**BANK LOAN ELIGIBILITY PREDICTION**

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

**Bachelor of Technology/Master of Technology**

In

**Computer Science and Engineering School of Engineering and Sciences**

Submitted by

**Candidate Name**

1. **MADHAV AP21110011260**

**P. GURU KIRAN AP21110010707**

**M SRI SAI KIRAN AP21110010698**

**Y Sai Aditya Maanas AP21110010573**



Under the Guidance of

**(Mr.Rajiv Senapati)**

**SRM University–AP**

**Neerukonda, Mangalagiri, Guntur Andhra Pradesh – 522 240**

**[December, 2023]**

# Certificate

Date: 28/11/2023

This is to certify that the work present in this Project entitled “**BANK LOAN ELIGIBILITY PREDICTION**” has been carried out by **N. MADHAV, P. GURU KIRAN, M SRI SAI KIRAN, Y. ADITYA MAANAS** under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology/Master of Technology in **School of Engineering and Sciences**.

**Supervisor**

(Signature)

Prof. / Dr. [Name] Designation,

Affiliation.

**Co-supervisor**

(Signature)

Prof. / Dr. [Name] Designation,

Affiliation.

# Acknowledgements

I express my sincere gratitude to my professor Mr.Rajiv senapati for providing me with the opportunity to delve into the complexities of the course Data warehouse and mining through this project. This endeavor has been a profound learning experience, offering insights into the fundamental concepts of dataset selection,preprocessing,normalization and applying different algorithms. Professor

Mr.Rajiv Senapati unwavering support and guidance have been instrumental in navigating the complexities of our project. Lastly, heartfelt thanks to my dedicated teammates for their collaborative efforts, crucial in the successful completion of this project.

# Table of Contents

Certificate i

[Acknowledgements iii](#_TOC_250008)

[Table of Contents v](#_TOC_250007)

Abstract vii

List of Publications ix

Statement of Contributions xi

Abbreviations xv

List of Tables xv

List of Figures xvii

[List of Equations xix](#_TOC_250006)

1. [Introduction 1](#_TOC_250005)
   1. Heading 2 1
      1. Heading 3 1
2. [Methodology 3](#_TOC_250004)
   1. Heading 2 3
      1. Heading 3 3
3. [Discussion 5](#_TOC_250003)
4. [Concluding Remarks 7](#_TOC_250002)
5. [Future Work 9](#_TOC_250001)

[References 11](#_TOC_250000)

Abstract:

A loan is basically borrowing money from the bank or other organization with a commitment of repaying it only on a monthly basis with interest added to it. Loans are a significant income for the banks. The repayment of the loan is very important for them so they have to verify the worthiness of the customer before accepting the loan. The project's main goal is to forecast whether the bank would accept loans for customers or not. Banks follow a set of standards and undertake background checks before lending money to customers. The rate of loan applications has increased significantly in recent years. The approval of loans is always highly risky. The necessity of customers repaying their loans is well understood by bank managers. Even after taking several measures and reviewing the loan application data, loan approval decisions are not always correct. As a result, the purpose of this study is to estimate loan eligibility using a decision tree , Gaussian Naive Bayes machine learning models.

List of Figures

Figure 1. Data set preprocessing 27

Figure 2. Normalization 28

Figure 3. Decision Tree 32

Figure 4. Naive Bayes Classifier 35

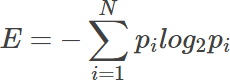
# List of Equations

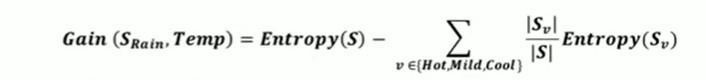
### Naive Bayes Classifier Formulae



1. **Decision tree Classifier**

Formula for Entropy



Formula for Gain for index

### Accuracy Formula from Confusion Matrix:

𝐴𝑐𝑐𝑢𝑟𝑎𝑐𝑦 = 𝑇𝑃 + 𝑇𝑁 / 𝑇𝑃 + 𝑇𝑁 + 𝐹𝑃 + 𝐹𝑁

# Introduction

Loan is one of the ways to get funding for the needs of the people. The banks will be flooded for the loan request but they have to go through various checks like the credit score of the person and all the factors because the repayment of loan is most important for the bank’s survival. Loan recovery is a major contributor in a bank's financials. It is quite difficult to predict if the customer will be able to repay the loan. So in this project, We predicted whether the bank can approve a loan for a customer or not based on a few parameters like customer education background, source of income, loan amount, and the credit history of the customer using a machine learning model. First, we preprocessed the data by applying data preprocessing and feature engineering techniques. Then we used the Decision Tree Machine Learning algorithm on the preprocessed data to forecast whether the loan would be approved or not, and we calculated the algorithm's accuracy.

Decision tree is a classification algorithm which constructs a tree based on the training data and makes predictions.The models are compared using a variety of measures, including Mean Square Error (MSE), Root Mean Square Error (RMSE), and Root Mean Square Error (RMSE) (RMSE).

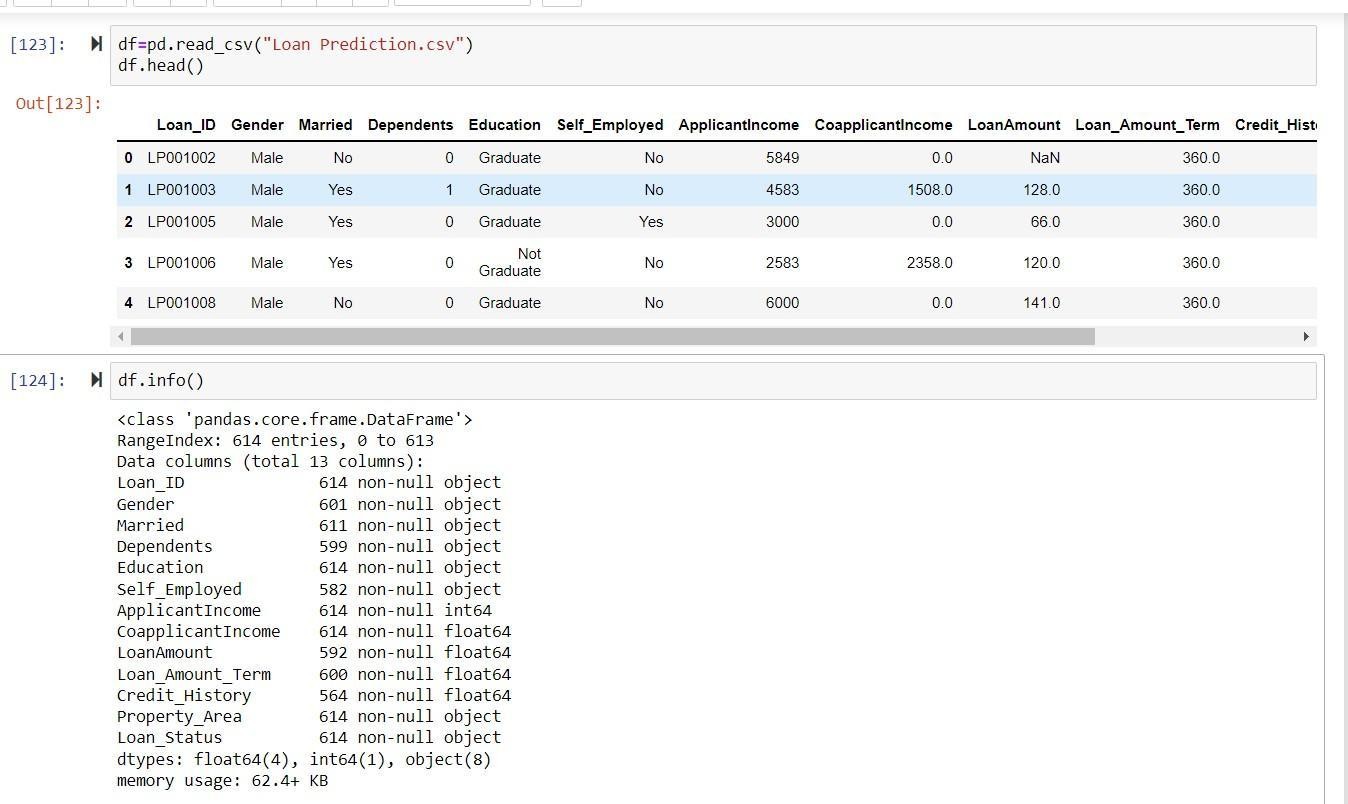
Steps involved:

* 1. DataSet selection
  2. Data Preprocessing
  3. Feature Selection
  4. Applying ML model
  5. Error calculation

# Methodology

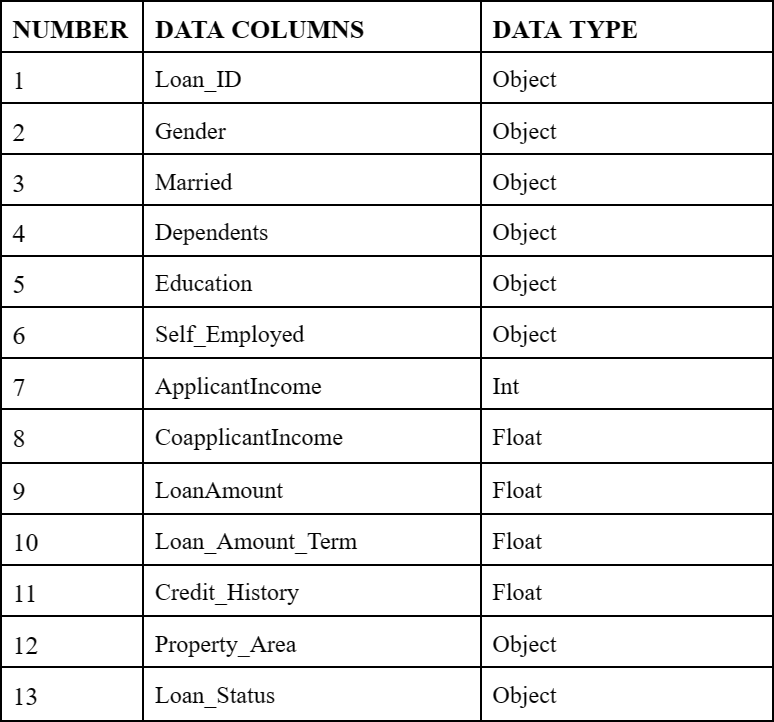
### DATASET SELECTION:

Finding a meaningful dataset for the topic we've chosen is the most important task in every machine learning research. The bank loan dataset was downloaded from the Kaggle website.



### DATASET DESCRIPTION:

The dataset we have taken from kaggle consists of a total of 614 rows and 13 columns.



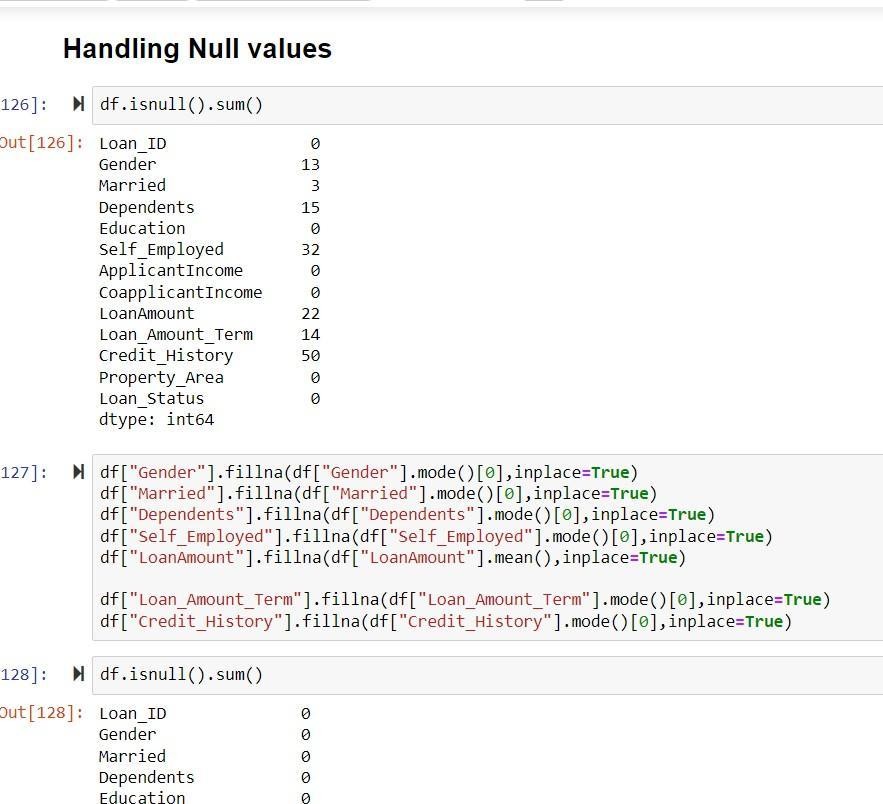
### DATA PREPROCESSING:

The real-world data is subjected to noise and null values. So we need to clean the dataset and apply a machine learning model for better results. First, we checked for any duplicates in the dataset and then went for Data Cleaning which includes removing noise and null values.

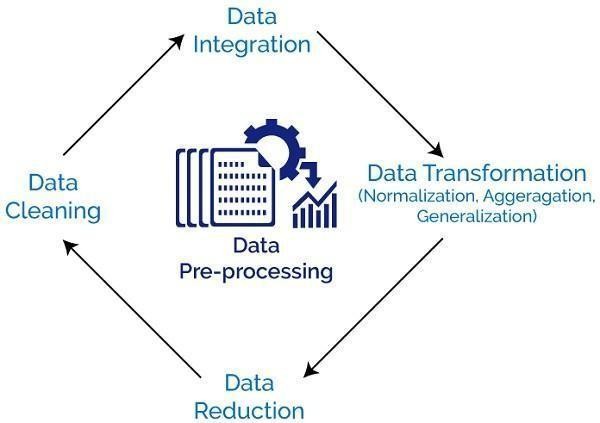
There are a few methods to handle the null values:

* + 1. Ignoring the tuple
    2. Filling by mean
    3. Filling by most repeated value
    4. Filling by a constant

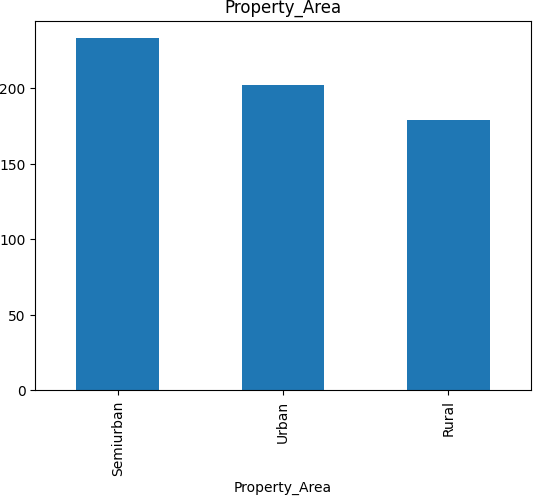
Here we used mean and mode methods to fill the null values.



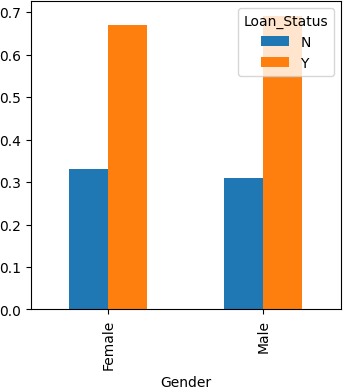
After dealing with null values, we need to convert categorical columns into numerical data columns. We cannot apply the machine learning model to categorical columns. So we Found the categorical columns from the dataset and used an inbuilt module from sklearn LabelEncoder to convert categorical data into numerical data.



**Some Data Visualization:**



This Above Bar Graph shows the Number of Semi Urban,Urban,Rural Mentioned in the data set



This above Bar Graph shows the Loan status of Male and Female.

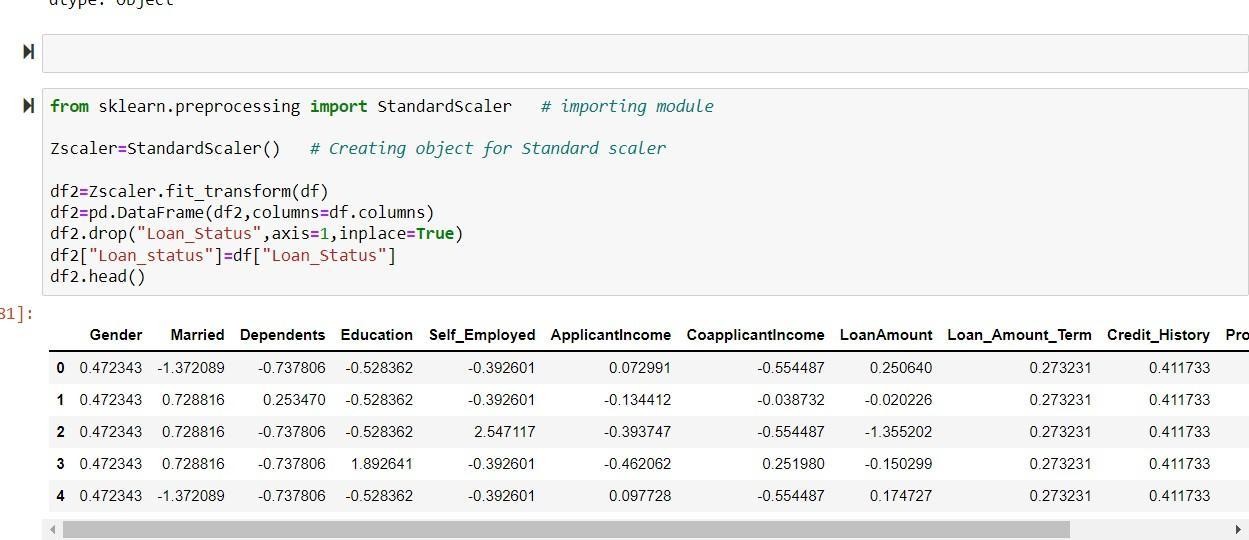
### NORMALIZATION:

After merging data from various sources, the data may also need to transform into forms appropriate for mining. The data transformation includes the following techniques:

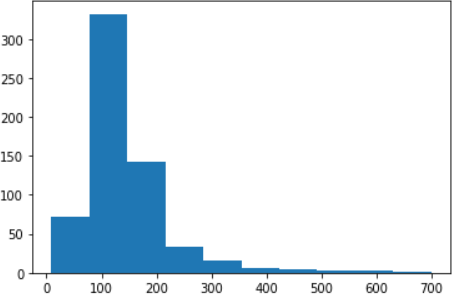
* + 1. Smoothing
    2. Aggregation
    3. Normalization
       1. Min-Max Normalization
       2. Z score normalization
       3. Decimal scaling

We need to normalize the data so that the accuracy of the model improves. There are various methods to normalize the data like Min-Max normalization, Z-score normalization, and also NumPy has an inbuilt function to normalize certain columns in

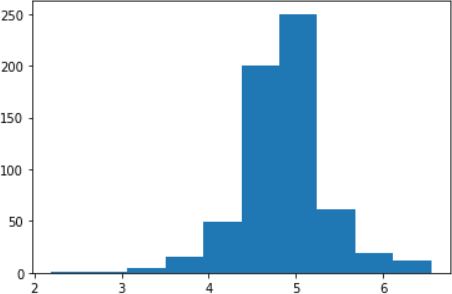
the data.



Before Normalization



After Normalization:



### Feature Selection:

Machine learning is based on a basic principle: if you put garbage in, rubbish will come out. When I say garbage, I'm referring to data noise. When there are a lot of features, this becomes much more significant. When constructing an algorithm, you don't have to employ every feature available to you. You may help your algorithm by providing only the most critical features into it.

Feature selection is the process of selecting required or most valued features from the dataset which contribute most to our predicted variable. Feature selection helps to improve accuracy and reduce the running time of the model since we are removing the columns which are less significant to our predicted variable. It reduces the overfitting of the model.

# 3. Discussion

### DECISION TREE:

**Algorithm:**

It is a greedy algorithm

The tree is constructed in a top-down recursive divide and conquer method At the start, all the training examples are the root.

Attributes are categorical(if continuous-valued they are discretized in advance) Examples are partitioned recursively based on selected attributes

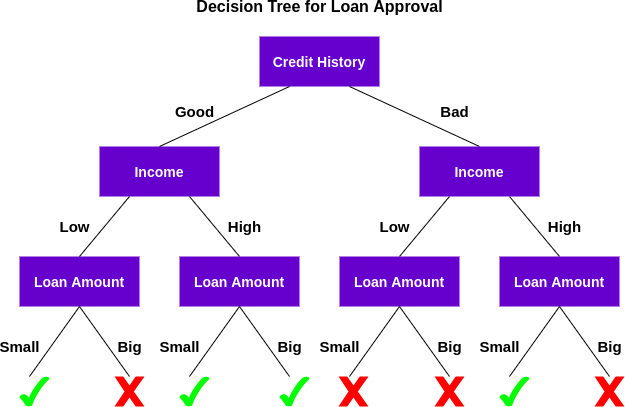
Test attributes are selected on the basis of heuristic or statistical measures. Conditions for stopping partitioning:

A node's samples all belong to the same class.

There are no remaining attributes for further partitioning There are no samples left.

### Advantages:

* + 1. Simple to understand and to interpret. Trees can be visualized.
    2. Able to handle multi-output problems.
    3. Even if the underlying model from which the data were created violates some of its assumptions, it still performs well.

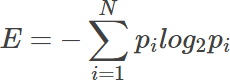


The determination of the characteristic for the root node in each level is a key difficulty in the Decision Tree. This process is known as attribute selection. We have two popular attribute selection measures:

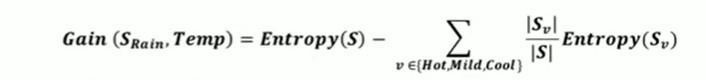
* + - 1. Information Gain
      2. Gini Index

Given the dataset, we need to find the root node for the tree. To do so, we must first determine the information gain of each column. The root node will be determined by the column with the highest information gain. To compute the information gain, we must first calculate the entropy of each possible outcome in each column, as well as the overall entropy. Similarly, the creation of the root node continues by picking other nodes in the appropriate columns.

### Entropy formula:



**Gain formula:**



We used the inbuilt module to implement the decision tree algorithm to our dataset.

### The Gaussian Naive Bayes (GNB)

This algorithm is a probabilistic classifier based on Bayes' theorem with the assumption of independence between features. It is specifically designed for continuous data, and it assumes that the features are normally (Gaussian) distributed.

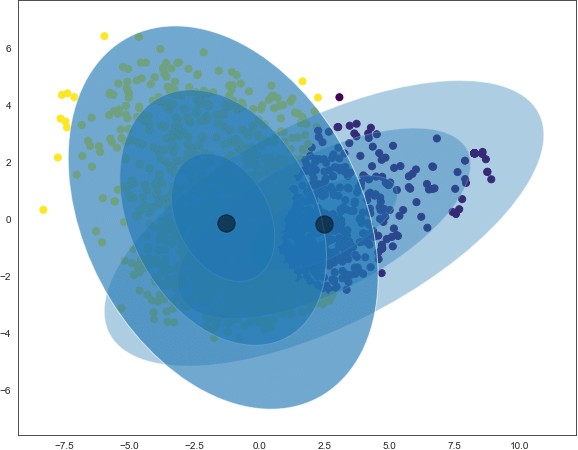
Gaussian Naive Bayes is the application of Naive Bayes on normally distributed data. Gaussian Naive Bayes assumes that the likelihood(P(xi|y )) follows the Gaussian Distribution for each xi within yk . Therefore,



To classify each new data point x the algorithm finds out the maximum value of the posterior probability of each class and assigns the data point to that class.

### Advantages of Naïve Bayes Classifier:

* Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
* It can be used for Binary as well as Multi-class Classifications.
* It performs well in Multi-class predictions as compared to the other Algorithms.
* It is the most popular choice for text classification problems.



### PERFORMANCE ANALYSIS:

The developed model is tested using the AUC curves and Confusion matrix.

The accuracy of the Decision Tree model is 74.79%.The Accuracy of Random forest model is 74.79% .The Accuracy of Gaussian Naive Bayes model is 82.92%

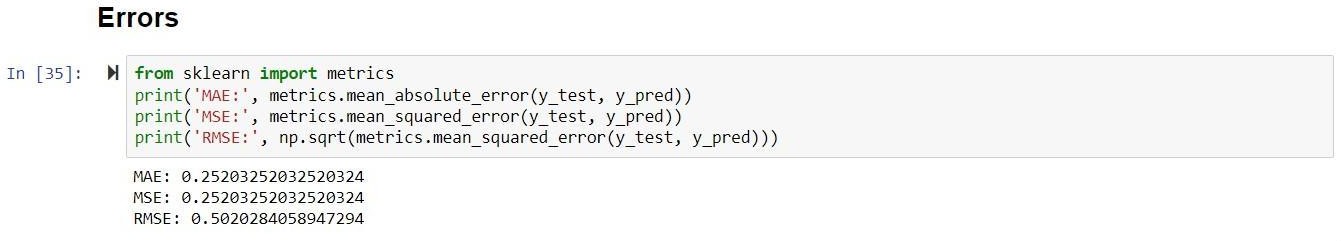
## Error calculations:

* + 1. **For Decision Tree :**

MAE: 0.25203252032520324

MSE: 0.2520325203252032

RMSE: 0.502028405894729



* + 1. **Gaussian Naive Bayes (GNB)** MAE: 0.17073170731707318 MSE: 0.17073170731707318 RMSE: 0.4131969352706687

## Confusion Matrix

1. **Decision Tree:**

[[19 14]

[17 73]]

1. **Gaussian Naive Bayes**

[[15 18]

[ 3 87]]

# Concluding Remarks

The model began with the data cleaning and processing followed by implementation. The decision tree model gave 74% accuracy for the dataset. .The Gaussian Naive Bayes model gave accuracy of 82.92% As a result applicants with poor credit history are rejected and applicants with higher income have more chances to get the loan as the chances to get back the loan amount are higher. Gender and marital status play less role in determining loan eligibility.

# References

* 1. https://[www.irjet.net/archives/V8/i4/IRJET-V8I4785.pdf](http://www.irjet.net/archives/V8/i4/IRJET-V8I4785.pdf) 2.[https://deepnote.com/%40rhishab-mukherjee/Loan-Prediction-Project-TermPaper-](https://deepnote.com/%40rhishab-mukherjee/Loan-Prediction-Project-TermPaper-54) [54](https://deepnote.com/%40rhishab-mukherjee/Loan-Prediction-Project-TermPaper-54)