# LOAN ELIGIBILITY PREDICTION

# A COMPARITIVE ANALYSIS

# REPORT

INT – 247


Name : **Nirujogi Satish Kumar**

Reg No. : **11903349**

Section **: K19KH**

Roll no. : **37**


Project Github Repository :- https://github.com/satish118/Loan-Eligibility-Data-Prediction/


LOVELY PROFESSIONAL UNIVERSITY

## ABSTRACT

To, Start a new venture or a business or purchase anything the cost is very high. So, people take loans for funds saving are not much enough for them, loans are easy funding and also easy to repay monthly installment. But here, the process of taking loans is very difficult and time consuming process, people needs to pass every stage of loan approval process it's completely based upon applicants eligibility as per banks terms of policy. So, here the Loan Eligibility prediction is introduced. The main aim of this project is to Compare the prediction for loan eligibility with different machine learning models. We applied four types of machine learning models which are Logistic Regression, K-Nearest Neighbours, Random Forest, Support Vector Machine(SVM). In this case K-Nearest Neighbours has most accuracy than other three models.


## TABLE OF CONTENTS

## INTRODUCTION

A Prediction model makes use of data mining, information and chance to forecast an outcome. every model has a few variables referred to as predictors that are possibly to persuade future outcomes. The facts that became amassed from various resources then a statistical version is made. it can use a easy linear equation or a sophisticated neural network mapped the usage of a complicated software program.

As greater statistics turns into to be had the version will become more delicate and the mistake decreases which means then it'll be capable of are expecting with the least hazard and ingesting as less time as it can.

The Prediction model allows the banks through minimizing the risk associated with the loan approval system and helps the applicant by means of reducing the time taken within the system. the main aim of the project is to evaluate the loan Prediction trends made implemented the usage of various machine learning algorithms and pick out the fine one with good accuracy out of them that may shorten the mortgage approval time and decrease the risk associated with it.

it's miles accomplished by predicting if the loan can be given to that man or woman on the idea of various parameters like credit score, income, age, marital status, gender, etc. The prediction model now not handiest allows the applicant but additionally enables the bank by way of minimizing the risk and lowering the wide variety of defaulters.

## LITREATURE REVIEW

[1] Li, S.T., Shiue, W., and M.H.Huang, "The evaluation of consumer loans using support vector machines.," Expert Systems with Applications, vol.30, no.4, 2006.

This real world dataset, which classifies credit applicants described by a attributes as good or bad credit risks, has been successfully used for credit scoring and evalution systems.

[2] Amira Hassan and Ajith Abraham, "Modeling Consumer Loan Default Prediction Using Ensemble Neural Networks", *International Coference on Computing Electrical and Electronic Engineering (ICCEEE)*, 2013.

It shows how the ensemble models works on consumer loan prediction models. It was briefly explained in it about usage of Ensemble Nueral networks like Random Forest.

[3] Archana Gahlaut, Tushar and Prince Kumar Singh, "Prediction analysis of risky credit using Data mining classification models", *28th International Conference on Computing Communication and Networking Technologies (ICCCNT),* 2017.

Analysing the credit scores of the consumer of loans using different types machine learning classification models for credit risk analyzation.

[4] M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020.

A breifiely explained way to how the loan approval prediction works using various machine learning models or algorithms.

[5] V. Singh, A. Yadav, R. Awasthi and G. N. Partheeban, "Prediction of Modernized Loan Approval System Based on Machine Learning Approach," 2021 International Conference on Intelligent Technologies (CONIT), 2021.

It algorithms like decision trees, binary classification and random forest also which are more useful giving best accuracy results for this loan approval prediction system.

## METHODOLGY

Data is the main component of machine learning to know. Predictive models use information for training which gives accurate correct outputs. without data we can't teach the model. Device gaining knowledge of involves building these models from statistics and uses them to are expecting new records. machine learning is a subset of artificial Intelligence. It gives system capability to learn wherein automatically learns and improves the performance without being explicitly programmed. The dataset used here is taken from Kaggle.

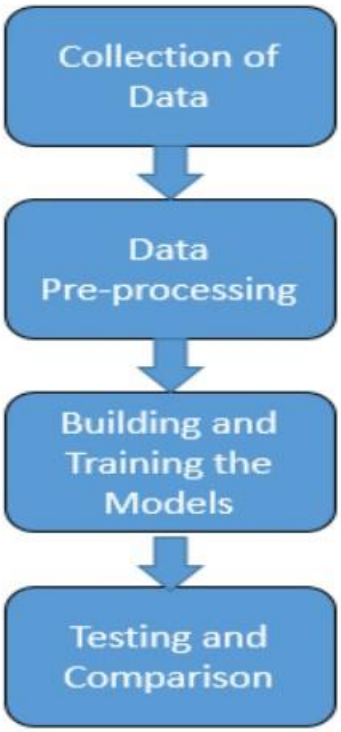



**Fig.1 - Methodolgy**

Our data has total of 13 various features and 614 rows which is collected from Kaggle the dataset have features like,

Loan_ID : it is an identification number for representing the loan application

Gender : which indicates whether the applicant is male or female

Married : which indicates whether the applicant is married or not married

Dependents : which shows how many dependents depending on that applicant

Education : which indicates whether the applicant is Graduted or not

Self_Employed : which indicates whether the applicant is self_emplyed or not

Applicants Income : which shows appliants income

Co-applicant Income : which shows the Co-Applicants income compared to the applicants income

Loan Amount : which indicates how much loan amount that applicants applied for

Loan Amount Term : which indicates the how much term does the applicants taking for.
Property Area : which indicates wether the applicants property is in urban or rural area's.

Loan_Status : which indicates whether the loan is approved or not as per above mentioned credentials.

## DATA PRE-PROCESSING

Data pre-processing is the process of cleaning our data set. Data is pre-processed in different phases. In first phase, the null values are filled by using the traditional mean and mode method.
In second phase, the data visualization is done by plotting different graphs between the attributes.
In the third phase, is engineered using the other features of the dataset and the correlation between the attributes is found using the different types of plots.
In the last phase, the categorical attributes are taken care of by using the Label Encoding Technique.

### Data Cleaning

The dataset needed to be cleaned and prepared before implementing into various machine learning algorithms. The values which are mentioned null values or missing are to be removed for the algorithms to run smoothly without any corruption. So, here every feature has to be red and remove the null values from that which is used to maintain the integrity of the datset.

### Relation between various attributes

the realtion between every attribute or feature makes the difference in output accuracy, so we have to aware about relation between every feature in statistically using correaltion**.** Here we compare the categorical – categorical variable and numerical – numerical variable, categorical – numerical variable and plot them using graphs.

The outliers are data point where they have extreme postions compared other one's which leads underfitting or overfitting of data. So, we have to remove them comparing the features where they have extreme positions.

In the categorical – numerical analysis we compare the features loan status as categorical and Applicant Income as numerical varibels have lot's of outliers and also the distribution is also positively skewed**.** Even in the co-applicant income has lot's of outliers.

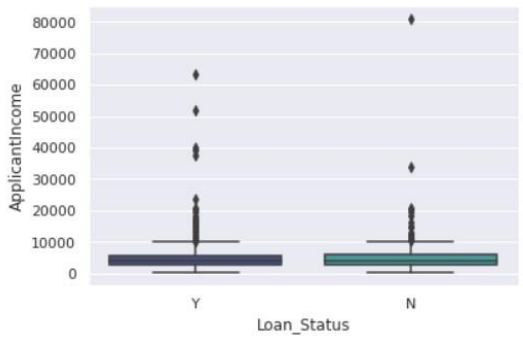



**Fig.2 – outliers in applicant income**

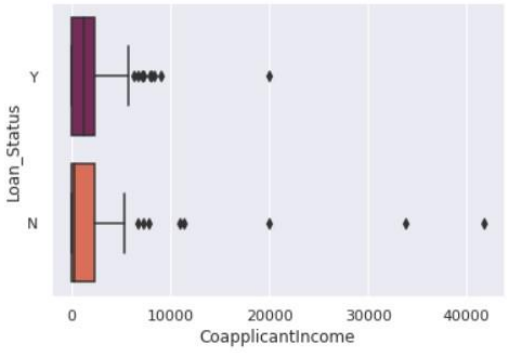



**Fig.3 – outliers in Co-applicant income**

By checking the correaltion between two features we can come to know about the linear association between them. Here, we are performing correaltion between the applicant income and co-applicant income.

There is negative correlation between Applicant income and Co-Applicant Income the pearson correlation is -0.11660 and the p-value is 1.46 from these we ecame to know that there is 95 percent confidence interval between both of them **.**

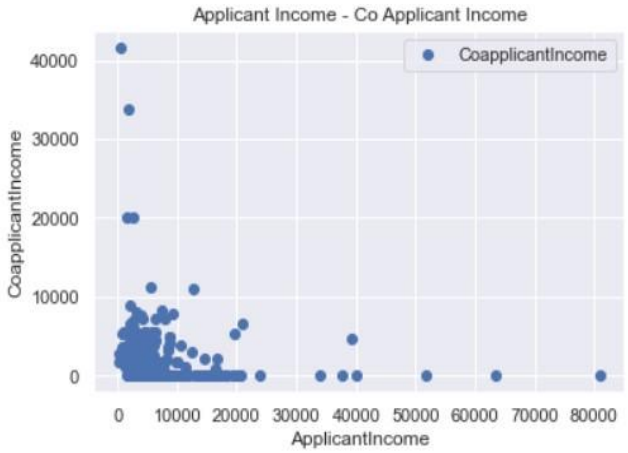



**Fig . 4 – Correlation between Applicant income & CoApplicant Income**

## Data Imputation

Imputation is a technique for substituting an estimated value for missing values in a dataset. In this section, the imputation will be performed for variables that have missing values.
Now, the variables with missing values have to filled with some values. So, here we have performed mode on the missing values in the data imputation form, the missing value also get filled with some data then it will work fine with machine learning models.

## ONE HOT ENCODING

The machine learning algorithms cannot perform good operations with categorical variables as good as with numerical variables.
Many machine learning algorithms cannot operate on label data directly. They require all input variables and output variables to be numeric. This means the categorical data must be converted into numerical form, then only the output variables can be used to perform in application.

Here, the categorical variables are Gender, married, education, self-employed, loan status, these variables be renamed with positive values of numerical one's.

## Removing Outliers

Outliers are Data points which are in extreme positions which leads to statistical changes in output whether it extreme low or extreme high. Removing outliers leads output to be statistically significant, here the outliers can be removed using quantile method .

## SKEWED DISTRIBUTION TREATMENT

Skewed distribution is the measure of symmetric probability distribution. The skewness comes with removing outliers, normalizing the data. The normalization of data is done using square root method here the applicant income and co-applicant income and loan Amount because these are independent variables

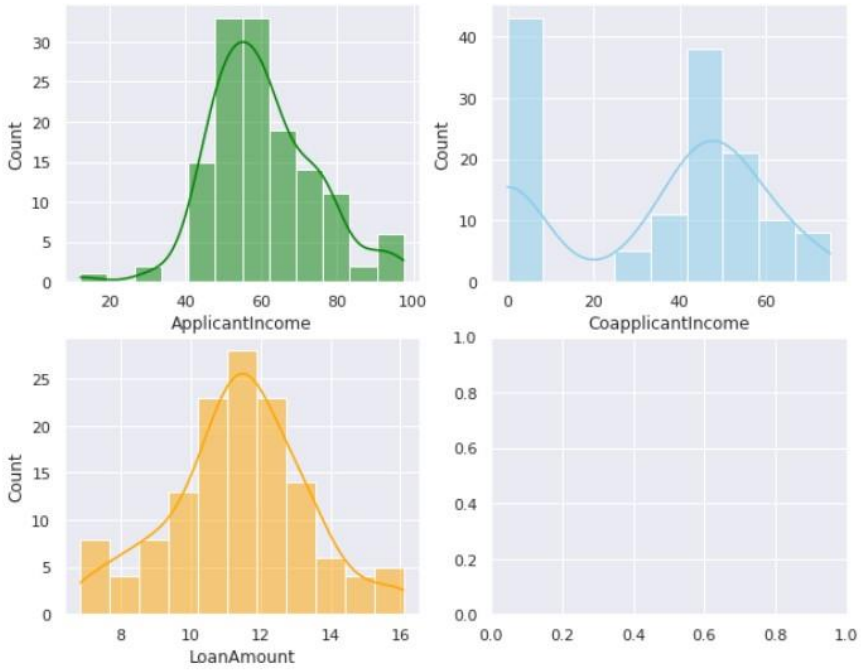



**Fig . 5 - Skewed Distribution**

## SMOTE TECHNIQUE

SMOTE is a synthetic minoirity oversampling technique it is used to increase the number of cases in the dataset in balanced way.it will generate new instances from existing minority cases that as from input.
here, the number between approved and rejected loans are in imbalanced form so we use smote technique to avoid overfitting. It takes some random sample from the feature and identify the k-nearest neighbours and then selecting the one neighbor and identify the vector between current data point and selected neighbour.

## DATA NORMALIZATION

Data normalizing is a technique to make data of every feature into one range. In this mixmax scaler is used to preserve the the shape of original distribution and to fit transform the data in to it's range whether it is negative or positive.

## DATA SPLITTING

The splitting of data is to make accurate predictions for upcoming data using these trained and tested models. So, the data is splits into two parts which are 80 percent for training and 20 percent for testing, and also mentioned some variables for training and testing purposes in the x, y form, which are X_train, X_test, y_train, y_test.

# MODEL IMPLEMENTATION

The models implemented in this project are logistic regression, K-Nearest neioghbours, Support Vector Machine(SVM), Random Forest(RF).

## LOGISTIC REGRESSION

It's a classification algorithm, it is used to classify the inputs into different classes. It should only be used when the target variables fall into discrete categories. It basically works according to a threshold, if the value crosses the threshold then it should be put into one class otherwise the other.
here, the logistic regression contains classification_report which cmention's the precision, recall F1 score and support.
the precision is the ration of true positives to the ration of false and true positives. The Recall is ratio of true negatives tom ratio of false and true negatives. The F1 score is weighted mean of precision and recall the closer the value is 1the better the performance happens. The support is actual occurences of the dataset to diagnose the evaluation process.
it shows the metrics of performance of logistic regression on this dataset. The accuracy from these attributes is **80.00%**

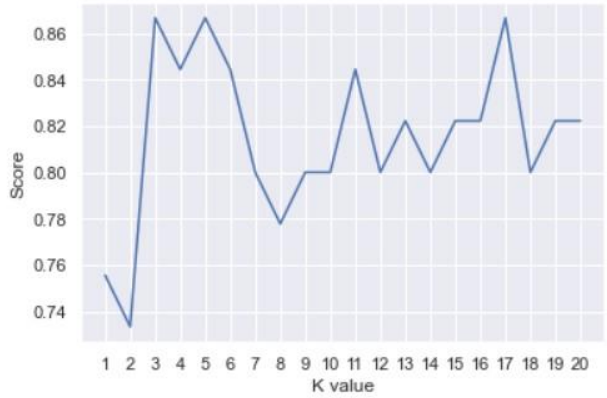
```
              precision    recall  f1-score   support

           0       0.82      0.78      0.80        23
           1       0.78      0.82      0.80        22

    accuracy                           0.80        45
   macro avg       0.80      0.80      0.80        45
weighted avg       0.80      0.80      0.80        45

[[18  5]
 [ 4 18]]
LR accuracy: 80.00%
```

**Fig. 6 – Logistic Regression Accuracy**

## K – NEAREST NEIGHBOURS (KNN)

The k-nearest neighbours (KNN) are a non-parametric algorithm that can be used to solve classification and regression problems. Non-parametric algorithms do not make any strong assumptions about the form of the mapping function, which makes the algorithm free to learn any functional form from the training data.
The k in the KNN algorithm, is the number of nearest neighbours, which are the main deciding factor in the algorithm.

Here, the K-Neighbour Classifier takes the neighbours in range of 20 from data and fit around test and train data splits into square root and normalize the data in the range of 20. So, that it can predict the accuracy.
In this data the scoreListKnn match the accuracy with **86.67%.**




```
KNN best accuracy: 86.67%
```

**Fig. 7 KNN Accuracy**

## SUPPORT VECTOR MACHINE (SVM)

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen.
Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Here, the classification_report is used to match accuracy which contains precision, recall, F1score, support and the accuracy with support vector machine in this data set is **86.67%**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.95 | 0.78 | 0.86 | 23 |
| 1 | 0.81 | 0.95 | 0.88 | 22 |
| accuracy | | | 0.87 | 45 |
| macro avg | 0.88 | 0.87 | 0.87 | 45 |
| weighted avg | 0.88 | 0.87 | 0.87 | 45 |

```
[[18  5]
 [ 1 21]]
SVC accuracy: 86.67%
```

**Fig . 8 – SVM Accuracy**

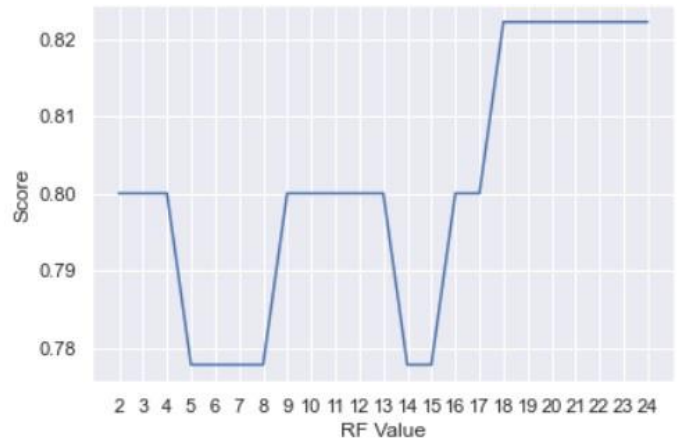
## RANDOM FOREST (RF)

Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression.

Random forest is a bagging technique and not a boosting technique. The trees in random forests are run in parallel. There is no interaction between these trees while building the trees.

It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

it takes estimator as 1000 and randomstate 1 , maximum leaf nodes are in range of 25 the accuracy it matches with this data set is **82.22%**




```
Random Forest Accuracy:  82.22%
```

**Fig . 9 – Random Forest Accuracy**

## COMPARISION

By performing the machine learning models on this data set the accuracy it predicted has almost nearset ones

Accuracy of all four models

| | Model | Accuracy |
|---|---|---|
| 1 | K Neighbors | 86.666667 |
| 2 | SVM | 86.666667 |
| 3 | RandomForest | 82.222222 |
| 0 | Logistic Regression | 80.000000 |

**Fig . 10 – Model Comparision**

## CONCLUSION

In this research paperwe have used machine learning models to predict the eligibility of an applicant for the loan. The data is pre-processed and fed to different regression models to determine the best model and classification metrics we compared different models.
So, according to above results we came to know that Logistic Regression is the most effective model with maximum accuracy and can be used as an effective model for predicting weather an applicant is eligible for loan or not, which should help banks to skip the tedious process of loan eligibility.
The highest accuracy is with KNN which **86.67%** and the overall average accuracy is **82.57%.**

## REFERENCES

[1] Li, S.T., Shiue, W., and M.H.Huang, "The evaluation of consumer loans using support vector machines.," Expert Systems with Applications, vol.30, no.4, 2006.

[2] Amira Hassan and Ajith Abraham, "Modeling Consumer Loan Default Prediction Using Ensemble Neural Networks", *International Coference on Computing Electrical and Electronic Engineering (ICCEEE)*, 2013.

[3] Archana Gahlaut, Tushar and Prince Kumar Singh, "Prediction analysis of risky credit using Data mining classification models", *28th International Conference on Computing Communication and Networking Technologies (ICCCNT),* 2017.

[4] M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020.

[5] V. Singh, A. Yadav, R. Awasthi and G. N. Partheeban, "Prediction of Modernized Loan Approval System Based on Machine Learning Approach," 2021 International Conference on Intelligent Technologies (CONIT), 2021.

[6] sklearn.model_selection.train_test_split — scikit-learn 0.24.2 documentation. (n.d.). Scikit-Learn. Retrieved June 28, 2021, from https://scikitlearn.org/stable/modules/generated/sklearn.model_selection.train_test_split.ht ml