A project report on

CUSTOMER CREDIBILITY PREDICTION FOR BANK LOAN AUTHENTICATION SYSTEM USING MACHINE LEARNING

Submitted in partial fulfilment for the award of the degree of

Master of Science

in

Data Science

by

GEETHA LAKSHMI.B-21MDT0057

Under the guidance of

Dr. G. MOKESH RAYALU

SCHOOL OF ADVANCED SCIENCE

VIT, Vellore.

April, 2023

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

ABSTRACT:

Banking sectors plays a leading role in the economic progress of all the nations of the world. One of the main sources of bank's revenue depends on the credit amount which has been approved to the people. Most of the bank's financial gain is attained from the gain earned in credits. With the rise the financial area, most of the individuals are seeking for bank credits however the organisation has its finite resources that should be granted for specific individuals. Now it's a big deal for the banks to take a secure choice in finding out the right individual to whom the credit can be granted. Even after the lengthy manual validation of credit applicants, it's uncertain for the employees to say whether the approved customer will repay the credit or not. Here in this project, in order reduce the burdensome of the bank employees and to save their resources, we are trying to decrease the vulnerability behind the approved individual. To accomplish this, the dataset can be taken out of the Kaggle library which contains the past information of credit applicants to whom the credit was sanctioned and with the help of those information the system will be built and trained under various machine learning models and predicting the appropriate model for the credit endorsement system which results in the exact outcomes. The primary goal of our work is to forecast the eligibility of the customers whether the individual is applicable for the loan or not. We have experimented with a wide variety of machine learning methods., to reduce the efforts of the employees which leads to a miscalculation in finding the eligible applicant for the credit authentication process. Here, we examine the various details of the customers, with the features including income, previous credit records, education status, and their asset details from previous records of credit applicants regarding their loan approval, and the ideal elements are chosen which have a straightforward influence on the outcome for our credit authentication system.

KEYWORDS-Credit authentication, Machine learning, Classification models, Testing and training set, Applicant details, Decision making.

1. INTRODUCTION

The primary revenue stream of most of the banks is loan disbursement. Loans given to applicants account for most of a bank's revenue. The banks here charge interest provided to applicants while on loan. Bank's main goal is to invest their resources in dependable clients. Various banks have been processing loans so far following a backwards process of vetting and verification. But, as of right now, no bank can guarantee whether the applicant who is selected for a loan application is secure or not. To avoid this circumstance, we developed the Bank Loan Authentication mechanism employing machine learning, a system for the approval of bank loans. A model called Loan Prediction System determines if a specific customer is qualified and capable of repaying a loan. This method examines a few factors, including the marital status, income, and spending of the user and different elements. Several users of the training sample use this approach. These elements are considered when creating the necessary model. To obtain the desired outcome, this model is applied to the test data set. The result will be presented as either yes or no. If a customer responds with a yes, it means they can repay the loan, while a no means they are not able to repay the loan. We can approve loans for clients based on these variables.

1.1 OBJECTIVE:

For both the customers and the bankers, credit authentication is quite advantageous. The aim of this study is to choose the verified candidates which makes it easy, quick, and effective technique for the credit approval system. Banks are competing for an alternative over each other to increase the overall business in the intense battle.

In order to maintain a stable business battle, banks now understand that retaining their consumers and avoiding fraud must be the main goals of any competitive plan of action. The two main goals which is to be achieved are as follows:

1.Determining the capability of a customer to payback the credit based on their essential qualities.

2. Deciding the best model to approve the customer's credit application.

3. Evaluating the applicants creditworthiness based on the customer segmentation.

1.2 MOTIVATION:

The motivation behind this “Customer credibility prediction for bank loan authentication system using machine learning” is driven by the desire to advance technical and research skills, gain practical experience in using machine learning methods to solve critical challenges, and possibly have a beneficial social effect. The key component of business and economics involves the management of money and financial resources, including financial analysis, investment management, risk management, and financial planning. A crucial part of corporations, governments, and people as well as the global economy is played by banking and finance. It utilises concepts from economics, accounting, statistics, and mathematics to assist individuals and businesses in making wise financial decisions. This fascinates me to do research on one fine topic called bank loan approval system. This system is helpful for both bankers and customers to prevent defaulters and to know their eligibility for bank loans. Automation can speed up processing times, save money, and increase accuracy by reducing the need for human interaction in the loan approval process.

1.3 BACKGROUND:

The project backdrop for a machine learning loan approval system is constructing an algorithm that can forecast the possibility of loan default based on numerous parameters such as the customer's credit rate, salary, past job records, debt-to-income ratio, and other pertinent information. Using previous loan data, the system will be trained to find patterns and correlations between loan repayment and numerous characteristics. The algorithm will then be evaluated with a different set of data to assess its accuracy and suggest areas for improvement. The chosen algorithm must strike a balance between maximising prediction accuracy and minimising incorrect positives and false negatives.

The algorithm can be included into the bank's loan approval procedure once it has been created and tested in order to produce choices on loan approval that are more precise and time effective.

The generation of models and programmes which is taken from the data and make predictions or judgements without explicit instructions constitutes the field of machine learning, a subset of artificial intelligence. Many industries, including finance, healthcare, transportation, and marketing, are using machine learning more and more to automate operations, make predictions, and derive insights from data.

The three major categories of machine learning algorithms are:

Supervised learning, unsupervised learning, and reinforcement learning are the three main subtypes found in machine learning. Unsupervised learning is the method of finding patterns and structure in unlabelled data as opposed to supervised learning, which includes training a model using labelled data. A decision-making agent is trained through reinforcement learning to consider feedback from the environment.

The merits of machine learning include its capacity to manage vast amounts of data, recognise intricate patterns and relationships, automate processes, and enhance choice. The necessity for high-quality data, the possibility of bias and discrimination, the interpretability of some models, and the requirement for constant update and preservation are some of the difficulties that machine learning must overcome.

Ultimately, machine learning is a rapidly developing field that has the potential to drastically change many different industries and greatly enhance our quality of life.

1.4 PROBLEM STATEMENT:

For banks, approving a customer’s credit is an important process. Retrieval of a credit makes up a large portion of bank’s revenue. It is extremely a big deal to identify whether an applicant will repay the credit or not. Thus, if the commercial banks need to streamline the credit authentication process, it can do so by leveraging the details submitted by the customers in an online application which includes their sexuality, relationship status, academic level, number of family members, revenue, credit amount and history, and other details. To focus on these specific customers, the challenge has been made to determine the ideal customer groups which make them eligible for the suitable credit type.

The primary mission of the problem is to forecast which customers are capable to return their credits and those who are not able to return the credits. Obviously, we must design a categorization framework based on the customers information. Usage of different kinds of ML algorithms like logistic regression, decision trees, random forests, and gradient boosting classifier in addition to that we’ll be categorize them into clusters by the customer segmentation technique so that we come to know that in which type of credit that the applicants are eligible. A precise model with less error proportion must be created. The major goal of the thesis is approving a credit to a specific customer is trustworthy or not and to determine the credit categories in which the customers are eligible.

2. RELATED WORK

2.1 LITERATURE SURVEY:

REVIEW:1

TITLE: Using partial least squares and support vector machines for bankruptcy prediction.

AUTHORS: Zijiang Yang, Wenjie You, Guoli Ji.

YEAR:2011

MODEL USED: Support Vector Machine, Partial least square

This approach fuses support vector machines (SVM) and partial least squares (PLS) based feature selection. PLS is effective at detecting complicated nonlinearities and correlations among financial data. The outcomes show its exceptional forecasting capacity. The high levels of accuracy rate in our model can result in significant financial and other advantages for enterprises through activities like credit approval, loan portfolio management, and security management.

REVIEW:2

TITLE: The bank loan approval decision from multiple perspectives, Expert Systems with Applications

AUTHORS: Karl D. Majeske, Thomas W. Lauer.

YEAR:2012

MODELS USED: Bayesian classifier and decision tree model

In this study, the authors proposed a random classifier to assess the constructive power of the two-way categorization systems in terms of actual customer credibility rank and the banking credit proposals. This probability model has been attained through Bayesian classifier and decision tree model. The Bayesian decision model offers a framework for finding categorization rules which result in the best classifications with the largest average profit or the lowest estimated cost. By comparing payoffs from multiple opinion, it is possible to spot situations where the different classifications produced by the views result in either profitability or liabilities.

REVIEW:3

TITLE: Predictive and probabilistic approach using logistic regression:
Application to prediction of loan approval.

AUTHORS: A. Vaidya

YEAR: 2017

In this study, the author Vaidya has taken the probabilistic approaches as a solution of credit approval process and delivered an accurate decision based on the machine learning model that he built. Choosing logistic regression algorithm as a perfect model to his study, he attained the accuracy of 0.791 so that this study deals with whether the client is applicable to assign a loan or not based on the past records of approved clients.

REVIEW:4

TITLE: Study on a prediction of P2P network loan default based on the machine
learning LightGBM and Xgboost algorithms according to different high
dimension data cleaning.

AUTHORS: Xiaojun Maa , Jinglan Shaa ,Dehua Wangb , Yuanbo Yuc , Qian
Yanga Xueqi Niua.

YEAR: 2018

This study reveals that the best classification prediction results are produced by the LightGBM classifier depends on numerous observational datapoints. The lightbgm Algorithm benefits the optimal multidimensional data set-based, with an error rate of 19.9% and an efficiency of 80.1%.

REVIEW:5

TITLE: Tree-Based Methods for Loan Approval

AUTHORS: M. Alaradi and S. Hilal.

YEAR: 2020

This study suggests to use a boosting-based decision-tree predictive model to make it easier to determine whether bank customers are eligible based on their attributes.In this area of research, multiple tree techniques have proven to be the best representative strategies. Boosting became superior method.

REVIEW:6

TITLE: An Approach for Prediction of Loan Approval using Machine Learning Algorithm.

AUTHORS: M. A. Sheikh, A. K. Goel and T. Kumar

YEAR: 2020

In this paper, performance metrics like sensitivity and specificity are utilized for differentiating the classifiers. The outcomes were demonstrated in which the model yields varied results. The model helps in suggest that a bank should not just give loans to wealthy customers but also consider other customer features that are crucial identifying credit risks.

REVIEW:7

TITLE: Credit risk evaluation model with textual features from loan descriptions for P2P lending.

AUTHORS: Weiguo Zhang, Chao Wang, Yue Zhang, Junbo Wang.

YEAR: 2020

This study employs modified data to acquire attributes which estimate the chances that debtors would fail on credits by utilizing deep learning-based methods. DL models perform better than CNN risk models, their use in financial field risk assessment was limited by this "black box" aspect.

REVIEW:8

TITLE: Credit Sanction Forecasting

AUTHORS: P. Kirubanantham, A. Saranya and D. S. Kumar

YEAR:2021

In this study, the authors developed a framework named credit sanction forecasting which is used to establish the debt returning capacity of the customers. Their aim is to forecast the qualified clients for their credit sanction framework by using various combination of min-max standardization, ML models including random forest and logistic regression and used the DL algorithm which involves tensorflow in their work. Among all, random forest classifier results in best accuracy to forecast the most reliable clients for their framework.

REVIEW:9

TITLE: Swindle: Predicting the Probability of Loan Defaults using CatBoost Algorithm

AUTHORS: S. Barua, D. Gavandi, P. Sangle, L. Shinde and J. Ramteke.

YEAR:2021

In this paper, the authors deals with the Swindling technique which is implemented by using the catboost algorithm to prevent the credit risk also they have executed document validation with the help of tesseract and 9amelot module which reduces the possibility of issuing the loans to defaulters and unapproved consumers by any financial institutions. Catboost algorithm results in best accuracy compared to random forest and decision tree. After establishing the credit risk probability, the applicants were also given recommendations for the customized loans.

REVIEW:10

TITLE: Design and Simulation of Loan Approval Prediction Model using AWS Platform.

AUTHORS: H. Ramachandra, G. Balaraju, R. Divyashree and H. Patil

YEAR:2021

This study involves in predicting whether the customers will repay the loan or not using the accuracy attained by the stated ML algorithms. This paper results in higher accuracy prediction for loan approval system. With this method, the banks can easily extract the required information from huge amounts of informative collections, which helps in accurate forecasting and reduces the number of bad credit concerns.

REVIEW:11

TITLE: Deep learning based bi-level approach for proactive loan prospecting.

AUTHORS: Justin Munoz, Ahmad Rezaei, Mahdi Jalili, Laleh Tafakori

YEAR:2021

This study discovers that DL methods, perform better than other ML methods when it comes to ranking customers. Also suggests K approaches as ideal model could still have a better precision. Resource plan could benefit from the use of multiple K cut-offs. Also, ranking models for revealed the benefit of choosing soft classifiers over hard classifiers.

REVIEW:12

TITLE: Predicting acceptance of the bank loan offers by using support vector machines

AUTHORS: Akça, Mehmet & Sevli, Onur.

YEAR:2022

In this study, the authors developed the machine learning classifier utilizing support vector method to assess the bank credit proposals to be approved or not. They utilized SVM method to evaluate outcomes using four SVM kernels, a grid search methodology to forecast accurate results and comparative evaluation for significantly more trustworthy outcomes. According to research, a poly kernel yields the ideal outcomes. In terms of approximation, SVM is one of the top statistical-machine learning algorithm. Their study is concluded from the comparison that SVM with a polynomial kernel is effective for tackling banking system classification issues because our study's accuracy and other metrics outperformed those of comparable studies.

REVIEW:13

TITLE:] Machine Learning Models for Predicting Bank Loan Eligibility

AUTHORS: U. E. Orji, C. H. Ugwuishiwu, J. C. N. Nguemaleu and P. N. Ugwuanyi

YEAR:2022

This study deals with the findings demonstrated the model performance accuracy, with 95.5% accuracy and Logistic Regression scoring 80% on average among all other models. They used batch ML methods (bagging and boosting) as well as additional methods like SMOTE which helps in financial authorities, firms, and debtors to succeed in streamline the loan approval process.

REVIEW:14

TITLE: A Novel Approach for Bank Loan Approval by Verifying Background Information of Customers through Credit Score and Analyze the Prediction Accuracy using Random Forest over Linear Regression

AUTHORS: Sandeep, C.V. and Devi, T

YEAR:2022

To evaluate the efficiency of novel random forest (RF) and linear regression (LR) algorithms for approving bank debts.The novel (RF) results in 70% higher accuracy than LR. It will be beneficial in the future for many applications with improved accuracy than other algorithms that don’t take into consideration the appropriate number of variables.

REVIEW:15

TITLE: Customer Loan Eligibility Prediction using Machine Learning Algorithms in Banking Sector

AUTHORS: C. Naveen Kumar, D. Keerthana, M. Kavitha and M. Kalyani

YEAR:2022

This study results in the findings which differentiates from rest of the developed models, the model decision tree using the adaboost technique yields higher precision. Both consumers and workers can utilise this application to see if they qualify for the credit or not.

REVIEW:16

TITLE: Analysis of customer segmentation model through k-means

AUTHORS: T. K. Bhatia, S. Gupta and A. Sharma

YEAR:2022

MODELS USED: K-MEANS CLUSTERING MODEL.

Customer segmentation is a strategy used by businesses to target specific, smaller groups of consumers with pertinent messages that will encourage them to make a purchase. This strategy is based on the idea that each and every customer is unique.

2.2 EXISTING SYSTEM:

Until today, most of the banks are processing the loans using paper and pen. When huge no of applicants seeks for bank loans, the banks' approval period might be very long. After the bank approves the credit, there is no assurance that the selected applicant will be able to pay it back.

The latest advancements of data mining and machine learning methods has sparked interest in applying these methods in a wide range of areas. The banking industry is not an exception, and the growing need for financial institutions which has effective threat management that sparked deep attraction in creating new techniques for risk estimation. The utilization of ML methods might improve the estimation of the financial distress to which banks are exposed. The Basel Standards, which establish outline for regulatory standards and threat reduction procedures as a ground rule for banks to handle and evaluate their risks, have undergone ongoing evolution in the field of credit risk. Two methods are offered from Basel II for calculating the minimum capital requirement, including the standardised method and the internal ratings-based method (IRB). Banks weigh a variety of risk factors when estimating the potential loss, they could face in the future. The estimated loss that the bank might incur in the event that a applicant declined is also from their metrics. The likelihood that a specific customer would default or not is one of the factors considered by EL estimation.

Applicants that are in default failed to uphold their end of the bargain and may be unable to repay their loans. So, acquiring a model that can forecast defaulting customers is desirable.

Logistic Regression which is considered as a better technique to be frequently utilized for calculating the likelihood of applicant default. In this article, a variety of machine learning approaches will be examined and explored to see if those models may displace the methods that are now in use.

Drawback of the existing system:

Existing loan authentication systems have a significant flaw in that they frequently rely largely on outdated credit scoring models, which can be prejudiced and do not accurately reflect a person's trustworthiness or capacity to repay a loan. These models frequently use variables like income, and employment status to assess applicant’s creditworthiness, but they could overlook other essential variables like credit history, education, savings, and possible future earnings. Furthermore, many loan approval processes take a long time and are ineffective, requiring borrowers to submit extensive paperwork and wait days or even weeks for a response. This can be particularly challenging for people who need money right now, such as those dealing with emergencies or unforeseen costs.

The lack of openness in the decision-making process with regard to current loan approval processes is another problem. In addition to not having access to the information or algorithms used to make the decision, borrowers could not completely comprehend why their loan request was approved or declined. Because of this lack of openness, the loan approval procedure may be viewed with suspicion and distrust.

For the bank loan authentication process, most of the banks have been used their self-commercial systems. The current system relies on data mining techniques, which is an outdated method for loan acceptance. A generalised dataset is created by combining many data sets, and various machine learning techniques are then applied to provide results. Yet these procedures are not up to the mark. Large banks are experiencing financial troubles as a result. To address this problem, we develop a new method for bank loan authentication.

3. PROJECT DESCRIPTION

This paper deals with the past records of the approved and unapproved applicants for the bank loan approval process which is acquired from Kaggle public repository. With the help of these records, the project undergoes various analysis of machine learning techniques to find which supervised ML algorithm shows better accuracy for predicting the loan eligibility after training the past records and by saving the model for our bank loan authentication system, this system makes easy to predict whether the upcoming new applicants are eligible for the loan or not. Added to that customer segmentation has been done which groups the applicants using unsupervised machine learning algorithm based on the most dominating features which has been extracted by feature engineering technique and by saving the cluster model to the bank loan authentication system. Along with the applicant eligibility prediction, this model allows us to predict to which cluster the upcoming applicant belongs to based on their monthly income, loan amount, credit history. The models have been streamlined to the bank loan authentication system to reduce the risk of default and to improve the customer satisfaction by providing fast and accurate loan decisions.

3.1 PROPOSED SYSTEM:

The proposed model seeks to estimate the applicant's eligibility for repaying their credit by analysing the applicant's features. These behavioural features have been feed to our bank loan authentication system as input. Based on the result of the best classification algorithm, the system can decide whether to approve or reject the applicant's loan request. Loan approval prediction and their complexity can be predicted using various data analytics technologies. To anticipate the type of loan, it is necessary to train the past data of approved and unapproved applicants using various algorithms before comparing it to test data. To identify the similarities in a training dataset of often granted loans, and then to develop a model based on these identified patterns. The training dataset has been passed on to the ML models then, the best classifier is found then built utilizing this training dataset. The details of each and every upcoming application form function as a test data. Following the evaluation, the system forecasts whether the newly submitted application is an appropriate customer for obtaining a loan or not depending on the inferences drawn from the training data and decides whether the applicant would be capable of returning his or her loan.

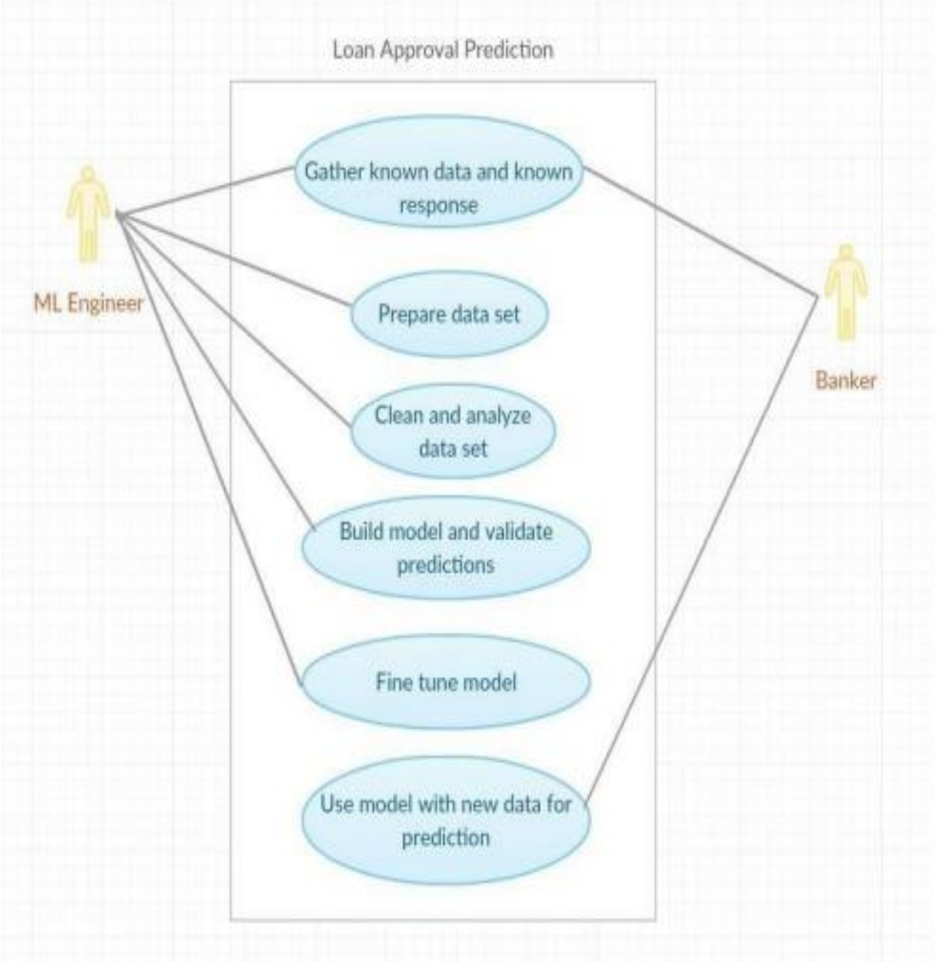

Loan Approval Prediction
ML Engineer
Banker
Gather known data and known response
Prepare data set
Clean and analyze data set
Build model and validate predictions
Fine tune model
Use model with new data for prediction

Fig 1. Work plan

Then, train the system utilizing two factors to accomplish our goal. Predictive characteristics are the first parameter, also known as independent variables, and the classification level over those factors is the second. The framework uses a forecasting model to categorise the loan applicant dataset as either Y for approval or N for denial. The information is gathered from a financial dataset that includes recent records of credit lenders whose loan applications have either been approved or denied. Based on the historical data, supervised machine learning techniques are used to develop a suggested model that may anticipate loan applicant payback. To visualise the transmission of data characteristics and create a categorization in our process, we utilised the scikit-learn module, which enables python operations. Since the model is related to supervised classification and prediction category, the different variety of ML algorithms such as the Logistic regression, Random forest, Decision tree, support vector classifier, Gradient boosting, GaussianNB classifier, KNN classifier then finally linear discriminant have been used to build the model. Among these algorithms the better models which gives higher accuracy have been induced to hyperparameter tuning to get accurate prediction results.

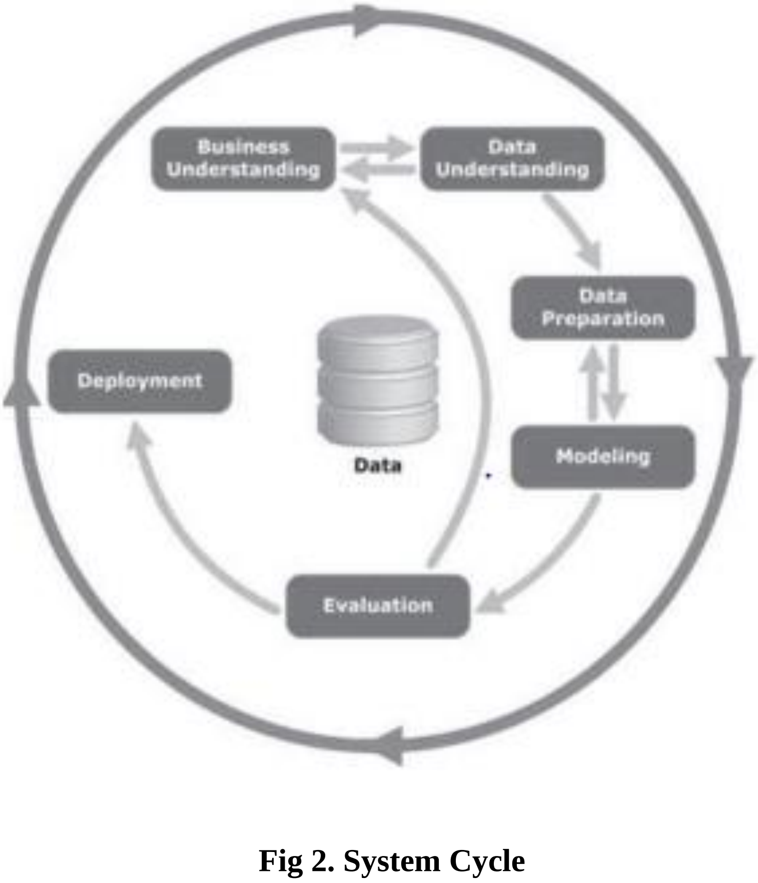


Fig 2. System Cycle

We will be using two major phases in our proposed system which is the Model training and Model forecasting. In the phase of training the model, there exists a purpose of creating a categorization regarding with the provided training data, a model is constructed utilising various classifier functions in model learning. Model prediction involves using the model to forecast the reaction towards the supplied records. We first train a dataset using a number of characteristics as a forecasting attribute and a loan status attribute as a target values. Classification algorithms are used to build and train all models for classification. Following the review of the prediction outcomes, the random forest classifier is ultimately taken as the ideal model towards classification as it shows the higher accuracy for prediction when compared with other machine learning models. Added to that customer segmentation has been done which groups the applicants using unsupervised machine learning algorithm based on the most dominating features which has been extracted by feature engineering technique and by saving the cluster model to the bank loan authentication system. Along with the applicant eligibility prediction, this model allows us to predict to which cluster the upcoming applicant belongs to based on their monthly income, loan amount, credit history. The models have been streamlined to the bank loan authentication system to reduce the risk of default and to improve the customer satisfaction by providing fast and accurate loan decisions.

3.2 BENEFITS OF THE PROPOSED SYSTEM:

Using machine learning techniques, they created an autonomous loan prediction system. We will train the computer using earlier datasets so that the system could analyse and interpret the procedure. Then the model would analyse for eligible applicants and gives result.

- The loan approval process will be speed up.
- Because the entire process will be automated, human mistake will be prevented.
- Loans will be granted to eligible applicants without delay.
- With the help of customer segmentation, the applicants will be able to find the reason of their loan approval and denial.
- Also, the bankers will be able to segment the applicants based on their credibility.

The implementation of machine learning into credit sanction systems has the chance of enhancing overall effectiveness, precision, and user experience while lowering risk and costs for lenders.

Improved Accuracy: Machine learning algorithms can evaluate huge quantities of data to identify patterns and predict outcomes with great precision. This can help lenders make more accurate and objective loan approval decisions.

Faster Processing: Traditional loan approval processes can be time-consuming and labour-intensive. Lenders can reduce the time it takes to process loan applications by automating much of the application review process with machine learning.

Increased Efficiency: Automated loan approval processes can reduce the need for human intervention, resulting in increased efficiency and cost savings for lenders.

Better Customer Experience: By shortening processing times and increasing accuracy, machine learning can provide borrowers with a better customer experience.

Advanced Management Of risk: By analysing data on loan applicants' behaviour, credit history, and other factors, machine learning models can benefit lenders detect and handle risk more effectively.

3.3 GOALS OF THE PROJECT:

- To collect information on loan applications from previous loan applicants and prepare it for machine learning.
- To choose the most essential features which utilizes variety of feature selection approaches that could affect loan acceptance.
- To choose and develop a machine learning model that can precisely forecast loan approval outcomes.
- To determine the effectiveness of the ML models using utilizing measures, comprises accuracy, precision, recall, and F1 score.
- To find various segments of applicants based on their essential features customer segmentation has to be done using unsupervised ML model.
- To deploy the ML model into a stream lit application so that the bankers can input loan application information and receive a decision on loan approval.
- To retrain the machine learning model on a regular basis, keep track of how it performs, and make any necessary adjustments to ensure accuracy.

3.4 DATASET DESCRIPTION:

The collection of data is the first step in the proposed model. In our study, we have used an open loan eligible dataset from the Kaggle public repository. The data is then cleaned and filtered in accordance with the specifications. Next, we divided the collection into training and testing samples. so that the machine learning approach could be trained and validated. At the last stage, we analyse the model's precision.

Due to the large number of applicants who apply for the loan, the data set is substantial. This data set is displayed in Table 1 and includes numerous criteria like income, education, loan amount, etc. Applicant information entered as applicants serves as test data after the model has been trained. determining the outcomes of fitting various algorithms to the training and test datasets To figure out whether or not the applicant is qualified for the credit.

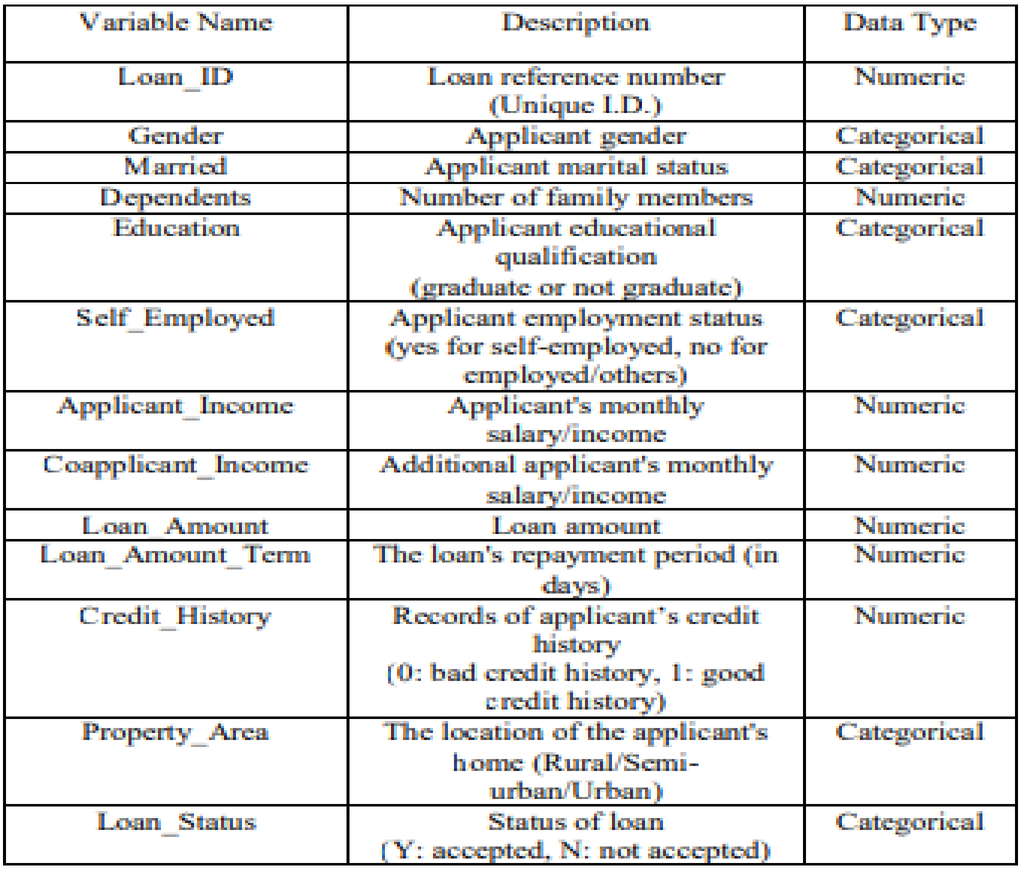
The dataset link is given below:

https://www.kaggle.com/datasets/vikasukani/loan-eligible-dataset

(i) DATA FORMAT:

The dataset consists of sample observations with 13 features and was obtained from a bank loan application. The dataset has been pre-processed in order to deal with missing values, encode categorical variables, and scale numerical variables.

The dataset is made available in CSV format, with one row per loan application and one column per feature. The first row represents the attribute names, followed by the selected feature values for each credit application.



Variable Name	Description	Data Type
Loan_ID	Loan reference number (Unique I.D.)	Numeric
Gender	Applicant gender	Categorical
Married	Applicant marital status	Categorical
Dependents	Number of family members	Numeric
Education	Applicant educational qualification (graduate or not graduate)	Categorical
Self_Employed	Applicant employment status (yes for self-employed, no for employed/others)	Categorical
Applicant_Income	Applicant's monthly salary/income	Numeric
Coapplicant_Income	Additional applicant's monthly salary/income	Numeric
Loan_Amount	Loan amount	Numeric
Loan_Amount_Term	The loan's repayment period (in days)	Numeric
Credit_History	Records of applicant's credit history (0: bad credit history, 1: good credit history)	Numeric
Property_Area	The location of the applicant's home (Rural/Semi-urban/Urban)	Categorical
Loan_Status	Status of loan (Y: accepted, N: not accepted)	Categorical

Table 1. Data description

(ii) FEATURES:

1. Loan ID – unique identifier for each loan application.
2. Gender –gender of the loan applicant(Male or Female).
3. Married – marital status of the loan applicant (Yes or No).

4. Dependants– number of dependents the loan applicant has (0,1,2, or 3+).
5. Education– education level of the loan applicant (A degree holder or not).
6. Self-employed –independent work status of the loan applicant (Yes or No).
7. Applicant Income–credit applicant’s revenue.
8. Co applicant salary–the applicant’s co-applicant’s revenue.
9. Loan Amount– the credit amount requested by the applicant.
10. Loan Amount term– term of the loan requested by the loan applicant.
11. Credit History– credit history of the loan applicant (1 for good history and 0 for bad history).
12. Property Area–location of the property for which the loan is being requested (city, village or sub-urban).
13. Loan Status –credit sanction status (Y for approved, N for not approved).

3 TECHNICAL SPECIFICATION

The software tools and libraries required for our bank loan authentication system is given below:

Software Specifications:

- Anaconda Version 3(64-bit)
    - Jupyter Notebook(for developing ML models and data analysis)
    - Spyder (for developing python web application)

Libraries Required:

- **Pandas -** Data analysis and manipulation.
- **Matplotlib -** Data Visualisation using plots.
- **Seaborn -** High level data visualisation for statistical graphs.
- **Numpy -** Array operations and mathematical functions.
- **Scikit-learn -** provides machine learning models and evaluation.
- **Pickle** -Retrieval of Machine learning models.
- **Streamlit** -Web application framework**.**

4 DESIGN APPROACH AND METHODOLOGY

Our proposed system begins from the gathering of dataset which has the past records of loan approved customers. Here, we have used a public dataset which is available in Kaggle named as loan eligible dataset for the analysis and exploration in our study.

With the help of these records, the project undergoes various analysis of machine learning techniques to find which supervised ML algorithm shows better accuracy for predicting the loan eligibility after training the past records and by saving the model for our bank loan authentication system, this system makes easy to predict whether the upcoming new applicants are eligible for the loan or not. Added to that customer segmentation has been done which groups the applicants using unsupervised machine learning algorithm based on the most dominating features which has been extracted by feature engineering technique and by saving the cluster model to the bank loan authentication system. Along with the applicant eligibility prediction, this model allows us to predict to which cluster the upcoming applicant belongs to based on their monthly income, loan amount, credit history.

5.1 MACHINE LEARNING:

The subfield of artificial intelligence (AI) known as "machine learning" involves teaching computer systems to recognise patterns in data, draw conclusions from it, and make predictions or judgements without being explicitly instructed to do so. In machine learning projects, big datasets are used to train machine learning models, which may then be applied to new data to generate predictions or choices.

In case of predictive data analytics, to analyse the past data and to give predictions about future outcomes depends upon the statistical and machine learning algorithms. Without being explicitly programmed in a computer system, the machine learning algorithms can able to study from the data and can perform various tasks. To make predictions, the ML algorithm helps to find patterns and relationships from data. It also involves building a powerful model that autonomously taught from data and may anticipate upcoming occurrences and offer outcomes lacking the need of human involvement.

Machine learning is the process of analysing data, seeing patterns, and making predictions or judgements based on that data using algorithms and statistical models. It is a potent tool that may be applied to several tasks to increase precision, effectiveness, and decision-making skills.

5.1.3 TYPES OF MACHINE LEARNING MODELS:

(i) SUPERVISED MACHINE LEARNING: -

In this type of learning the system learns from the past ideas. It takes marked/classified datapoints to build an algorithm to accurately categorise the datasets or to forecast the outcomes is what is meant by supervised learning. The datasets for the input and output are both labelled.

In this type of machine learning, a model is trained using a labelled dataset with known inputs (features) and known desired outputs (labels). When working on a loan approval project, the labelled dataset might be made up of previous loan applications. The input features might comprise applicant's credit rate, monthly revenue, previous job records, and other pertinent data. The model can then forecast the chance of loan approval for fresh loan applications using this labelled dataset.

The two primary types of supervised learning are as follows:

Classification: In classification, the output variable is discrete or categorical and might take the form of "yes" or "no," "spam" or "not spam," etc. To accurately categorise fresh input data into one of the predefined categories, a model must be learned. Support vector classifiers, decision trees, random forests, and logistic regression are a few examples of categorization techniques.

Regression: In regression, the output variable is a continuous or numerical value, such a house's price or a room's temperature. The objective is to generate a system which can forecast the outcome variable's numerical value from fresh input datapoints. Regression techniques with neural networks, polynomial regression, and linear regression as examples.

(ii) UNSUPERVISED MACHINE LEARNING: -

The programs have been set free to examine also to cluster the unlabelled dataset in unsupervised learning. Without requiring any assistance from humans, the programme finds the hidden groupings or patterns. There are no predetermined right answers; instead, the computer itself figures them out.

With this type of machine learning, a model is trained on a dataset that has no labels but has known inputs and unknown outputs. Unsupervised learning could be used in a loan approval project to find trends among the data, such as grouping related credit application forms based on their attributes.

Unsupervised learning falls into two basic categories:

Clustering: The purpose of clustering is to collect related data points into categories depending on a similarity or distance metric. Each cluster's data points ought to be more comparable to one another than they are to those in other clusters. K-means, hierarchical clustering, and DBSCAN are a few examples of clustering methods.

Dimensionality Reduction: The objective of dimensionality reduction is to decrease the number of input features while maintaining the most crucial pieces of data. By lowering the chance of overfitting, this can be helpful for visualising high-dimensional data, discovering hidden correlations between variables, and enhancing the effectiveness of machine learning models. Principal component analysis (PCA), t-SNE, and autoencoders are a few examples of dimensionality reduction algorithms.

Association mining: A method used in unsupervised machine learning to find interesting connections or patterns between variables in a big dataset is association mining, also known as association rules learning or market bundle analysis. It is widely utilized in finding frequently occurring goods that can be utilised for cross-selling, targeted marketing, and product recommendations in the context of transactional data, such as retail sales or online shopping baskets.

(iii) REINFORCEMENT MACHINE LEARNING: -

In Reinforcement learning, the algorithms concentrate primarily on agents and the actions that agents should perform in each environment. When the programme moves through its area of investigation, information is provided in the form of rewards and penalties given. The machine is instructed to make decisions using this algorithm. It operates in the following manner: the system is subjected to a setting in which it is continuously trained by the method of experimentation and failures.

With this kind of machine learning, a model is taught to make choices based on incentives or penalties. By rewarding the model for accurately predicting loan approvals and punishing it for inaccurate predictions, reinforcement learning could be used to improve the loan approval process in a project.

5.2 ALGORITHMS/MODELS USED:

For Banks, predicting whether a loan will be approved is essential since it allows them to assess the risk involved with each loan application. For our proposed model, we will be using 2 models for our bank loan authentication system to predict loan approval with accuracy.

MODEL 1- Best model for classification using supervised algorithms to forecast the credit sanction accurately.

A group of machine learning models known as supervised algorithms produce predictions using labelled data. A supervised algorithm can be trained on previous loan data to predict whether bank loans will be granted or rejected for each loan application. Using this historical data, the algorithm can then forecast upon the fact that a fresh loan application would be approved or denied depending on its specific characteristics, such as credit score, income, and loan amount.
The supervised algorithms used for classification purpose are given below:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Classifier
- Gradient Boosting
- Gaussian Naïve Bayes
- K-nearest neighbour
- Linear Discriminant

MODEL 2- Customer Segmentation using unsupervised algorithms to group applicants into different segments based on their features.

Customer segmentation, which enables financial organisations such as banks to classify clients according to their traits and determine which loan products are suitable for each segment, is a successful strategy for forecasting loan approvals. Here are some typical methods for predicting loan approvals based on client segmentation:

Demographic Segmentation: This segmentation strategy divides the consumer base into categories according to personal details involving age, gender, salary, academic status, and profession. Banks can customise their loan packages to match the unique demands and preferences of each customer segment by understanding the demographics of those categories.

Behavioural segmentation: Segmenting customers based on their behaviour, such as their purchasing patterns, saving tendencies, and credit card usage, is known as behavioural segmentation. Banks can find the best lending programmes by examining consumer behaviour.

Psychographic segmentation: This divides clients into categories according to their values, attitudes, and way of life. Banks can customise their loan products to match the unique demands and preferences of each customer segment by having a thorough understanding of the psychographic traits of those groups.

Customer Lifetime Value (CLV) Segmentation: This strategy to segmenting customers is based on how valuable each one may be to the bank over their possible lifespan. Banks can identify high-value clients who are more likely to use a variety of bank goods and services, including loans, by examining customer data such as transaction history and credit score.

The unsupervised algorithms used for clustering purpose are given below:

- K- Means Clustering algorithm.

Banks can forecast loan approvals more accurately and learn more about their customers by combining supervised algorithms and customer segmentation models. The customer segmentation models can be used to customise loan offerings to certain client segments, while the supervised algorithms can be used to anticipate loan approvals based on individual loan applications.

5.3 METHODOLOGY:

The proposed system of our Bank Loan Authentication System involves various steps to determine whether a online Loan requesting form to be granted or rejected based on various factors. Following are the outline of the process involved in our Bank Loan Authentication system:

- ➢ Hypothesis Generation.
- ➢ Data Collection and Data Understanding.
- ➢ Data Pre-processing (missing value and outlier treatment)
- ➢ Data Exploration (Exploratory Data Analysis)
    - Univariate Analysis
    - Bivariate Analysis
- ➢ Feature Selection (based on the relation with the loan approval system).
- ➢ Model Building and model selection
    - ✓ For Classification model 1:
        - Logistic Regression
        - Decision Tree
        - Random Forest
        - Support Vector Classifier
        - Gradient Boosting
        - Gaussian Naïve Bayes
        - K-Nearest Neighbour
        - Linear Discriminant
    - ✓ For Customer Segmentation model 2:
        - K-Means Clustering
- ➢ Model Training.
    - ✓ For Classification Model 1
    - ✓ For Customer Segmentation model 2:
- ➢ Model Tuning
- ➢ Model Evaluation (Testing the model with best accuracy on the testing data.)
- ➢ Selecting and saving the best model for classification and customer segmentation.
- ➢ Model Deployment (Coding a website using stream lit).
- ➢ Predicting the approval status for the applicants.

1. HYPOTHESIS GENERATION:

The first step is to generate hypothesis for our proposed system. It involves predicting about certain factors which may affect the decision to approve or reject the online loan application. This provides framework for understanding the relationship between different features and loan approval decision.

The factors mentioned below are said to have great impact on the loan approval process.

Credit History – Applicants with good credit history are said to have high chance of Loan Approval.

Applicant's Income- Applicants with good/moderate monthly income have a Possibility of getting loan approved.

Loan Amount- Applicants with reasonable loan amount with respective to their income have chances getting loan approved.

Loan Duration – Applicant's with the less/ moderate duration of Loan repayment will have chances getting loan approved.

Property Area - Applicants whose asset at urban/semi urban have chances of getting loan approved.

Job status – Applicants whose job status is self-employed/Business have chances getting loan approved.

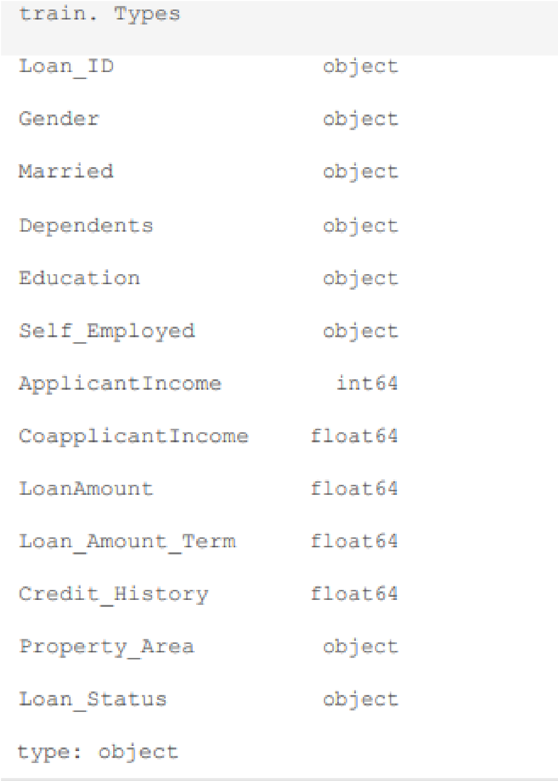
Education status- Applicants with good education status have chances of getting loan approved.

According to the loan approval process , I believe that the above mentioned factors have great impact on the loan approval process. These factors generate the understanding of our hypothesis in our Bank Loan Authentication System.

2. DATA COLLECTION:

The dataset has been acquired from the Kaggle open access data source for our loan authentication process. The sample data consists of 614 observations including 13 features. The collected data includes both the input and output labels.
The features(input labels) includes Loan number, Sexuality, spouse status, Dependants, Educational status, work status, Applicant revenue, Co-applicant revenue, credit Amount and Duration ,Credit history and asset details .
The output labels indicate the decision of loan approval process made by the bankers.
Typically, these labels are binary, with 1 denoting an accepted loan and 0 denoting a denied loan.

UNDERSTANDING THE DATA:



```
train. Types
Loan_ID                  object
Gender                   object
Married                  object
Dependents               object
Education                object
Self_Employed            object
ApplicantIncome           int64
CoapplicantIncome       float64
LoanAmount              float64
Loan_Amount_Term        float64
Credit_History          float64
Property_Area            object
Loan_Status              object
type: object
```

There are three types of variables present in our dataset.

(i) CATEGORICAL VARIABLES- Loan ID, Gender, Spouse status, Dependents, Education, Work Status, Asset Area, Loan State.

(ii) INTEGER VARIABLES- Applicant Income.

(iii) FLOAT VARIABLES- Co-applicant Salary, Loan Amount, Loan Amount Term, and Credit History.

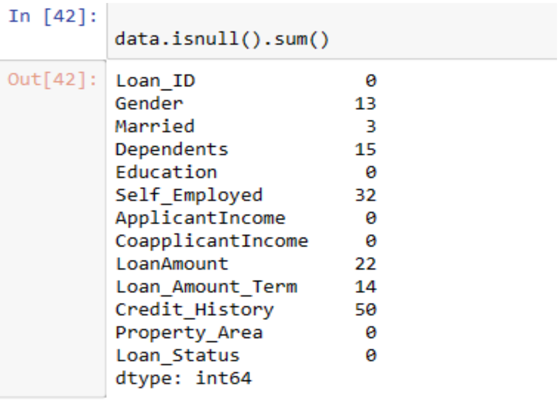
3. DATA PREPROCESSING:

Data Pre-processing is the most important part of building a machine learning-based bank loan authentication system. Data which is used to train an algorithm for machine learning model must be cleaned, transformed, and prepared. The steps involved in preparing data for a machine learning-based loan approval system are as follows:

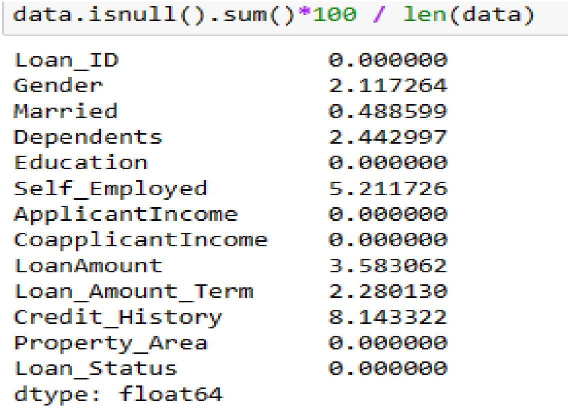
(i) DATA CLEANING:

This include addressing missing numbers, outliers, and duplicates in addition to deleting irrelevant or unnecessary data. Outliers can be found and eliminated using statistical techniques or domain expertise, while missing data can be filled in utilising imputation techniques like mean, median, or mode.

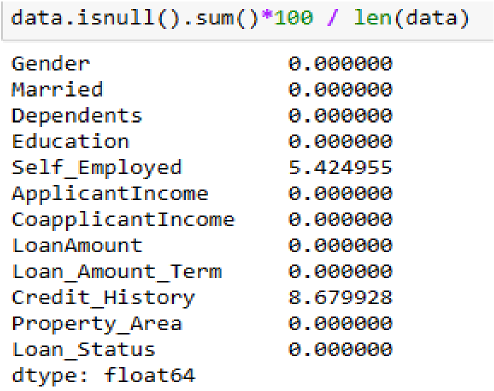
Here in our process, we will be undergoing checking for null values and the missing values are found in the columns of Self Employed and Credit History will be imputed by using the statistical mode technique.



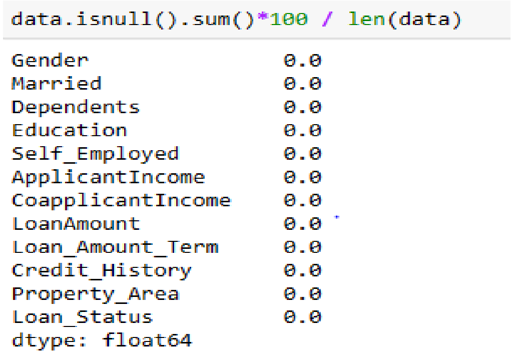
```
In [42]:  data.isnull().sum()

Out[42]:  Loan_ID               0
          Gender               13
          Married               3
          Dependents           15
          Education             0
          Self_Employed        32
          ApplicantIncome       0
          CoapplicantIncome     0
          LoanAmount           22
          Loan_Amount_Term     14
          Credit_History       50
          Property_Area         0
          Loan_Status           0
          dtype: int64
```



```
data.isnull().sum()*100 / len(data)

Loan_ID              0.000000
Gender               2.117264
Married              0.488599
Dependents           2.442997
Education            0.000000
Self_Employed        5.211726
ApplicantIncome      0.000000
CoapplicantIncome    0.000000
LoanAmount           3.583062
Loan_Amount_Term     2.280130
Credit_History       8.143322
Property_Area        0.000000
Loan_Status          0.000000
dtype: float64
```



```
data.isnull().sum()*100 / len(data)

Gender               0.000000
Married              0.000000
Dependents           0.000000
Education            0.000000
Self_Employed        5.424955
ApplicantIncome      0.000000
CoapplicantIncome    0.000000
LoanAmount           0.000000
Loan_Amount_Term     0.000000
Credit_History       8.679928
Property_Area        0.000000
Loan_Status          0.000000
dtype: float64
```



```
data.isnull().sum()*100 / len(data)

Gender               0.0
Married              0.0
Dependents           0.0
Education            0.0
Self_Employed        0.0
ApplicantIncome      0.0
CoapplicantIncome    0.0
LoanAmount           0.0
Loan_Amount_Term     0.0
Credit_History       0.0
Property_Area        0.0
Loan_Status          0.0
dtype: float64
```

(ii) DATA TRANSFORMATION:

To work with machine learning algorithms, the data must be transformed into an appropriate format so that the machine learning algorithms can use. Using methods like one-hot encoding or label encoding, for illustration, categorical variables can be transformed into numerical values.

The encoding we have used in our model is label encoding. In this type of encoding, categorical variable is being mapped to numerical value using dictionary mapping technique.

- In Gender variable, ‘Male’ is mapped to 1 and ‘Female’ is mapped to 0.
- In Married variable, ‘Yes’ is mapped to 1 and ‘No’ is mapped to 0.
- In Education variable, ‘Graduate’ is mapped to 1 and ‘Not Graduate’ is mapped to 0.
- In Self Employed variable, ‘Yes’ is mapped to 1 and ‘No’ is mapped to 0.
- In Property Area variable, ‘Semiurban’ is mapped to 2, ‘urban is mapped to 1 and ‘Rural’ is mapped to 0.
- In Loan status variable, ‘Y’ is mapped to 1 and ‘N’ is mapped to 0.

(iii) FEATURE SCALING:

The method of adjusting the data to a similar range is known as feature scaling, and it has been done to prevent certain attributes from over dominating in the model's training process. This over dominating phase can be prevent by using the techniques of standardisation or normalisation while training our model.

Here in our model we have used standardization method for feature scaling process.

STANDARDIZATION- the mean of each attribute is subtracted from the data, and the resulting number is then divided by the standard deviation of that attribute. This guarantees that the converted data maintains mean as 0 then, deviation as 1.

Standardization can be carried out in model training in a variety of methods, including by manually applying the standardisation formula to the data or by using built-in functions in machine learning libraries like Scikit-learn's StandardScaler. To enhance the ability of the model, the datapoints are subjected to the feature selection technique(ANOVA).

(iv) DATA SPLITTING:

To assess the model's performance on new data, the data is divided into training, validation, and testing sets. The test set is used to fine-tune the model's hyperparameters while the training set is used to train the machine learning model. The performance of the model on untested data is assessed using the test set.

4. DATA EXPLORATION

Data exploration is the method of analysing different variables in the loan eligible dataset, helps in finding patterns and insights that might help increase the precision of the machine learning model.

- UNIVARIATE ANALYSIS: A univariate analysis is carried out to understand the shape of distribution and the overall statistics of each variable in the dataset independently.

- BIVARIATE ANALYSIS: A bivariate analysis is carried out to understand the relationship between two variables. The analysis is carried out between each input variable with the target variable to get the understanding of the distribution and overall statistics of that input variable with respective target variable.

For Categorical variables, the table of frequencies or bar charts, which determine the proportion of each category in each variable, will be used to calculate the frequency of categorical features (Loan ID, Gender, Married, Dependents.).

For numerical variables, Probability density plots, box plots are a useful tool for visualising the distribution of a variable's numerical properties.

HEAT MAPS: we use heat maps to visualise the link across two categorical variables, such as loan approval status and property area.

CORRELATION MATRIX: By using the heatmap, the findings of the correlation matrix point out the elements that are most closely related to the target variable. To increase the accuracy in the ML classifier, these variables can be employed to features.

We can find patterns and insights that can assist increase the precision of the machine learning model by conducting univariate and bivariate analysis. By adding these factors as to the ML model, for instance, we may be able to determine that the applicant's income and property area have a strong correlation with the loan acceptance status.

5. FEATURE SELECTION:

The feature selection process is used to choose attributes that are more suitable for predicting a customer's loan eligibility. To enhance ability of the classifier, the datapoints are subjected to the feature selection technique (ANOVA).
The most crucial qualities that will have the greatest influence on the decision to approve the loan must be chosen. Techniques for feature selection can include correlation analysis or feature importance ratings.

6. MODEL BUILDING:

An objective of a bank loan approval system is to forecast whether a credit application would be accepted or denied depending on different factors, including income, credit score, employment status, and loan amount. There are two classes for the target variable in this binary classification problem: approved (1) and rejected (0).

- **MODEL SELECTION:**

We require two models for our proposed system, bank loan authentication process.

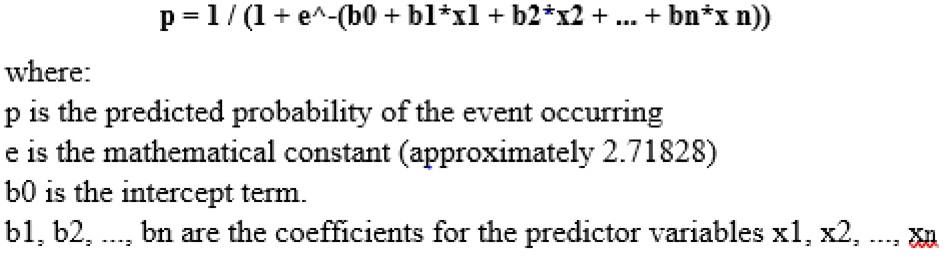
MODEL 1- For Classification problem (using supervised algorithm).
MODEL 2- For Customer segmentation (using unsupervised clustering algorithm).

ALGORITHM SELECTION FOR MODEL 1:

When choosing a model for a loan approval system, the type of problem (classification or regression), the quantity of the dataset, and the performance indicators should all be considered.

Some typical algorithms for a binary classification problem (loan acceptance or denial) are:

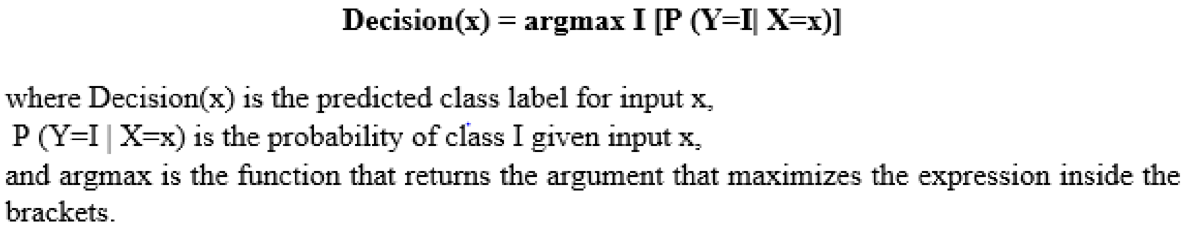
(i) **Logistic regression**: This linear model is easy to understand, quick to train, and simple. It performs effectively especially when there is a roughly linear relationship between the attributes and the target variable. In LR, the connection among the characteristics and the target variable's log-odds is modelled by a linear equation. The chance of the loan application being granted given the attributes is then calculated from the log-odds using the logistic function.



p = 1 / (1 + e^-(b0 + b1*x1 + b2*x2 + ... + bn*x n))

where:
p is the predicted probability of the event occurring
e is the mathematical constant (approximately 2.71828)
b0 is the intercept term.
b1, b2, ..., bn are the coefficients for the predictor variables x1, x2, ..., xn

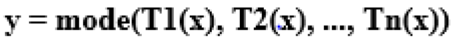
(ii) **Decision Tree**: It can handle both category and numerical features because it is a non-parametric model. It is interpretable and capable of capturing intricate correlations between characteristics. When using the decision tree approach, a tree-like structure is formed, with each node denoting a feature and each edge denoting a potential value of the feature. Recursively dividing the data into the features that most effectively divide the target variable results in the tree (Approved or rejected). After that, the tree is clipped to prevent overfitting.



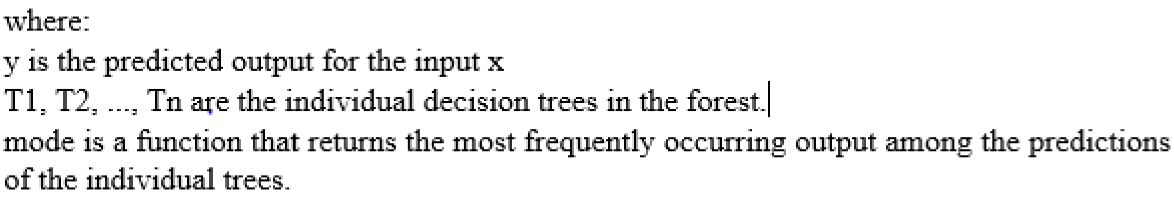
Decision(x) = argmax I [P (Y=I| X=x)]

where Decision(x) is the predicted class label for input x,
P (Y=I | X=x) is the probability of class I given input x,
and argmax is the function that returns the argument that maximizes the expression inside the brackets.

(iii) **Random Forest**: This ensemble model mixes various decision trees to improve the prediction accuracy. It can handle noisy data and performs well with large datasets. Overall, the random forest algorithm is a strong and adaptable method that may be used to solve classification issues in loan approval systems. The random forest algorithm builds a collection of decision trees using a set of randomly selected data and feature subsets. An arbitrary sample of the training data and a random subset of the characteristics are used to construct each decision tree. Combining all these distinct decision trees' projections yields the best results.



y = mode(T1(x), T2(x), ..., Tn(x))

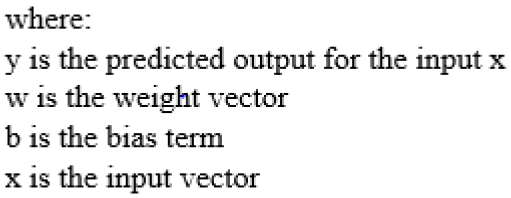


where:
y is the predicted output for the input x
T1, T2, ..., Tn are the individual decision trees in the forest.
mode is a function that returns the most frequently occurring output among the predictions of the individual trees.

(iv) **Support Vector Machine**: This non-parametric model performs well on datasets of small to medium size. It can handle interactions between characteristics that are both linear and non-linear.

Finding a hyperplane that best divides the data into distinct classes is how SVM operates (approved or rejected). The margin—the separation between the hyperplane and the nearest data points from each class—is maximised by selecting the hyperplane. When the data cannot be divided linearly, SVM employs a method known as the kernel strategy which move the datapoints into a higher-dimensional region where a hyperplane may divide it.

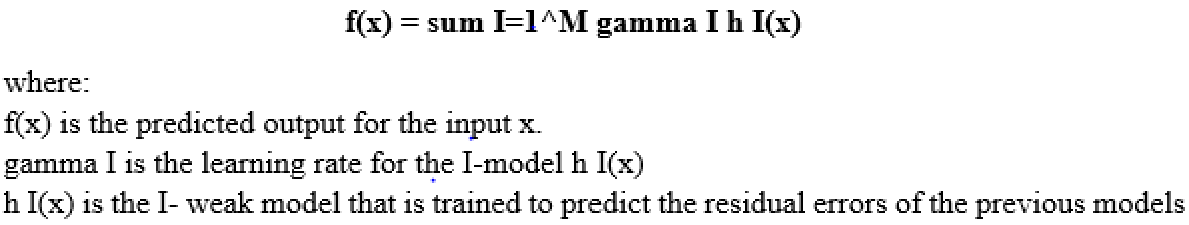
y = sign (w ^T x + b)



where:
y is the predicted output for the input x
w is the weight vector
b is the bias term
x is the input vector

(v) **Gradient Boosting Classifier**: This ensemble model also combines a number of weak learners to increase prediction accuracy. It can handle noisy data and performs well with large datasets.

When utilising gradient boosting, the algorithm is learned repeatedly on the training data, with each iteration concentrating on the data points that were incorrectly classified in the previous iteration. In an effort to fix the mistakes caused by the earlier trees in the ensemble, the algorithm builds a new decision tree.

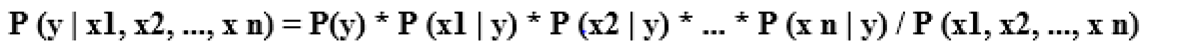


f(x) = sum I=1^M gamma I h I(x)

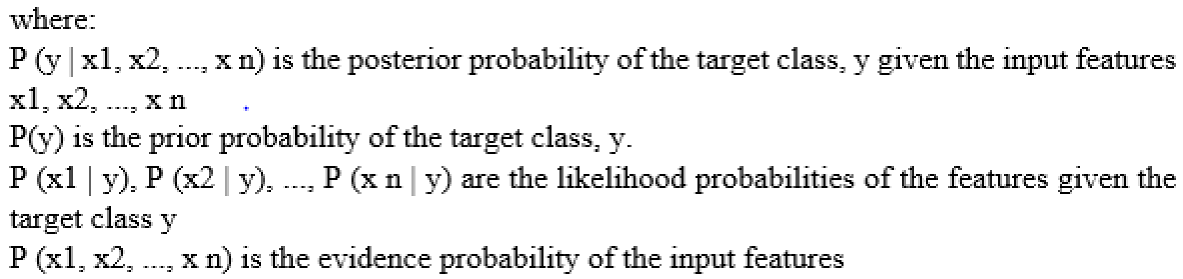
where:
f(x) is the predicted output for the input x.
gamma I is the learning rate for the I-model h I(x)
h I(x) is the I- weak model that is trained to predict the residual errors of the previous models

(vi) **Gaussian Naive Bayes**: It is a random method that determines, the fact that the values of each feature, how likely a given sample is to belong to each class.

The features inside each class in Gaussian Naive Bayes are thought to be independent and regularly distributed. From the training data, the algorithm learns the means and variances of each characteristic for each class. Using the Bayes theorem, it then utilises these parameters to predict the likelihood that a given sample belongs to each class. However, it may not perform well when the independence assumption is violated or when the features have non-Gaussian distributions.



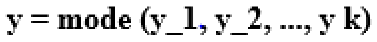
P (y | x1, x2, ..., x n) = P(y) * P (x1 | y) * P (x2 | y) * ... * P (x n | y) / P (x1, x2, ..., x n)



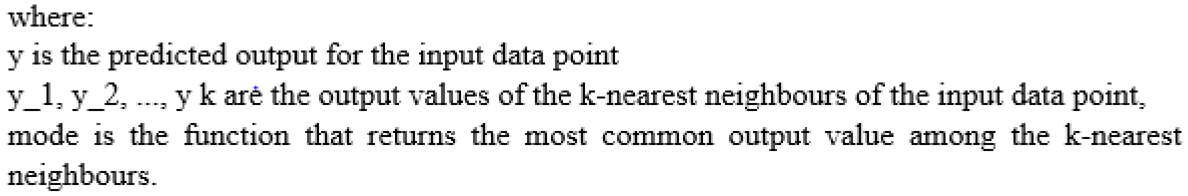
where:
P (y | x1, x2, ..., x n) is the posterior probability of the target class, y given the input features x1, x2, ..., x n
P(y) is the prior probability of the target class, y.
P (x1 | y), P (x2 | y), ..., P (x n | y) are the likelihood probabilities of the features given the target class y
P (x1, x2, ..., x n) is the evidence probability of the input features

(vii) **K-Nearest Neighbour**: The fundamental concept behind KNN is to categorise a new sample using the class labels of its k-nearest neighbours in the training set, where the distance between the new sample and the neighbours is calculated using a distance metric, such as Euclidean distance.

KNN uses the number k as a hyperparameter, which the user must select. Smoother choice borders can result from higher values of k, whereas more complex decision boundaries can result from lower values of k. However, it may not perform well when the dataset is imbalanced or whenever there are many characteristics, this will result in the curse of dimensionality.



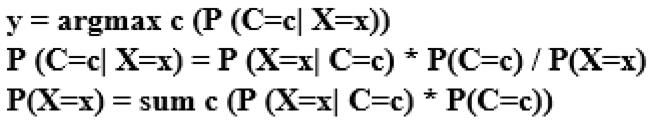
y = mode (y_1, y_2, ..., y k)



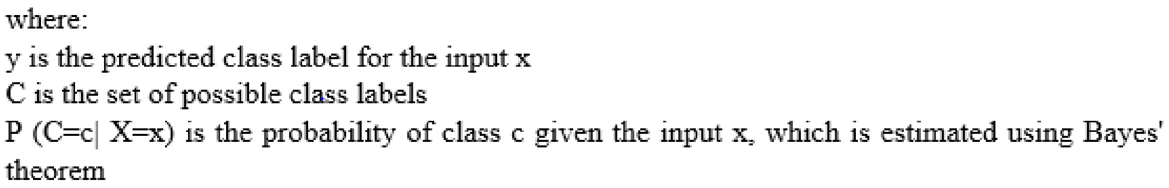
where:
y is the predicted output for the input data point
y_1, y_2, ..., y k are the output values of the k-nearest neighbours of the input data point,
mode is the function that returns the most common output value among the k-nearest neighbours.

(viii) **Linear Discriminant**: Finding an ordered set of attributes which most effectively distinguishes between different categories of loan approval outcomes is the fundamental goal of LDA. (Approved or rejected).

For each class of loan approval outcomes, the LDA method first produces the mean and covariance matrix. The linear combination of features that maximises the ratio of between-class variance to within-class variance is then determined by computing the within-class scatter matrices and the between-class scatter matrix. The discriminant function, which is a linear combination, can be used to categorise fresh loan applications according to their features. However, it may not perform well when the dataset is imbalanced or whenever there are many characteristics, this will result in the curse of dimensionality.



y = argmax c (P (C=c| X=x))
P (C=c| X=x) = P (X=x| C=c) * P(C=c) / P(X=x)
P(X=x) = sum c (P (X=x| C=c) * P(C=c))



where:
y is the predicted class label for the input x
C is the set of possible class labels
P (C=c| X=x) is the probability of class c given the input x, which is estimated using Bayes' theorem

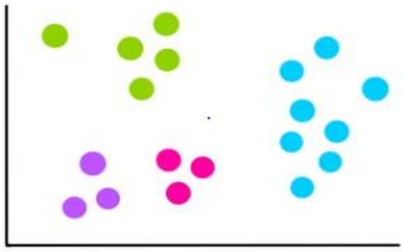
All the above mentioned 8 algorithms can be used binary classification problems. However, they differ in their assumptions about the data and the underlying relationships between features and the target variable. It is essential to assess each algorithm's performance using the proper criteria and choose the one that solves the particular problem at hand.

ALGORITHM SELECTION FOR MODEL 2:

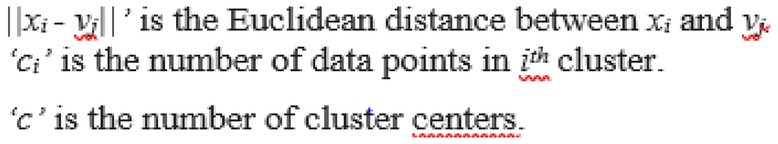
Customer segmentation is an essential phase in creating a system for bank loan authentication since it enables lenders to recognise and group borrowers according to their traits, actions, and credibility. Also, it enables applicants to recognize the reason for being approved and unapproved. To better meet the demands of each sector and reduce their vulnerability to default risk, lenders can achieve this by adjusting their lending methods and risk assessment methodologies.

When building a customer segmentation model, it is necessary to extract the most important features to do clustering on applicants. The clustering algorithm which we will using to build the customer segmentation model is given below:

K-Means Clustering: A dataset is divided into k clusters by the K-means cluster method, and each data point is given to the cluster with the nearest centre. The technique aims to maximise variation among clusters and minimise variance of data points inside each cluster.



$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\| x_i - v_j \| \right)^2$$



$||x_i$ - $v_j||$ ’ is the Euclidean distance between x_i and v_j.
‘c_i’ is the number of data points in i^{th} cluster.
‘c’ is the number of cluster centers.

7. MODEL TRAINING:

- **For Classification model:**

The model training starts with a phase of partitioning the datapoints. A training set and a testing set were used to partition the dataset into two halves. The testing set is utilized in evaluating the performance of the ML models after they have been taught with the training dataset. The problem at hand and the data that are accessible determine the ratio used to separate the dataset into training and testing sets.

The process of selecting the most essential features or predictors that have the strongest correlations with the target variable comes after model training. (i.e, the loan approval status). It can also be essential to use feature engineering to convert the raw data into a format that the machine learning models can use.

Predictive models can be created using a different kinds of ML algorithms, comprising logistic regression, decision trees, random forests, support vector classifier, gradient boosting, k-nearest neighbour, linear discriminant and gaussian naïve Bayes. Once the pertinent features have been found. Following training on the training set, these models are tested on the testing set by utilizing with various evaluation metrices, comprising accuracy, precision, recall, F1 score, and ROC curve.

Finally, the classifier which performed better on the testing dataset is chosen, then it is applied to future data forecasts. To prevent overfitting, it is crucial to remember that model selection should be based on the testing set performance rather than the training set performance.

Building a predictive model that accurately forecasts the loan approval status of new applicants based on their features and history is the overall objective of model training.

- **For Customer segmentation (clustering model):**

For training a cluster model , it is necessary to extract the most important features to do clustering on applicants.

The data is first prepared by selecting three relevant features: Applicant Income, Credit History, and Loan Amount. Two subsets of data are then created based on the Loan Status column: one subset for approved applicants and one subset for unapproved applicants.

K-means clustering is performed separately on each subset using the KMeans class which is present in the scikit-learn module. The n_clusters value is set to 3, which specifies that the algorithm should attempt to identify three clusters in the data. The random state parameter is set to 0 to ensure that the results are reproducible.

The clustering results are visualized using scatter plots. Each point in the scatter plot represents an applicant, and the color of the point indicates which cluster the applicant belongs to. The scatter plots are created using the matplotlib library.

Finally, the clusters are visualized in 3D scatter plots for both the approved and unapproved customers, where each point in the scatter plot represents a customer, and the color indicates the cluster that the customer belongs to. The red points in the scatter plots represent the centroids of each cluster.

K-means clustering can be used to identify patterns in loan approval data that could inform loan approval decisions. The trained K-means clustering models can be used to predict which cluster new loan applicants belong to, which could be useful in forecasting whether to accept their request or not.

8. MODEL TUNING:

Model tuning, sometimes referred to as hyperparameter tuning, is the process of choosing the appropriate collection of hyperparameters for a machine learning model to get the best results on a particular dataset. A range of probable values for each hyperparameter is chosen, and the model's effectiveness is evaluated using each plausible setting of the hyperparameters.

Numerous techniques, involving grid, random search, and Bayesian analysis, can be used to do this. Here, we prefer to use randomized search to tune our model to enhance the model's functionality.

(i) **Grid Search-** It includes evaluating each combination of hyperparameters that is likely to be within a certain range.

(ii) **Bayesian optimization-** It predicts the ideal set of hyperparameters based on past evaluations using probabilistic models.

(iii) **Randomized Search-** We first specify a range of values for each hyperparameter we want to modify in randomised search. After that, we choose a set of hyperparameters at random from this range, and we use cross-validation to assess the model's functionality. This method is repeated a predetermined number of times, trying a different random combination of hyperparameters each time. We choose the set of hyperparameters at the end of the process based on its performance on the validation set.
For hyperparameter tuning, randomised search is a valuable method, particularly in high-dimensional hyperparameter spaces where grid search would be computationally impractical. It can enhance the effectiveness of machine learning models and offers a good trade-off between efficiency and search quality.

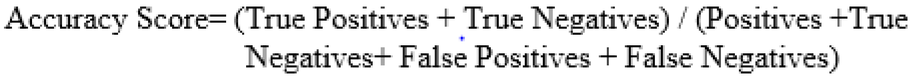
9. MODEL EVALUATION:

Evaluation metrices plays a huge role in evaluating a model's performance and comparing the performance of several models. It's necessary to select appropriate measurements based on the current problem and to interpret the metrics in light of the problem.

- **For Classification Model 1:**

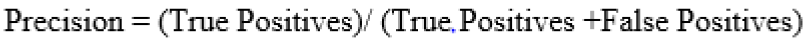
The model Evaluation can be done by some of the popular metrices which are listed below **:**

(i) **Accuracy:** Accuracy score is defined as how many occurrences were accurately categorised as a percentage of all the instances. Accuracy of the model can be found by the below given equation.



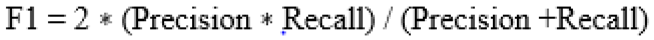
Accuracy Score= (True Positives + True Negatives) / (Positives +True Negatives+ False Positives + False Negatives)

(ii) **Precision:** The ratio of real positive results (positive instances that were correctly recognised) to the total number of anticipated positive results is known as the precision score. The provided algorithm can be used to determine the model's accuracy.



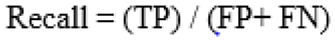
Precision = (True Positives)/ (True Positives +False Positives)

(iii) **F1 Score:** The balance between recall and precision is provided by the harmonic average of precision and recall. F1 score of the model can be found from the equation given below.



F1 = 2 * (Precision * Recall) / (Precision +Recall)

(iv) **Recall:** Recall can be defined as the ratio of real positives to all occasions where a result was truly positive. Recall of the model can be found by the following equation.



Recall = (TP) / (FP+ FN)

(v) **AUC-ROC:** The region that lies below the receiver operating characteristic graph assesses the trade-off within real positives and invalid positives.

(vi) **Cross Validation Score**: It is a a technique that involves folding the data into numerous folds, training the model on various fold combinations, then testing the model on the remaining fold.

CV Score = 1/k * ∑I =1 to k (score I)

where k is the number of folds in the cross-validation, and scorei is the evaluation metric for the ith fold of the data.

(vii) **Confusion Matrix**: A table that displays the real positives, invalid positives, real negatives, and false negatives for each classifier to indicate how well it performed.

When the set of data is short or the model needs to be adjusted for a large number of hyperparameters, cross-validation score is especially helpful. Since cross-validation is based on numerous training/testing splits of the data, it allows you to acquire more approximate estimate of the model's functionality.

So, in our Evaluation we consider cross validation score as the most important metric to compare the performance of various models.

- **For Customer Segmentation Model 2:**

Based on the analysis's specific objectives and problem domain, evaluation criteria for clustering models are chosen. The silhouette score, Calinski-Harabasz index, and Davies-Bouldin index are a few often utilized measures. These metrics are useful for comparing various clustering algorithms or hyperparameter settings as well as to assess the quality and coherence of the generated clusters.

The Evaluation metric we used in our model 2 is given below:

Silhouette Score: It is a measure of an object's unity with its own cluster in comparison to other clusters. (separation). By giving each data point a score based on how far away from the next closest cluster it is from its own, it provides an estimate of the quality of clustering. This score can be found from the equation given below.

Silhouette Score= (p-q) / max (p, q)

Here, p = mean distance to the points in the nearest cluster

q = mean intra-cluster distance to all the points.

Values range from -1 to 1, with higher scores indicating better-defined clusters. In general, a score above 0.5 is considered a good result for clustering.

Score +1 – indicates that the sample data is far away from the nearest cluster.

Score 0 - indicates that the sample data is on or closer to the boundary of two nearest.

cluster.

Score -1 – indicates that the sample data is employed to the wrong cluster.

10. MODEL DEPLOYMENT

After the model is evaluated, the ideal model is selected and saved for deploying our proposed bank loan authentication system. For our proposed system , I have chosen to deploy my model by creating web-based interface using stream lit.

STREAM LIT:

For creating engaging data science apps in Python, I consider using Streamlit. It offers a straightforward method for quickly creating web applications. By using Streamlit we can design a web-based interface for applicants to apply for loans and for loan officers to assess and approve those applications in the context of a loan approval system.

Once the stramlit application has been deployed, it can be tested to ensure everything is operating as it was expected. Also need to to be monitored to check any tuning needed to the model.

Using Streamlit to implement a loan approval system can be an outstanding method to make your machine learning model more widely used. Designing a user-friendly interface with Streamlit makes it simple for users to submit loan applications and for loan officers to examine and approve those applications.

5.4 UML DIAGRAMS FOR OUR PROPOSED SYSTEM

(i) Overall methodology:

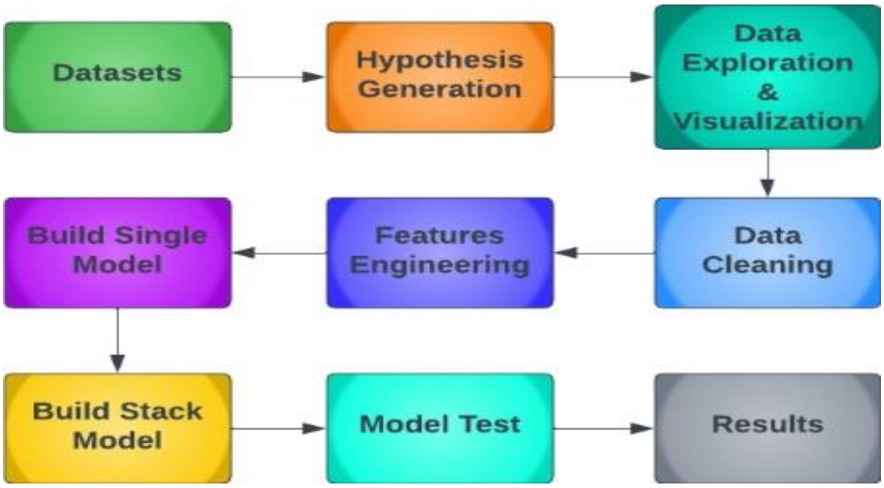


Fig 3 Methodology (Outline)

(ii) Model Building:

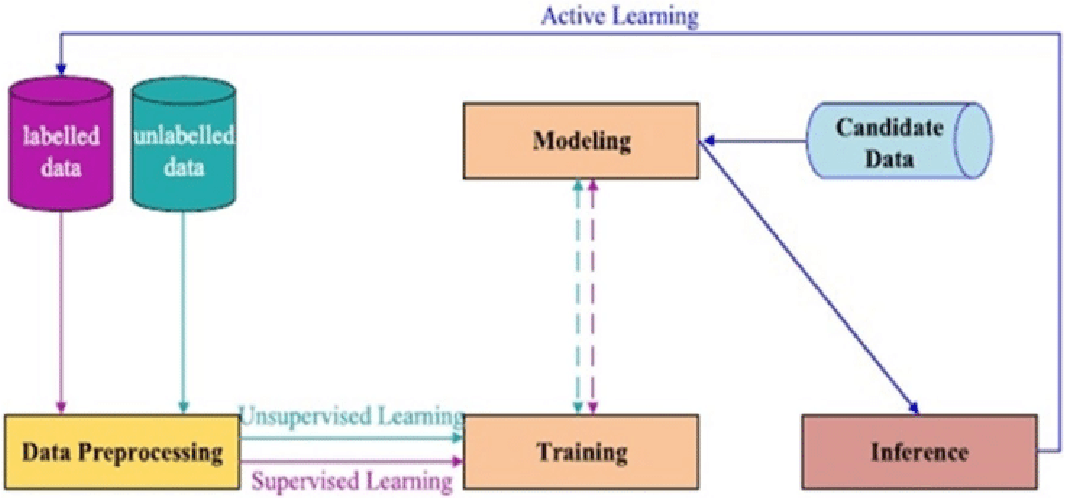


Fig 4. Model Building

(iii) Model training process:

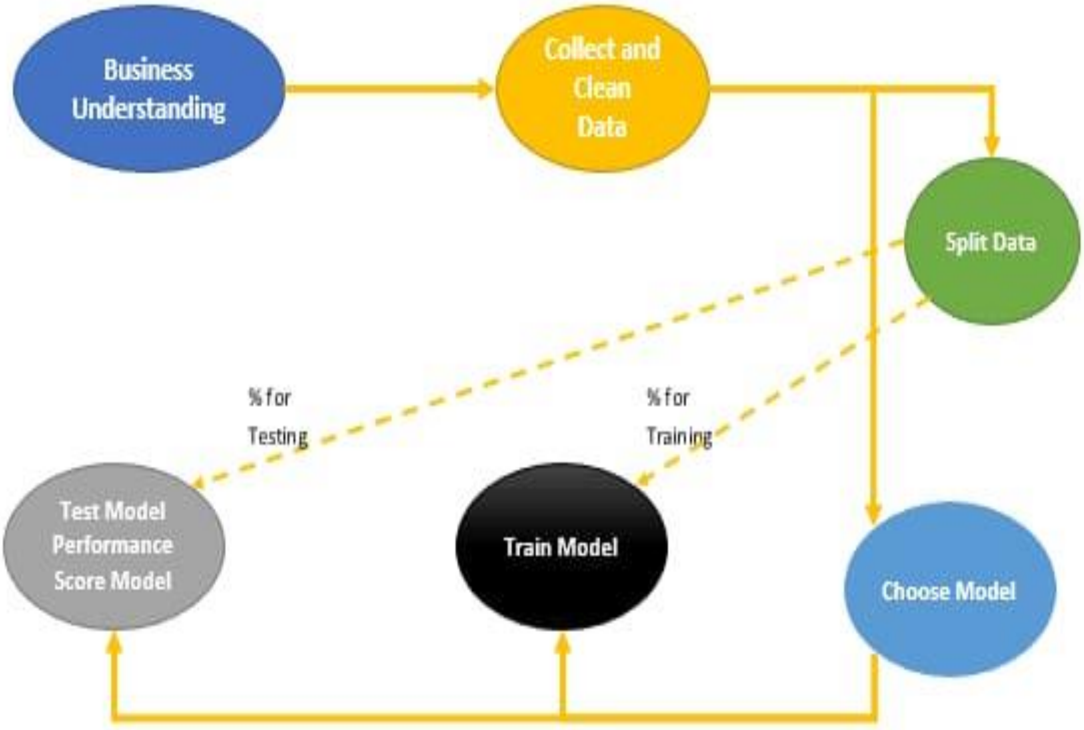


Fig.5. Model Training

(iv) Use Case Diagram :

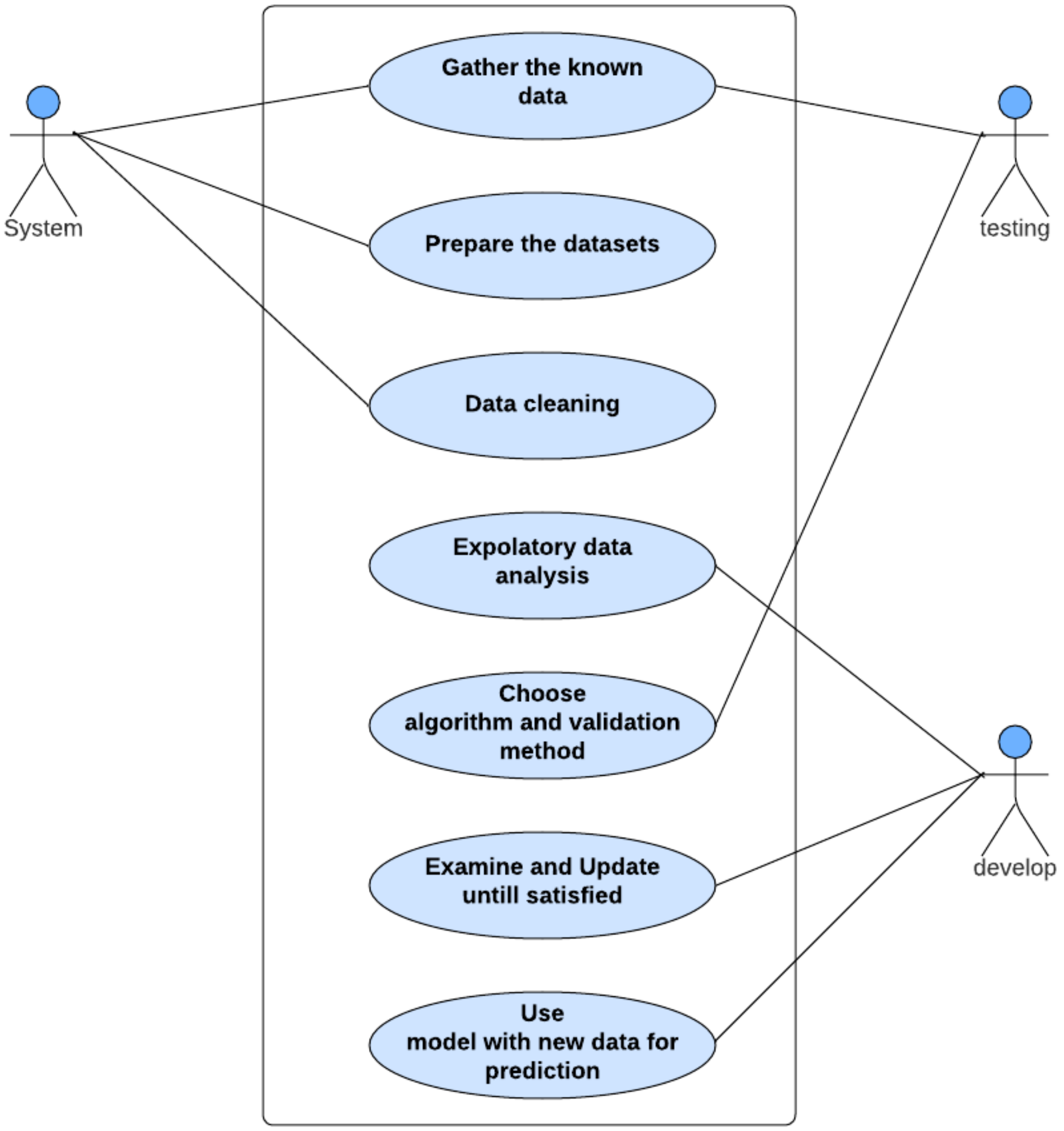
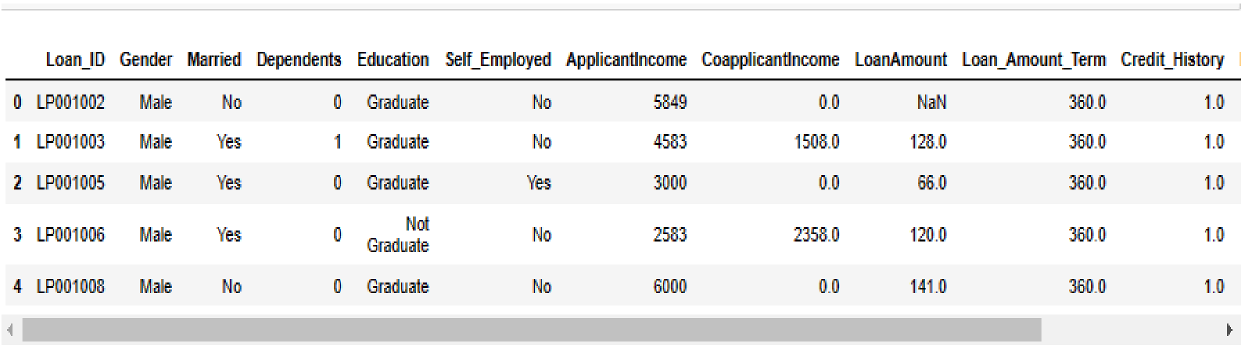

System
Gather the known data
Prepare the datasets
Data cleaning
Expolatory data analysis
Choose algorithm and validation method
Examine and Update untill satisfied
Use model with new data for prediction
testing
develop

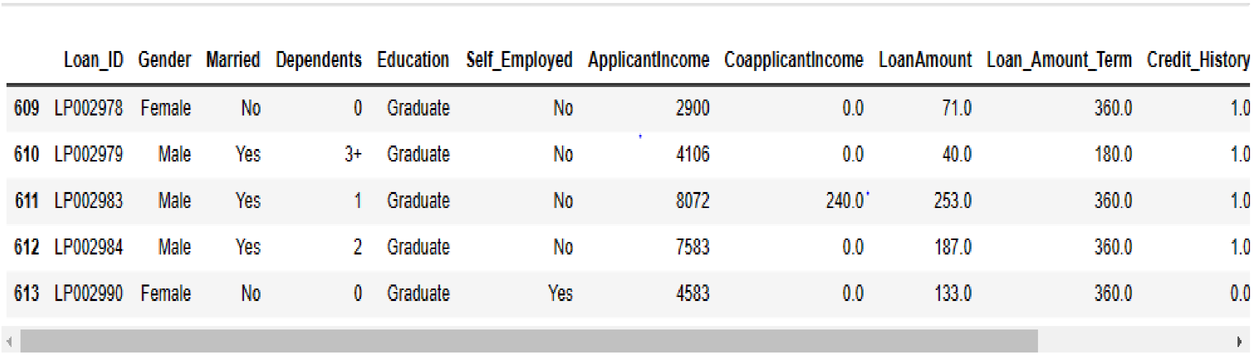
Fig 5. Use Case Diagram

7. PROJECT OUTPUTS

1. Reading the dataset (First and Last 5 rows)

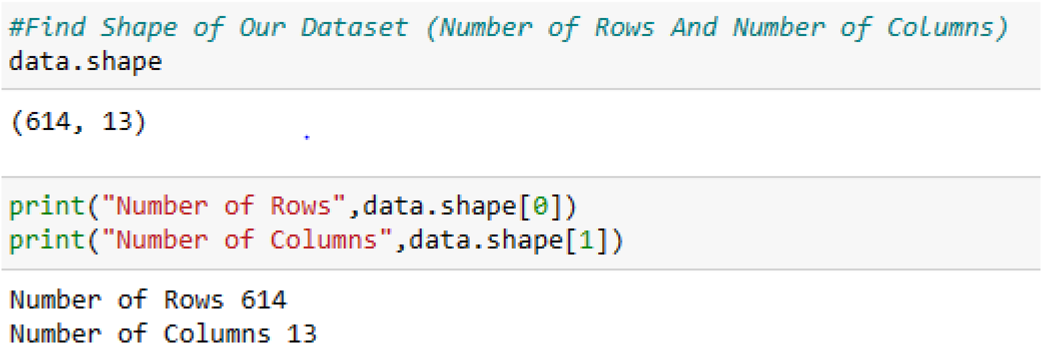


	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0



	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
609	LP002978	Female	No	0	Graduate	No	2900	0.0	71.0	360.0	1.0
610	LP002979	Male	Yes	3+	Graduate	No	4106	0.0	40.0	180.0	1.0
611	LP002983	Male	Yes	1	Graduate	No	8072	240.0	253.0	360.0	1.0
612	LP002984	Male	Yes	2	Graduate	No	7583	0.0	187.0	360.0	1.0
613	LP002990	Female	No	0	Graduate	Yes	4583	0.0	133.0	360.0	0.0

2. Finding Shape of the dataset:



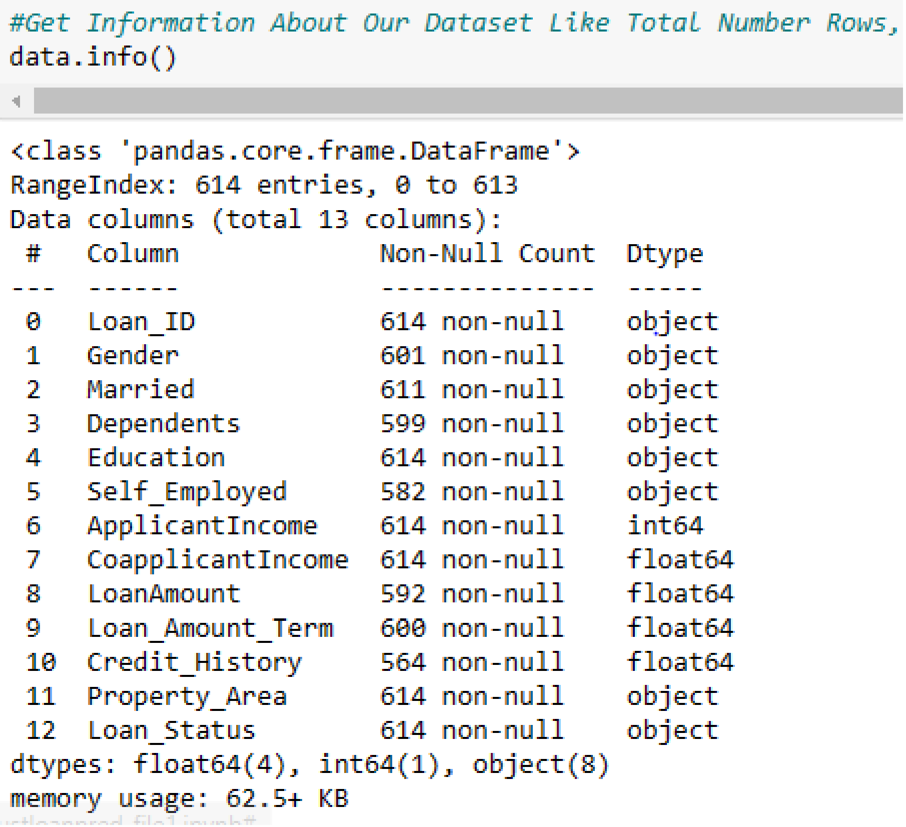
```
#Find Shape of Our Dataset (Number of Rows And Number of Columns)
data.shape
```

```
(614, 13)
```

```
print("Number of Rows",data.shape[0])
print("Number of Columns",data.shape[1])
```

```
Number of Rows 614
Number of Columns 13
```

3. Getting Dataset Information:

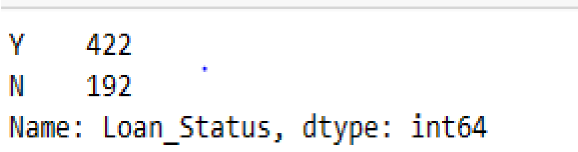


```
#Get Information About Our Dataset Like Total Number Rows,
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 614 entries, 0 to 613
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Loan_ID            614 non-null    object 
 1   Gender             601 non-null    object 
 2   Married            611 non-null    object 
 3   Dependents         599 non-null    object 
 4   Education          614 non-null    object 
 5   Self_Employed      582 non-null    object 
 6   ApplicantIncome    614 non-null    int64  
 7   CoapplicantIncome  614 non-null    float64
 8   LoanAmount         592 non-null    float64
 9   Loan_Amount_Term   600 non-null    float64
 10  Credit_History     564 non-null    float64
 11  Property_Area      614 non-null    object 
 12  Loan_Status        614 non-null    object 
dtypes: float64(4), int64(1), object(8)
memory usage: 62.5+ KB
```

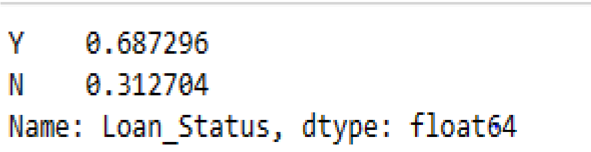
4. Univariate Analysis for the target variable (Loan _status):

Loan approved and unapproved counts:



```
Y    422
N    192
Name: Loan_Status, dtype: int64
```

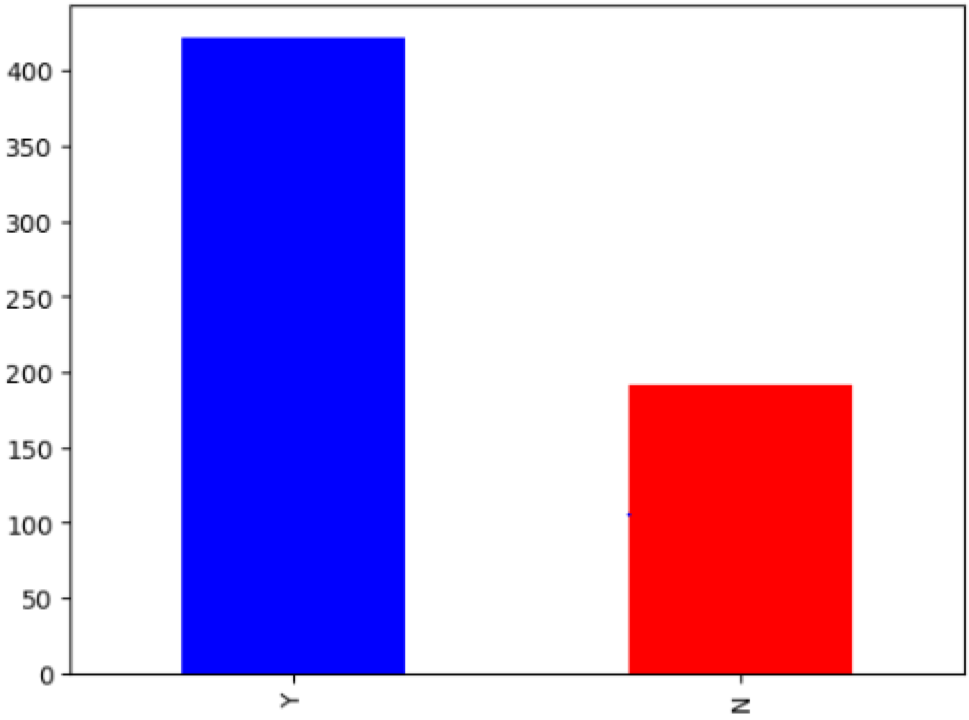
Normalized counts:



```
Y    0.687296
N    0.312704
Name: Loan_Status, dtype: float64
```

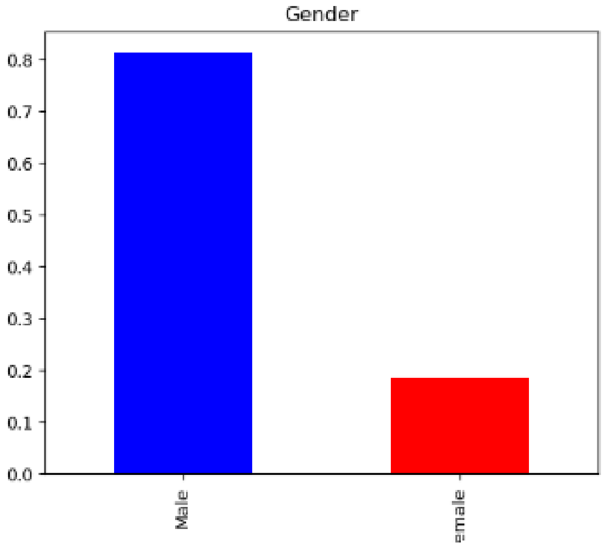
Bar plot for the target variable(Loan_status):

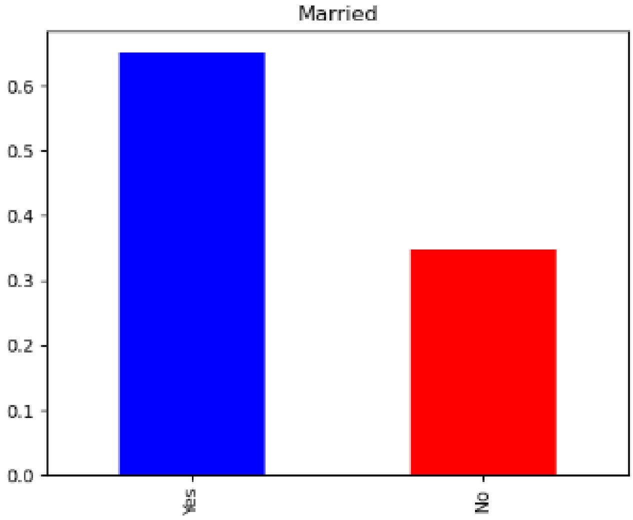
```
<AxesSubplot:>
```



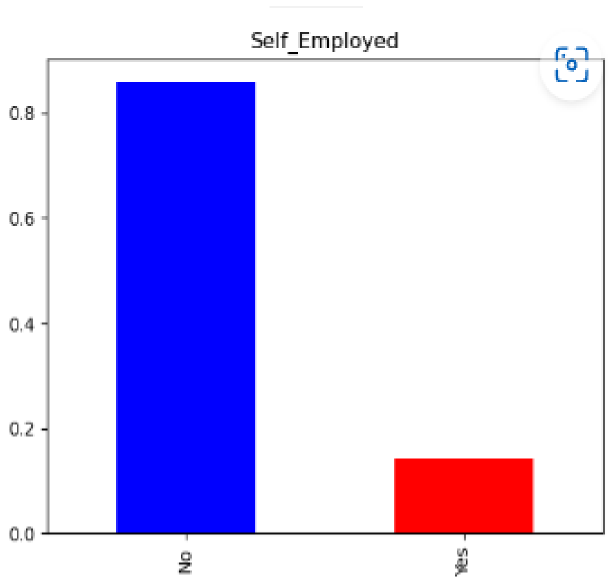
Univariate Analysis for the Categorical variable (Gender, Married, Self_Employed, Credit_History, Education):

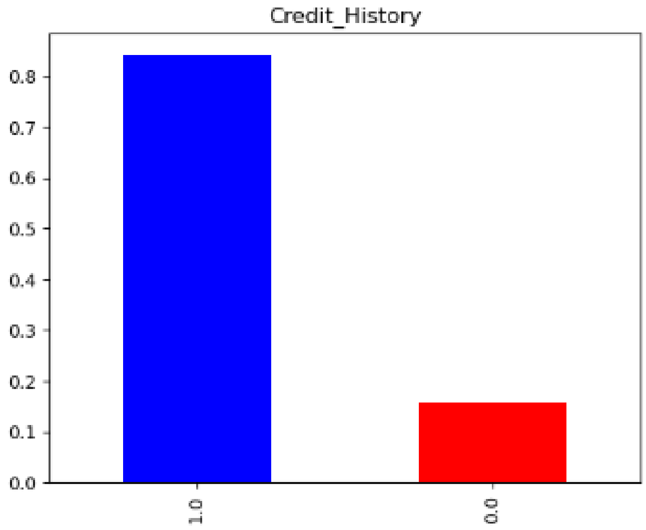
(i) Gender **(ii) Married:**



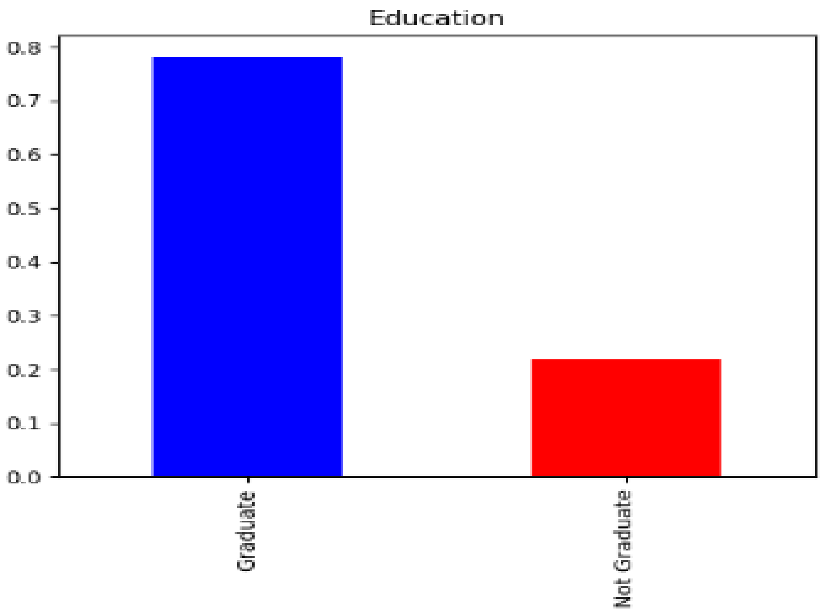


(iii)Self_Employed **(iv) Credit_History**



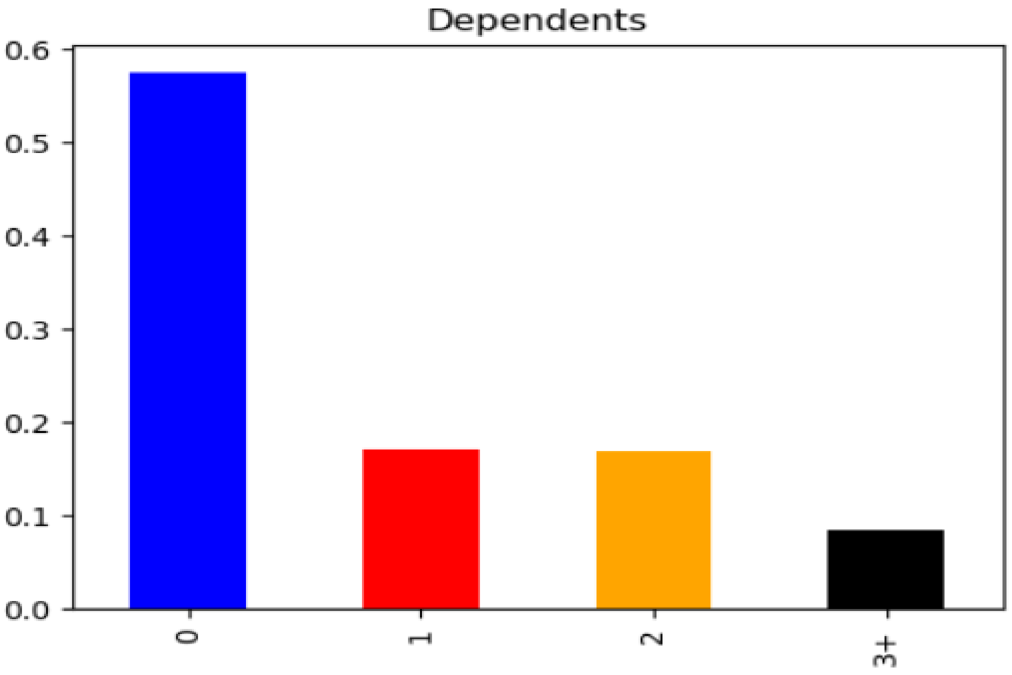


(v)Education

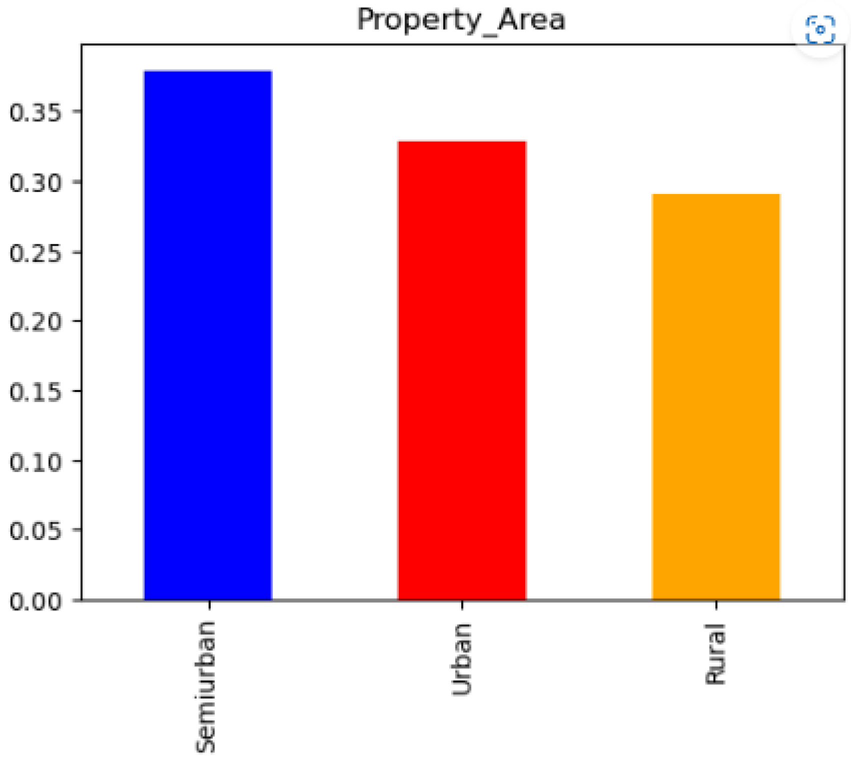


ORDINAL VARIABLE (Dependants, Property_Area)

(vi)Dependants

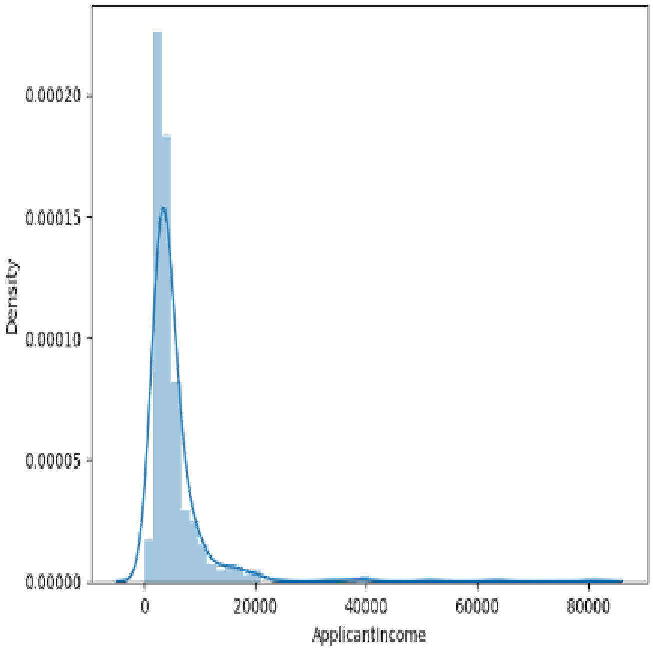


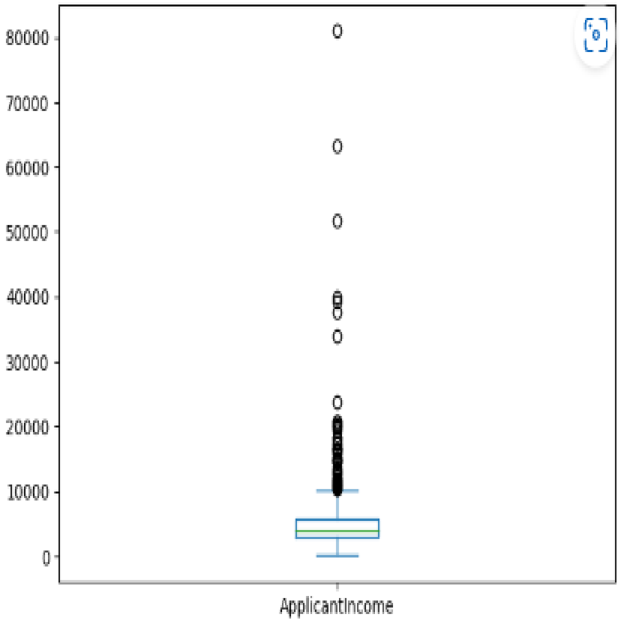
(vii)Property_Area



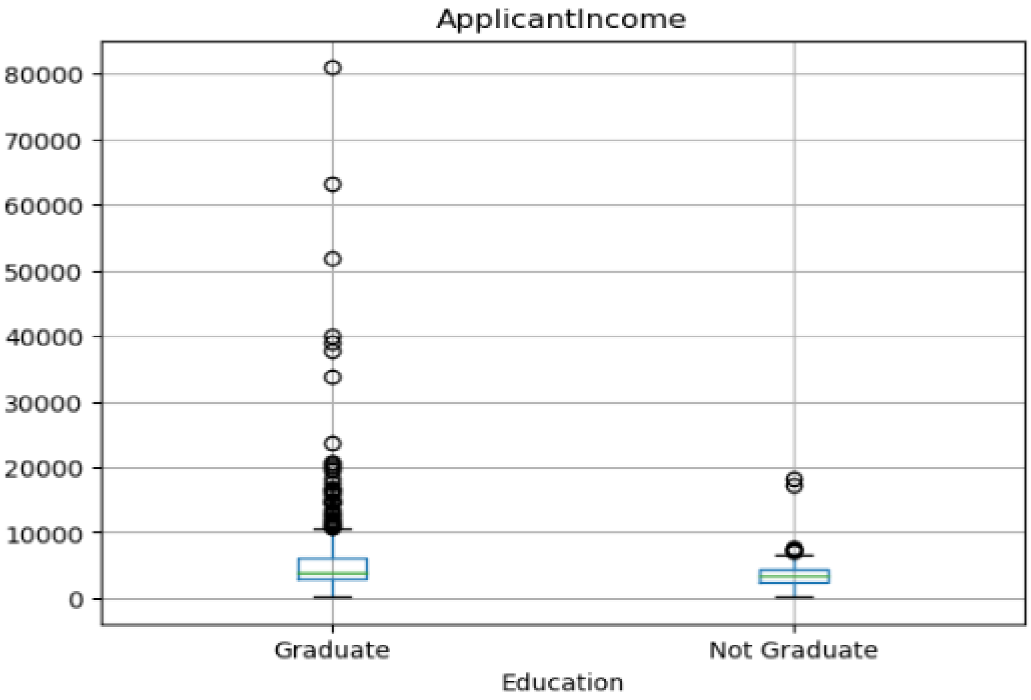
NUMERICALVARIABLE(ApplicantIncome,LoanAmount,Loan_Amount_Term):

(viii)ApplicantIncome

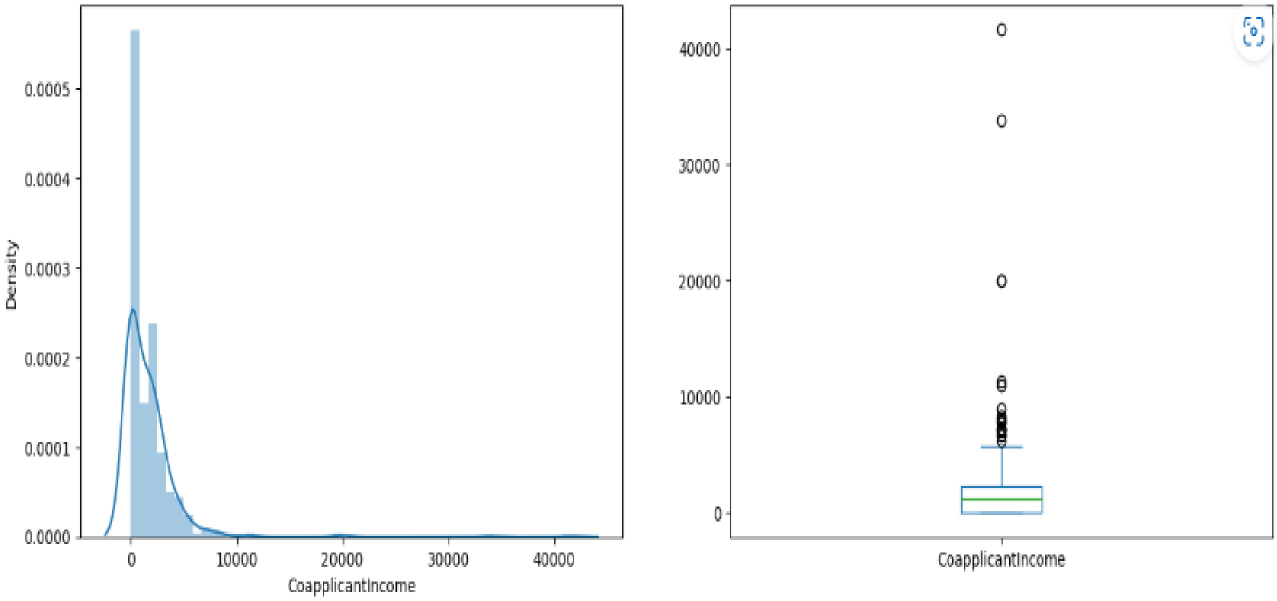




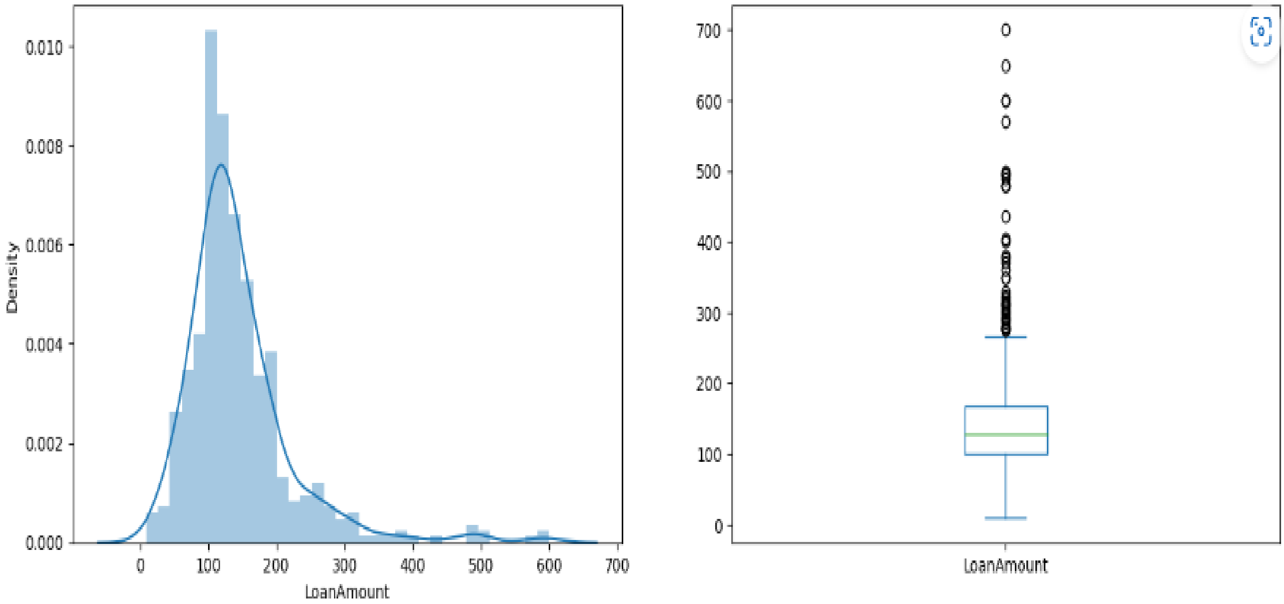
(ix)Applicant income by Education



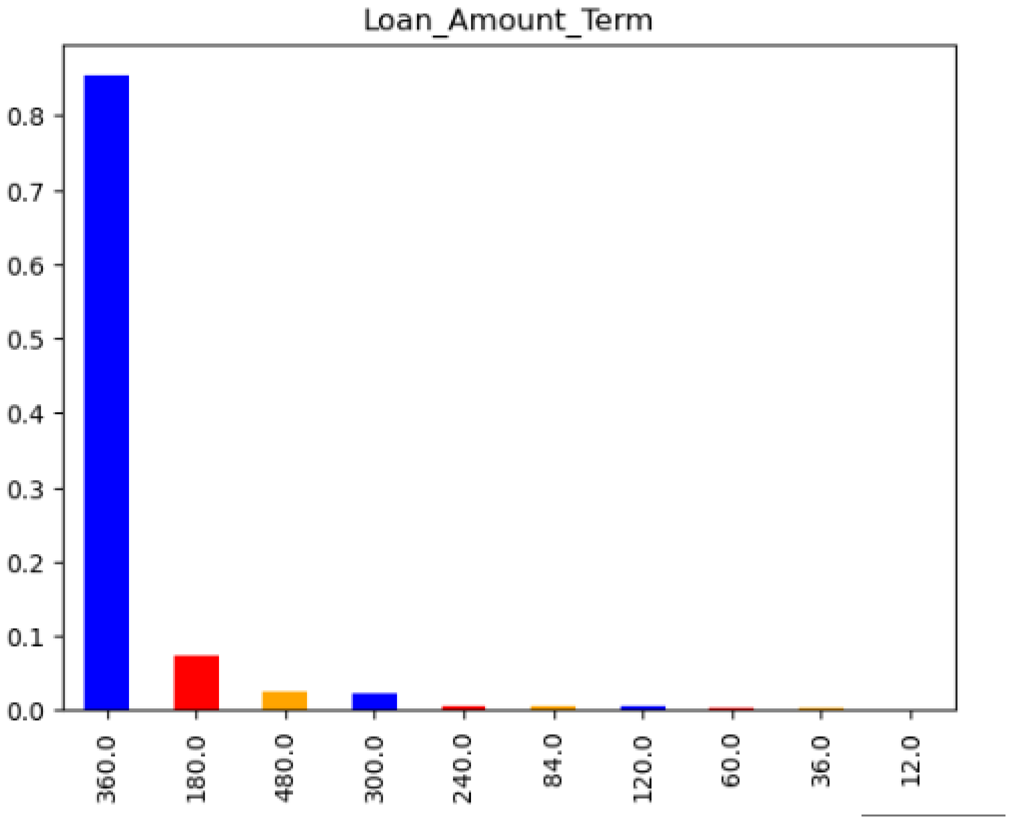
(x)CoapplicantIncome:



(x)LoanAmount:

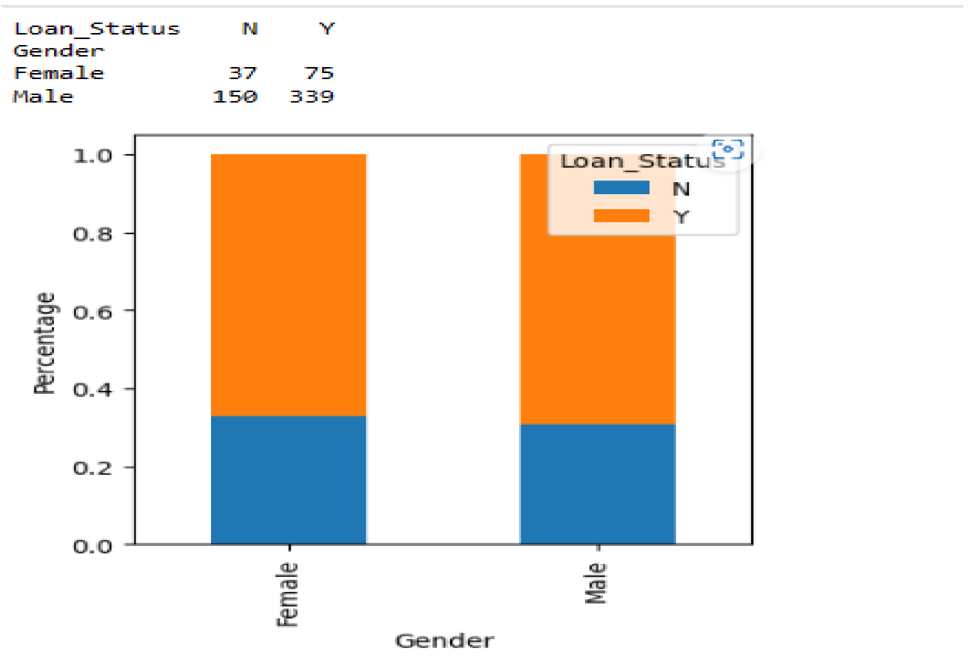


(xi)Loan_Amount_Term:



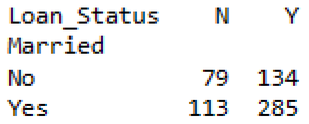
5. BIVARIATE ANALYSIS (Categorical Vs Target Variable):

(i) Gender Vs Loan_Status

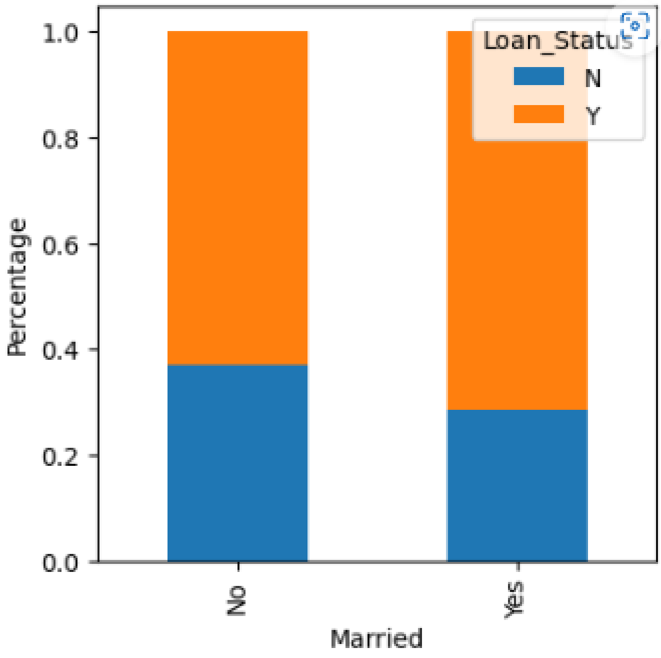


```
Loan_Status     N    Y
Gender
Female         37   75
Male          150  339
```

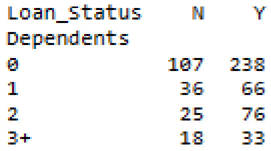
(ii) Married Vs Loan_Status



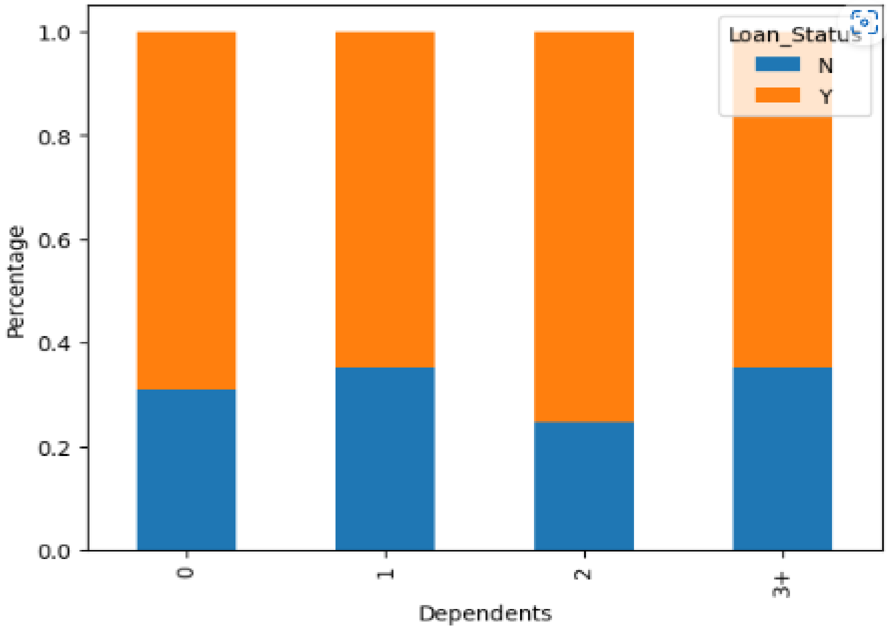
```
Loan_Status    N    Y
Married
No            79  134
Yes          113  285
```



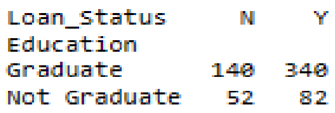
(iii) Dependants Vs Loan_Status



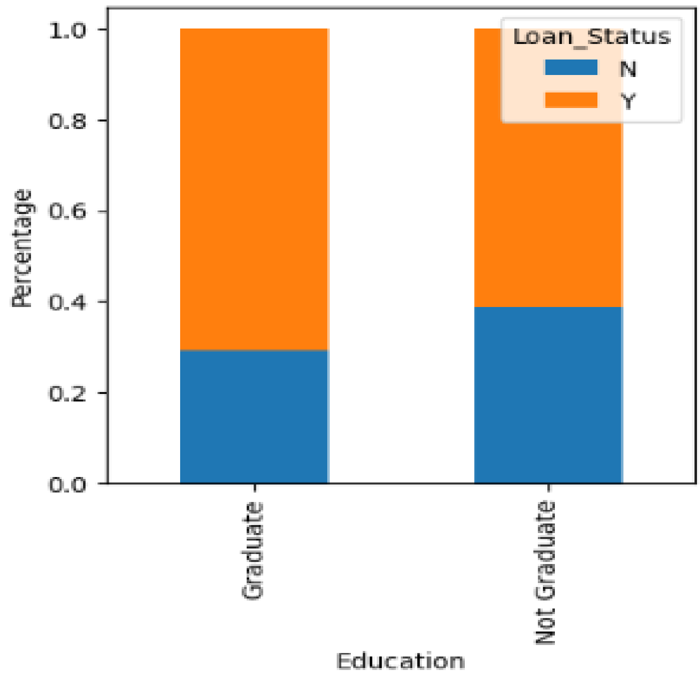
```
Loan_Status    N    Y
Dependents
0            107  238
1             36   66
2             25   76
3+            18   33
```



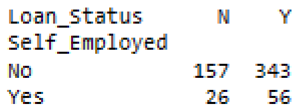
(iv) Education Vs Loan_Status



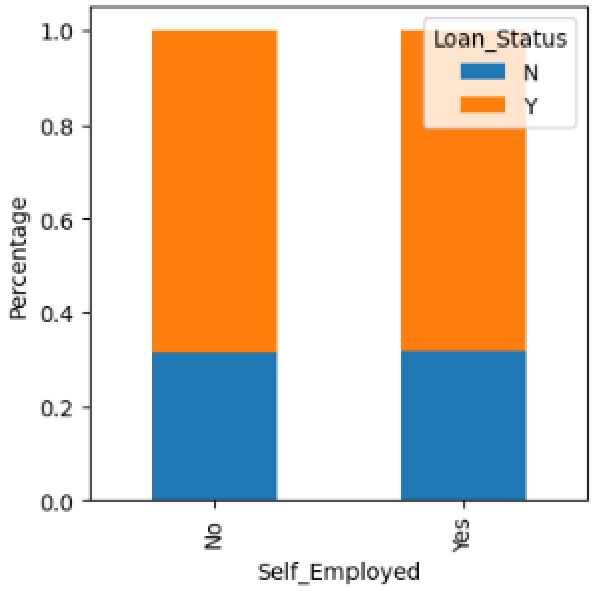
```
Loan_Status     N    Y
Education
Graduate      140  340
Not Graduate   52   82
```



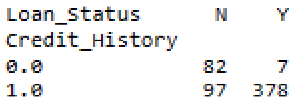
(v) Self_Employed Vs Loan_Status



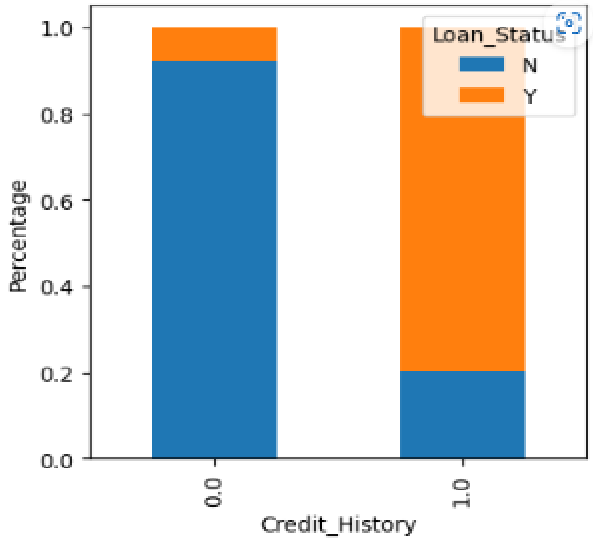
```
Loan_Status      N    Y
Self_Employed
No             157  343
Yes             26   56
```



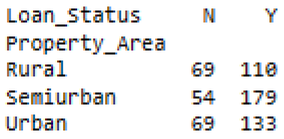
(vi) Credit_History Vs Loan_Status



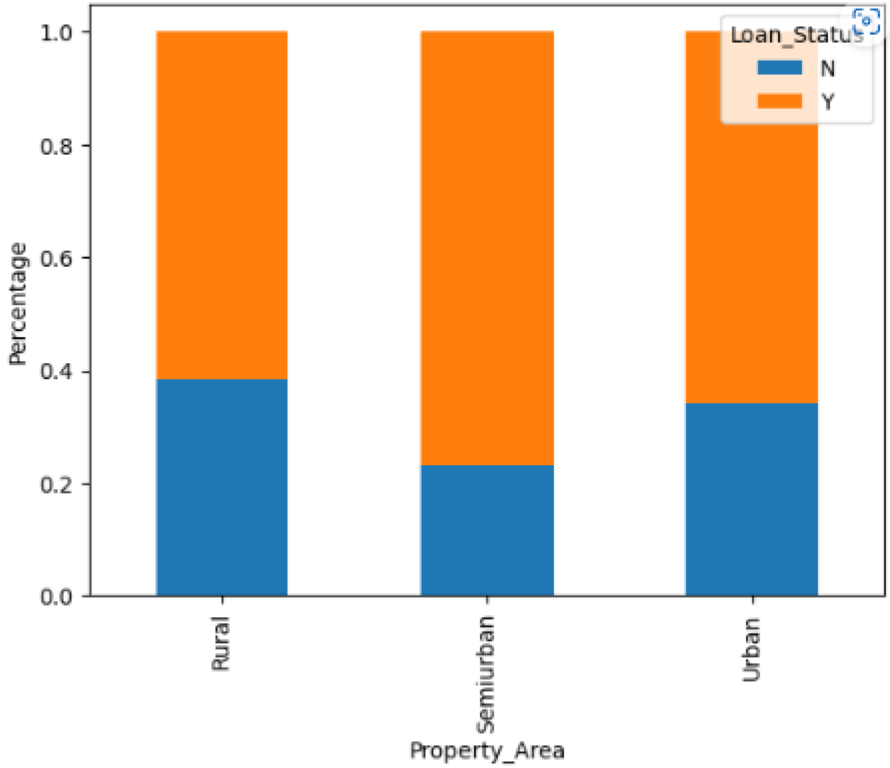
```
Loan_Status      N    Y
Credit_History
0.0             82    7
1.0             97  378
```



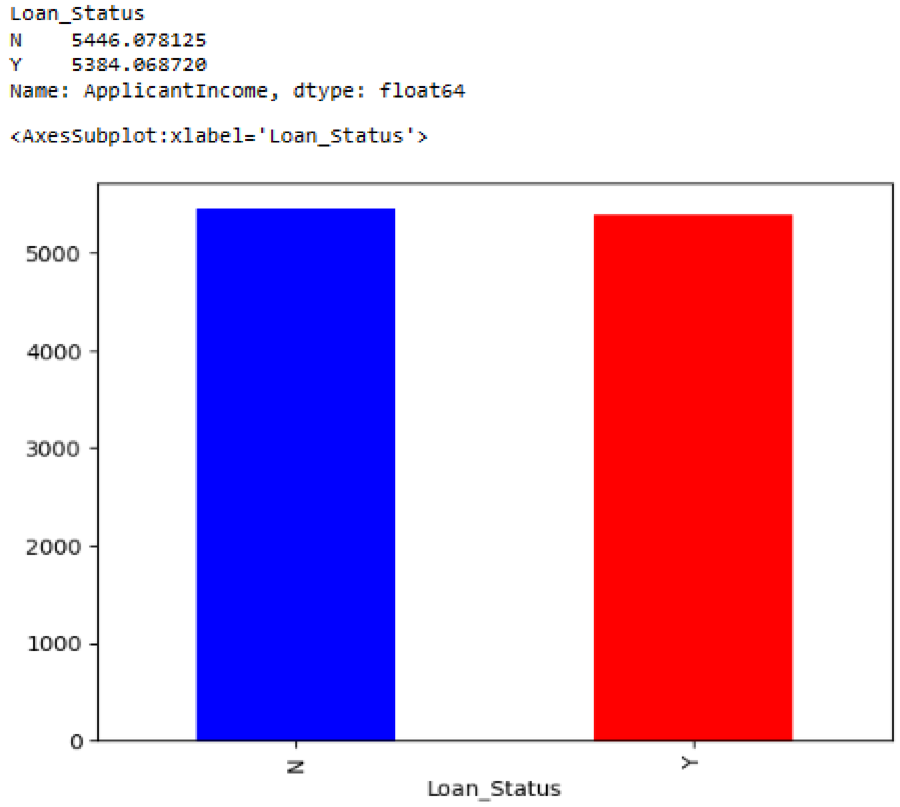
(vii) Property_Area Vs Loan_Status



```
Loan_Status    N    Y
Property_Area
Rural         69  110
Semiurban     54  179
Urban         69  133
```



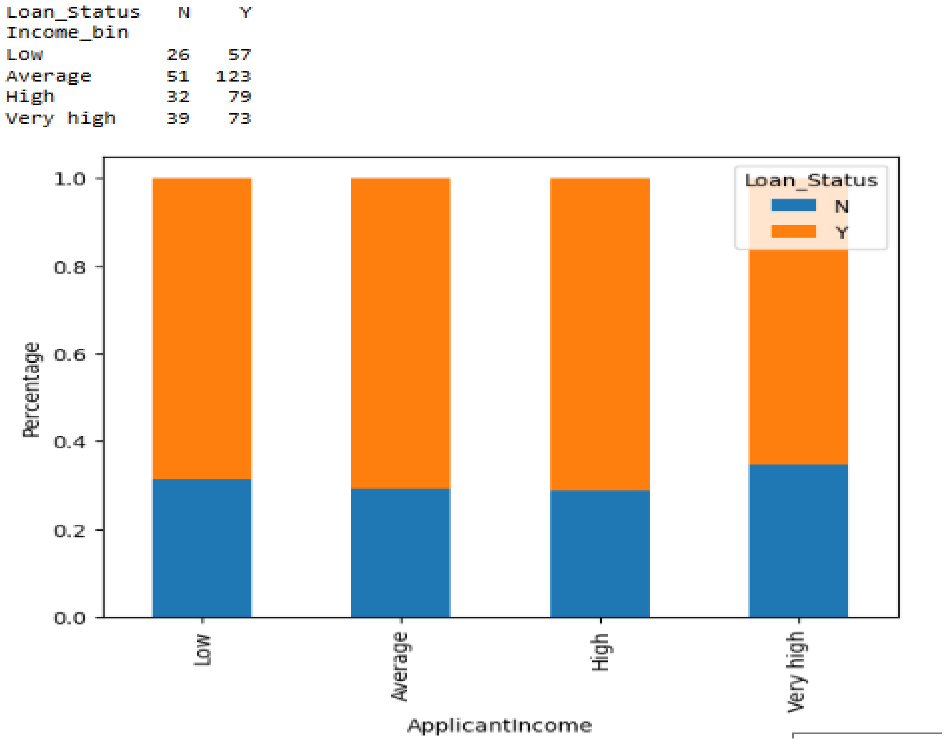
(viii) Numerical Vs Target Variable(ApplicantIncome Vs Loan_Status)



```
Loan_Status
N    5446.078125
Y    5384.068720
Name: ApplicantIncome, dtype: float64
```

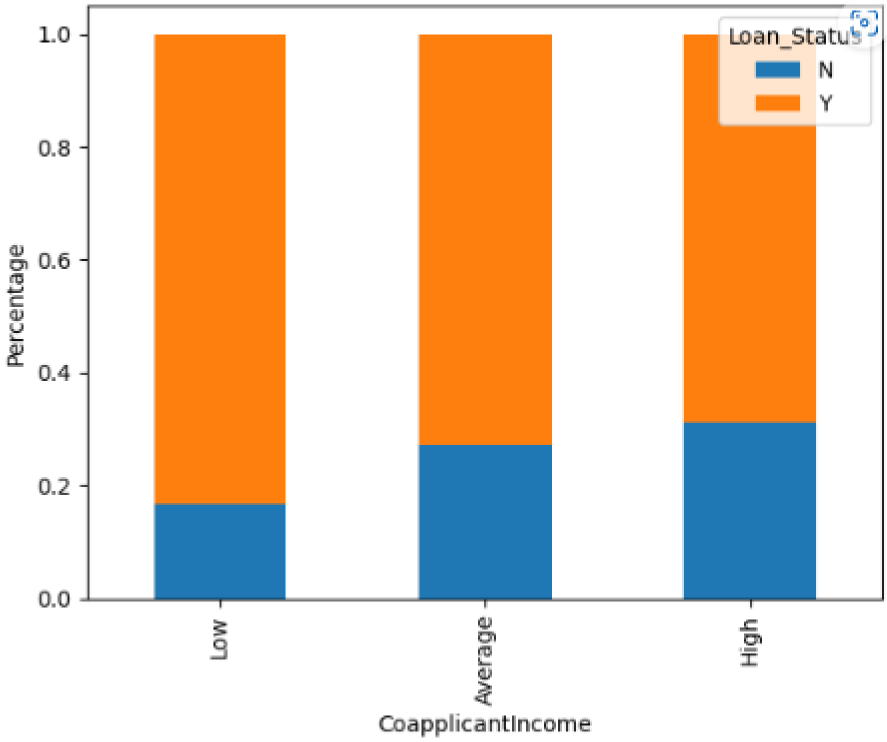
```
<AxesSubplot:xlabel='Loan_Status'>
```

(ix) BINNING PROCESS: (Income_bin Vs Loan_Status)

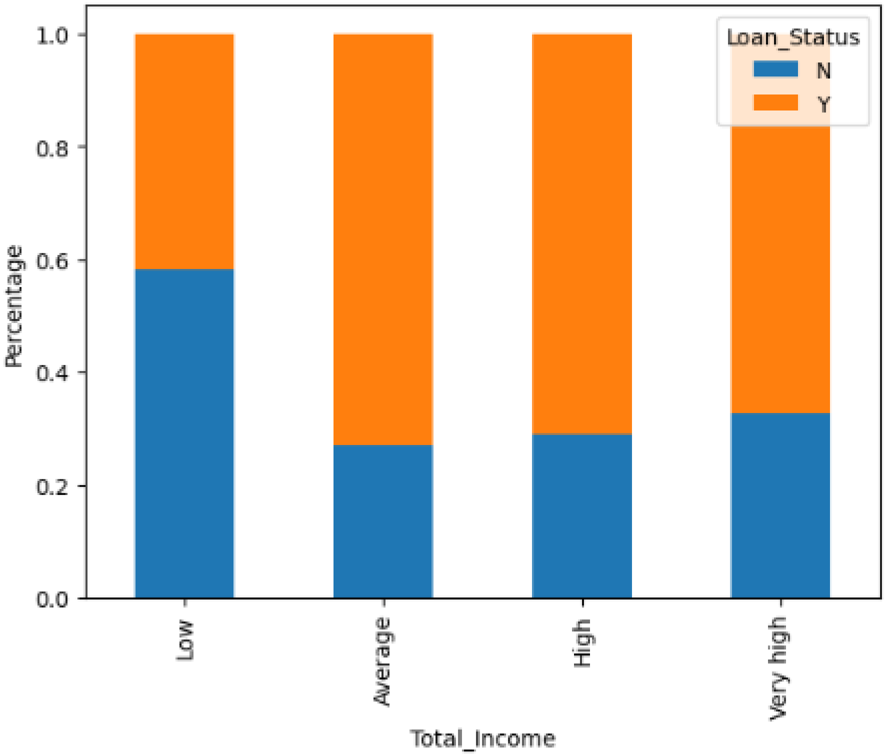


```
Loan_Status    N    Y
Income_bin
Low           26   57
Average       51  123
High          32   79
Very high     39   73
```

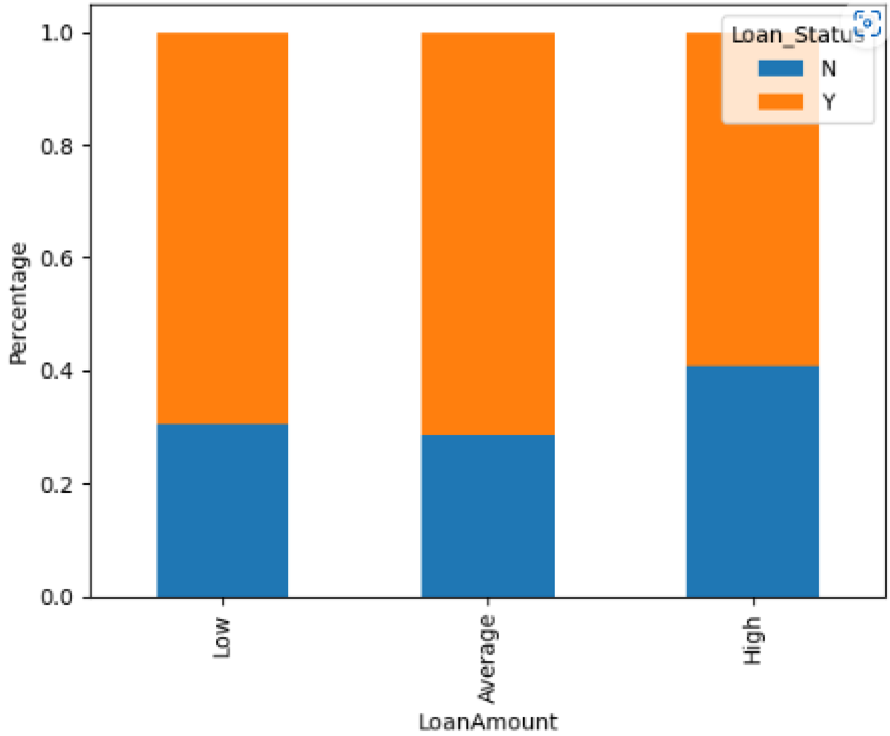
(x) Coapplicant_Income_bin Vs Loan_Status



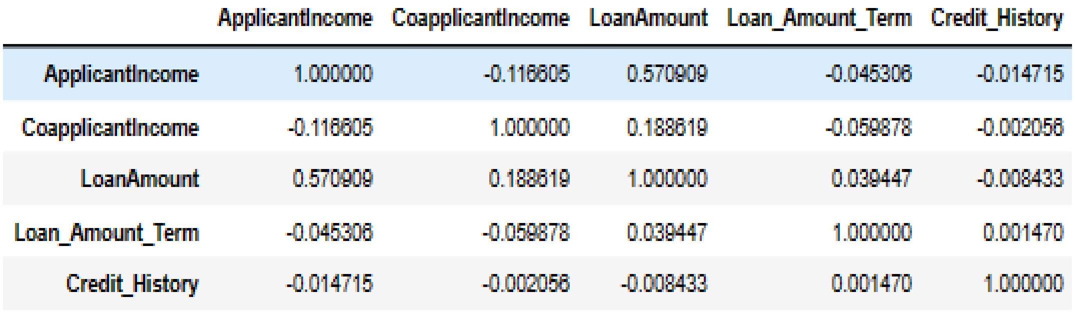
(xi) Total_Income _bin Vs Loan_Status;



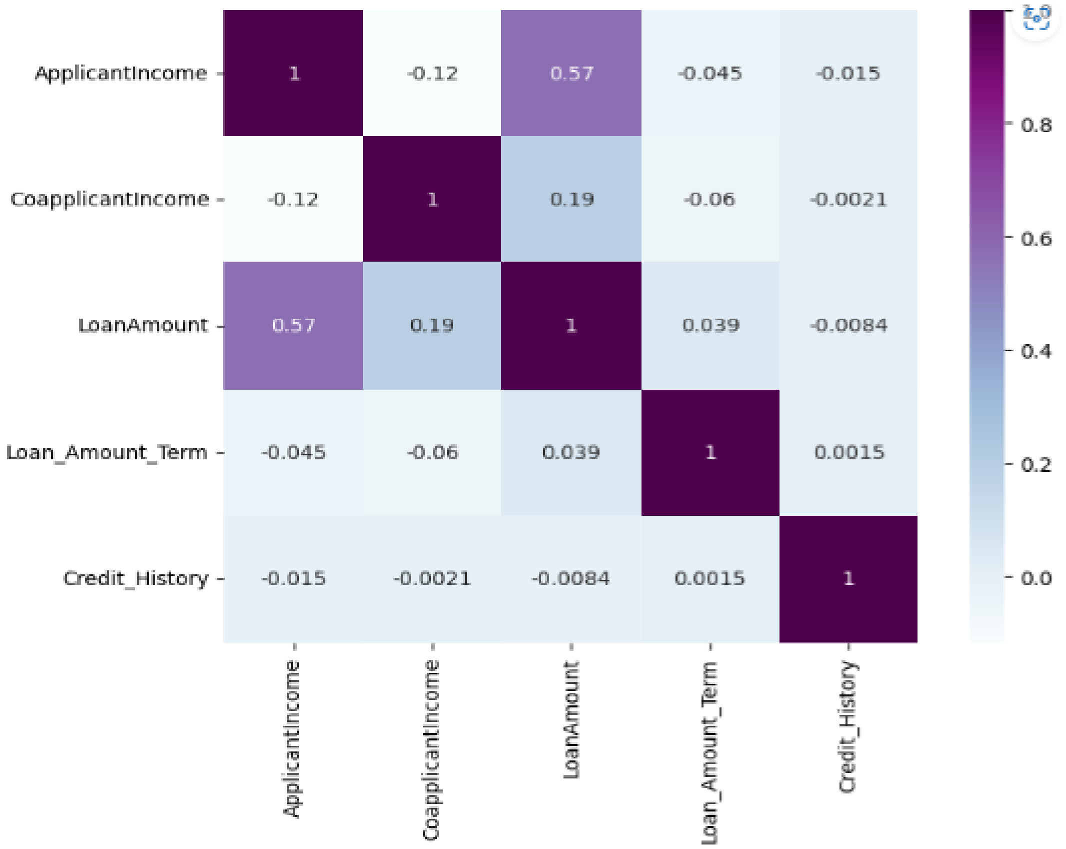
(xii) Loan_Amount_bin Vs Loan_status:



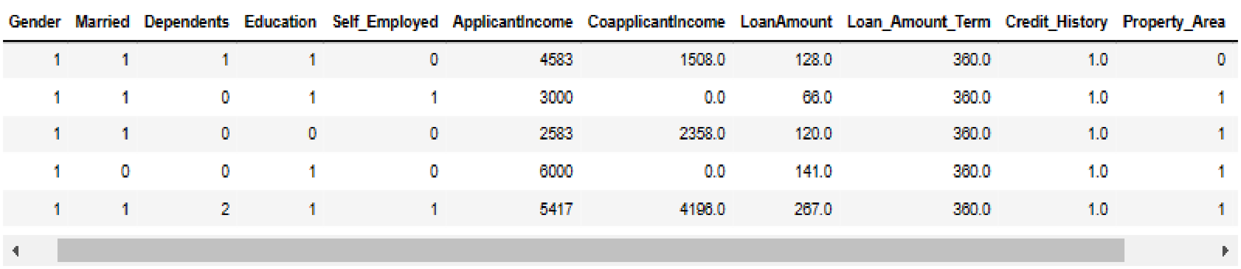
6. CORRELATION MATRIX:



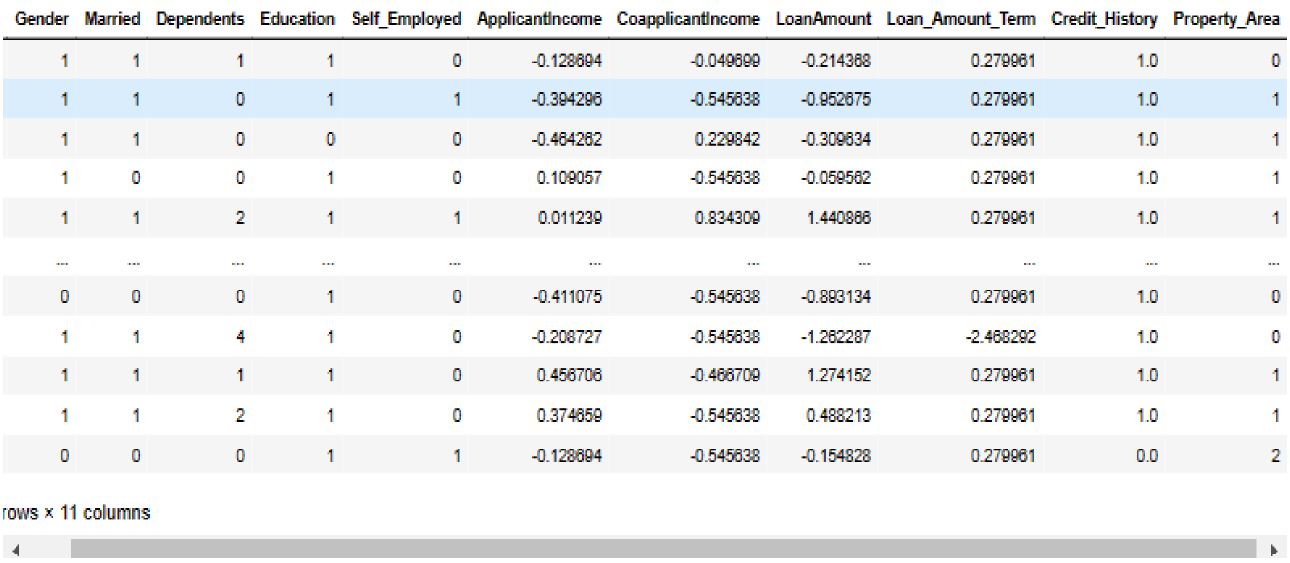
	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
ApplicantIncome	1.000000	-0.116605	0.570909	-0.045306	-0.014715
CoapplicantIncome	-0.116605	1.000000	0.188619	-0.059878	-0.002056
LoanAmount	0.570909	0.188619	1.000000	0.039447	-0.008433
Loan_Amount_Term	-0.045306	-0.059878	0.039447	1.000000	0.001470
Credit_History	-0.014715	-0.002056	-0.008433	0.001470	1.000000



7. FEATURE SCALING:



Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
1	1	1	1	0	4583	1508.0	128.0	360.0	1.0	0
1	1	0	1	1	3000	0.0	66.0	360.0	1.0	1
1	1	0	0	0	2583	2358.0	120.0	360.0	1.0	1
1	0	0	1	0	6000	0.0	141.0	360.0	1.0	1
1	1	2	1	1	5417	4196.0	267.0	360.0	1.0	1

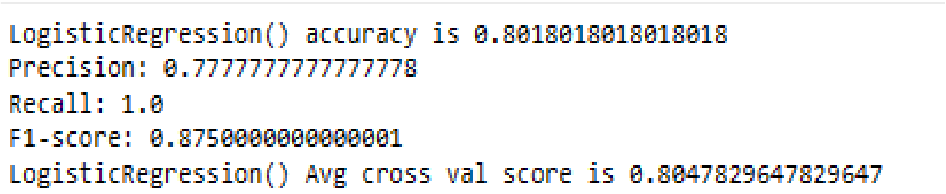


Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
1	1	1	1	0	-0.128694	-0.049699	-0.214368	0.279961	1.0	0
1	1	0	1	1	-0.394296	-0.545638	-0.952675	0.279961	1.0	1
1	1	0	0	0	-0.464262	0.229842	-0.309634	0.279961	1.0	1
1	0	0	1	0	0.109057	-0.545638	-0.059562	0.279961	1.0	1
1	1	2	1	1	0.011239	0.834309	1.440866	0.279961	1.0	1
...	...	...	...	...	...	...	...	...	...	...
0	0	0	1	0	-0.411075	-0.545638	-0.893134	0.279961	1.0	0
1	1	4	1	0	-0.208727	-0.545638	-1.262287	-2.468292	1.0	0
1	1	1	1	0	0.456706	-0.466709	1.274152	0.279961	1.0	1
1	1	2	1	0	0.374659	-0.545638	0.488213	0.279961	1.0	1
0	0	0	1	1	-0.128694	-0.545638	-0.154828	0.279961	0.0	2

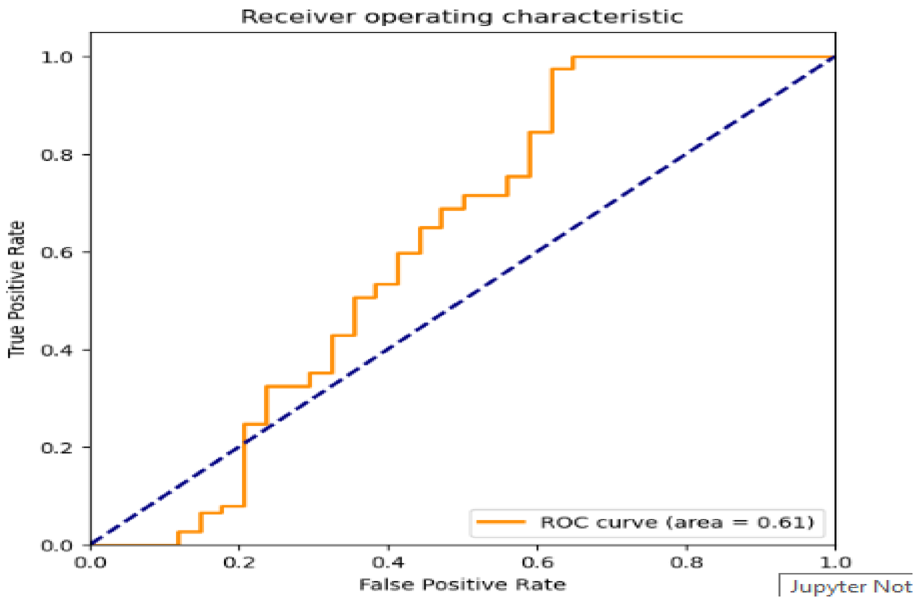
rows × 11 columns

8. MODEL BUILDING

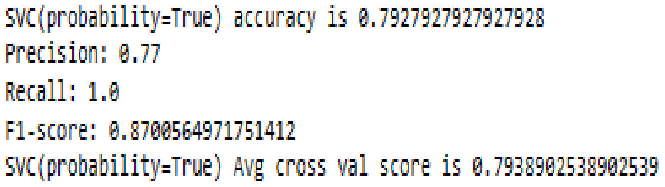
(i) Logistic Regression:



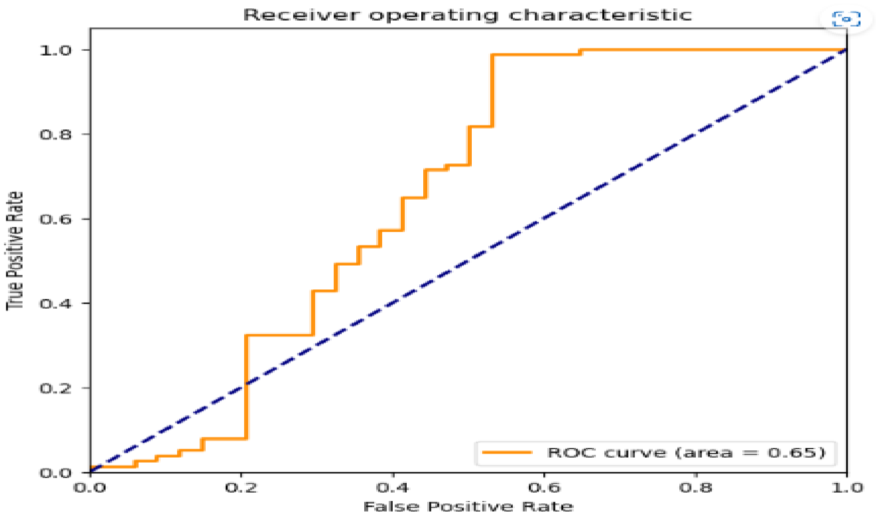
```
LogisticRegression() accuracy is 0.8018018018018018
Precision: 0.7777777777777778
Recall: 1.0
F1-score: 0.8750000000000001
LogisticRegression() Avg cross val score is 0.8047829647829647
```



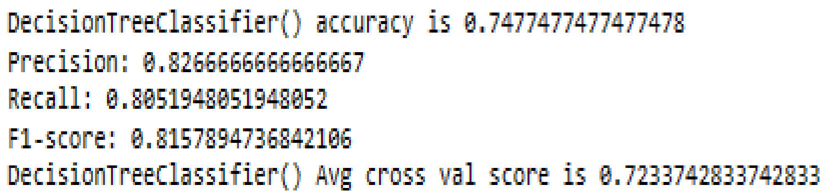
(ii) Support Vector Machine:



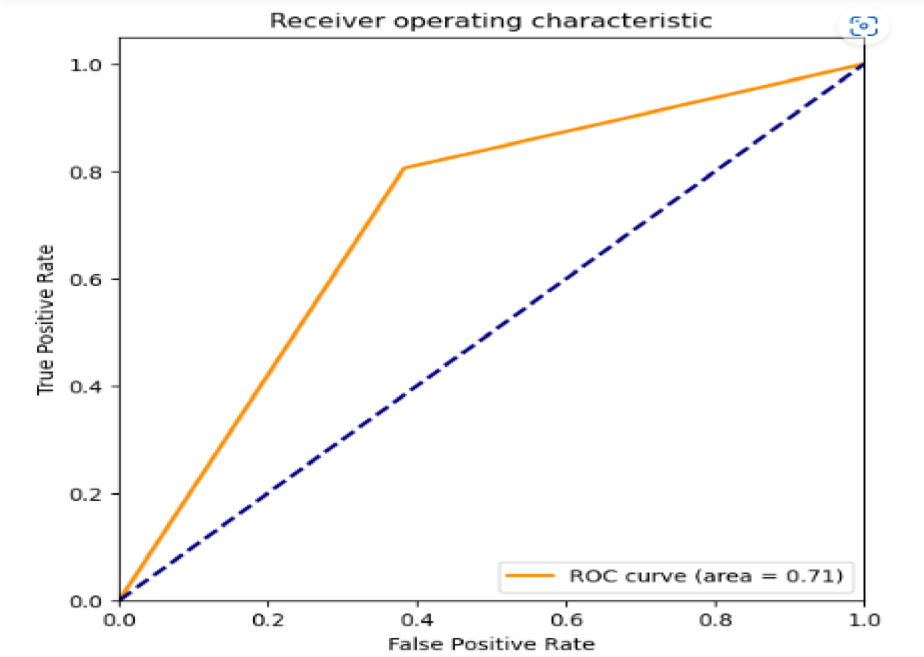
```
SVC(probability=True) accuracy is 0.7927927927927928
Precision: 0.77
Recall: 1.0
F1-score: 0.8700564971751412
SVC(probability=True) Avg cross val score is 0.7938902538902539
```



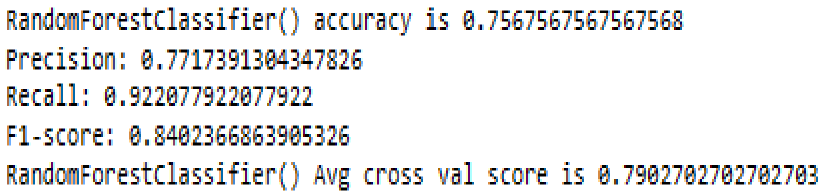
(iii) Decision Tree:



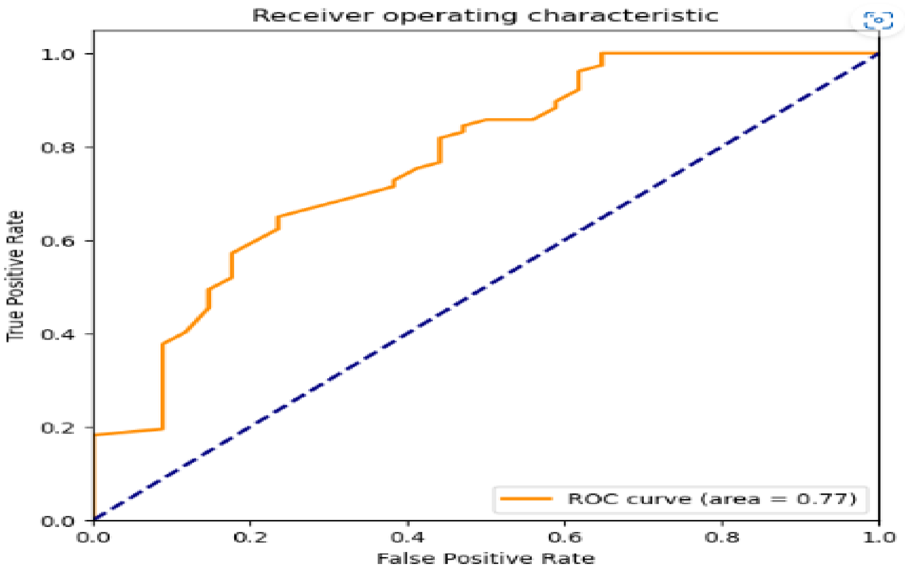
```
DecisionTreeClassifier() accuracy is 0.7477477477477478
Precision: 0.8266666666666667
Recall: 0.8051948051948052
F1-score: 0.8157894736842106
DecisionTreeClassifier() Avg cross val score is 0.7233742833742833
```



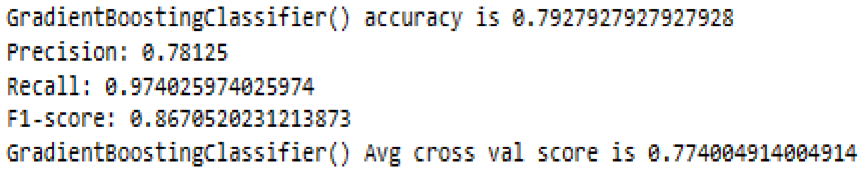
(iv) Random Forest:



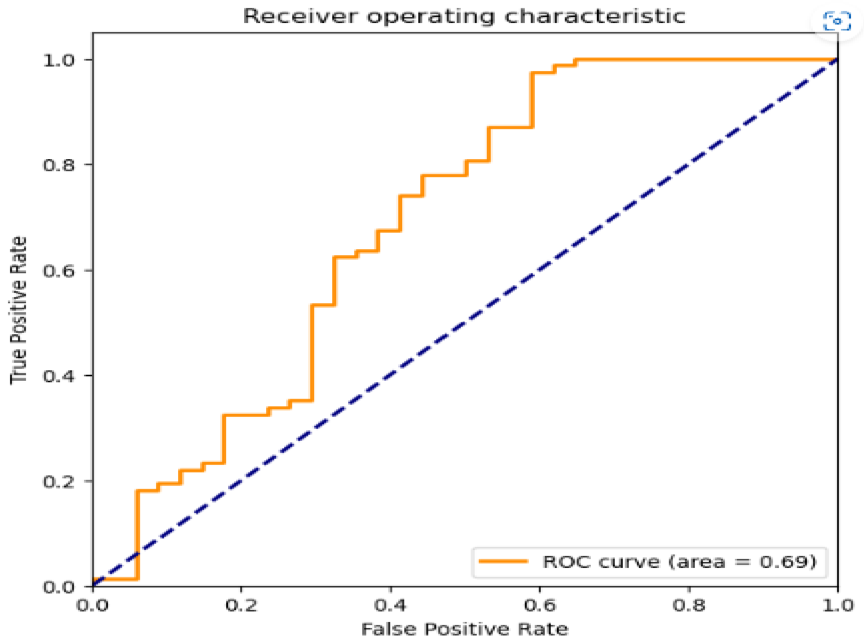
```
RandomForestClassifier() accuracy is 0.7567567567567568
Precision: 0.7717391304347826
Recall: 0.922077922077922
F1-score: 0.8402366863905326
RandomForestClassifier() Avg cross val score is 0.7902702702702703
```



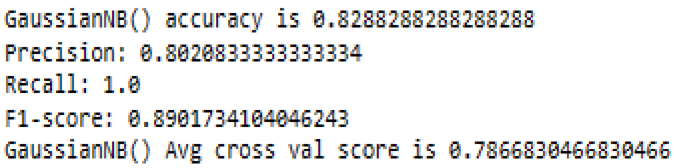
(v) Gradient Boosting:



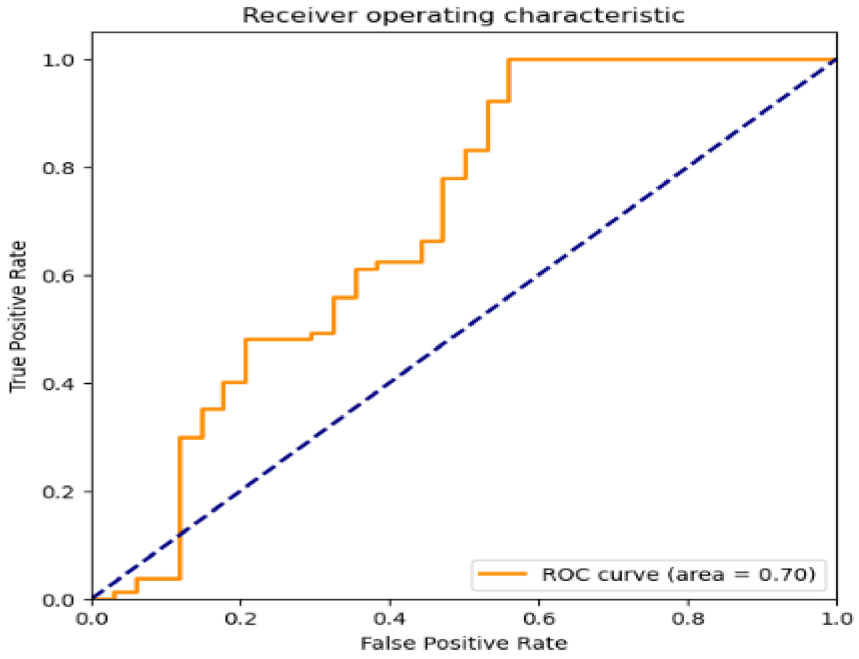
```
GradientBoostingClassifier() accuracy is 0.7927927927927928
Precision: 0.78125
Recall: 0.974025974025974
F1-score: 0.8670520231213873
GradientBoostingClassifier() Avg cross val score is 0.774004914004914
```



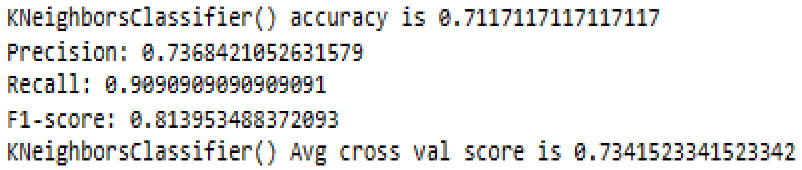
(vi) Gaussian Naïve Bayes:



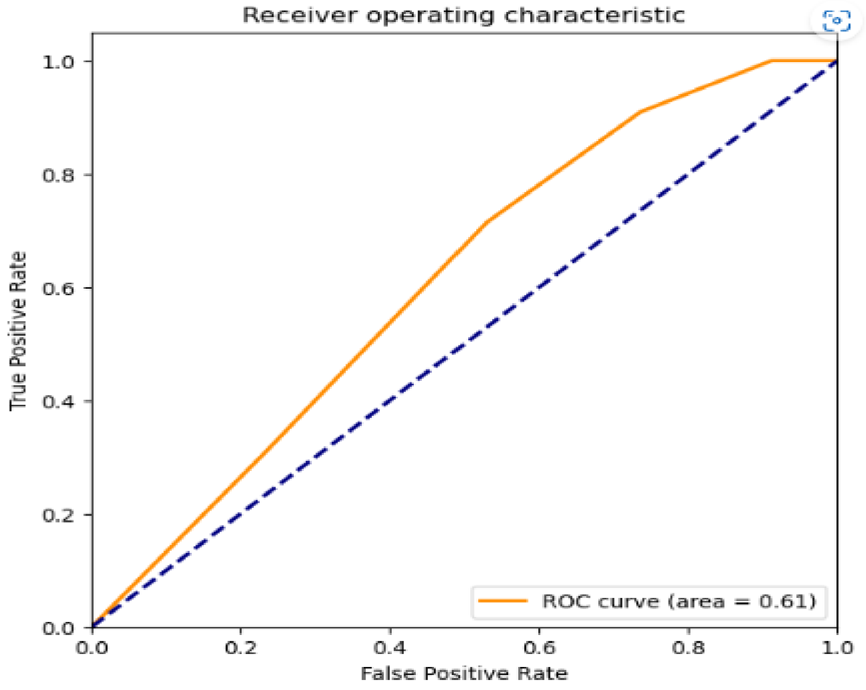
```
GaussianNB() accuracy is 0.8288288288288288
Precision: 0.8020833333333334
Recall: 1.0
F1-score: 0.8901734104046243
GaussianNB() Avg cross val score is 0.7866830466830466
```



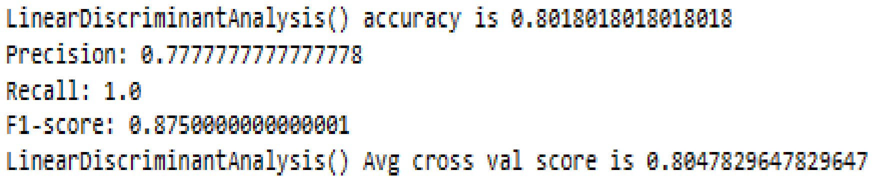
(vii) K-Nearest Neighbour:



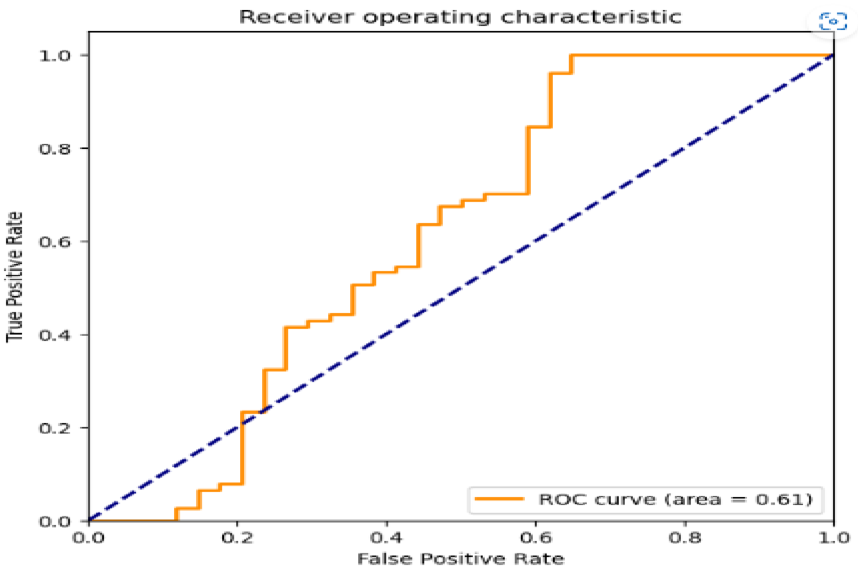
```
KNeighborsClassifier() accuracy is 0.7117117117117117
Precision: 0.7368421052631579
Recall: 0.9090909090909091
F1-score: 0.813953488372093
KNeighborsClassifier() Avg cross val score is 0.7341523341523342
```



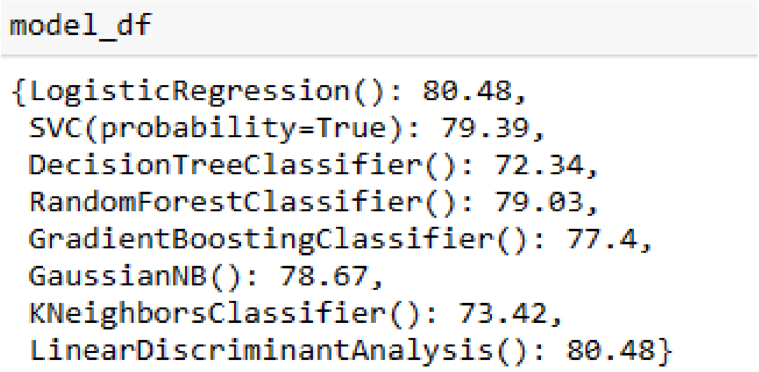
(viii) Linear Discriminant:



```
LinearDiscriminantAnalysis() accuracy is 0.8018018018018018
Precision: 0.7777777777777778
Recall: 1.0
F1-score: 0.8750000000000001
LinearDiscriminantAnalysis() Avg cross val score is 0.8047829647829647
```



9. OVERALL CROSS VALIDATION SCORE



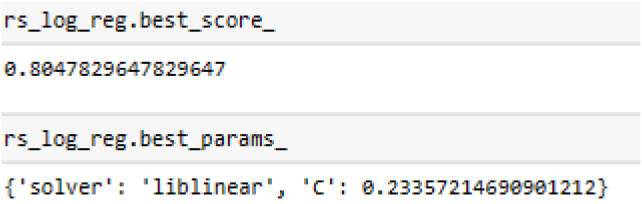
```
model_df
```

```
{LogisticRegression(): 80.48,
 SVC(probability=True): 79.39,
 DecisionTreeClassifier(): 72.34,
 RandomForestClassifier(): 79.03,
 GradientBoostingClassifier(): 77.4,
 GaussianNB(): 78.67,
 KNeighborsClassifier(): 73.42,
 LinearDiscriminantAnalysis(): 80.48}
```

10. HYPERPARAMETER TUNING (for best 3 models)

Cross Validation score and Best hyperparameters of the model

(i) Logistic Regression:



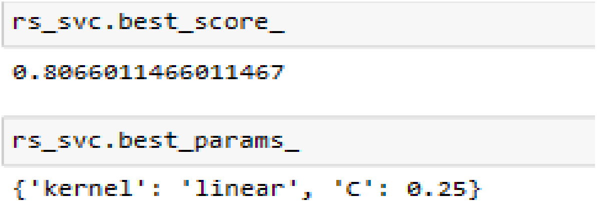
```
rs_log_reg.best_score_
```

```
0.8047829647829647
```

```
rs_log_reg.best_params_
```

```
{'solver': 'liblinear', 'C': 0.23357214690901212}
```

(ii) Support Vector Classifier:



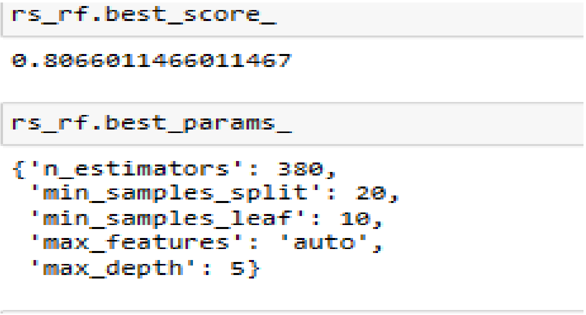
```
rs_svc.best_score_
```

```
0.8066011466011467
```

```
rs_svc.best_params_
```

```
{'kernel': 'linear', 'C': 0.25}
```

(iii) Random Forest:

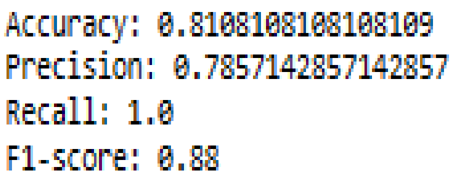


```
rs_rf.best_score_
```

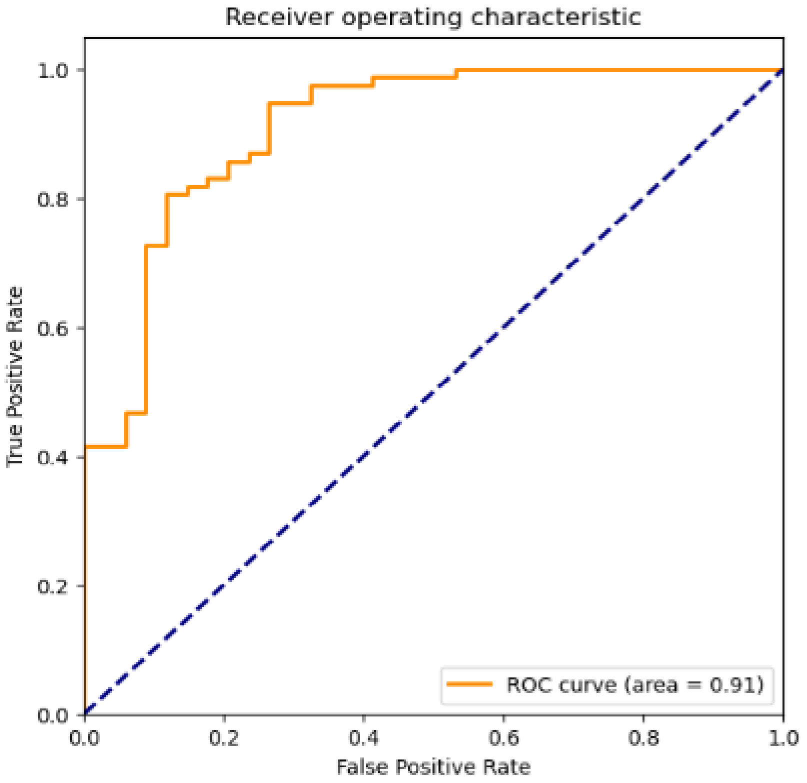
```
0.8066011466011467
```

```
rs_rf.best_params_
```

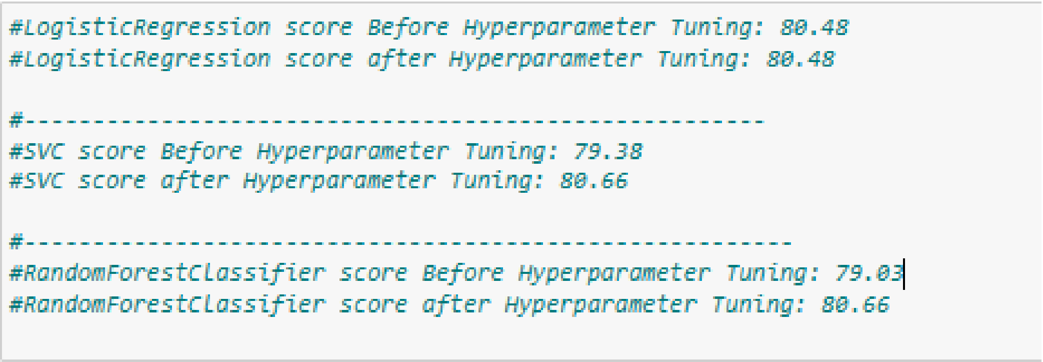
```
{'n_estimators': 380,
 'min_samples_split': 20,
 'min_samples_leaf': 10,
 'max_features': 'auto',
 'max_depth': 5}
```



```
Accuracy: 0.8108108108108109
Precision: 0.7857142857142857
Recall: 1.0
F1-score: 0.88
```

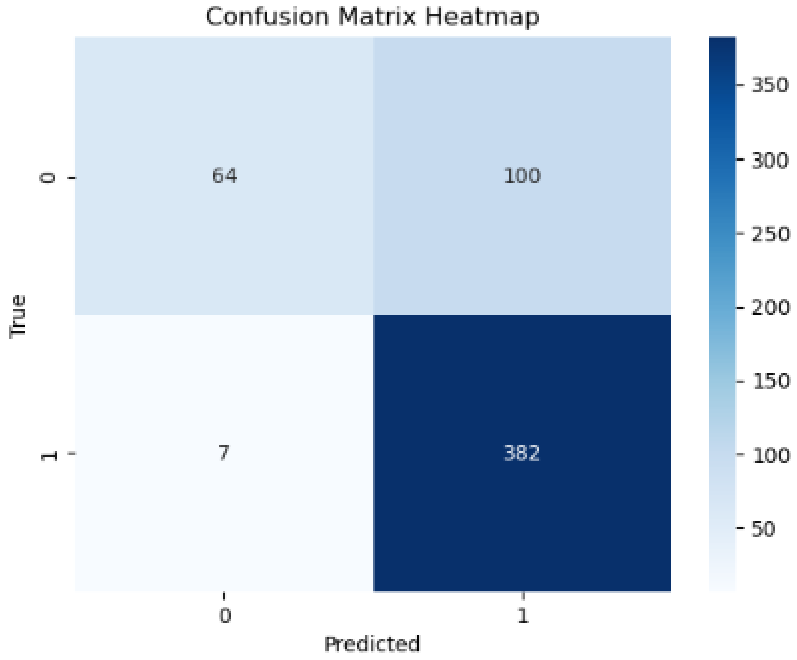


11.BEST CLASSIFICATION MODEL (RANDOM FOREST)

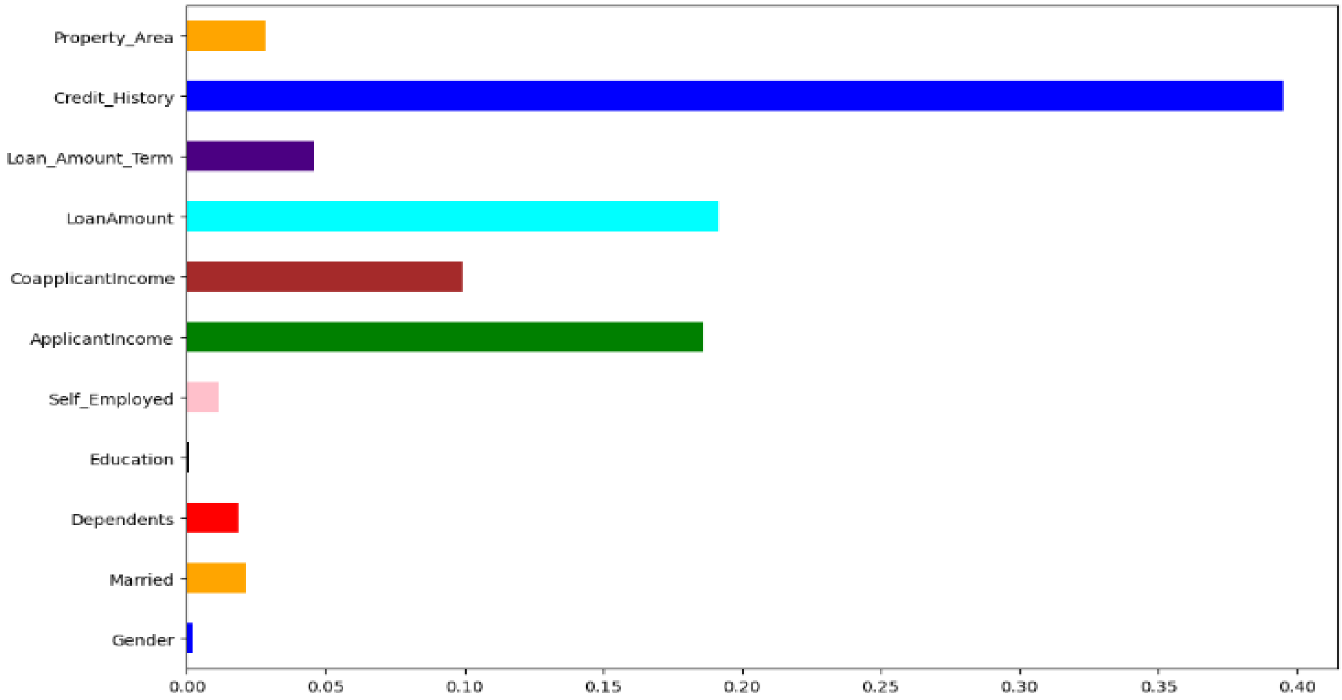


```
#LogisticRegression score Before Hyperparameter Tuning: 80.48
#LogisticRegression score after Hyperparameter Tuning: 80.48

#-------------------------------------------------------
#SVC score Before Hyperparameter Tuning: 79.38
#SVC score after Hyperparameter Tuning: 80.66

#---------------------------------------------------------
#RandomForestClassifier score Before Hyperparameter Tuning: 79.03
#RandomForestClassifier score after Hyperparameter Tuning: 80.66
```

12.CONFUSION MATRIX:



13.FEATURE SELECTION:

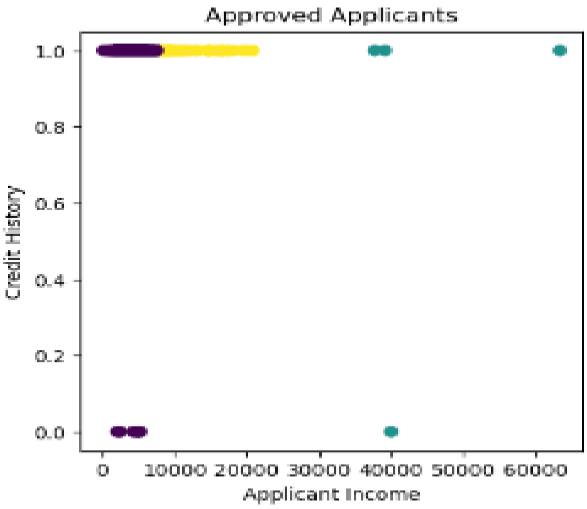


14. CUSTOMER SEGMENTATION :

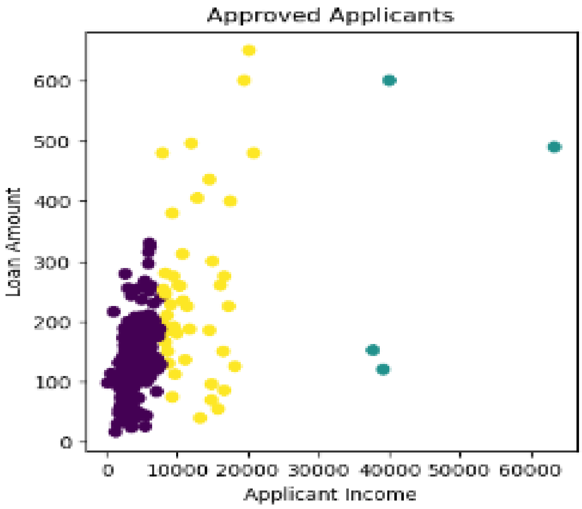
Using K-means clustering algorithm, the dataset divided into 2 categories approved and unapproved applicants on the 3 most important featurs (credit_history, Loan_amount, Loan_Amount_Term)

(i) Clustering on approved Applicants:

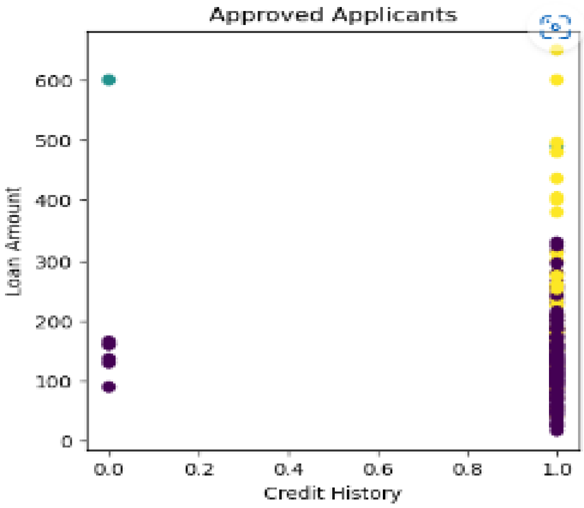
Applicant_Income VS Credit_History



Applicant_Income Vs Loan_Amount

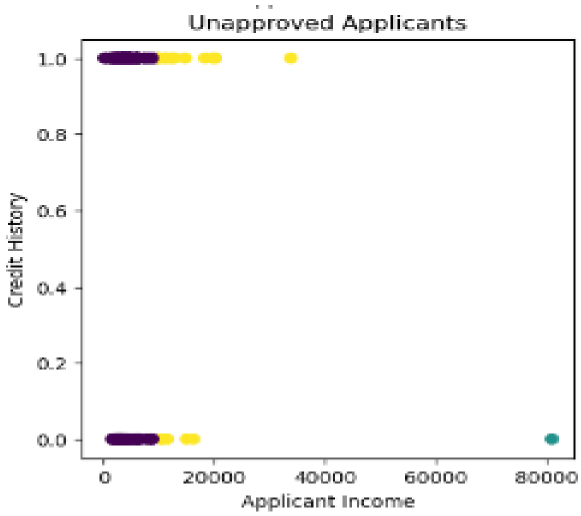


Credit_History Vs Loan_Amount

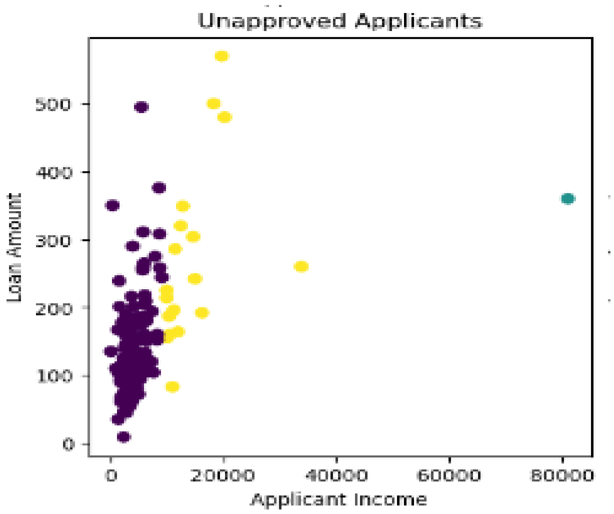


(ii) Clustering on Unapproved Applicants:

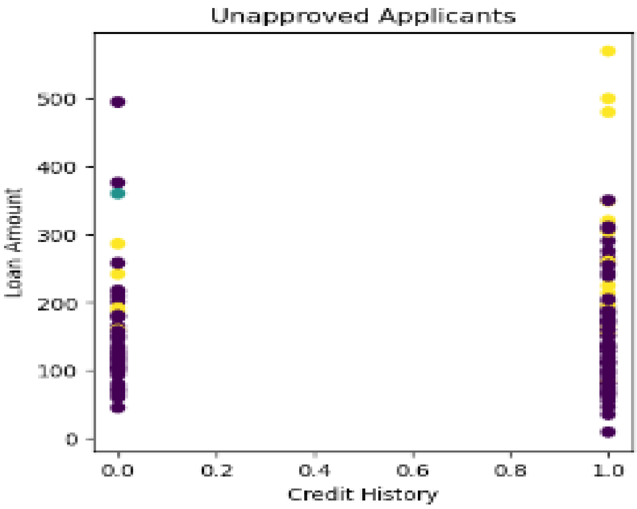
Applicant_Income VS Credit_History



Applicant_Income Vs Loan_Amount

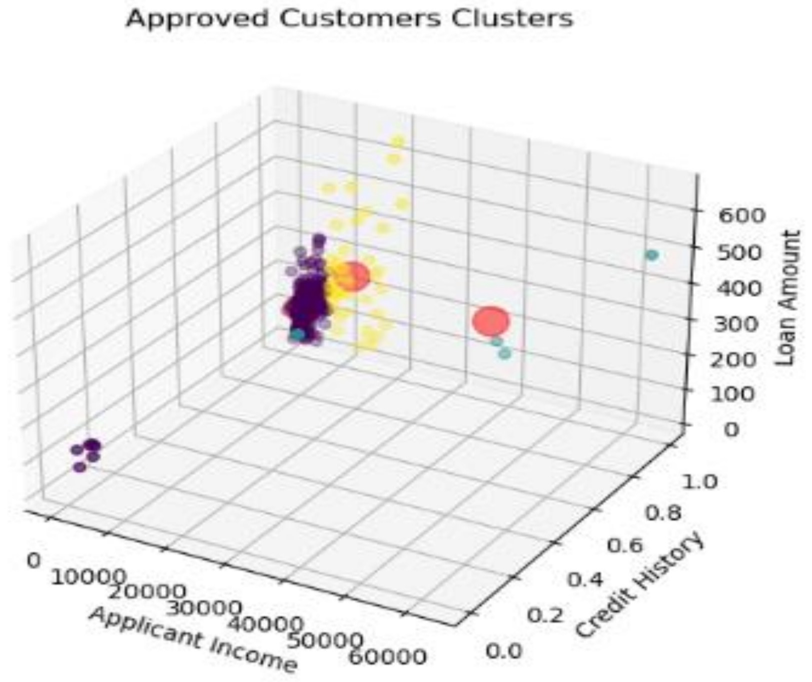


Credit_History Vs Loan_Amount

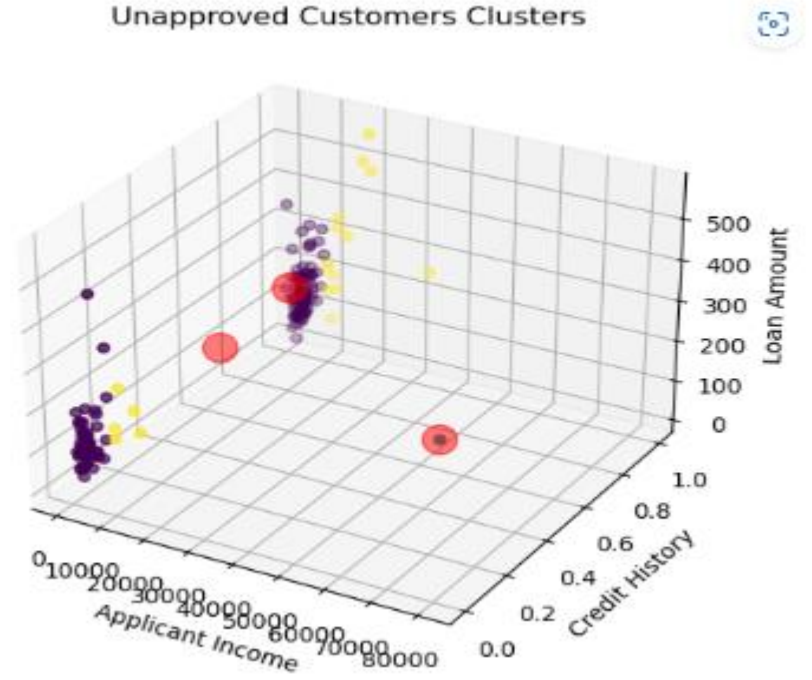


15.3d Visual of customer segmentation on approved and unapproved applicants based on the 3 dominant features. (Credit_History, Loan_Amount, Loan_Amount_Term)

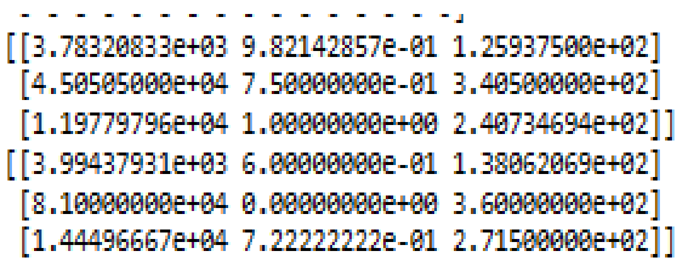
(i) **Approved applicants cluster:**



(ii) **Unapproved applicants cluster:**



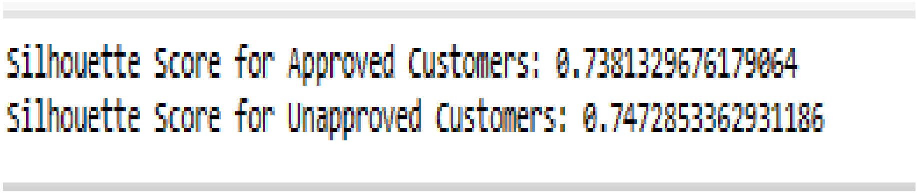
(iii) **Cluster Centroids:**



```
[[3.78320833e+03 9.82142857e-01 1.25937500e+02]
 [4.50505000e+04 7.50000000e-01 3.40500000e+02]
 [1.19779796e+04 1.00000000e+00 2.40734694e+02]]
[[3.99437931e+03 6.00000000e-01 1.38062069e+02]
 [8.10000000e+04 0.00000000e+00 3.60000000e+02]
 [1.44496667e+04 7.22222222e-01 2.71500000e+02]]
```

16. Cluster Model Evaluation:

Silhoutte Metric Score:

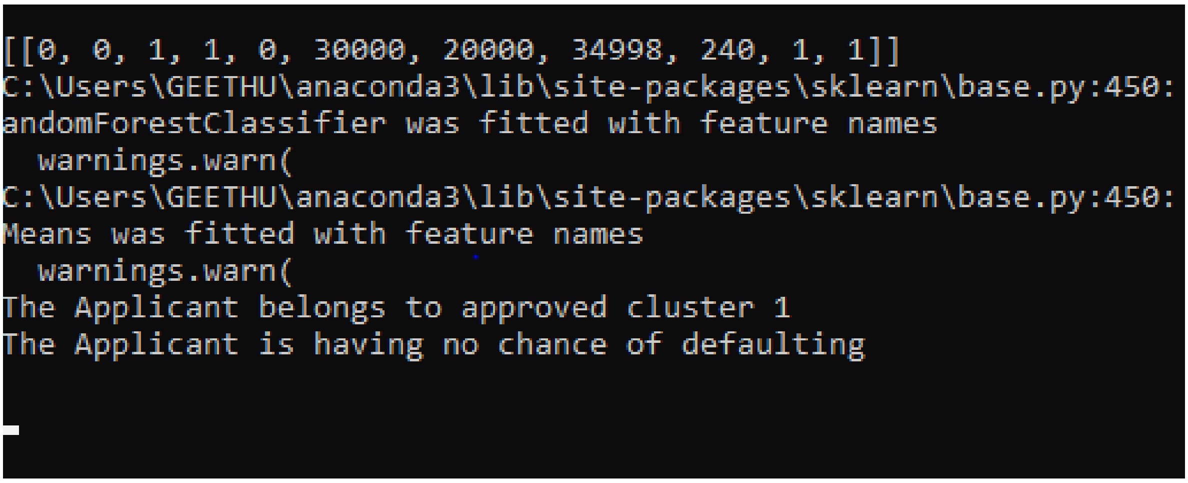


```
Silhouette Score for Approved Customers: 0.7381329676179064
Silhouette Score for Unapproved Customers: 0.7472853362931186
```

17.Streamlit Model Deployment (Bank Loan authentication System)

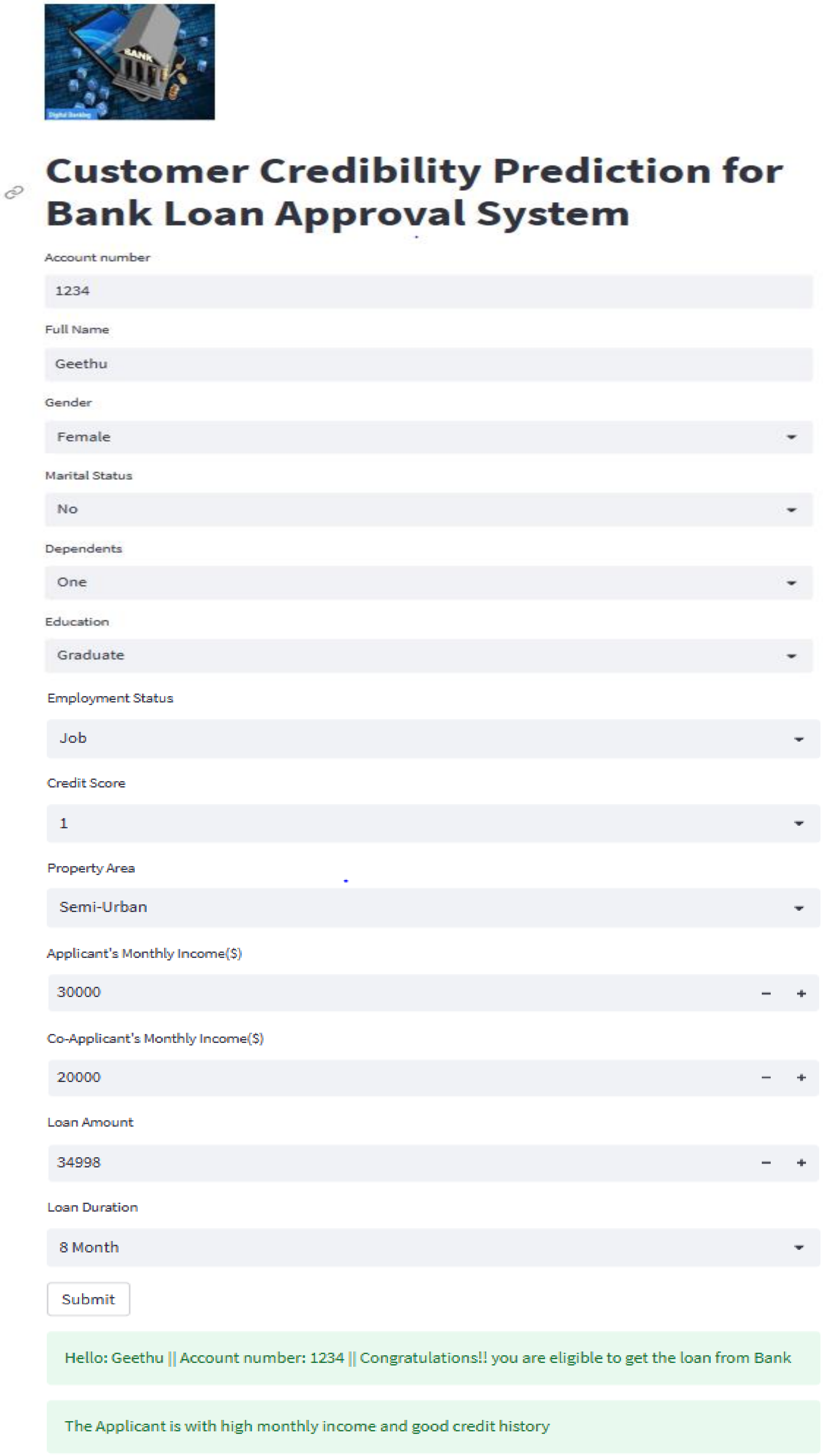
(i) **For an approved applicant**

Model result at bankers side:



```
[[0, 0, 1, 1, 0, 30000, 20000, 34998, 240, 1, 1]]
C:\Users\GEETHU\anaconda3\lib\site-packages\sklearn\base.py:450:
andomForestClassifier was fitted with feature names
  warnings.warn(
C:\Users\GEETHU\anaconda3\lib\site-packages\sklearn\base.py:450:
Means was fitted with feature names
  warnings.warn(
The Applicant belongs to approved cluster 1
The Applicant is having no chance of defaulting
```

At User Side:



Customer Credibility Prediction for Bank Loan Approval System

Account number

1234

Full Name

Geethu

Gender

Female

Marital Status

No

Dependents

One

Education

Graduate

Employment Status

Job

Credit Score

1

Property Area

Semi-Urban

Applicant's Monthly Income($)

30000

Co-Applicant's Monthly Income($)

20000

Loan Amount

34998

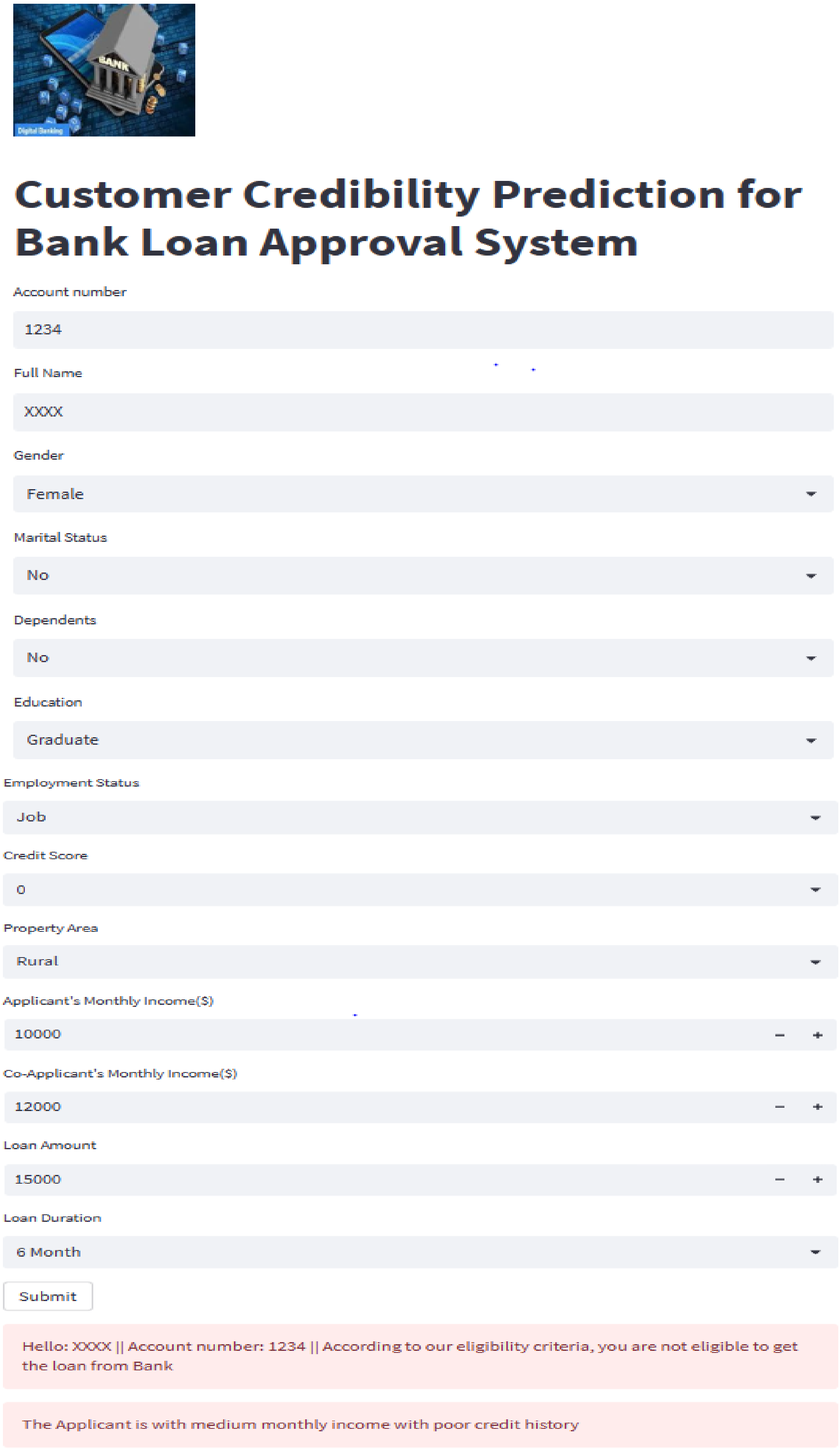
Loan Duration

8 Month

Hello: Geethu || Account number: 1234 || Congratulations!! you are eligible to get the loan from Bank

The Applicant is with high monthly income and good credit history

(ii) For an unapproved applicant (User side)



Customer Credibility Prediction for Bank Loan Approval System

Account number

1234

Full Name

XXXX

Gender

Female

Marital Status

No

Dependents

No

Education

Graduate

Employment Status

Job

Credit Score

0

Property Area

Rural

Applicant's Monthly Income($)

10000

Co-Applicant's Monthly Income($)

12000

Loan Amount

15000

Loan Duration

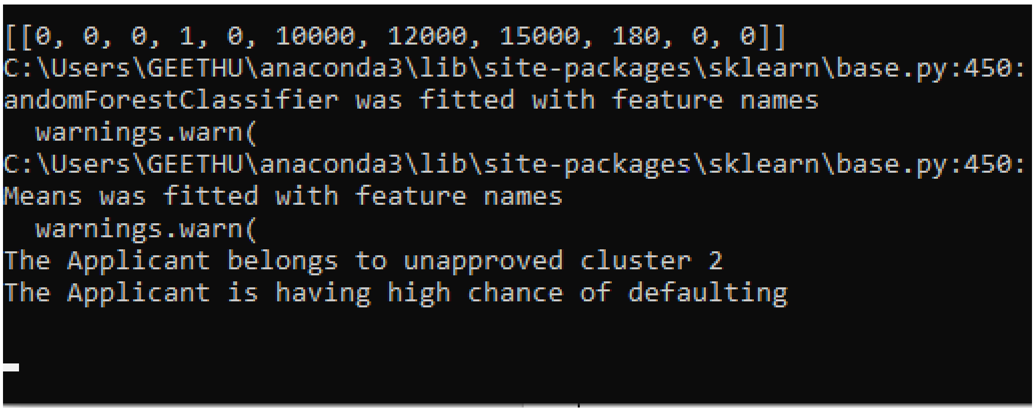
6 Month

Submit

Hello: XXXX || Account number: 1234 || According to our eligibility criteria, you are not eligible to get the loan from Bank

The Applicant is with medium monthly income with poor credit history

At banker side:



```
[[0, 0, 0, 1, 0, 10000, 12000, 15000, 180, 0, 0]]
C:\Users\GEETHU\anaconda3\lib\site-packages\sklearn\base.py:450:
andomForestClassifier was fitted with feature names
  warnings.warn(
C:\Users\GEETHU\anaconda3\lib\site-packages\sklearn\base.py:450:
Means was fitted with feature names
  warnings.warn(
The Applicant belongs to unapproved cluster 2
The Applicant is having high chance of defaulting
```

6 RESULT AND DISCUSSION

As our proposed system comprises of 2 models one is for classification and another one is for customer segmentation purpose. Now, after successful completion of our proposed methodology in bank loan approval system, we come up with the following results for both the classification and customer segmentation model.

8.1 Findings of Classification Model 1:

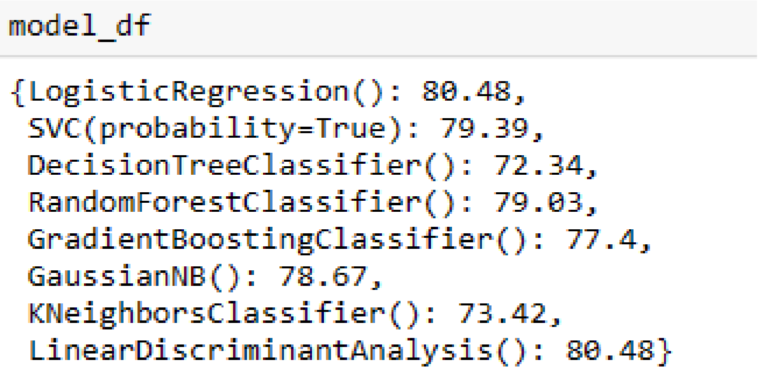
As a result of training the Eight supervised ML algorithms comprising logistic regression, support vector classifier, random forest, decision tree, gradient boosting, gaussianNB classifier, KNN classifier and finally linear discriminant analysis for our binary classification problem, we come up with the results after evaluating with several metrices such as cross validation score, accuracy, precision, recall, f1 score and roc curve .

The findings of the models are tabulated below:

Model Name	Cross Validation Score	Accuracy	Precision	Recall	F1 Score	ROC Area
Logistic Regression	0.80478	0.80180	0.77777	1.0	0.87500	0.61
Support Vector Machine	0.79389	0.79279	0.77	1.0	0.87005	0.65
Decision Tree	0.72337	0.74774	0.82666	0.80519	0.81578	0.71
Random Forest	0.79027	0.75675	0.77173	0.92207	0.84023	0.77
Gradient Boosting	0.77400	0.79279	0.78125	0.97402	0.86705	0.69
Gaussian Naïve Bayes	0.78668	0.82882	0.80208	1.0	0.89017	0.70
K-Nearest Neighbour	0.73415	0.71171	0.73684	0.90909	0.81395	0.61
Linear Discriminant Analysis	0.80478	0.80180	0.77777	1.0	0.87500	0.61

Table 3. Model Results

From the results we found, we consider the cross-validation metric as most important metric for our model evaluation as we trained the model based on cross fold combinations. So, the overall performance summary of the various models in cross validation metric is given below:

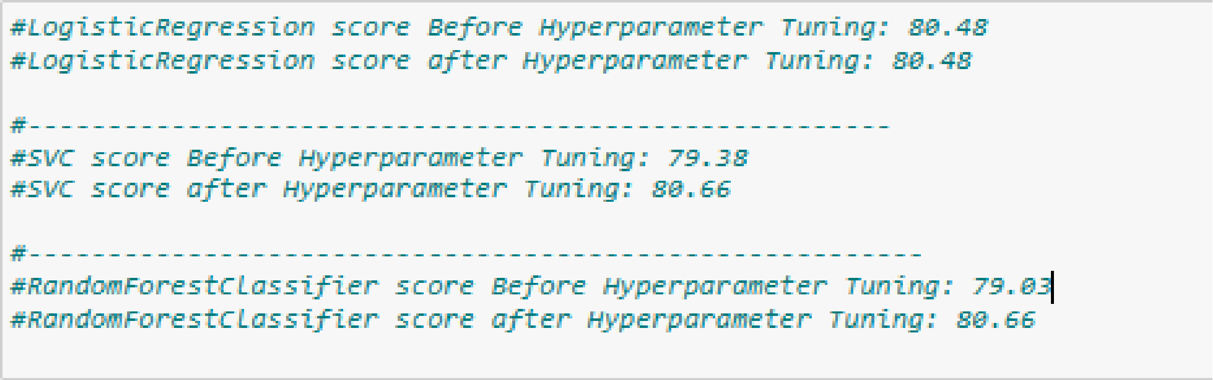


```
model_df
```

```
{LogisticRegression(): 80.48,
 SVC(probability=True): 79.39,
 DecisionTreeClassifier(): 72.34,
 RandomForestClassifier(): 79.03,
 GradientBoostingClassifier(): 77.4,
 GaussianNB(): 78.67,
 KNeighborsClassifier(): 73.42,
 LinearDiscriminantAnalysis(): 80.48}
```

From this summary, we tend to pick the top 3 most models which have higher cross validation score. Later, the 3 models (Logistic Regression with the score 80.48, Support Vector with the score of 79.39, Random Forest with the score of 79.03) is induced for hyperparametric tuning using randomized search to get a better collection of hyperparameters which results in finding the best model of our proposed system.

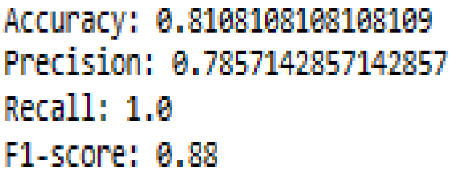
The results of the models after hyperparametric tuning is given below:



```
#LogisticRegression score Before Hyperparameter Tuning: 80.48
#LogisticRegression score after Hyperparameter Tuning: 80.48

#-----------------------------------------------------
#SVC score Before Hyperparameter Tuning: 79.38
#SVC score after Hyperparameter Tuning: 80.66

#-------------------------------------------------------
#RandomForestClassifier score Before Hyperparameter Tuning: 79.03
#RandomForestClassifier score after Hyperparameter Tuning: 80.66
```

Based on the cross validation score after hyperparametric tuning , Random Forest and Support Vector tends to give same result, among these Random Forest Classifier is chosen as the best model for our Bank loan approval system.

Then, the model is evaluated with various metrics which gives better results among all other models and the model output is given below for the following metrics.



```
Accuracy: 0.8108108108108109
Precision: 0.7857142857142857
Recall: 1.0
F1-score: 0.88
```

With the hyperparameter tuning, the Random Forest Classifier improved its score from 79.03% to 80.66%, which is a significant improvement. This indicates that the hyperparameters were able to optimize the model's performance and make it better suited for the loan approval system.

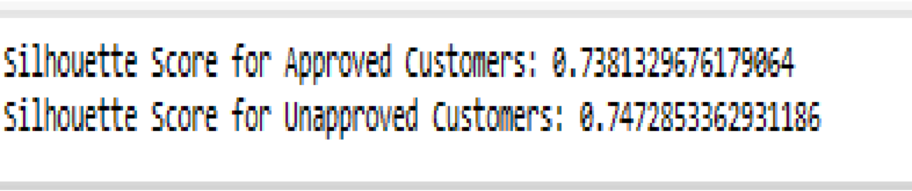
By selecting and saving the model to our Bank Loan Authentication System, it makes predictions on whether to accept or deny the credit applications regarding to the features included in the dataset. These features might include information such as the applicant's credit rating, earnings, job status, loan amount, etc.

8.2 Findings of Customer Segmentation Model 2:

For segmenting the applicants based on the most essential features, the dataset have been divided into 2 categories of approved and unapproved applicants. To find trends in applicant income, credit history, and loan amount as they were extracted as most essential features for our model from the feature selection, KMeans clustering is used with three clusters for the approved applicants. The same three clusters of KMeans clustering are used for the applicants who were rejected. Scatter plots are used to display the clustering findings. The clustering results demonstrate that the data exhibits distinct patterns that can be used to guide a decision regarding whether to approve a loan application.

In addition to this, kmeans clustering results are visualized in 3D scatterplots. The centroids of each cluster are also shown in the plots. The 3D visualization helps to understand the relationship between the variables more easily. This Kmeans clustering model is evaluated using the metric called Silhouette score which finds how well the clustering algorithm is performed on the trained dataset. It compares an item's similarity to its cluster itself with the rest of the groups.

The Silhouette rating of kmeans clustering model for 2 categories of approved and unapproved applicants is given below:



```
Silhouette Score for Approved Customers: 0.7381329676179064
Silhouette Score for Unapproved Customers: 0.7472853362931186
```

Since the silhouette score for the categories is between -1 to 1 which indicates that the clusters are well defined and placed far away from the other clusters which is at the right distance from one another. Also, the score is above 0.5 which indicates that the model attains good results for our clustering model.

And the K-means model is saved to predict the upcoming applicant’s cluster based on the most dominating features we selected. This helps both bankers and the applicants to get the knowledge about loan approval and loan denial.

Findings of the Model’s performance:

Based on the two chosen model, Random Forest for classification and k-means for customer segmentation the model is deployed to a Bank Loan Authentication System using streamlit. The system takes various inputs such as name, account number, marital status, dependants, education, type of the property area, credit history, loan amount , applicant income, co applicant income and loan amount duration. By the process of feature selection, we get to know the most dominating features are credit history, applicant income and loan amount, so the prediction will be mostly based on this dominating features.

Based on the details given by the user(applicant):

Model 1- predicts that the applicant is classified for the credit sanction or not.

Model 2- predicts the reason to the borrowers, in what way the borrower is eligible or not eligible for the credit, also to the bankers it gives info of the particular applicants segments they belongs to which helps them to easily understand about the applicant’s segmentation.

8 LIMITATION AND FUTURE WORK

9.1 DRAWBACKS OF THE PROPOSED SYSTEM:

(i) **Lack of Transparent:** Since machine learning algorithms seems to be complicated and challenging to grasp, it might be a great challenge to understand about how the decisions are being made. Due to this lack of transparency, it may be difficult to defend our decisions in front of the authorities or debtors.

(ii) **Bias:** If the data which we used in our model for training, does not represents the population the machine learning algorithms continued to be biased. This results to exhibits the biasness towards certain group of borrowers.

(iii) **Limited data:** In order for machine learning algorithms to train efficiently, they need a lot of data. The algorithm's accuracy, however, may be impacted in some circumstances by the availability of little data.

(iv) **Data quality:** The quality of the dataset which is utilized to train the algorithm for our model has a significant effect on its accuracy. Decisions may be made incorrectly if the data is insufficient, wrong, or out of date.

(v) **Over rely on historical data:** Machine learning-based loan approval system frequently rely their decisions on the past data. But it's possible that this strategy may not be good at spotting future market patterns or adjustments.

(vi) **Lack of human overview:** Automated decision-making systems without human supervision is possible using machine learning algorithms. this may result in errors or bad conclusions, depending on how well the algorithm has been taught or tested.

Even though machine learning can be an effective tool in automated loan approval systems, it is essential to be aware of its limitations and to put the right safeguards in place to address them.

9.2 FUTURE WORK:

Machine learning-based automated loan approval systems have a promising future as technology advances. While the Random Forest Classifier appears to be the best model based on these findings, it is crucial to carefully assess the model's performance on a holdout dataset before implementing it in a real-world scenario. This guarantees the model's robust performance and its ability to generalise to new data. Additionally, it's critical to think about the moral implications of using machine learning models to approve loans and to make sure the process is impartial and fair to all applicants.

The following are some potential future applications for this technology:

(i) **Explainable:** Transparency and explainability are becoming more and more important as machine learning algorithms become more complicated. Future loan approval systems might include methods to help explain how judgements are made, making it simpler for lenders to comprehend and inform borrowers of decisions.

(ii) **Integration of other data sources:** Other data sources instead of past loan approval trained credit bureau data, such as other additional testing data, social media activity, payment history, and employment records, can improve loan approval systems. (Lenders could make better decisions if they had a more complete picture of a borrower's creditworthiness.

(iii) **Blockchain use:** The application of blockchain technology to loan approval system can enhance data security, lower fraud, and promote transparency in the lending process. Blockchain can speed up the loan approval process and lessen the possibility of mistakes or errors by establishing a distributed system where all stakeholders have access to the same information.

(iv) **Better borrower experience:** As automated loan approval systems develop, there is a chance to enhance the borrower experience by delivering more transparency, quicker turnaround times, different categories of EMI duration and tailored offers). This might include functions like real-time feedback, fast loan pre-approvals, and customised loan alternatives based on a borrower's credit history and financial position.

(v) **Use of machine learning in risk assessment:** The ML models are used to analyse huge amounts of data also spot patterns as well as trends that conventional underwriting models might overlook. Lenders can lower the risk of default and make better loan decisions by utilising machine learning for risk assessment.

In general, the incorporation of new technology, the acceptance of alternative data sources, and an emphasis on enhancing the borrower experience while ensuring compliance with rules are likely to characterise loan approval processes in the future.

10. CONCLUSION

The efficiency and precision our automation bank loan authentication system might be greatly increased with the use of a machine learning-based loan approval system. Machine learning algorithms can analyse and spot trends that are suggestive of creditworthiness or the possibility of default by using historical loan data and other pertinent information. This can lower the risk of fraud and help lenders make more informed decisions about the applicant segmentation which applicants to accept or reject. Also, the borrowers get to know about their reason of loan approval and denial. It's crucial to understand that machine learning cannot solve every problem relating to loan approval. It is still vital to preserve the privacy and security in a sensitive client data and to make sure that lending decisions are impartial and fair. Additionally, machine learning models are not perfect and may be subject to biases or errors that lead to poor judgement.

11. REFERENCES

[1] Zijiang Yang, Wenjie You, Guoli Ji,Using partial least squares and support vector ma chines for bankruptcy prediction,Expert Systems with Applications,Volume 38, Issue 7,2011,Pages 8336-8342,ISSN 0957-4174,https://doi.org/10.1016/j.eswa.2011.01.021.

[2] A. Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, India, 2017, pp. 1-6, doi: 10.1109/ICCCNT.2017.8203946.

[3] A. Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, India, 2017, pp. 1-6, doi: 10.1109/ICCCNT.2017.8203946.

[4] Xiaojun Ma, Jinglan Sha, Dehua Wang, Yuanbo Yu, Qian Yang, Xueqi Niu, Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning,Electronic Commerce Research and Applications,Volume 31,2018,Pages 24-39,ISSN 1567-4223,

[5] M. Alaradi and S. Hilal, "Tree-Based Methods for Loan Approval," 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), Sakheer, Bahrain, 2020, pp. 1-6, doi: 10.1109/ICDABI51230.2020.9325614.

[6] 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), Coimbatore, India, 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.

[7] Weiguo Zhang, Chao Wang, Yue Zhang, Junbo Wang,Credit risk evaluation model with textual features from loan descriptions for P2P lending,Electronic Commerce Research and Applications,Volume 42,2020,100989,ISSN 15674223,https://doi.org/10.1016/j.elerap.2020.100989.

[8] P. Kirubanantham, A. Saranya and D. S. Kumar, "Credit Sanction Forecasting," 2021 4th International Conference on Computing and Communications Technologies (ICCCT), Chennai, India, 2021, pp. 155-159, doi: 10.1109/ICCCT53315.2021.9711790.

[9] S. Barua, D. Gavandi, P. Sangle, L. Shinde and J. Ramteke, "Swindle: Predicting the Probability of Loan Defaults using CatBoost Algorithm," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1710-1715, doi: 10.1109/ICCMC51019.2021.9418277.

[10] H. Ramachandra, G. Balaraju, R. Divyashree and H. Patil, "Design and Simulation of Loan Approval Prediction Model using AWS Platform," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2021, pp. 53-56, doi: 10.1109/ESCI50559.2021.9397049.

[11] Justin Munoz, Ahmad Asgharian Rezaei, Mahdi Jalili, Laleh Tafakori,Deep learning based bi-level approach for proactive loan prospecting,Expert Systems with Applications,Volume 185,2021,115607,ISSN 0957-4174

https://doi.org/10.1016/j.eswa.2021.115607.

[12] Akça, Mehmet & Sevli, Onur. (2022). Predicting acceptance of the bank loan offers by using support vector machines. International Advanced Researches and Engineering Journal. 6. 142-147. 10.35860/iarej.1058724.

[13] U. E. Orji, C. H. Ugwuishiwu, J. C. N. Nguemaleu and P. N. Ugwuanyi, "Machine Learning Models for Predicting Bank Loan Eligibility," 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON), Lagos, Nigeria, 2022, pp. 1-5, doi: 10.1109/NIGERCON54645.2022.9803172.

[14] Sandeep, C.V. and Devi, T., 2022. A Novel Approach for Bank Loan Approval by Verifying Background Information of Customers through Credit Score and Analyze the Prediction Accuracy using Random Forest over Linear Regression Algorithm. *Journal of Pharmaceutical Negative Results*, pp.1748-1755.10.47750/pnr.2022.13.S04.211

[15] C. Naveen Kumar, D. Keerthana, M. Kavitha and M. Kalyani, "Customer Loan Eligibility Prediction using Machine Learning Algorithms in Banking Sector," 2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2022, pp. 1007-1012, doi: 10.1109/ICCES54183.2022.9835725.

[16] T. K. Bhatia, S. Gupta and A. Sharma, "Analysis of Customer Segmentation Model through K-Means Clustering," 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2022, pp. 1-6, doi: 10.1109/ICRITO56286.2022.9965157.