adds no useful information to the dataset.



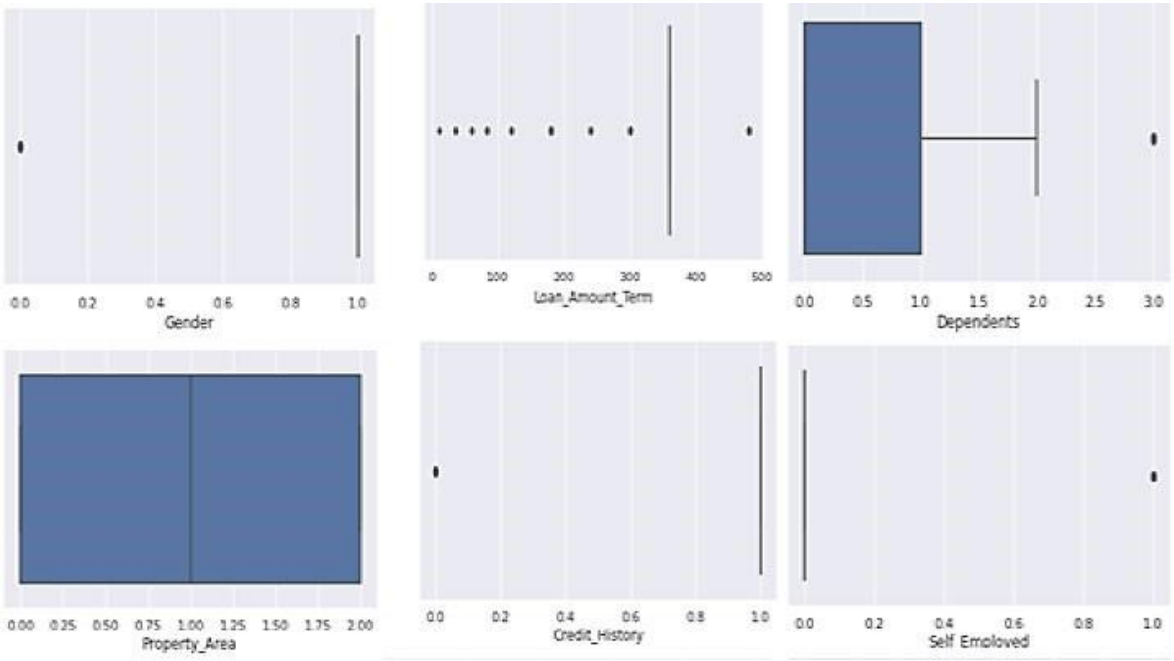
We will use the Mean Imputation technique to estimate the missing values for the few missing data. In the gender feature, for example. The dataset contains 489 males and 112 females. Furthermore, we discovered 13 missing data in the train section and 11 missing values in the test section. We discover that males have a high connection with Loan status when we calculate the correlation between gender and Loan status. As a result, the male category was used to fill in the gaps. To ensure that the model functions effectively, the gender values were changed to numeric values.

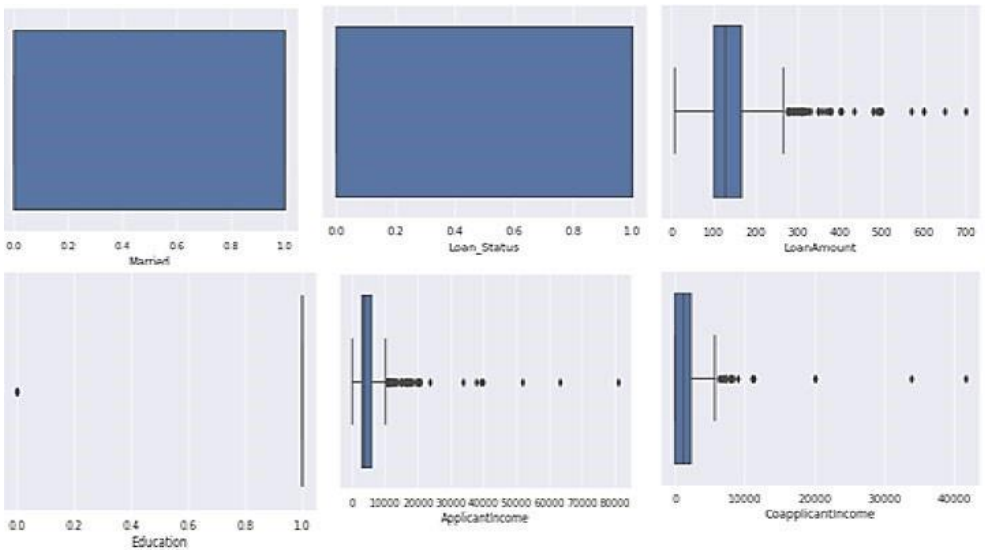
	Gender	Loan-Status
0	Female	0.66
1	Male	0.69



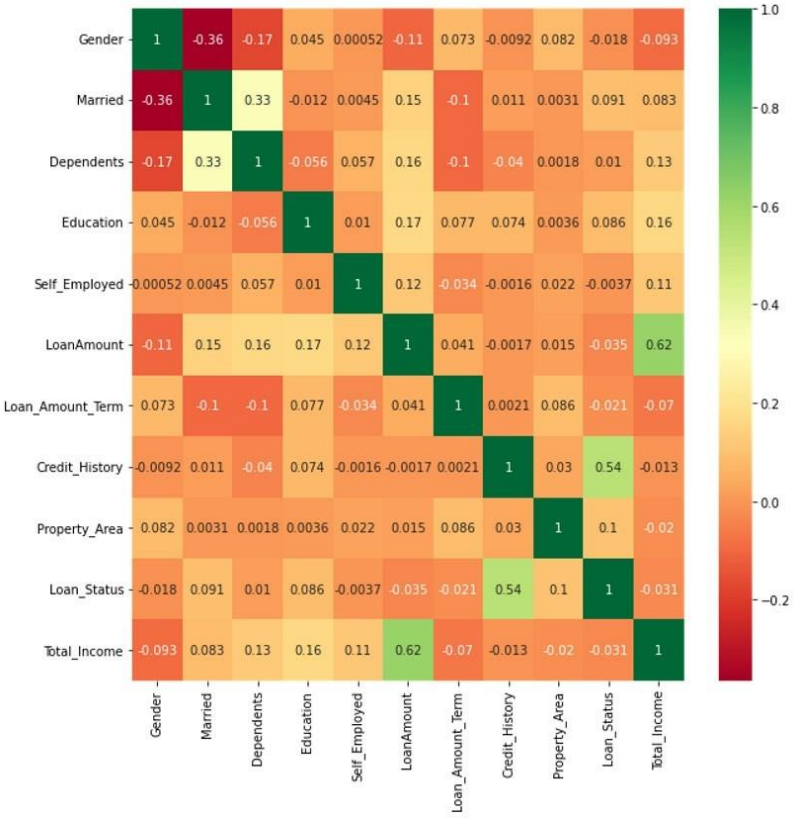
5. 2. Outliers Treatment:

After checking for missing values, the following step is to look for any outliers in the obtained data. As shown in Figure 8, we employ the box plots technique to identify outliers. The box plots figure reveals four variables with outliers: applicant income, Coapplicant income, loan amount, and loan amount term. As a result, the discovered outliers were eliminated using the univariate technique.





The final stage in pre-processing is to test the correlation between data properties to determine which is the most significant and notable feature of the prediction method Because of this We utilize a heat map to visualize the correlation for this purpose. among the variables The heat map for the is depicted in Figure. Data attributes gathered Figure shows a heat map. We can quickly identify the most essential loan attribute prediction. Notably, Loan ID has been removed from the list. The heat map because it has no bearing on the prediction process.

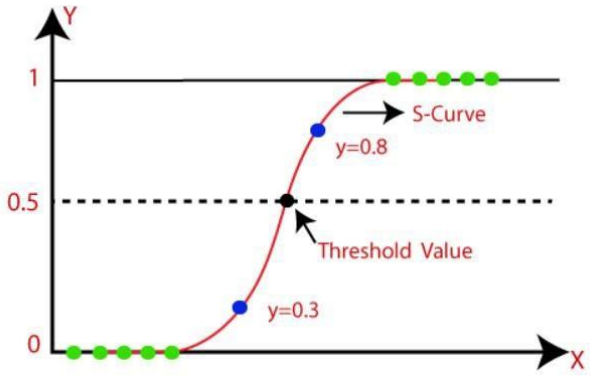


6. ALGORITHM USED

6.1. Logistic Regression

The Supervised Learning technique includes the popular Machine Learning algorithm of logistic regression. It predicts the categorical dependent variable from a set of independent variables.

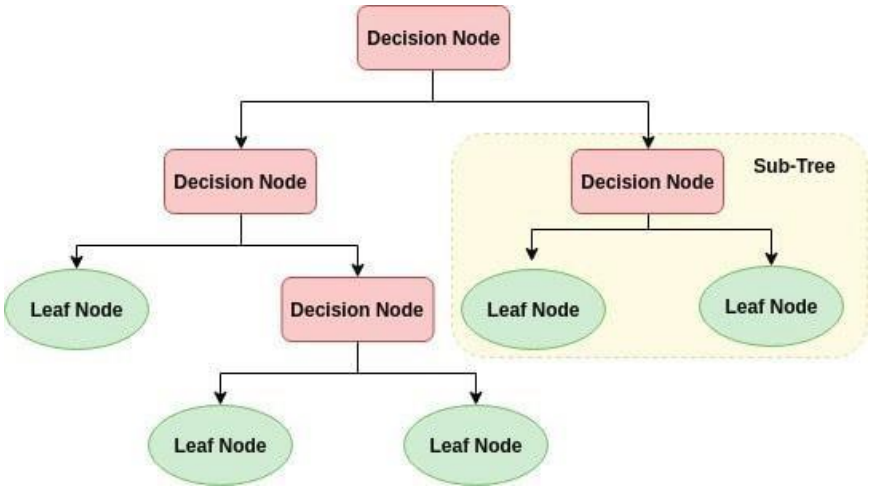
Instead of using a regression line, we used a "S" shaped logistic function that predicts two maximum values (0 or 1). Logistic regression is used to predict the outcome of a categorical dependent variable. It can be Yes or No, 0 or 1, true or false, and so on, but rather than offering exact values like 0 or 1, it delivers probabilistic values that lie between 0 and 1.



6.2. Decision Tree

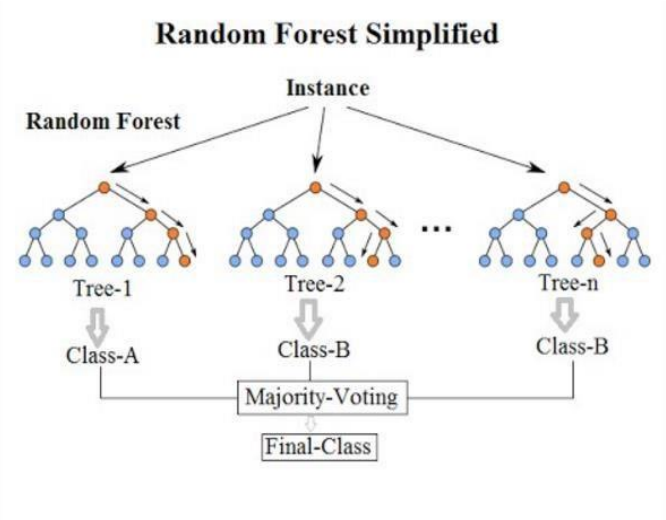
The Decision Tree algorithm belongs to the supervised learning algorithm family. By following the rules in the training data set, a Decision Tree can be used to create a training model that can be used to predict the class or value of the target variable (training data). We start at the bottom of the tree when using Decision Trees to estimate a record's class label. The root attribute values are compared to the record attribute values. Following the branch corresponding to that value based on the comparison,

we go on to the next node.



6.3. Random Forest Classifier

Random Forest is an algorithm for supervised learning. It produces a forest out of a group of decision trees trained using the bagging approach. The main idea of the bagging method is that mixing many learning models enhances the final output. The random forest analyses each tree's projections and forecasts the final model based on the majority of votes. The more trees in the forest, the higher the precision and the lower the likelihood of overfitting.



6.4. The Prediction Module

We trained and evaluated three algorithms on the data after completing pre-processing. Logistic regression is the first algorithm. Logistic regression is a statistical technique for predicting the likelihood of binary answer variables. It is utilized when our label(y) is a binary answer variable with values of 1 or 0, yes or no, and so on. It is a simple and widely used algorithm for solving classification problems. A Decision Tree classifier is the second algorithm. A decision tree is a popular approach for creating classification models. The models are constructed in the shape of a tree.

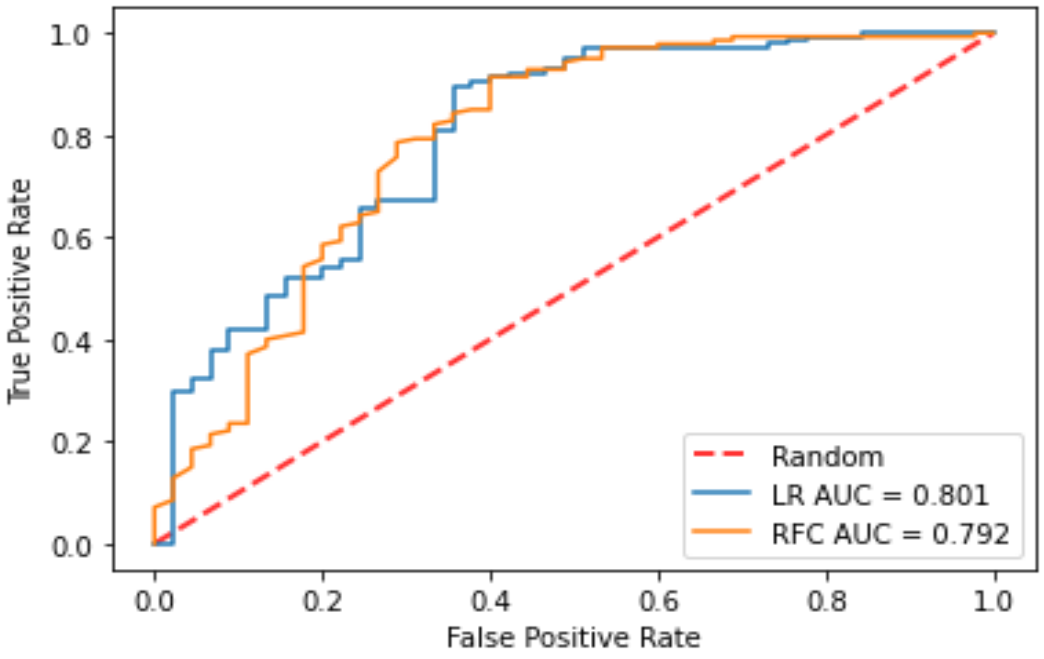
Each node in the tree represents a variable test, and each branch that descends from that node represents one of the possible values for that characteristic. Random Forest Classifier is the third algorithm. Random Forest is a classification system comprised of several decision trees. Using feature randomness and Bagging, it attempts to construct an uncorrelated forest of trees. An uncorrelated forest of trees predicts more accurately than any individual tree.

7. RESULT ANALYSIS

This section records the experimental findings of the suggested prediction model utilizing the three machine algorithms (logistic regression, decision tree, and random forest). The table below summarizes the findings of the three algorithms.

Algorithms	Precision	Recall	F1	Accuracy	AUC
Logistic Regression	0.79	0.98	0.88	0.91	0.80
Decision tree	0.77	0.83	0.80	0.82	0.75
Random forest	0.78	0.93	0.84	0.86	0.79

Precision, recall, F1, accuracy, and area under the curve are the performance metrics employed. AUC provides information about the performance measure over all conceivable classification levels. The greater the AUC number, the better the model at determining whether or not the loan should be authorized. As demonstrated in the table, logistic regression outperforms the other two techniques. Furthermore, receiver operating characteristic (ROC) values have been utilized to demonstrate the diagnostic ability of binary classifiers. The prediction model that is closer to the top left corner performs better than the others. The image demonstrates the ROC for the three models and indicates that linear regression outperforms the others in various threshold classifications.



8. CONCLUSION

One of the most important tasks for banks is loan application processing. Many approaches to loan prediction are proposed in the literature. Machine-learning methods are presented among these ways to anticipate loan status based on various inputs and criteria. As a result, we trained and evaluated a dataset to predict loan approval in this study. The dataset contains 13 variables, and we discovered that Credit History is the most essential feature for loan prediction. The preprocessing process begins with data comprehension, data cleaning, outlier discovery, and elimination.

In the suggested prediction model, three machine-learning algorithms were trained and tested on the data: linear regression, decision tree, and Random Forest. Logistic regression outperformed the others, with 84% accuracy, while Decision Tree and Random Forest achieved 72% and 81% accuracy, respectively. The ROC curve was used to validate the reported results. Furthermore, the proposed model was compared to related publications. This work's potential next directions include acquiring a realistic dataset with more prediction features to improve prediction. Furthermore, forecast accuracy must be increased. This can be accomplished by utilizing feature extraction and hybrid machine learning algorithms.

Code hosted on Kaggle: https://www.kaggle.com/code/shivkumarmehmi/loan-approvalprediction

9. REFERENCES

1. ZAMANI, S., and MOGADDAM, (2016) A. Natural Customer Ranking of Banks in Terms of

   Credit Risk by Using Data Mining: A Case Study: Branches of Mellat Bank of Iran. Journal of UMP Social Science Technology Management.
2. BAE, J.K., and KIM, (2018) J. A personal credit rating prediction model using data mining in ubiquitous smart environments. International Journal of Distributed Sensor Networks.
3. HAMID, A.J., and AHMED, T.M. (2016) Developing prediction model of loan risk in banks using data mining. Machine Learning Appliances an International Journal.
4. 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).
5. Sudhamathy G.-"Credit Risk Analysis and Prediction Modelling of Bank Loans Using R", International Journal of Engineering and Technology (IJET), Vol. 8, No. 5, pp. 1954-1966, Oct-Nov 2016.