Machine Learning model to predict if the client is high risk or low risk if we were to provide them loan.

A PROJECT REPORT

***Submitted by***

Sandayyagari Akshay Reddy

16 JUNE 2024

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| Content | Page no |
| Problem statement | 3 |
| Overview | 4 |
| SOFTWARE SPECIFICATION | 5 |
| IMPLEMENTATION  data LoADING  DATA EXPLORATION  DATA VISUALIZATION  DATA TRANSFORMATION  FEATURE Engineering  DATA SPLITtING  MODEL TRANING & PREDICTION | 7  8-10  10-11  12  13  13  14-16 |
| Result and analysis | 17 |
| Conclusion | 18 |
| code file link | 18 |

**Problem Statement**

Based on the given financial data create a ML model to predict if the client is high risk or low risk if we were to provide them loan. We need to predict the column Risk\_Flag and it contains value 1 if the client is high risk else it will be 0.

Perform all the various steps of machine learning like data exploration, feature engineering and model building.

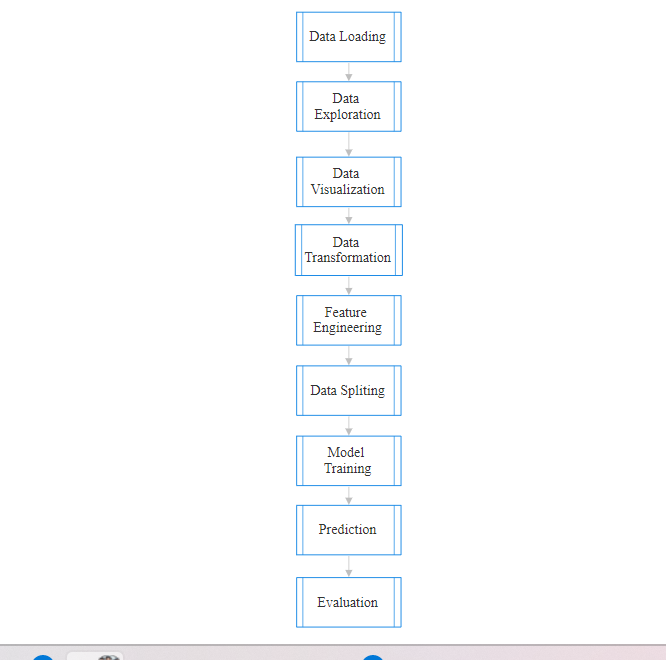
**Overview**

In this project we are going to develop a Machine Learning Model to predict the client is high risk or low risk if we were to provide them loan. Mainly we need to predict the column Rick\_flag in the dataset given. So, for the implementation I used Google Colab. First, I started with different data exploration techniques like checking the information, checking missing values etc of the dataset and then I have done data visualization to explore the data within deeper and know what data exactly in the dataset. There is some of data which may affect the prediction of the model to control that I have some data transformation techniques like changing the categorical values to numerical values and after that I have done feature engineering which is the most important part of the model where we should select the feature which are important for prediction. After this we should split our dataset into two data frames one is testing and other is for training and then we should select a machine learning model which best suits our problem statement I took the logistic regression model to predict the out and the last step is evaluation of our model with different performance metrics like F1-Score, Recall etc.

**Software Specification**

* Platform: Google Colab
* Libraries used: Pandas, Matplotlib, Seaborn, Scikit-Learn

**Flowchart**

****

**Implementation**

1. Data Loading

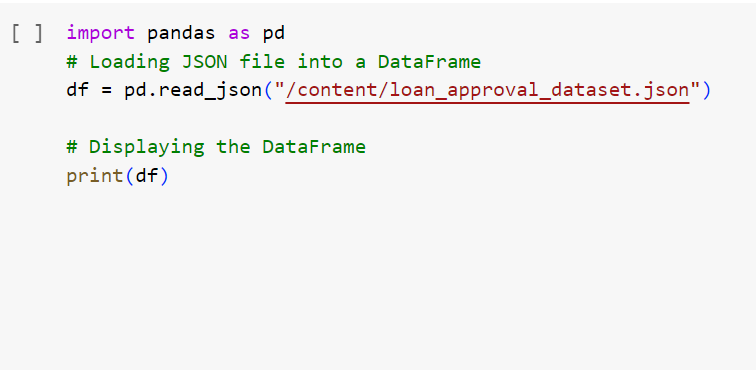


Fig1: Data Loading

Explanation: This is the first step of our project where we load our dataset file into colab and display the data frame.

1. Data Exploration

In Data Exploration I have done 5 different techniques they are

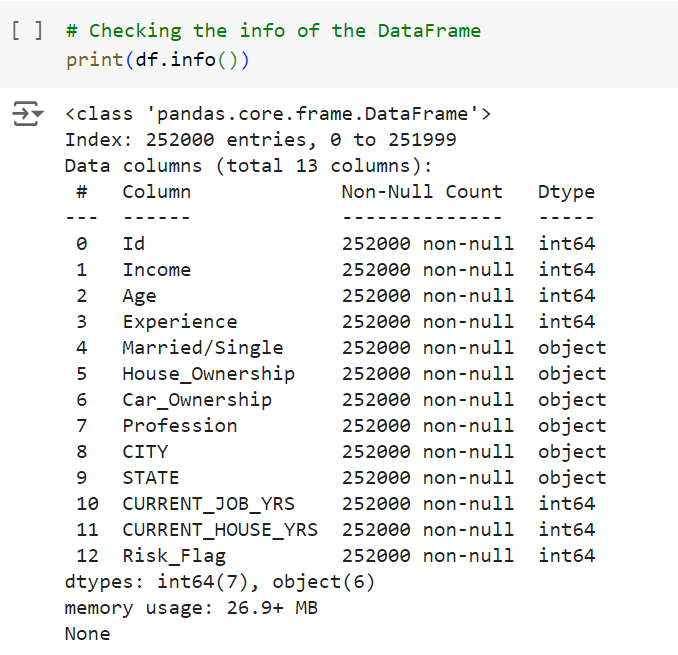


Fig2: Info of Data Frame

To get the rough idea of the feature data types and no of values in the dataset I have performed this.

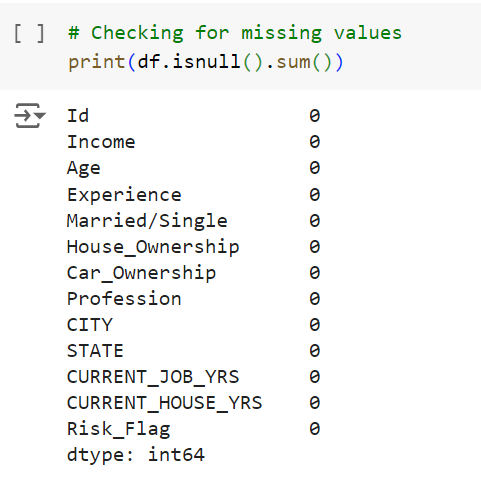


Fig2.1: Missing Values

Here to check whether there are any missing values in our dataset I performed this.

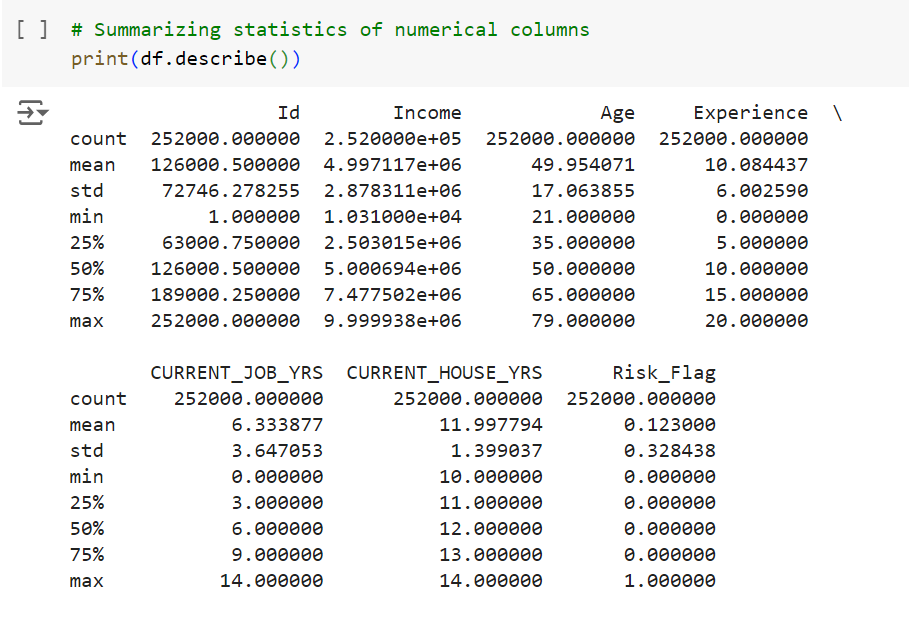


Fig2.2: Describing Dataset

Here I have done because to know more about the data set and the values.

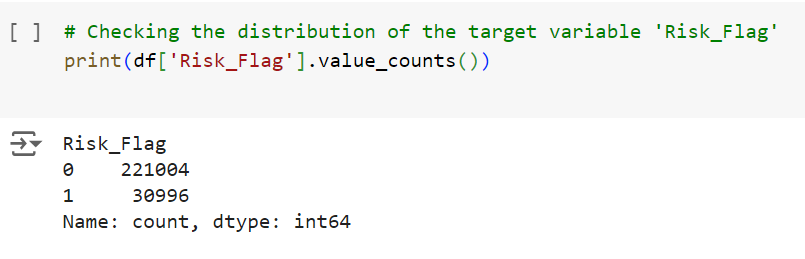


Fig2.3: Risk\_Flag column

The project we are doing mainly turns around the dataset Risk\_Flag so I have to know how many classes are there and how many of them are there in that column.

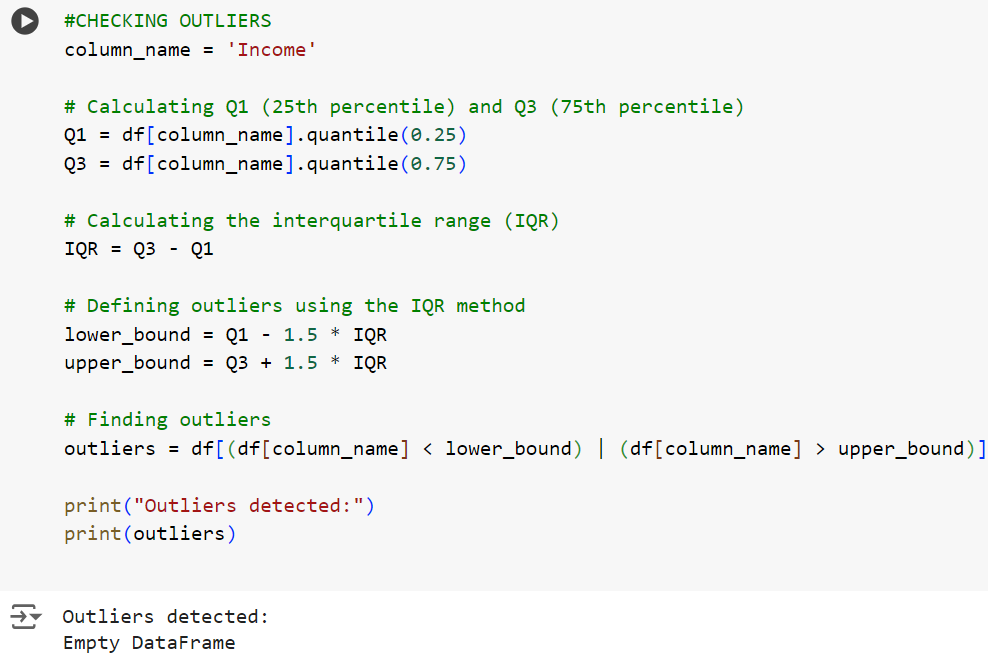


Fig2.4: Outliers

I want to whether in our dataset the outliers are there or not because it may affect our prediction.

1. Data Visualization

I Have three visualization techniques why only there I will explain further

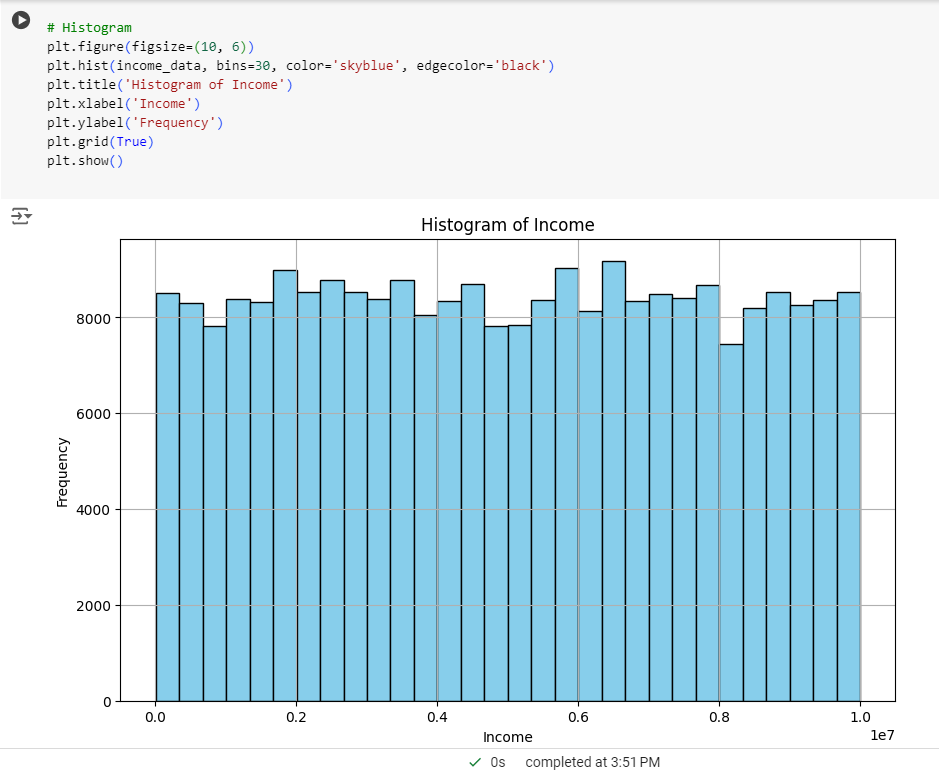


Fig3: Histogram for feature “Income”

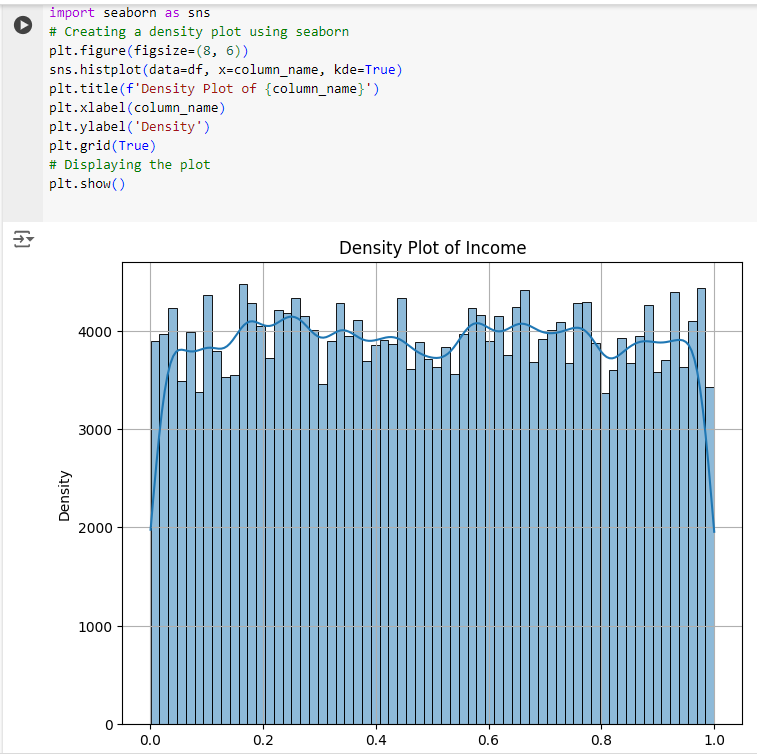


Fig3.1: Density Plot for feature “Income”

Why I have done these plots is in the dataset after the encoding of the features which are with categorical values only feature “Income” is with higher values. so, to normalize it and in normalization for which technique I should use for this I have done these plots.

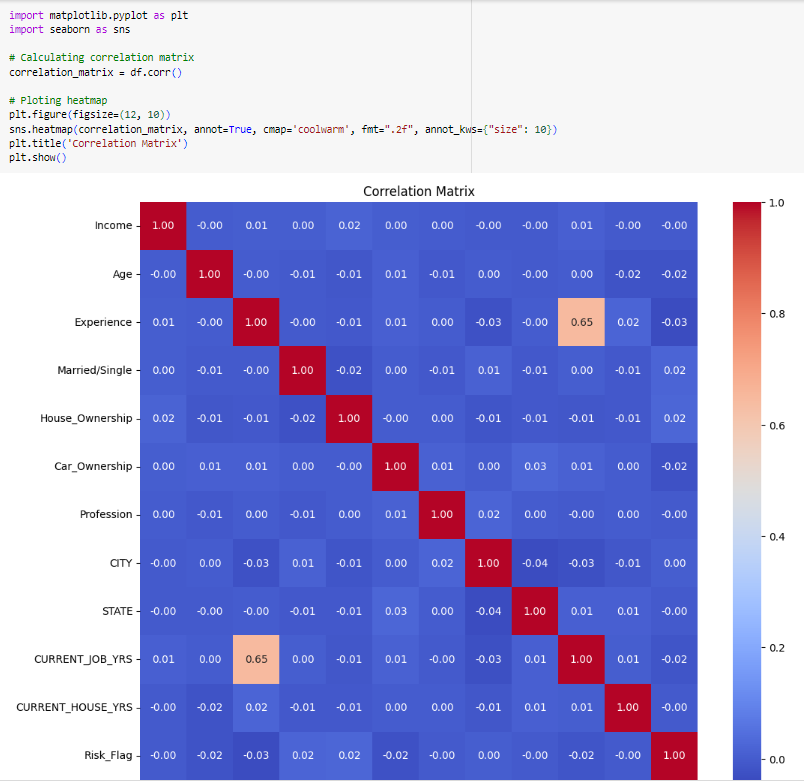


Fig3.2: Correlation Matrix

I have done feature selection by using correlation values so for the values I have generated a correlation matrix

1. Data Transformation



Fig4: Removing a column

In the dataset I have a feature “Id” so I am removing this because at the time of training the model it may affect our prediction

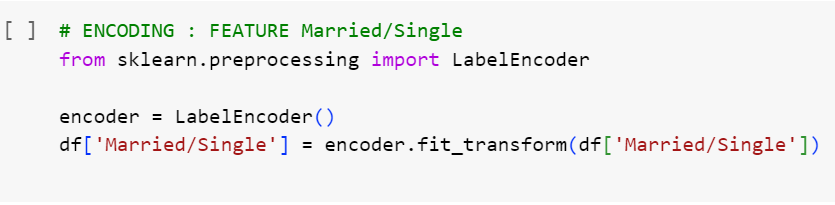
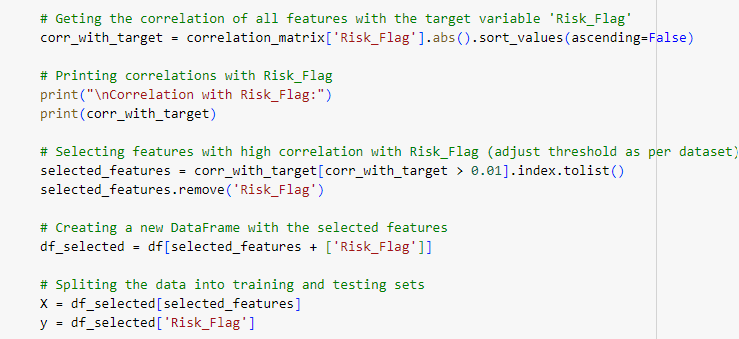


Fig4.1: Encoding

In our dataset we have some feature with categorical values so at the time of training it is difficult to process the categorical values that why we are converting categorical values to numerical values using encoding. the above image shows only for the column “Married/Single” I have done it for all the columns with the categorical values.

1. Feature Engineering



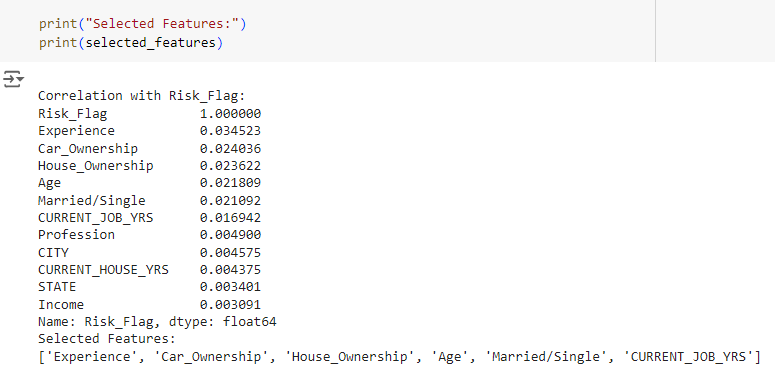


Fig5: Feature Engineering

Here we have selected the main features to take for the predicting the output.

1. Data Splitting

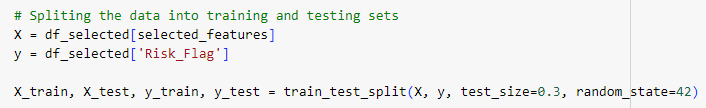


Fig6: Data Splitting

Here we have done two datasets one is for testing and one is for training. We have divided into 30:70 ratio

1. Model Training & Prediction

Using Logistic Regression

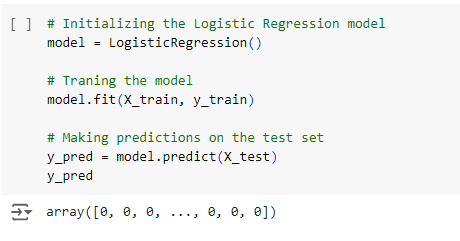


Fig7: Training & Prediction LR

We are training the model using logistic Regression after training we are predicting the values.

Using Random Forest

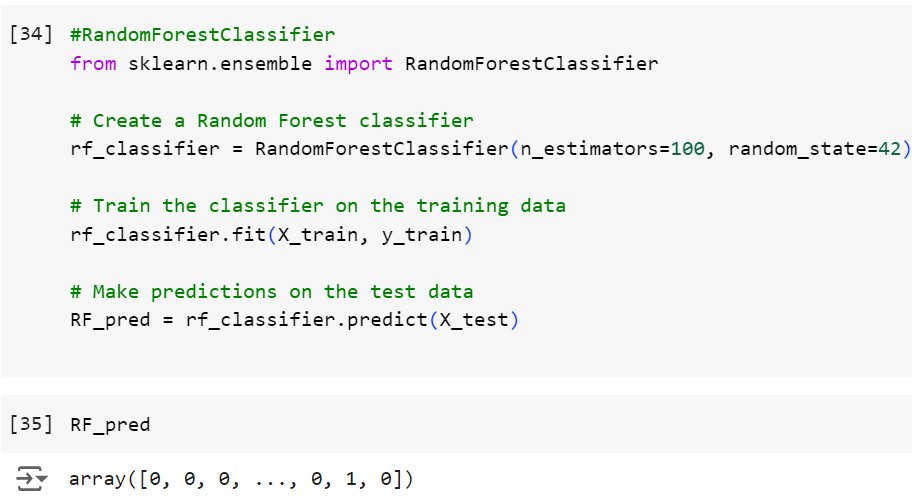


Fig7.1: Training & Prediction RF

We are training the model using Random Forest after training we are predicting the values.

Using KNN

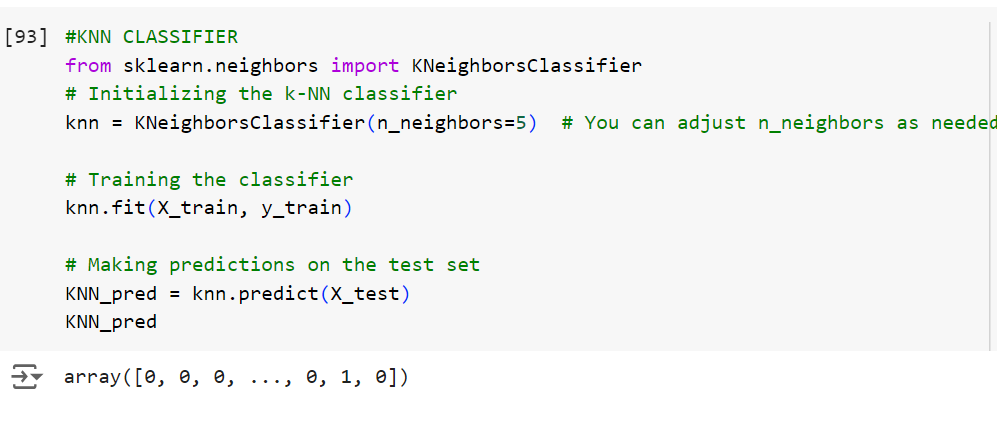


Fig7.2: Training & Prediction KNN

We are training the model using knn after training we are predicting the values.

Using Decision Tree

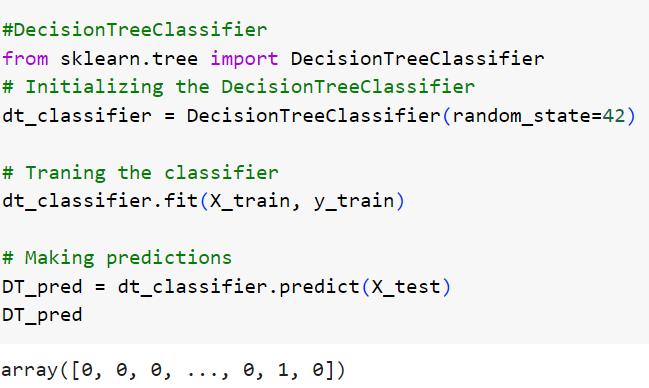


Fig7.3: Training & Prediction DT

We are training the model using Decision Tree after training we are predicting the values.

Using Naïve bayes

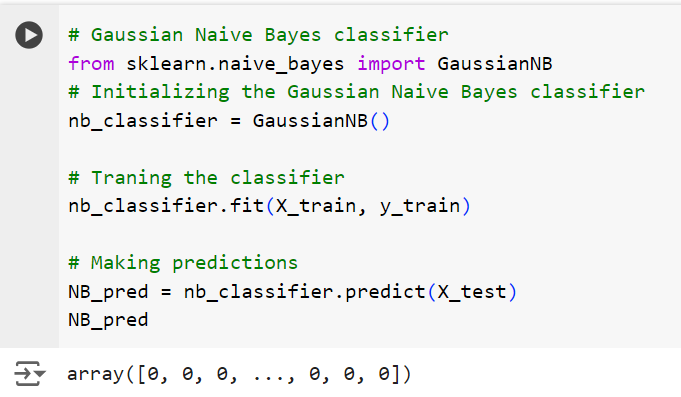


Fig7.4: Training & Prediction NB

We are training the model using Naïve bayes after training we are predicting the values.

**Result and Analysis**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Accuracy** | **Precision call** | **Recall** | **F1-Score** |
| Logistic Regression | 0.8773 | 0.44 | 0.50 | 0.47 |
| Random Forest | 0.8786 | 0.71 | 0.63 | 0.66 |
| KNN | 0.8710 | 0.69 | 0.63 | 0.65 |
| Decision Tree | 0.8758 | 0.70 | 0.63 | 0.65 |
| Naïve Bayes | 0.8773 | 0.44 | 0.50 | 0.47 |

Fig: Evaluation Metrics

The above table is consisting of different types of evaluation metrics of different machine learning models for our project. As we can see in all the evaluation metrics the random forest model performed better than other models and then followed Decision Tree, KNN. Here we can observe another thing that both logistic regression and Naïve bayes got same values in all the evaluation metrics. And these are the low performed models than the other models.

**Conclusion**

For the problem statement to predict if the client is high risk or low risk if we were to provide them loan the best model as per the condition, I have taken is Random Forest for predicting the risk level of clients applying for the loan.

**Code Link:**

<https://colab.research.google.com/drive/1u0Ag9FwDdiQQOJkFJsfTi0hKz2tzFV8T?usp=sharing>