*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661,p-ISSN: 2278-8727*
*PP 18-21*
*www.iosrjournals.org*

# Loan Approval Prediction based on Machine Learning Approach

Kumar Arun, Garg Ishan, Kaur Sanmeet
*(sh.arun.rana@gmail.com , CSED, Thapar University, India)*
*(ishangarg9292@gmail.com, CSED, Thapar University, India)*
*(sanmeetkbhatia@gmail.com,CSED,Thapar University,India)*

***Abstract:*** *With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So in this paper we try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. The main objective of this paper is to predict whether assigning the loan to particular person will be safe or not. This paper is divided into four sections (i)Data Collection (ii) Comparison of machine learning models on collected data (iii) Training of system on most promising model (iv) Testing*

***Keywords -*** *Loan, Machine Learning, Training, Testing, Prediction.*

## I. INTRODUCTION

Distribution of the loans is the **core business part** of almost every banks. The main portion the bank's assets is directly came from the profit earned from the loans distributed by the banks. The prime objective in banking environment is to invest their assets in safe hands where it is. Today many banks/financial companies approves loan after a regress process of verification and validation but still there is no surety whether the chosen applicant is the deserving right applicant out of all applicants. Through this system we can predict whether that particular applicant is safe or not and the whole process of validation of features is automated by machine learning technique. The disadvantage of this model is that it emphasize different weights to each factor but in real life sometime loan can be approved on the basis of single strong factor only, which is not possible through this system.

Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this Paper is to provide quick, immediate and easy way to choose the deserving applicants. It can provide special advantages to the bank. The Loan Prediction System can can automatically calculate the weight of each features taking part in loan processing and on new test data same features are processed with respect to their associated weight .A time limit can be set for the applicant to check whether his/her loan can be sanctioned or not. Loan Prediction System allows jumping to specific application so that it can be check on priority basis. This Paper is exclusively for the managing authority of Bank/finance company, whole process of prediction is done privately no stakeholders would be able to alter the processing. Result against particular Loan Id can be send to various department of banks so that they can take appropriate action on application. This helps all others department to carried out other formalities.

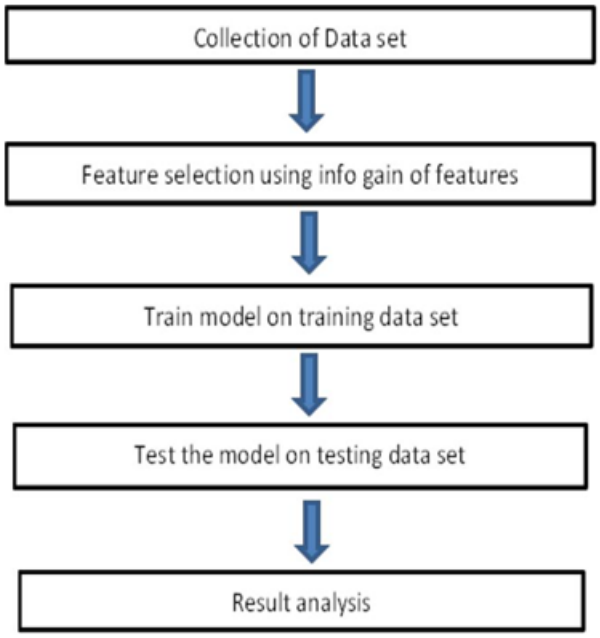
## II. Data Set

The training data set is now supplied to machine learning model, on the basis of this data set the model is trained. Every new applicant details filled at the time of application form acts as a test data set. After the operation of testing, model predict whether the new applicant is a fit case for approval of the loan or not based upon the inference it conclude on the basis of the training data sets.

| Variable Name | Description | Type |
|---|---|---|
| Loan_ID | Unique Loan ID | Integer |
| Gender | Male/ Female | Character |
| Maritial_Status | Applicant married (Y/N) | Character |

| Variable Name | Description | Type |
|---|---|---|
| Dependents | Number of dependents | Integer |
| Education_Qualification | Graduate/ Under Graduate | String |
| Self_Imployed | Self Imployed (Y/N) | Character |
| Applicant_Income | Applicant income | Integer |
| Co_Applicant_Income | Coapplicant income | Integer |
| Loan_Amount | Loan amount in thousands | Integer |
| Loan_Amount_Term | Term of loan in months | Integer |
| Credit_History | credit history meets guidelines | Integer |
| Property_Area | Urban/ Semi Urban/ Rural | String |
| Loan_Status | Loan Approved(Y/N) | Character |

### 2.1 Loan Prediction Methodology




### 2.2 MACHINE LEARNING METHODS:

Six machine learning classification models have been used for prediction of android applications .The models are available in R open source software. R is licensed under GNU GPL. The brief details of each model is described below.

2.2.1 Decision Trees (C5.0):

The basic algorithm of decision tree [7] requires all attributes or features should be discretized. Feature selection is based on greatest information gain of features. The knowledge depicted in decision tree can represented in the form of IF-THEN rules. This model is an extension of C4.5 classification algorithms described by Quinlan.

2.2.2 Random Forest (RF):
Random forests [8] are a group learning system for characterization (and relapse) that work by building a large number of Decision trees at preparing time and yielding the class that is the mode of the classes yield by individual trees.

2.2.3 Support Vector Machine (SVM):
Support vector machines are administered learning models that uses association r learning algorithm which analyze features and identified pattern knowledge, utilized for application classification. SVM can productively perform a regression utilizing the kernel trick, verifiably mapping their inputs into high-dimensional feature spaces [9].

2.2.4 Linear Models (LM):
The Linear Model [10] is numerically indistinguishable to a various regression analysis yet burdens its suitability for both different qualitative and numerous quantitative variables.

2.2.5 Neural Network (Nnet):
Neural networks [14] are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs, to find patterns in data, or to capture the statistical structure in an unknown joint probability distribution between observed variables.

2.2.6 Adaboost (ADB):
**Adaboost** short for " Adaptive Boosting ". It is delicate to noisy information data and outliers. It is different from neural systems and SVM because Adaboost preparing methodology chooses just those peculiarities known to enhance the divining power of the model, decreasing dimensionality and conceivably enhancing execution time as potentially features don't have to be processed.[14]

**III. Parameter setting for machine learning models**

| Model | Parameter Setting |
|---|---|
| Decision Trees | Min Split = 20, Max Depth = 30, Min Bucket = 7 |
| Random Forest | Number of tree = 500,Number of variables=8 |
| Support Vector Machine | Kernel Radial Basis |
| Linear Model | Multinomial |
| Ada boost | Min Split = 20, Max Depth = 30, Number of tree = 50 |
| Neural network | Hidden layer nodes=10 |

*Loan Prediction*

## IV. CONCLUSION

From a proper analysis of positive points and constraints on the component, it can be safely concluded that the product is a highly efficient component. This application is working properly and meeting to all Banker requirements. This component can be easily plugged in many other systems.

There have been numbers cases of computer glitches, errors in content and most important weight of features is fixed in automated prediction system, So in the near future the so –called software could be made more secure, reliable and dynamic weight adjustment .In near future this module of prediction can be integrate with the module of automated processing system. the system is trained on old training dataset in future software can be made such that new testing date should also take part in training data after some fix time.

## REFERENCES

[1]. Rattle data mining tool: available from http://rattle.togaware.com/rattle-download.html
[2]. Aafer Y, Du W &Yin H 2013, DroidAPIMiner: 'Mining API-Level Features for Robust Malware Detection in Android', in: Security and privacy in Communication Networks Springer, pp 86-103 .
[3]. Ekta Gandotra, Divya Bansal, Sanjeev Sofat 2014**, '**Malware Analysis and Classification: A Survey'available from http:// www.scirp.org/journal/jis

[4]. K. Hanumantha Rao, G. Srinivas, A. Damodhar, M. Vikas Krishna: Implementation of Anomaly Detection Technique Using Machine Learning Algorithms: Internatinal Journal of Computer Science and Telecommunications (Volume2, Issue3, June 2011).
[5]. *J. R. Quinlan. Induction of Decision Tree. Machine Learning, Vol. 1, No. 1. pp. 81-106., 1086.*
[6]. *Mean Decrease Accuracy https://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.hml*
[7]. *J.R. Quinlan. Induction* of decision trees. MachinelearningSpringer, 1(1):81–106, 1086.
[8]. Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. R News( http://CRAN.R-project.org/doc/Rnews/ ), 2(3):9–22, 2002.
[9]. S.S. Keerthi and E.G. Gilbert. Convergence of a generalizeSMO algorithm for SVM classifier design. Machine Learning, Springer, 46(1):351–360, 2002.
[10]. J.M. Chambers. Computational methods for data analysis. Applied Statistics, Wiley, 1(2):1–10, 1077.

# A Model to Predict Loan Defaulters using Machine Learning

R. B. Saroo Raj [1], Gurpartap Singh [2], Balaji S [3], K. H. Ajit Baskar [4]
[1] Asistant Professor, Computer Science Department, SRM Institute of Science and Technology, Chennai,India
[2, 3, 4] Computer Science Department, SRM Institute of Science and Technology, Chennai,India

**Abstract – With the upgrade in the managing an account part bunches of individuals are applying for bank loans however the bank has its restricted resources which it needs to give to constrained individuals just, so discovering to whom the loan can be allowed which will be a more secure alternative for the bank is a regular procedure. So, in this paper we endeavour to lessen this hazard factor behind choosing the protected individual to spare heaps of bank endeavours and resources. This is finished by mining the Data of the past records of the general population to whom the loan was allowed previously and based on these records/encounters the machine was prepared utilizing the machine learning model which give the most precise outcome. The primary target of this paper is to anticipate in the case of allotting the loan to specific individual will be protected or not. This paper is separated into four segments (i)Data Collection (ii) Comparison of machine learning models on gathered data (iii) Training of framework on most encouraging model (iv) Testing.**

**Index Terms – loan, machine learning, data set, prediction.**

## 1. INTRODUCTION

Conveyance of the loans is the core business part of every banks. The principle partitions of bank's profit is straightforwardly originated from the benefit earned from the loans distributed by the banks. The prime goal in managing an account domain is to put their benefits in safe hands where it is. Today numerous banks/monetary organizations favour loan after a relapse procedure of confirmation and approval yet at the same time there is no surety whether the picked candidate is the meriting right application out everything being equal. Through this framework we can anticipate whether that specific candidate is protected or not and the entire procedure of approval of highlights is computerized by machine learning method. The burden of this model is that it underscores diverse weights to each factor yet in genuine at some point loan can be affirmed based on single solid factor just, which isn't conceivable through this framework. Loan Prediction is extremely useful for representative of banks and also for the candidate moreover. The point of this Paper is to give brisk, quick and simple approach to pick the meriting candidates. It can give extraordinary preferences to the bank. The Loan Prediction System can consequently figure the heaviness of every element participating in loan handling and on new test data same highlights are prepared as for their related weight. A period breaking point can be set for the candidate to check whether his/her loan can be authorized or not. Loan Prediction System enables hopping to particular application with the goal that it very well may be keep an eye on need premise. This Paper is solely for the overseeing expert of Bank/fund organization, entire procedure of expectation is done secretly no partners would have the capacity to modify the handling. Result against specific Loan Id can be send to different division of banks so they can make proper move on application. This encourages all others office to did different customs

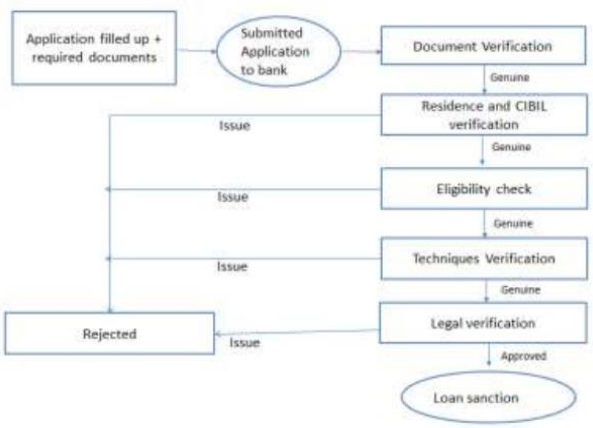



Figure 1 Process of Loan Sanction

## 2. RELATED WORK

Many researches have been conducted based on data mining and data analysing in the field of financial and banking sector. This section presents briefly some of these techniques which are used in loans management and their finding Sudhakar et al focused on specifying the data mining applications usefulness, these applications are using several machine learning algorithms such as decision trees and Radial Basis Neural Networks. This study came with in which way to apply these applications in a loan approval assessment field. McLeod presents Neural networks properties and their fitness for the credit granting process. [1]

Barney et al made a comparison of the performance of prediction algorithms to identify the farmers who will default on the loans of their Home Administration and those farmers who return back the credits as in the appointment. By using an

unstable data, this study proofed that neural networks regarding better logistic regression to classify farmers into two groups, those who pay back on time and those who default to return their loans [2].

## 3. PORPOSED MODELLING

We mean by loan assessment process, the grouping of steps that are taken to decide about giving a loan to the client or not. At the point when the client applies for a loan conceding application, the bank officer must research about what called 5 C's which are Character (or Credit History), Cash Flow (or Capacity), Collateral, Capitalization and Conditions. It is useful for assessment loan application and it viewed as a supportive system for gauge the credit hazard identified with a plausible bank.

## 4. DATA SET

The raw data set contains 75 fields for each loan begun. In any case, not the majority of the fields are helpful for our learning models, for example, the loan ID and the month in which last instalment was received, and in this way we evacuated such fields. We additionally evacuated fields for which more noteworthy than 10% of the loans were missing data for. Clear cut highlights, for example, address state (for instance, California), were ventured into Boolean sections, one segment for each particular esteem that the highlight could take. At long last, we evacuated any loans that were missing data for any field (around 3% of the loans in our dataset). To name the dataset, we characterized any loan that defaulted, were charged off, or were late on instalments was ordered as negative precedents, while we arranged any loan that was completely paid or current was named positive precedents.

| No | The attribute | Description | Data type |
|---|---|---|---|
| 1 | Credit_history | Previous history of customer credit | Nominal |
| 2 | Purpose | The loan purpose | Nominal |
| 3 | Gender | Male or female | Nominal |
| 4 | Credit_amount | The amount of credit | Numeric |
| 5 | Age | Customer Age | Numeric |
| 6 | Housing | Rent, own or for free | Nominal |
| 7 | Job | Is the customer has a job | Nominal |
| 8 | Class | The class of loan good/bad | Nominal |

Table 1 Data set Description

## 5. MODEL IMPLEMENTATION

The process of classification crowds the data set into groups of classes according to their similarity. There are several classification algorithm or classifiers like Naïve Bayes Classifier, Neural Network Classifier, decision Tree Classifier. There are several algorithms in each of this technique which used to produce a model to predict the class of unknown class tables. The major goal of this algorithm is the provision by a model for predicting the class of unknown records

Every classifier algorithm consists of these few steps:

- Prepare the training set, a record that are already have known class label
- Build model by applying one of learning algorithm to learning data set
- Apply model to unknown test data set
- Evaluate the accuracy of model

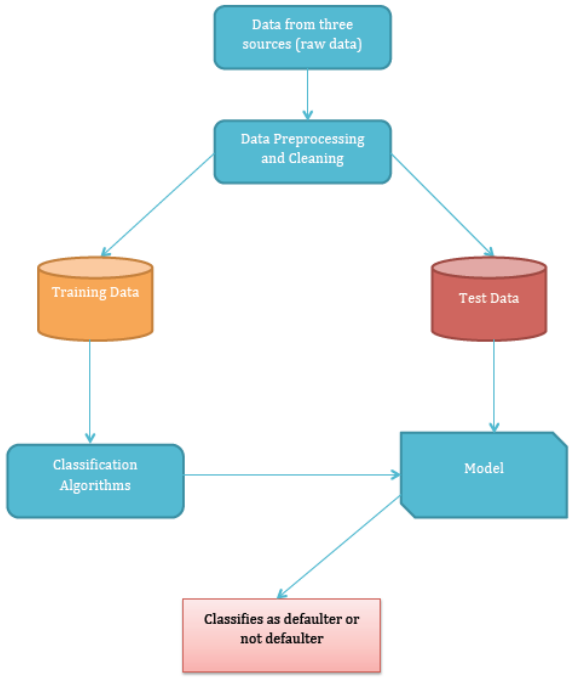



Figure 2 Workflow Diagram

In our research we use three different classification algorithms to build two different models. These algorithms are j48 decision tree, and naïve Bayes. Original data set has been divided into two parts of 20% and 80%. These are used for testing and training data respectively.

a. *J48 Classification Algorithm* -J48 is the upgraded version of C4.5 algorithm or can be seen as C4.5 implementation. J48 takes as an info the arrangement of tables and produce a decision tree as a yield. The created decision tree is indistinguishable to the structure of tree. It comprises of root, halfway and leaf hub. The hubs in the created tree contain a decision which manual for the outcome. It split the information data set into totally unrelated sets, each set with a name. Part measure is connected to figure out which credit prompt the ideal part like data gain foundation [3]

b. *Naïve Bayes*- The Bayesian is a supervised learning method. It portrayed with it is style,

straightforwardness, and robustness. Therefore, it is turned out to be broadly utilized in prediction or classification purposes. It guesses that properties of a class are self-determining in real life. [4]

## 6. RESULTS AND DISCUSSIONS

Naive Bayes, we acquired the outcomes from the two analyses in table 2 and in the wake of looking at the effectively grouped occurrence percent we find that the best algorithm for loan arrangement is j48 algorithm. J48 algorithm is best since it has high exactness and low mean supreme mistake as appeared in the outcome. Additionally, it is able to characterize the cases effectively than alternate strategies. Perplexity grid of the two algorithms demonstrated that the j48 algorithm is the best one. The tests have been completed a few times and, in each time, the preparing and test sets estimate have been changed (80% preparing 20% test set,60% preparing 40% test and 70% preparing 30% test) and we acquire a similar outcome which is J48 algorithm is best in arranging loans to great and terrible loan. this model help bank administrator to acknowledge or dismiss loan applications by anticipating that if the exchange will lead bank to chance or not and bolster decision producer to settle on a compose decisions.

| **Techniques** | **Correctly classified instance percent** |
|---|---|
| J48 | 78.3784% |
| Naïve Bayes | 73.8739% |

Table 2 Results from algorithms

## 5. CONCLUSION

In this paper, two algorithms - j48, and naive Bayes algorithms were used to build a predictive model that can be used to predict and classify the applications of loans that introduced by the customers to good or bad loan by investigating customer behaviors and previous pay back credit. The model has been implemented by using python and machine learning. After applying classification's data mining techniques algorithms which are j48, and naive Bayes, we find that the best algorithm for loan classification is j48 algorithm. J48 algorithm is best because it has high accuracy and low mean absolute error

## ACKNOWLEDGEMENT

This work was supported by computer science department of SRM Institute of science and technology, Chennai. All faculties of department helped us in making this paper. We are greatly thankful to all of them. We would like to thank Mr. R. B. Saroo Raj for guiding us in this project.

## REFERENCES

[1] Ogawa, Ms Sumiko, et al. Financial Interconnectedness and Financial Sector Reforms in the Caribbean. No. 13-175. International Monetary Fund, 2013.

[2] Strahan, Philip E. "Borrower risk and the price and nonprice terms of bank loans." FRB of New York Staff Report 90 (1999).Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining pproaches for Healthcare." International Journal of Bio-Science and Bio-Technology 5.5 (2013): 241-266

[3] AboobydaJafar Hamid and Tarig Mohammed Ahmed, "DEVELOPING PREDICTION MODEL OF LOAN RISK IN BANKS USING DATA INING", Machine Learning and Applications: An International Journal (MLAIJ) Vol.3, No.1, March 2016.

[4] vTomar, Divya, and Sonali Agarwal. "A survey on Data Mining pproaches for Healthcare." International Journal of Bio-Science and Bio-Technology 5.5 (2013): 241-266.

[5] Sharma, Poonam, and Gudla Balakrishna. "PrefixSpan: Mining Sequential Patterns by Prefix-Projected Pattern." International Journal of Computer Science and Engineering Survey 2.4 (2011):111.

[6] Chitra, K., and B. Subashini. "Data Mining Techniques and its Applications in Banking Sector."International Journal of Emerging Technology and Advanced Engineering 3.8 (2013): 219-226.

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 08 | Aug 2020 www.irjet.net p-ISSN: 2395-0072

# Loan Prediction using Decision Tree and Random Forest

**Kshitiz Gautam[1], Arun Pratap Singh[2], Keshav Tyagi[3], Mr. Suresh Kumar[4]**

*[1-3]BTech student, Dept. of IT, Galgotias College of Engineering and Technology, Greater Noida, U.P*
*[4]Assistant Professor, Dept. of IT, Galgotias College of Engineering and Technology, Greater Noida, U.P*

---

**Abstract -** *In India, the number of people or organization applying for loan gets increased every year. The bank employees have to put in a lot of work to analyse or predict whether the customer can pay back the loan amount or not (defaulter or non-defaulter) in the given time. The aim of this paper is to find the nature or background or credibility of the client that is applying for the loan. We use exploratory data analysis technique to deal with the problem of approving or rejecting the loan request or in short loan prediction. The main focus of this paper is to determine whether the loan given to a particular person or an organization shall be approved or not.*

***Key Words***: Loan, Prediction, Machine Learning, Training, Testing.

## 1. INTRODUCTION

The term banking can be referred to as receiving and protecting money that is deposited by an individual or an entity. It also includes lending money to people and businesses which has to be paid back within the given amount of time without failing. Banking is a sector that is regulated in most of the countries as it is an important factor in determining the financial stability of the country. The prime goal in banking sector is to invest their assets in safe hands where there are less chances of failure. Today many banks and financial companies approve loan after a stressful, long and weary process of verification but still there is no surety whether the chosen applicant is credible or not or in other words if he is able to return the amount with interest in the given time. The purpose of the loan can be anything based on the customer needs. Loans are broadly divided as open ended and close-ended loans.

Examples of open-end loans are credit cards and a home equity line of credit (HELOC).

Close-ended loans decreases with each payment. It means the amount is reduced after an instalment.

In other words, it is a legal term that cannot be modified by the borrower. Personal loans, mortgages, auto payments, EMI and student loans are the most common examples of close-ended loans.

Secured or collateral loan are those loans that are protected by an asset. Houses, Vehicles, Savings accounts are the personal properties used to secure the loan.

## 2. DATA SET

A collection of data is taken from the banking sector. The Data set is in ARFF (Attribute-Relation File Format) format that is acceptable by Weka. ARFF file is composed of tags that include the name, types of attributes, values and data itself. For this paper we are using 12 attributes like gender, marital status, qualification, income, etc.

The table below represents the data set that we have used:

**Table-1:** Data set variables along with description and type

| Variable Name | Description | Type |
|---|---|---|
| Loan_ID | Unique ID | Integer |
| Gender | Male/Female | Character |
| Marital_Status | Applicant married(Y/N) | Character |
| Dependents | Number of Dependents | Integer |
| Education_Qualification | Graduate/Under Gradute | String |
| Self_Employed | Self-employed(Y/N) | Character |
| Applicant_Income | Applicant income | Integer |
| Co_Applicant_Income | Co-applicant income | Integer |
| Loan_Amount | Loan amount in thousands | Integer |
| Loan_Amount_Term | Term of loan in months | Integer |
| Credit_History | Credit history meets guidelines | Integer |
| Property_Area | Urban/Semi urban/Rural | String |
| Loan_Status | Loan Approved(Y/N) | Character |

Now in machine learning model, we first apply the training data set, in this data set the model is trained with known examples. The entries of new applicants will act as a test data which are to be filled at the time of submitting the application. After performing such tests, model can determine whether the loan approved to the person is safe or not basically about the loan approval on the basis of the various training data sets.

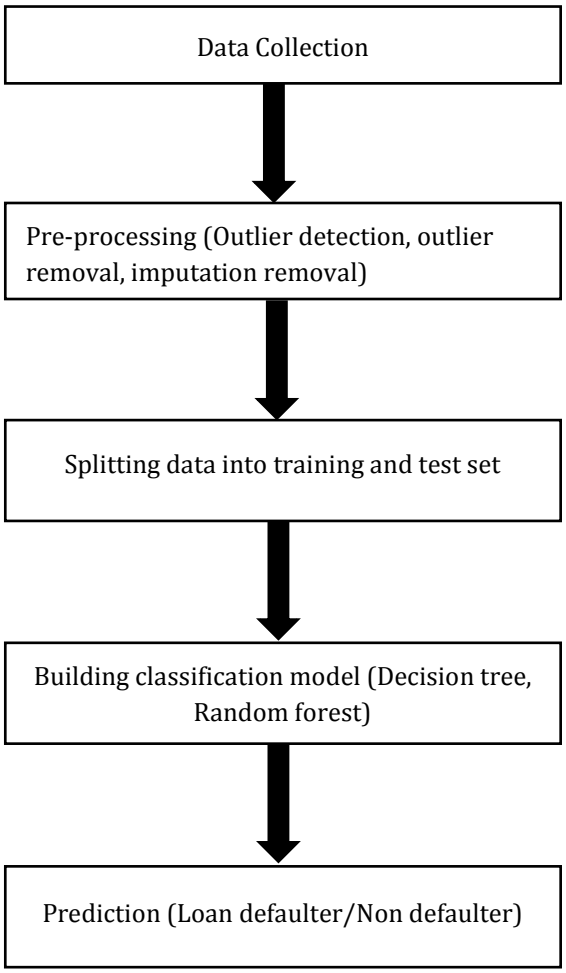



**Fig-1:** Chronology of Data

The diagram above gives us an outline on how data is used in this machine learning process or model.

Basically, it is divided into four parts in which we use data to predict the outcome of the whole process. First, we use training data set to train our model. After the model is trained, then we test it with unknown examples from the same scenario.

Another process that we use before testing and training data is data pre-processing. In data pre-processing we remove all sorts of values that can cause an error like redundant values, incomplete values, missing data, etc.

## 3. LOAN PREDICTION METHODOLOGY

The diagram 2 represents the working of our model. It basically gives us a rough idea on how the loan prediction system works. After collecting data, we use feature selection process on data. Feature selection can be defined as a process of reducing number of input variables when we develop a predictive model.

Feature selection is divided into two parts i.e. supervised method and unsupervised method. Supervised method is divided into three parts which are wrapper, filter and intrinsic. In supervised method we use target variable to remove discrepancies in data. While in unsupervised method we do not use target variable to remove discrepancies. Unsupervised method uses the process of correlation.

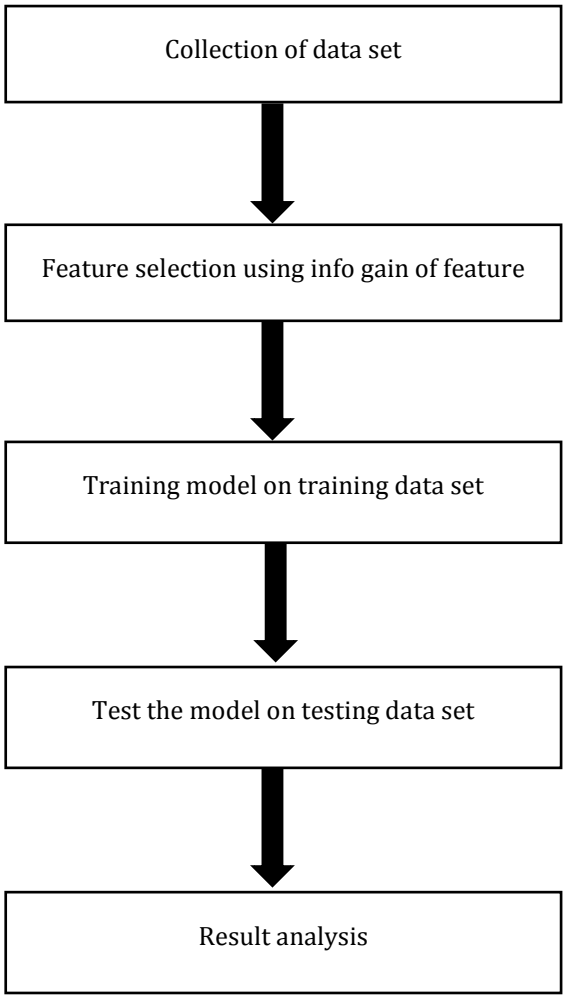



**Fig-2:** Loan Prediction Methodology

## 4. WORKING OF THE MODEL

We have represented the working of the model through a use case diagram. The figure below represents the attributes, process of the model that we have built.

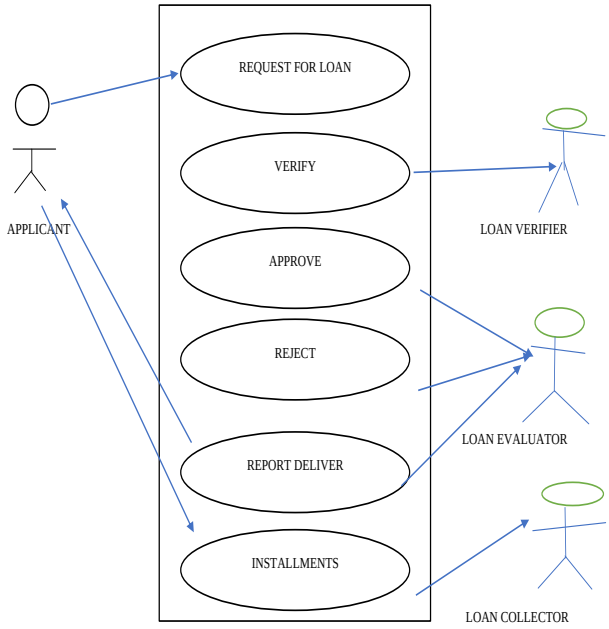



**Fig-3:** Use case diagram

**Table-2:** Use case diagram variable and description

| | |
|---|---|
| **Actor** | Applicant, Loan Verifier, Loan Evaluator, Loan Collector |
| **Description** | An applicant requests for a loan. After request loan verifier verified its document and transfer to loan evaluator may approve or reject the loan. |
| **Data** | Applicant personal information and its documents. |
| **Stimulus** | User command issue by online loan and application. |
| **Response** | Loan may be approved or may be rejected. |
| **Comments** | Improve installment policy. |

## 5. EXPLORATORY DATA ANALYSIS

1. The one whose salary is more can have a greater chance of loan approval.
2. The one who is graduate has a better chance of loan approval.
3. Married people would have an upper hand than unmarried people for loan approval.
4. The applicant who has a smaller number of dependents have a high probability for loan approval.
5. The lesser the loan amounts the higher the chance for getting loan.

## 6. Model used for training and testing

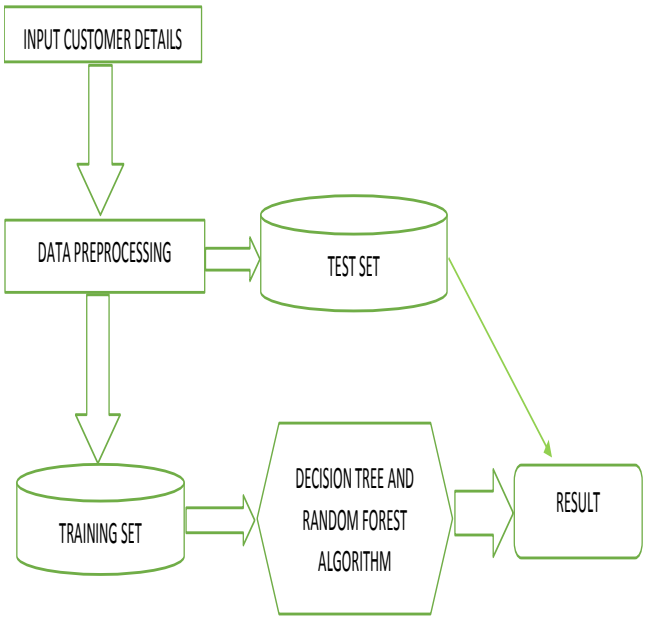



**Fig- 4:** Training and testing model

## 7. MACHINE LEARNING METHODS

Two machine learning classification models are used for the prediction of application that can be used in android applications. These models can also be accessed in the open source software R, which is licensed under GNU GPL. The brief description of each model is explained below.

### 7.1 Decision tree

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin toss comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

This model is an extension of C4.5 classification algorithms. We experimented with J48 Decision Tree classifier which is an implementation of C4.5 Decision Tree. In case of this classifier, the lower the confidence factor, the more pruning is done. For this we have used different confidence factors and analysed them with higher confidence factor and with the increase of confidence factor the accuracy has increased in each case. With the confidence factor of 0.15 the best accuracy is 62.12% and with a confidence factor of 0.25 it is 63.39%. It means that when less pruning is done the accuracy improves.

### 7.2 Random forest

Random forest or random decision forests are an ensemble learning method used for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

We have done several trials with Random Forest with different parameters: executions with supervised and unsupervised discretization's (equal-frequency and equal-width), with all attributes. In the experiments without attribute selection the best result was 85.75% and it was achieved with unsupervised equal-frequency 5 bins discretization with 450 trees and seed equal to 4.

**Table-3:** Parameter setting for machine learning models

| Model | Parameter Setting |
|---|---|
| Decision Tree | Min Split=20, Max Depth=30, Min Bucket=7 |
| Random Forest | Number of trees=450, number of variables=8 |

## 8. CONCLUSIONS

The main purpose of the paper is to classify and analyze the nature of the loan applicants. From a proper analysis of available data and constraints of the banking sector, it can be concluded that by keeping safety in mind that this product is much effective or highly efficient. This application is operating efficiently and fulfilling all the major requirements of Banker. Although the application is flexible with various systems and can be plugged effectively.

This paper work can be extended to higher level in future so the software could have some better changes to make it more reliable, secure, and accurate. Thus, the system is trained with a present data sets which may be older in future so it can also take part in new testing to be made such as to pass new test cases.

There have been numbers cases of computer glitches, errors in content and most important weight of features is fixed in automated prediction system. So, in the near future the so – called software could be made more secure, reliable and dynamic weight adjustment. In near future this module of prediction can be integrated with the module of automated processing system.

## REFERENCES

[1] J. R. Quinlan. Induction of Decision Tree. Machine Learning, Vol. 1, No. 1. pp. 81-106., 1086.

[2] A. Goyal and R. Kaur, "A survey on Ensemble Model for Loan Prediction", International Journal of Engineering Trends and Applications (IJETA), vol. 3(1), pp. 32-37, 2016.

[3] G. Shaath, "Credit Risk Analysis and Prediction Modelling of Bank Loans Using R".

[4] A. Goyal and R. Kaur, "Accuracy Prediction for Loan Risk Using Machine Learning Models".

[5] Hsieh, N. C., & Hung, L. P. (2010). A data driven ensemble classifier for credit scoring analysis. Expert systems with Applications, 37(1), 534-545.

[6] https://en.wikipedia.org/wiki/Exploratory_data_analysis

[7] https://www.experian.com/blogs/ask-experian/credit-education/score- basics/what-is-a-good-credit-score/

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

# PREDICTING LOAN APPROVAL USING ML

**Nikhil Bansode[*1], Adarsh Verma[*2], Abhishek Sharma[*3],**
**Varsha Bhole[*4]**

[*1,2,3,4]University Of Mumbai Department Of Information Technology, A.C Patil College Of Engineering, Navi Mumbai, India.

## ABSTRACT

Taking out loans from banks has become very common in today's world. Banks' main business is lending money. The primary source of benefit is the interest on the loan. However, because the bank has limited funds to distribute to a limited number of people, determining who the loan can be given to and who would be a better choice for the bank is standard procedure. Credit firms issue a loan after a lengthy period of authentication and confirmation. They are also concerned about whether the borrower will be able to repay the loan without difficulty. Many researchers have been exploring systems for determining loan acceptance in recent years. Machine learning can bring an additional reliable predictive modeling method to the banking business, which is still needed. The primary aim of this paper is to determine if a loan granted to some organization or a specific person would be accepted.

**Keywords:** Classification, Exploratory Data Analysis, Loan, Loan Approval, Machine Learning, Prediction, Python.

## I. INTRODUCTION

For banking institutions, loan approval is a crucial move. The loan applications were either accepted or rejected by the system. Loan recovery is a significant contributor to a bank's financial statements. Because data in banks is growing at such a rapid pace today, bankers must analyze a person's data before approving a loan. It is extremely difficult to predict whether the customer will be able to pay back the loan. Methodologies of Machine Learning (ML) are highly beneficial for determining what will happen when dealing with massive amounts of information. Machine learning assists specifically data learning from its own experiences, as well as data prediction and decision-making. The Python programming language is chosen in our research since it comes with all of the essential tools and libraries. Python is one of the most commonly used and preferred languages in Machine learning and artificial intelligence. As data is the most valuable resource on the globe, it has sparked a revolution in computer science. Machine Learning algorithms have provided numerous data analysis solutions. We would use five machine learning algorithms to predict customer loan approval: the Decision Tree algorithm, the Random Forest algorithm, the Logistic Regression algorithm, SVM, and K Neighbors algorithm. Our primary goal is to use machine learning concepts to calculate a customer's loan status and predict an immediate and precise outcome that assists the lender in analyzing the situation, providing better services, and reducing risk by selecting the appropriate person, saving the lender time and money. We would also test various machine learning algorithms and select the best among them.

The following is a breakdown of the paper's structure: An overview of relevant literature surveys on Research articles on Loan Prediction is included in Section II. Data Description can be found in Section III. The study's methodology for generating results is covered in Section IV. The research results are in Section V. Section VI brings the paper to a conclusion and the future work on this project.

## II. LITERATURE SURVEY

Rajiv Kumar and Vinod Jain, in their research paper [1] used the Python programming language to implement the logistic tree, decision tree, and random forest algorithms. They chose the Decision tree method as the most efficient after comparing the evaluation of three kinds of machine learning approaches in terms of prediction accuracy. They didn't fill in the blanks or properly categorize the data however, this can be fixed by filling in the blanks and properly categorizing the data.

Pidikiti Supriya and Myneedi Pavani, in their research paper [2] claim to have pre-processed the data to remove inconsistencies in the dataset. They've also compiled a list of Correlating Characteristics that were found to make people more likely to repay their debts. To divide the dataset into training and testing operations, the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

80:20 rule was used. The Python platform's corplot and boxplot are used to find the correlation between attributes. However, they haven't utilized any other method to compare accuracy results other than a decision tree. This may be avoided by training datasets with multiple algorithms and comparing their efficiency.

Kumar Arun and Garg Ishan, in their research paper [3] tested a total of six different machine learning approaches, including neural networks, support vector machines, random forests, decision trees, linear models, and Adaboost. There are four sections to this study. (i) Gathering of data (ii) Model evaluation using ML on the collected information (iii) System training using the most feasible model (iv) After the system has been trained on the most promising model, it is put to the test. R programming language was used to create this system. They didn't represent the data results for easier comprehension and comparison, but this problem can be solved by offering data visualization in the form of graphs or other matrix forms.

Authors in [4]. Initially, the information was cleansed. The next step was exploratory data analysis and feature engineering. They had done visualization through graphs. For loan prediction, four models are used. Decision Tree, Naive Bayes, Support Vector Machines, and Logistic Regression methods are the four methods. They determined confidently showing the Naive Bayes model is very capable of delivering superior results to other models after thoroughly studying positive attributes and constraints.

Authors in [5] said a set of data was obtained from the banking sector. The data set is in the ARFF (Attribute-Relation File Format) format, which Weka understands. They used exploratory data analysis to solve the challenge of granting or rejecting loan requests, as well as short-term loan projection. In their research, they did an exploratory data analysis. For prediction, two machine learning classification models are used Decision Tree and Random Forest. In their analysis, they chose the random forest method.

## III. DATA DESCRIPTION

We have got the loan data set through Kaggle [14]. The redundant and identical entries were deleted once the dataset was normalized. There is a chance that the data received possibly involve some null values, which could cause inconsistencies. Data must have been pre-processed to boost the algorithm's efficiency. Outliers must be eliminated, and variable conversion must be performed. The dataset gathered for forecasting loan default customers is divided into two groups: training and testing. Our data set includes a total of 13 columns. The response variable is Loan Status, and the remaining variables/factors determine whether the loan would be approved or not.

The characteristics are as follows:

**Table 1.** Loan Prediction Parameters

| Variable | Description | Category | Type |
|---|---|---|---|
| Loan_ID | Loan ID is unique | Qualitative | Integer |
| Gender | Man/ Woman | Categorical | Character |
| Married | Applicant married status (Y/N) | Categorical | Character |
| Dependents | Dependents count | Qualitative | Integer |
| Education | Education of the applicant | Categorical | String |
| Self_Employed | Self-employed person (Y/N) | Categorical | Character |
| ApplicantIncome | Applicants' earnings | Qualitative | Integer |
| CoapplicantIncome | Co-applicant's earnings | Qualitative | Integer |
| LoanAmount | Amount of the loan in thousands | Qualitative | Integer |
| Loan_Amount_Term | The loan's duration in months | Qualitative | Integer |
| Credit_History | Credit history complies with rules | Qualitative | Integer |
| Property_Area | Rural/Urban/Semi-Urban | Categorical | String |
| Loan_Status | Approval of the loan (Y/N) | Categorical | Character |

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

## IV. METHODOLOGY

**WORKING OF THE MODEL**

Based on the data provided by the borrower, an organization must automate the loan qualifying method (in real-time). Data such as Loan Amount, Gender, Marital Status, Income, Credit History, Education, Number of Dependents, and a few other details while completing a request form. As shown in Table 1. To make things simple, they created a system that allows them to identify types of applicants, who are qualified for a loan amount and approach them specifically. Since we need to classify everything before determining if the loan status is Yes or No, therefore this is considered as a classification issue. The system can quickly determine if a loan application is likely to be granted or rejected. Figure. 1, shows the working of the proposed model step by step.

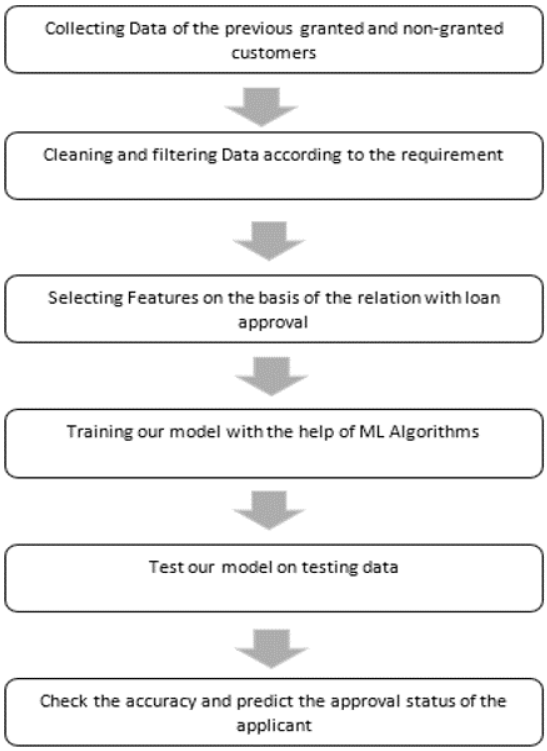



**Figure 1:** Proposed methodology

**ARCHITECTURE OF PROPOSED MODEL**

Our project makes use of a variety of algorithms to help us achieve a precise result. The Python programming language, which is among the most often used and popular languages in AI and ML because it comes with all of the necessary tools and libraries has been used in our project. It has several libraries, like pandas for the filtering process, matplotlib for plotting the data, data visualization, and exploratory data analysis. We have also used sklearn which is Scikit-learn which includes several clustering, regression, and classification algorithms that are commonly used in AI and machine learning. Numpy is used to deal with the multidimensional array and data structures.

Seaborn library has been used for data visualization. The model then applies this technique to pre-defined data set including all the information about our customers. In a linear pattern, the algorithms are executed one after the other. The data is then analyzed, segregated, and provided into the model to train it. As shown in Figure. 2. After each algorithm, the precision rate is displayed. We have trained our model with many algorithms to get a precise result. The Random Forest Algorithm, Decision Tree Algorithm, Logistic Regression Algorithm, SVM, and K Neighbors algorithm will all be used, with a 70% training set and a 30% testing set. We have discovered that logic regression, decision trees, and random forests have superior precision. Following the testing procedure, the model predicts if the current candidate based on the conclusion is a good candidate for getting a loan acceptance, it draws from the training data sets. As a result, the better we are in determining the capable borrower, the more beneficial it is to the organization.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

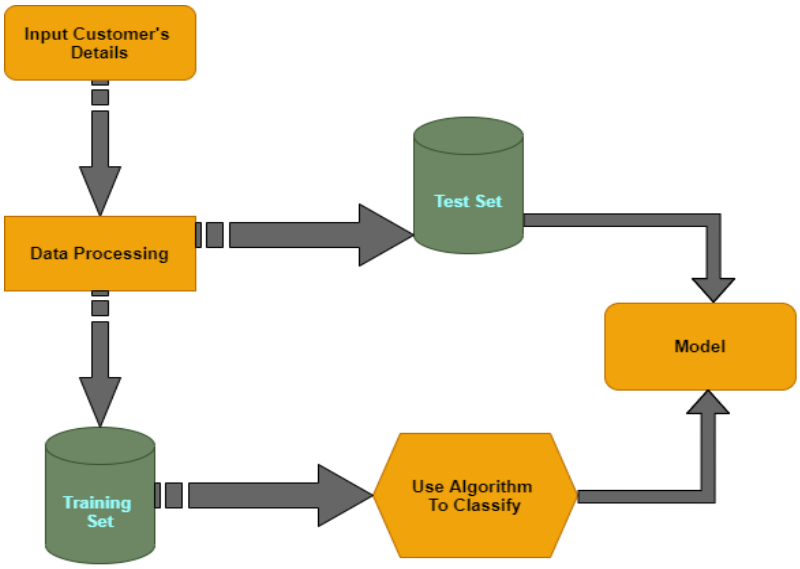



**Figure 2:** The proposed model's architecture.

This design is based on the category of Supervised learning, with the classification problem. A confusion matrix is used to visualize the data. It displays the results of the prediction model that was created. The model includes a variety of graphs and diagrams that aid in visualizing the model's data flow. Each graph illustrates a collection of procedures carried out on the data set.

**ALGORITHMS**

- Decision Tree- It's a supervised non-parametric machine learning technique. It can be used for regression as well as classification. On both input and output variables, it works in categorical and continuous modes. All attributes or features must be discretized by the basic algorithm of the decision tree. The most detailed features are chosen for inclusion in the feature set. IF-THEN rules can be used to interpret the knowledge represented in a decision tree.
- Logistic Regression- The LR approach is a popular way to solve binary classification problems. Binary LR makes use of binary dependent variables. The variables used should be relevant. Many of the model's independent variables should be self-contained. The sample size for LR should be large.
- Random Forest- It is most commonly used in classification and regression analysis disciplines. During training, the RF algorithm builds a large number of decision trees. An approach to characterization involves constructing an enormous number of Decision trees over the duration and obtaining the class that would be the mode of the classes produced by the independent tree.
- SVM- Support vector machines (SVM) are a class of supervised learning methods used for classification, regression, and outlier detection. A discriminative classifier is another name for an SVM classifier. In high-dimensional spaces, it is beneficial. When the number of dimensions exceeds the number of samples, the method serves its purpose. It is well-known for its kernel trick for dealing with nonlinear input spaces. SVM discovers an optimal hyperplane, which assists in the classification of new data points.
- K Neighbors- K-Nearest Neighbour is a method for Supervised Learning. It is a non-parametric algorithm, which appears to mean it makes no assumptions about the underlying data. It assumes a correlation between the new case/data and existing cases and places the new case in the category that is most similar to the existing categories.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

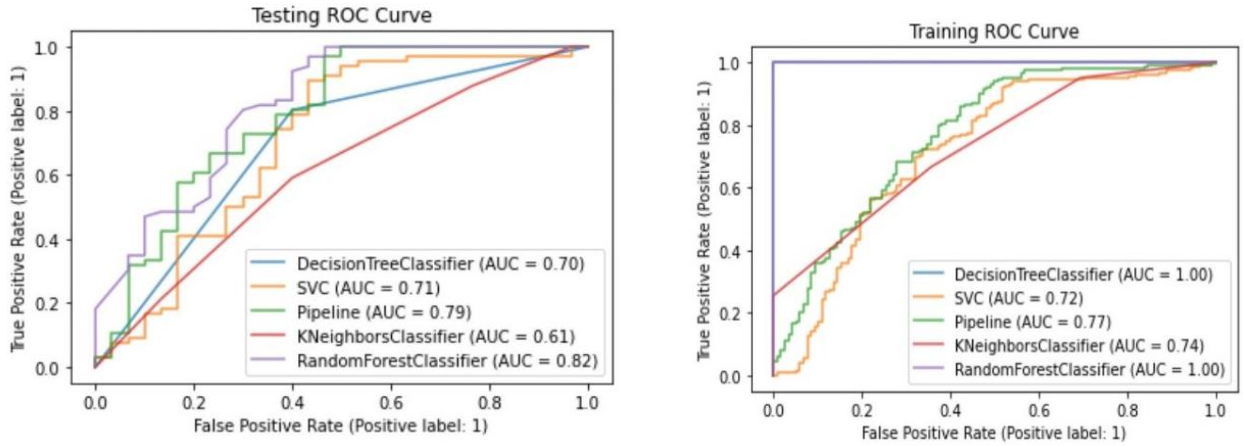



**Figure 3:** ROC Curves.

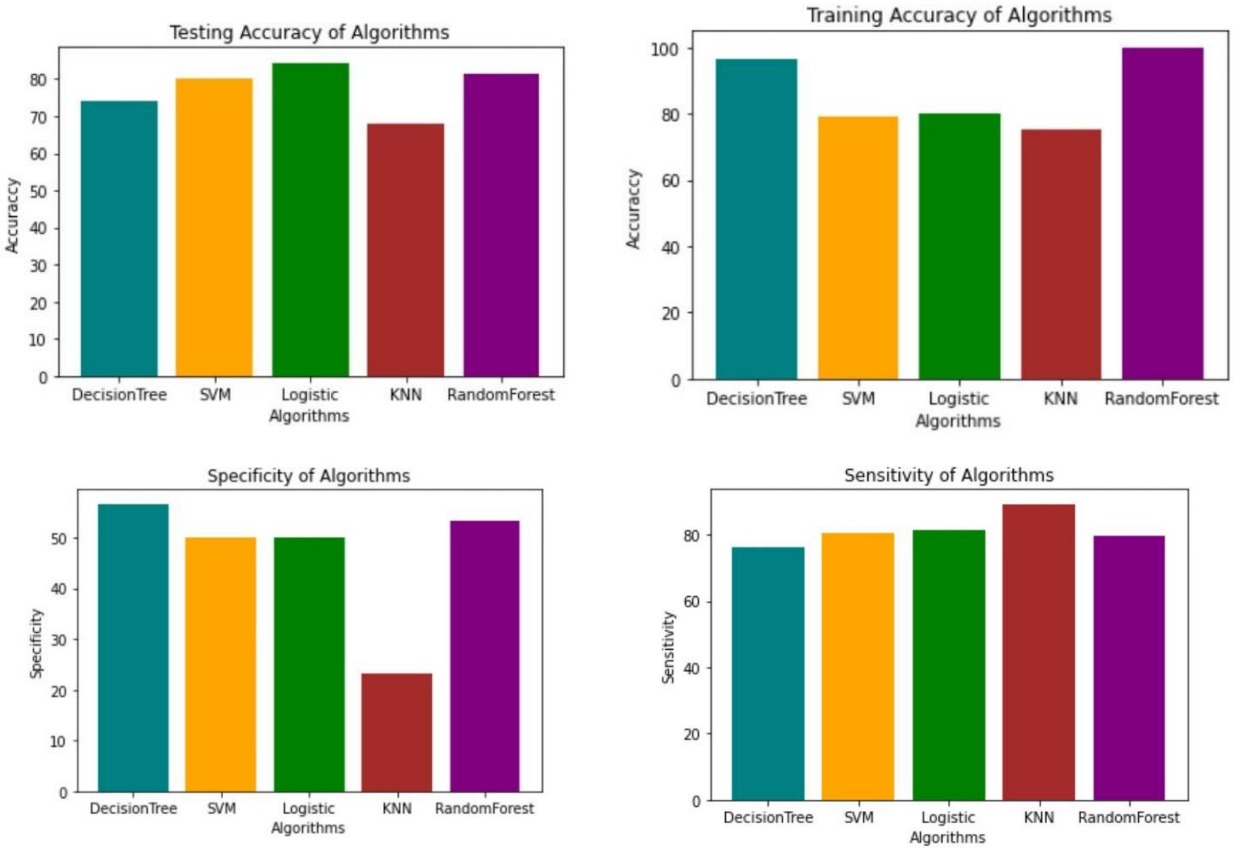



**Figure 4:** Algorithms Bar Graph Comparison.

## V. SCOPE

The scope of the project includes:

- Assists the lender in analyzing the situation.
- Gives better services for use.
- Reduce the risk factor by choosing the right person.
- Save time and money for the lender.

## VI. RESULTS AND DISCUSSION

To acquire an accurate result, we used a variety of strategies to train our model. With a 30 percent testing set and 70 percent training set, the Logistic Regression Algorithm, the Decision Tree Algorithm, Random Forest Algorithm, SVM, and K Neighbors were implemented. Logistic regression, on the other hand, has the highest accuracy of all the algorithms.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

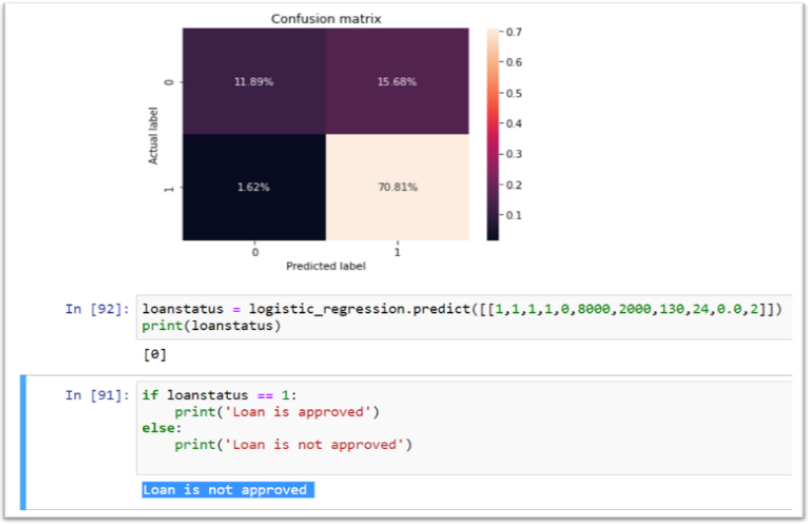



**Figure 5:** Rejected state.

In Figure. 5, after giving the input we can see the loan status as rejected. As the customer was not eligible.


**Figure 6:** Accepted state.

In Figure. 6, after giving the input we can see the loan status as accepted. Because the customer was an eligible candidate. Our model has a prediction accuracy of 84.376 %, indicating that it can predict defaulters. A heatmap was used to analyze everything. The correlation matrix is represented visually as a heatmap. It greatly helps in the speedy identification and verification of relationships between columns.

**Table 2:** Confusion Matrix

| | Predicted No: | Predicted Yes: | |
|---|---|---|---|
| Actual No: | 22 | 29 | 51 |
| Actual Yes: | 3 | 131 | 134 |
| | 25 | 160 | |

**Table 3:** Accuracy Table

| **Algorithms** | **Accuracy** |
|---|---|
| Logistic regression | 84.376 % |
| Random forest | 82.293 % |
| SVC | 80.200 % |

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
( Peer-Reviewed, Open Access, Fully Refereed International Journal )
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

| | |
|---|---|
| K Neighbors | 76.700 % |
| Decision tree | 71.873 % |

Accuracy is just a part of the whole model. Other things are taken into consideration when training the model and analyzing the results. For example, the sensitivity and specificity of a particular algorithm are also evaluated from these two terms. These two play a crucial role in evaluating and plotting the ROC Curve.

$$Specificity = \frac{TN}{TN + FP} \quad Sensitivity = \frac{TP}{TP + FN}$$

To calculate the sensitivity and specificity, one should first evaluate True Positive, True Negative, False Positive, and False Negative. These values can be calculated from the confusion matrix, as shown in Table 2. And accuracy of every algorithm is mentioned in Table 3.

## VII. CONCLUSION

This system would be able to determine the status of the loan whether it would get approved or denied swiftly in real-time. Displays accuracy with various algorithms. We have compared the Logistic regression algorithm to two other algorithms, random forest, and decision tree Table III. However, of all the algorithms, Logistic regression has the highest accuracy. Also, it can fill the missing values of the datasets, treat categorical values, scalability problems, overfitting problems, and provide a good visualization of the data using a confusion matrix. Applicants who have a poor credit history are likely to be rejected, especially to the risk of not repaying the loan. Applicants with high income who request low-interest loans have a stronger chance of being accepted, which is logical because they have a strong chance to repay their debts. Few essential characteristics, such as marital status and gender, appear to be overlooked by the organization, but the number of dependents is taken into consideration. The libraries are used professionally and are sufficient for now because we chose the Python programming language, but many aspects require additional exploration. Many areas of our project are left unexplored and might be studied and explored further.

For further research, applicants' Age, past health records, as well as the type of occupation they have will be utilized to evaluate the ambiguity factor of paying debts, and possible defaults of corporate loans for businesses and startups can be forecasted. Another method could be developed to forecast defaulters on different types of loans as well. We used a medium-sized data set to train our model, which may have influenced the outcome; therefore, a big and well-defined data set is required for more accurate results. This paperwork could be expanded to a higher level in the future, allowing the software to be improved to make it more dependable, secure, and accurate. The system has been trained using current data sets that may become older in the future, allowing it to participate in fresh testing to pass new test cases.

## VIII. REFERENCES

[1] Rajiv Kumar, Vinod Jain, Prem Sagar Sharma- Prediction of Loan Approval using Machine Learning- International Journal of Advanced Science and Technology Vol. 28, No. 7, (2019).

[2] Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma- Loan Prediction by using Machine Learning. International Journal of Engineering and Techniques - Volume 5 Issue 2, Mar-Apr 2019

[3] Kumar Arun, Garg Ishan, Kaur Sanmeet- Loan Approval Prediction based on Machine Learning Approach- IOSR Journal of Computer Engineering, p-ISSN: 2278-8727 PP 18-21.

[4] E. Chandra Blessie, R. Rekha- Exploring the Machine Learning Algorithm for Prediction the Loan Sanctioning Process- (IJITEE) ISSN: 2278-3075, Volume-9 Issue-1, November 2019.

[5] Kshitiz Gautam, Arun Pratap Singh, Keshav Tyagi - Loan Prediction using Decision Tree and Random Forest- (IRJET) Volume: 07 Issue: 08 | Aug 2020.

[6] Sujoy Barua, Divya Gavandi- Swindle: Predicting the Probability of Loan Defaults using CatBoost Algorithm- ICCMC 2021.

[7] Bhoomi Patel, Harshal Patil, Jovita Hembram- Loan Default Forecasting using Data Mining- (INCET) Belgaum, India. Jun 2020.

[8] Sourav Kumar, Amit Kumar Goel- Prediction of Loan Approval using Machine Larning Technique- International Journal of Advanced Science and Technology Vol. 29, pp. 4152 – 4161 (2020).

[9] Soni PM, Varghese Paul- Algorithm For the Loan Credibility Prediction System-(IJRTE) Volume-8, June 2019.

[10] X. Francis Jency, V.P.Sumathi, Janani Shiva Sri-An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients, -(IJRTE) Vol. 7, pp. 176-179, No. 48, 2018.

[11] S. Vimala, K. C. Sharmili - Prediction of Loan Risk using NB and Support Vector Machinel, vol. 4, no. 2, pp. 110-113 (ICACT 2018).

[12] K. Ulaga Priya, S. Pushpa- Exploratory analysis on prediction of loan privilege for customers using random forest- International Journal of Engineering & Technology, 7 (2018) 339-341.

[13] Deepak Ishwar Gouda, Ashok Kumar, Anil Manjunatha Madivala- Loan Approval Prediction Based On Machine Learning- e-ISSN: 2395-0056 Volume: 08 Issue: 01 | Jan 2021 (IRJET).

[14] https://www.kaggle.com/burak3ergun/loan-data-set

[15] A. Ibrahim, R. L., M. M., R. O. and G. A., "Comparison of the CatBoost Classifier with other Machine Learning Methods", International Journal of Advanced Computer Science and Applications, vol. 11, no. 11, 2020. Available: 10.14569/ijacsa.2020.0111190.

[16] J. Tejaswini, M. Kavya, D. Ramya, S. Triveni and V. Rao Maddumala, "ACCURATE LOAN APPROVAL PREDICTION BASED ON MACHINE LEARNING APPROACH", Journal of Engineering Sciences, vol. 11, no. 0377-9254, 2020.

[17] A. Gupta, V. Pant, S. Kumar and P. Kumar Bansal, "Bank Loan Prediction System using Machine Learning", Ieeexplore.ieee.org, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9336801. [Accessed: 08- Apr- 2022].

[18] A. Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval", Ieeexplore.ieee.org, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/8203946. [Accessed: 08- Apr- 2022].

[19] M. Ahmad Sheikh, A. Kumar Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm", Ieeexplore.ieee.org, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9155614/authors#authors. [Accessed: 08- Apr- 2022].

[20] V. Singh, A. Yadav and R. Awasthi, "Prediction of Modernized Loan Approval System Based on Machine Learning Approach", Ieeexplore.ieee.org, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9498475/authors#authors. [Accessed: 08- Apr- 2022].