Using Decision Trees and Random Forests to Predict Loans

# 1st Yash Mistry

*Devang Patel Institute of Advance Technology and Research Charotar University of Science and Technology*

D19DIT081

# 2nd Kushagra Patel

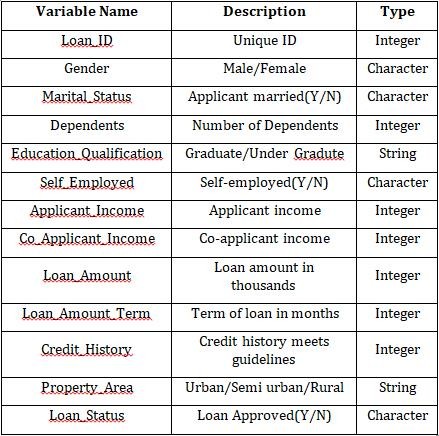
*Devang Patel Institute of Advance Technology and Research Charotar University of Science and Technology*

18DIT053

***Abstract*—Once a year, the amount of people or organisations applying for a loan in India are multiplied. The bank worker must be expected to expend a lot of energy to analyse or anticipate whether the client would reimburse the credit limit (defaulter or non-defaulter) within the specified time frame. The aim of this paper is to investigate the character, history, and quality of the shopper who is seeking a loan. We frequently use preliminary information data analysis to have an influence on the choice to approve or reject a loan request, and to make an easy loan forecast. The aim of this research is to establish whether or not a loan given to a specific independent organisation would be authorized.**

***Index Terms*—Loan, Prediction, Machine Learning, Training, Testing.**

Format) format. The name, types of attributes, values, and the data itself are all tags in an ARFF format. In this article, we use 12 attributes such as gender, legal status, qualification, salary, and so on. The table below shows the knowledge processing method we used:

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

1. INTRODUCTION

The expression banking is explained in relation to receiving security cash that has now been deposited by a party or an entity. It also includes giving money to people and organi- zations that must be paid back in a certain length of days without failing. Banking is also a sector that is controlled in most countries because it is a vital believe determinant of the country’s monetary stability. The primary aim of the banking industry is to ensure that their investments are in safe hands where there are less chances of loss. Several banks and money companies now accept loans after a nerve-wracking, lengthy, and exhausting screening process, but there is still no guarantee if the selected applicant is legitimate or not, or, in other words, whether he will be able to repay the loan with interest within the specified time frame. The purpose of the loan is usually something that the client wants. Open-ended and closed-ended loans are the two types of loans. Examples of open-end loans are credit cards and a home equity line of credit (HELOC). With each interest, the amount owed on a closed-ended loan decreases. After an installment, the amount is decreased. In other words, it’s a legal concept that the creditor can’t alter. Closed-ended loans include personal loans, mortgages, car payments, EMIs, and student loans, to name a few. A secured loan, also known as a collateral loan, is one that is secured by an asset. Houses, automobiles, and other types of property.

1. DATASET

The banking industry offers a wealth of information. For data processing, Weka uses the ARFF (Attribute-Relation File

The training data collection is now added to the machine learning algorithm, and the model is trained using established examples from this data set. The most recent applicants’ submissions will be used as assessment results, and will be filled in when the application is submitted. Following the completion of these assessments, the model would determine whether or not the loan issued to the applicant is secure, based on the loan authorization.

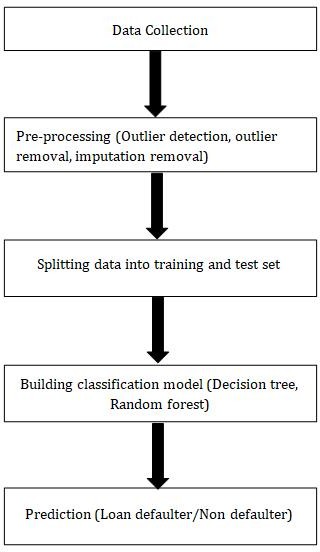
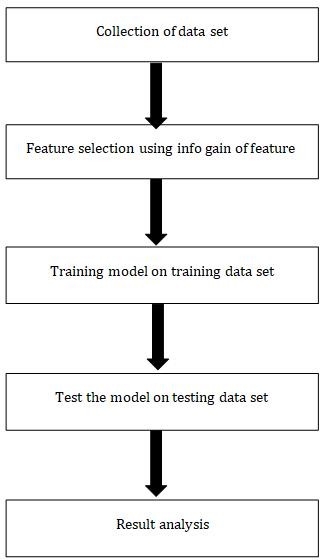


Fig-1: Data in chronological order.

Data is used in this machine learning model or process, as seen in the diagram above. It’s broken down into four parts, each of which uses data to forecast the operation’s outcome. To start, we use a training data set to train our model. After conditioning the model, we test it with uncertain examples from a similar situation. Another technique we use before analysing and training data is data pre-processing. ”We have a tendency to eliminate all forms of values that may lead to an error in information pre-processing, such as redundant values, insufficient values, missing data, and so on.” supervised and unsupervised feature selection approaches are the two forms of feature selection methods. The three parts of the supervised method are wrapper, filter, and intrinsic. In the supervised method, we use the goal variable to minimise data anomalies. The unsupervised solution to eliminating discrepancies would not involve the target variable. In the unsupervised method, the association mechanism is used.

Fig-2: Loan Prediction Methodology

1. EXPLORATORY DATA ANALYSIS
2. The person with the higher salary has a better chance of being approved.
3. Those who have completed their education have a higher chance of being approved.
4. Married people would have a better hand in obtaining approval than single people.
5. The person with the smallest number of dependents has a high chance of being approved.
6. The higher the chances of getting a loan, the smaller the loan sum.

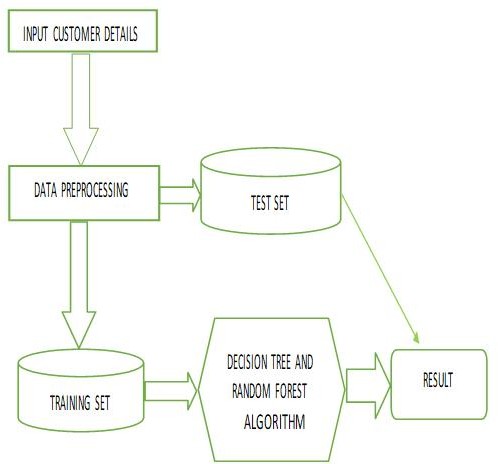
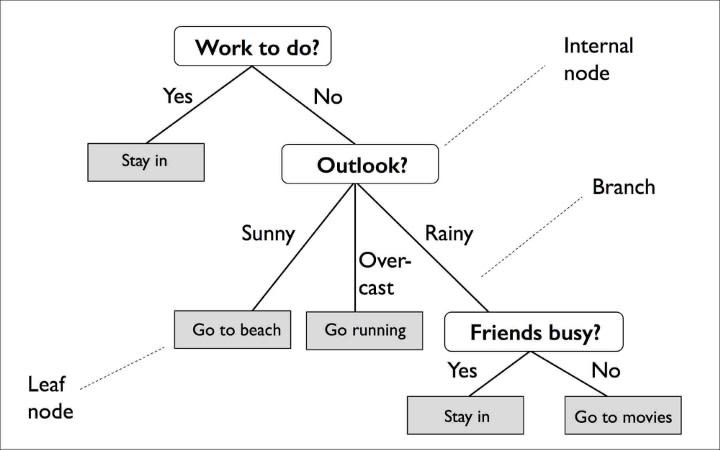


Fig-3: Training and testing model

1. MACHINE LEARNING METHODS

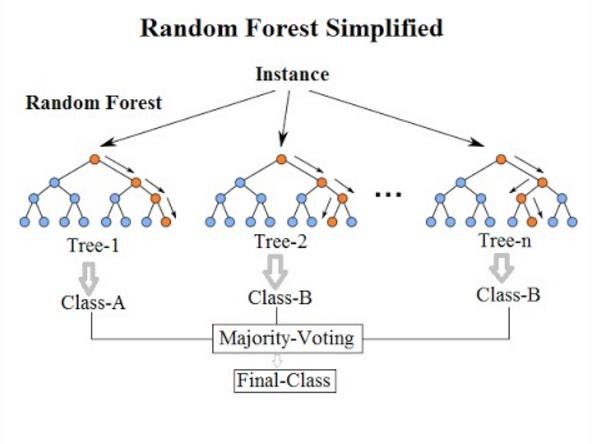
Two machine learning classification models are used to estimate the programmes that will be used in Android appli- cations. These templates can also be accessed using the open source programme R, which is licenced under the GNU GPL. A brief description of each model is given below.

1. DECISION TREE



A decision tree is a flowchart-like structure in which each internal node represents an attribute ”test,” each branch rep- resents the test’s outcome, and each leaf node represents a community mark (for example, whether a coin toss comes up heads or tails) (decision made after all attributes have been computed). The C4.5 classification algorithms are improved in this model. ”We used the J48 call Tree classifier, which is built on the C4.5 call Tree,” says the researcher. The lower the arrogance factor, the more pruning is completed in this classifier. For this, we used various confidence factors and analysed them with higher confidence factors, with the accuracy increasing in each case as the confidence factor increased. The easiest precision is 78 percent with a confidence factor of 0.77. It means that as the amount of pruning done decreases, the accuracy increases.

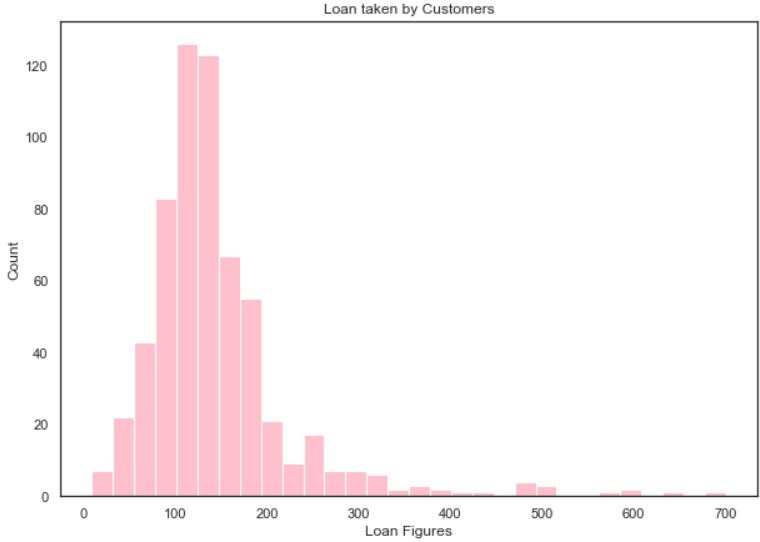
1. RANDOM FOREST



Random forest, also known as random call forests, is an ensemble learning method for grouping, regression, and alternative tasks that operates by training a large number of call trees and then predicting the class that is the mode of the categories or the mean predictor of the individual trees. The random forest chooses the supreme call, which is supported by the majority of the trees.

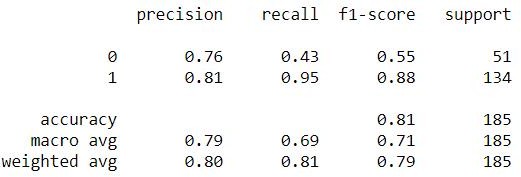
We used Random Forest to run many experiments, including supervised and unsupervised discretizations (equal-frequency and equal-width), as well as all attributes. In the tests without attribute filtering, the easiest result was 80.50 percent.

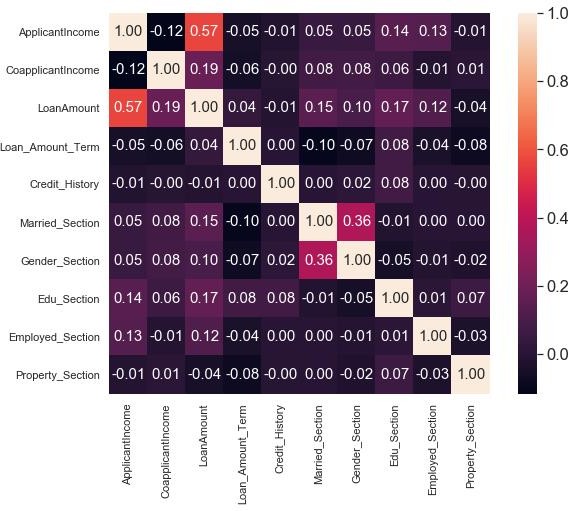
1. IMPLEMENTATION
2. *HISTOGRAM REPRESENTATION*



1. *CORRELATION MATRIX*

A table is a matrix that displays correlation coefficients among variables. The association between two variables is seen in every cell of the graph. A matrix is used to summarise data, as an input to several advanced analyses, and as a diagnostic tool for advanced analyses.



Fig- 4: Correlation Matrix

1. *TRAIN-TEST SPLIT*

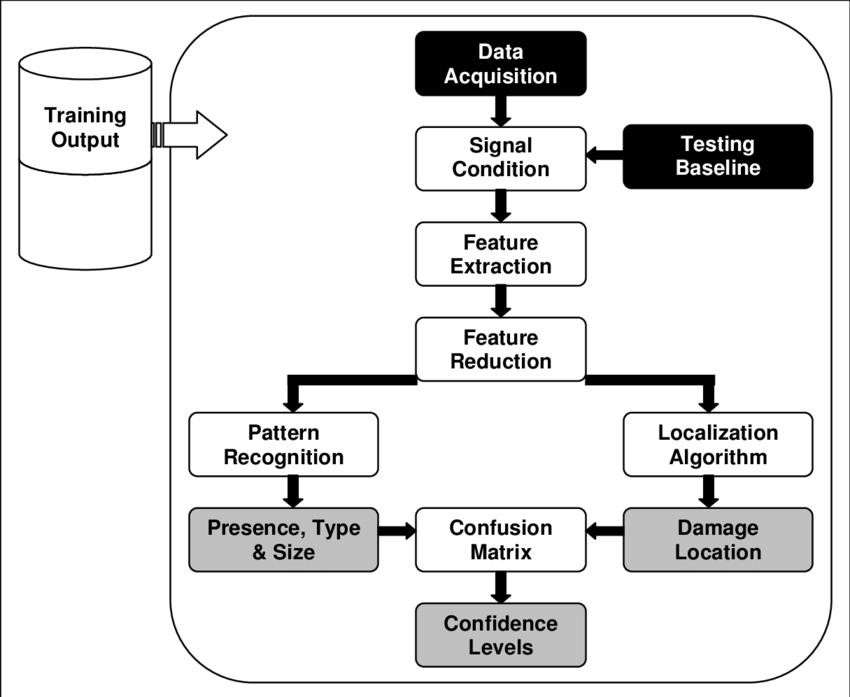
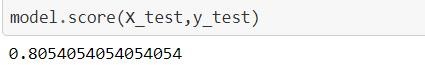
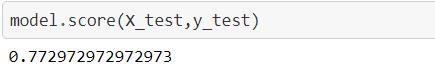
For measuring the success of a supervised machine learning algorithm, the train-test split protocol is used. You divided the data into two groups. The ”train” subset will be used to suit the machine learning model, while the ”test” subset will be used to test the model’s output. This is a crucial distinction to make. You shouldn’t train and compare the model with the same data so it would be cheating and the model would have already seen the data it’s supposed to predict. Using a train- test split protocol prevents overfitting, which occurs when a model has near-perfect precision when dealing with training results.

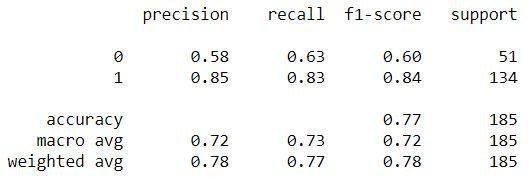
Fig-5: Train-Test

1. *RANDOM FOREST*

Accuracy in the testing model.

1. *DECISION TREE*





1. CONCLUSION

The key aim of the paper is to determine and assess the personalities of the loan applicants. Based on a detailed review of available evidence and banking sector restrictions, it is possible to assume that this product is highly successful or cost-effective when protection is taken into consideration. This app works admirably and satisfies all of Banker’s key criteria. Despite the fact that the programme is adaptable to a range of applications and can be easily disabled, it is also a security risk.

In the future, this paper work should be extended, and the software kit should be revised to make it more reliable, robust, and correct. As a result, the machine is trained with gift data sets that could become outdated in the future, enabling it to participate in new training, such as passing new test cases.

There have been many incidents of laptop failures and content failures, and the most significant weight of options is based on a machine-driven prediction system. As a result, a software package for safer, more reliable, and dynamic weight change could be developed in the not-too-distant future. In the near future, this prediction module may be merged with the machine-driven process system module.

1. REFERENCES [https://www.myperfectwords.com/blog/research-paper](http://www.myperfectwords.com/blog/research-paper-)-

guide/research-paper-example https://iopscience.iop.org/article/10.1088/1757-

899X/1022/1/012042/pdf https://en.wikipedia.org/wiki/Exploratory*dataanalysis* [https://www.researchgate.net/publication/259235118](http://www.researchgate.net/publication/259235118RandomF)*RandomF orestsandD* https://medium.com/swlh/using-decision-trees-and-random-

forests-to-predict-lending-profitability-c208b8106cb5 https://ieeexplore.ieee.org/document/9155614 [https://www.iosrjournals.or](http://www.iosrjournals.org/iosr-jce/papers/conf.15013/Volume)g/iosr[-jce/papers/conf.15013/](http://www.iosrjournals.org/iosr-jce/papers/conf.15013/Volume)V[olume](http://www.iosrjournals.org/iosr-jce/papers/conf.15013/Volume) [https://www.researchgate.net/publication/335966806](http://www.researchgate.net/publication/335966806AnEmpiricalStudyonL)*AnEmpir*[*ical*](http://www.researchgate.net/publication/335966806AnEmpiricalStudyonL)*S*[*tudyonL*](http://www.researchgate.net/publication/335966806AnEmpiricalStudyonL)<http://www.ijetajournal.org/volume-3/issue-1/IJETA-V3I1P6.pdf> https://aircconline.com/mlaij/V3N1/3116mlaij01.pdf