**LOAN ELIGIBILITY PREDICTION**

A Summer Internship Project Report Submitted in partial fulfillment of the requirements for the award of the degree of

**BACHELOR OF TECHNOLOGY IN**

**CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)**

Submitted by

D Aravind 21071A6616

G Jyothi Swaroop 21071A6622

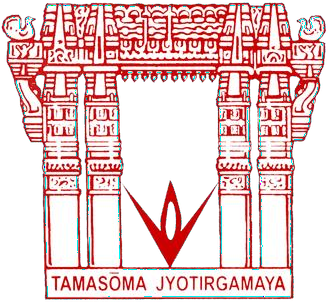
K Akshara Jyothi 21071A6628

P Lahari 21071A6650

Under the guidance of

**Dr.N.Sandhya**

**(Professor &Head, Department of CSE-AIML & IoT, VNR VJIET)**



**DEPARTMENT OF CSE(ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)**

**VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY**

An Autonomous Institute, NAAC Accredited with ‘A++’ Grade NBA Accredited for CE, EEE, ME, ECE, CSE, EIE, IT B. Tech Courses Approved by AICTE, New Delhi,

Affiliated to JNTUH Recognized as “College with Potential for Excellence” by UGC ISO 9001:2015 Certified, QS I GUAGE Diamond Rated

Vignana Jyothi Nagar, Pragathi Nagar, Nizampet (S.O), Hyderabad – 500 090, TS, India

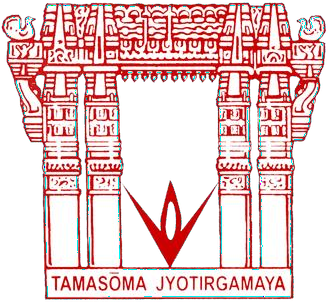
**VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY**

An Autonomous Institute, NAAC Accredited with ‘A++’ Grade NBA Accredited for CE, EEE, ME, ECE, CSE, EIE, IT B. Tech Courses Approved by AICTE, New Delhi,

Affiliated to JNTUH Recognized as “College with Potential for Excellence” by UGC ISO 9001:2015 Certified, QS I GUAGE Diamond Rated

Vignana Jyothi Nagar, Pragathi Nagar, Nizampet (S.O), Hyderabad – 500 090, TS, India

**DEPARTMENT OF CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)**



**CERTIFICATE**

This is to certify that the project report entitled “**LOAN ELIGIBILITY PREDICTION”** is a bonafide work done under our supervision and is being submitted by **D Aravind (21071A6616),G Jyothi Swaroop(21071A6622) , K Akshara Jyothi(21071A6628) ,P Lahari (21071A6650)** in partial fulfillment for the award of the degree of Bachelor of Technology in CSE(Artificial Intelligence and Machine Learning), of the VNRVJIET, Hyderabad during the academic year 2023-2024. Certified further that to the best of our knowledge the work presented in this thesis has not been submitted to any other University or Institute for the award of any Degree or Diploma.

**Dr. N. Sandhya**

**Professor& HOD**

**Dept. of CSE (AIML&IoT )**

**VNR VJIET**

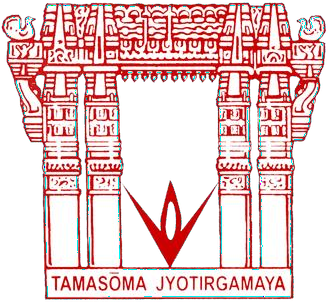
**VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY**

An Autonomous Institute, NAAC Accredited with ‘A++’ Grade NBA Accredited for CE, EEE, ME, ECE, CSE, EIE, IT B. Tech Courses Approved by AICTE, New Delhi,

Affiliated to JNTUH Recognized as “College with Potential for Excellence” by UGC ISO 9001:2015 Certified, QS I GUAGE Diamond Rated

Vignana Jyothi Nagar, Pragathi Nagar, Nizampet (S.O), Hyderabad – 500 090, TS, India

**DEPARTMENT OF CSE (ARTIFICIAL INTELLIGENCE & MACHINE LEARNING)**



**DECLARATION**

We declare that the major project work entitled “**LOAN ELIGIBILITY PREDICTION**” submitted in the department of CSE-Artificial Intelligence and Machine Learning, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, in partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology** in **CSE-Artificial Intelligence and Machine Learning** is a bonafide record of our own work carried out under the supervision of **Dr.N.Sandhya, Professor & Head, Department of CSE(AIML & IoT), VNRVJIET**. Also, we declare that the matter embodied in this thesis has not been submitted by us in full or in any part thereof for the award of any degree/diploma of any other institution or university previously.

Place: Hyderabad

|  |  |  |  |
| --- | --- | --- | --- |
| **D Aravind** | **G Jyothi Swaroop** | **K Akshara Jyothi** | **P Lahari** |
| (21071A6616) | (21071A6622) | (21071A6628) | (21071A6650) |

**ACKNOWLEDGEMENT**

Firstly, we would like to express our immense gratitude towards our institution VNR Vignana Jyothi Institute of Engineering and Technology, which created a great platform to attain profound technical skills in the field of Computer Science, thereby fulfilling our most cherished goal.

We are very much thankful to our Principal, **Dr. Challa Dhanunjaya Naidu,** and our Head of Department, **Dr. N. Sandhya**, for extending their cooperation in doing this project within the stipulated time.

We extend our heartfelt thanks to our guide, **Dr.N.Sandhya** for her enthusiastic guidance throughout the course of our project.

Last but not least, our appreciable obligation also goes to all the staff members of the CSE (AIML & IoT) and to our fellow classmates who directly or indirectly helped us.

**D Aravind 21071A6616**

**G Jyothi Swaroop 21071A6622**

**K Akshara Jyothi 21071A6628**

**P Lahari 21071A6650**

**ABSTRACT**

The rapid advancement of machine learning techniques has paved the way for innovative applications in various domains, and one such area is the prediction of loan eligibility. In the financial sector, accurately assessing an individual's eligibility for a loan is crucial for minimizing risks and optimizing lending practices. This project proposes a machine learning-based approach to predict loan eligibility by leveraging historical data and relevant features.

Banks are making major part of profits through loans. Though lot of people are applying for loans. It’s hard to select the genuine applicant, who will repay the loan. While doing the process manually, lot of misconception may happen to select the genuine applicant. Therefore, we are developing loan prediction system using machine learning, so the system automatically selects the eligible candidates. This is helpful to both bank staff and applicant. The time period for the sanction of loan will be drastically reduced.

The proposed loan eligibility prediction system aims to assist financial institutions in automating the decision-making process, thereby reducing manual efforts and improving efficiency. The results of this project contribute to the growing body of knowledge on the application of machine learning in finance and offer practical insights for implementing predictive models in real-world lending scenarios.

**INDEX**

**1.Title Page i**

**2.Certificate ii**

**3.Declaration iii**

**4.Acknowledgement iv**

**5.Abstract v**

**6.Index vi**

**1.**[**Introduction**](#_TOC_250059)  **01**

2.Literature Survey 02

3.Existing system 07

4. Software Requirement 08

5.Software Design 09

* 1. [**UML Diagrams**](#_TOC_250037) 
     1. [**Use Case Diagram**](#_TOC_250036)
     2. [**Activity Diagram**](#_TOC_250034)

6.Proposed System 12

7.Coding/Implementation 15

8.Results 23

9.Conclusion and Future Scope 25

10.References dd

# 1. INTRODUCTION

In the dynamic landscape of financial services, the accurate assessment of loan eligibility is a critical aspect that significantly influences the success and stability of lending institutions. Traditional methods of evaluating loan applications often involve complex, time-consuming processes that are prone to human error and subjectivity. The advent of machine learning technologies provides an unprecedented opportunity to revolutionize this process by automating decision-making and enhancing the precision of loan eligibility predictions.

The significance of this project lies in its potential to address key challenges faced by lending institutions, including risk management, operational efficiency, and customer satisfaction. The utilization of machine learning algorithms allows for a data-driven approach to decision-making, enabling financial institutions to make informed choices based on historical patterns and real-time data.

The project will involve the exploration and analysis of diverse data attributes such as assets, credit scores, employment type,educational qualification, income per annum, and other relevant factors that contribute to the determination of loan eligibility. Various machine learning techniques, ranging from decision trees to sophisticated ensemble methods, will be employed to build and optimize the predictive model.

**1.1 Problem statement**

The assessment of loan eligibility poses significant challenges for lending institutions. The reliance on traditional, manual processes results in inefficiencies, prolonged processing times, and increased operational costs. Moreover, the inherent subjectivity in human judgment introduces biases and errors, leading to inconsistent decisions and potential compliance issues. The underutilization of available data, including historical applicant information and external economic indicators, further hinders the comprehensive understanding of applicant profiles. Additionally, the absence of transparency in decision-making processes poses challenges in meeting regulatory standards and explaining the rationale behind loan approval or denial decisions. To address these issues, there is a critical need for the development and implementation of an advanced loan eligibility prediction system, leveraging machine learning techniques to automate and optimize decision-making processes. This project aims to enhance the accuracy, efficiency, and transparency of loan eligibility assessments, contributing to improved risk management and responsible lending practices in the financial sector.

# 2. LITERATURE SURVEY

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S.No** | **Title of the Paper** | **Year** | **Reference with DOI** | **Objective of the paper** | **Focus of the paper(discuss methods used)** | **Summary** |
| 1 | Prediction for Loan Approval using Machine Learning Algorithm | 2021 | <https://doi.org/10.3390/w10030264> | The primary objective is to boost banking profits by employing predictive analytics to predict loan defaulters. Through data collection, cleaning, and performance evaluation, the aim is to automate and streamline the loan eligibility process, facilitating quick and informed decisions for both the bank and applicants. |  | The paper focuses on optimizing loan approval processes in banking through predictive analytics, employing SVM and Naïve Bayes algorithms. The study emphasizes data collection, cleaning, and model evaluation. Naïve Bayes outperforms other models. The proposed system automates loan eligibility based on customer details, enhancing efficiency and accuracy. |
| 2 | Machine Learning Techniques for recognizing the Loan Eligibility. | 2022 | https://www.irjmets.com/uploadedfiles/paper/volume\_3/issue\_12\_december\_2021/17758/final/fin\_irjmets1641311178.pdf | The objective is to streamline and automate the loan approval process by predicting eligibility through machine learning algorithms. Parameters like marital status, credit history, and gender are considered. The study aims to find the most efficient algorithm utilizing data analysis and statistics to enhance accuracy and reduce processing time. | For the model training the mean accuracy scores of three machine learning algorithms: logistic regression, decision tree classification, random forest classification are compared. | On training the model with 3 different algorithms, the mean accuracy scores are obtained as follows:  Logistic regression – 80.78%  Random Forest – 79.79% Decision tree – 70.51%.  Thus, this paper gives a general idea that we can prefer logistic regression for loan eligibility |



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| 3 | Loan eligibility prediction using adaptive hybrid optimization driven-deep neuro fuzzy network. | 2023 | https://www.sciencedirect.com/science/article/abs/pii/S0957417423004049 | A significant issue in predicting loan eligibility is, making precise loan predictions using risk and evaluation analysis. The research's objective is to offer an optimization algorithm as a means of predicting loan eligibility. | They use a unique methodology called Adaptive Social Border Collie Optimization-based Deep Neuro Fuzzy Network (ASBCO-based DNFN). To train a deep neuro fuzzy network to predict loan eligibility, the suggested adaptive SBCO is used. | The Adaptive SBCO-based deep neural fuzzy network is a unique technique for predicting loan eligibility that is proposed in this research. The project's goal is to develop an optimization system for predicting loan eligibility. Here, the adaptive box-cox approach is used to transform the incoming loan data into a format appropriate for further processing. When compared to the current method, the suggested method performs better, with maximum accuracy, sensitivity, and specificity of roughly 95%, 95.4%, and 97.3%, respectively. |
| 4 | Deep learning based loan eligibility prediction with Social Border Collie Optimization | 2022 | https://doi.org/10.1108/K-10-2021-1073 | This aim of the research technique is to present a new method for precise loan prediction with risk and assessment analysis | A new method, namely Social Border Collie Optimization (SBCO)-based deep neuro fuzzy network for loan eligibility prediction. | The designed method performs with the highest accuracy of 95%, sensitivity and specificity of 95.4 and 97.3%, when compared to the existing methods such as fuzzy neural network (Fuzzy NN), multiple partial least squares regression model (Multi\_PLS), instance-based entropy fuzzy support vector machine (IEFSVM), deep recurrent neural network (Deep RNN). |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 5 | LOAN ELIGIBILITY PREDICTOR | 2021 |  | The project aims to minimize risk in loan approval by using machine learning on historical loan data, training a model to predict the safety of granting loans. By automating the selection process, the objective is to efficiently identify deserving applicants, benefiting both banks and applicants. The model's precision is crucial in maximizing the effectiveness of loan prediction. | Logistic regression is used in this paper. Logistic regression is a simple yet very effective classification algorithm so it is commonly used for many binary classification tasks. Customer churn, spam email, website or ad click predictions are some examples of the areas where logistic regression offers a powerful solution. | The product can easily help in identifying the deserving applicants thus producing quick results for the customer by banks. Also it can help the banks in determining whether assigning a loan to a person is safe or not thus answering how risky is the borrower. It is necessary for the banks to identify the risk factor while granting a loan, thus our prediction model can be very useful in such cases. |
| 6 | Loan Approval Prediction System Using Machine Learning | 2020 |  | The objective is to enhance loan approval processes in the banking sector by implementing a Loan Prediction System using Python. The system employs the random forest algorithm to classify data and assess the eligibility of customers based on various parameters. The goal is to identify safe customers who can reliably repay loans, optimizing the bank's asset investment strategy. |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **S.No.** | **Title of the paper** | **Year** | **Reference with DOI** | **Objective of the paper** | **Focus of the paper(discuss methods used)** | **Summary** |
| 7 | An Approach For Prediction Of Loan Approval Using Machine Learning Algorithm | 2021 | https://ijcrt.org/papers/IJCRT2106313.pdf | This paper will explore the process and result on formulating a new machine learning model that could predict a loan default; but more importantly, the model will focus on minimizing the overall loss in investment of bad loans in order to lessen the burden passed onto individual investors. As a side note, the paper will also explore privacy-preserving mechanism on sensitive information provided from the borrow’s credit report. The end goal is to evaluate a simplified version of RAPPOR (Randomized Aggregately PrivacyPreserving Ordinal Response) and determine whether data that have been hashed by this algorithm could still be use to predict loan default as stated previously. | This paper employs the Decision Tree Machine Learning Algorithm which efficiently performs both classification and regression tasks | Data cleaning, imputation, exploratory analysis led to model with 0.811 accuracy. Credit history crucial; high income, low sanction likely approved. Gender, marital status less impactful. |
| 8 | Loan Eligibility prediction using Machine Learning algorithms | 2022 | https://www.jetir.org/papers/JETIRFP06097.pdf | The objective is to develop a predictive model, using machine learning techniques ,to assess and automate the loan approval process in banking. The focus is on evaluating the credit risk of individual loan applicants, predicting whether they are likely to repay or default. The model aims to provide a quick, efficient, and automated method for selecting qualified applicants, helping banks make informed decisions and manage risks associated with loan distribution. | Support Vector Machine and Random Forest algorithms are used. | The proposed system for Bank loan credibility prediction may help the organizations in making the right decision to approve or reject the loan request of the customers. This can definitely help the banking industry to open up efficient de- livery channels and the huge financial losses. |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| S.No. | Title of the paper | Year | Reference with DOI | Objective of the paper | Focus of the paper | Summary |
| 9 | Loan Approval Prediction using Machine Learning Algorithms Approach | 2021 | https://ijirt.org/master/publishedpaper/IJIRT151769\_PAPER.pdf | To create a machine learning model for banking loan approval prediction, emphasizing credit risk reduction. Test Logistic Regression, Decision Tree, Support Vector Machine, and Random Forest, identifying Support Vector Machine as the most accurate method for informed decision-making and loss reduction. | The models used are Logistic Regression, Decision Tree, Support Vector Machine, and Random Forest. | The prediction method begins with data pre-processing, filling the missing values, experimental data analysis. After evaluating model on test dataset, each of these algorithms obtained a precision rate between 70% and 80%. Although here it can be concluded with certainty that the Support Vector Machine model is very efficient and produces superior results than other models. |

## 

## 3.EXISTING SYSTEM

Traditional loan eligibility assessments are known for their time-consuming nature, involving meticulous

document verification, credit checks, and manual evaluations. Human underwriters must gather and verify numerous documents like proof of income, employment records, bank statements, and credit reports from multiple bureaus.

Each document is scrutinized for accuracy, and any discrepancies can trigger additional investigations, causing delays. Additionally, the process often includes multiple layers of review within the bank, creating bottlenecks and leading to extended waiting times for loan applicants.

The loan eligibility checking process seamlessly integrates both manual and automated procedures to holistically evaluate a borrower's qualifications. Starting with personalized interactions for information gathering, lenders engage in direct conversations and interviews, gaining nuanced insights into the borrower's financial situation. Meticulous document verification by human reviewers ensures the authenticity of submitted materials. Credit checks, while facilitated electronically, often involve manual analysis by financial experts to interpret the intricacies of a borrower's credit history. Human underwriters play a pivotal role in assessing the debt-to-income ratio, considering factors that automated systems might overlook. Collateral assessments involve on-site inspections by experts, adding a personalized touch to the evaluation process. Throughout decision-making, a personalized approach is maintained as human experts review all aspects of the borrower's profile.

# 4. SYSTEM REQUIREMENTS

**Functional Requirements**

* The system must be able to use a variety of machine learning algorithms to combine the features and make predictions.
* The system must be able to use a large and diverse dataset to train the machine learning models.
* The system must be able to evaluate its own performance and identify areas for improvement.
* The web page must display the machine learning algorithm's prediction for the outcome.

# Non Functional Requirements

* The system must be scalable to handle a large number of users and predictions.
* The system must be reliable and available 24/7.
* The system must be easy to use and navigate.
* The system must be well-documented so that users and developers can understand how it works.
* The machine learning algorithm must be reliable and produce accurate predictions consistently.
* The web page must be responsive and load quickly on all devices.
* The web page must be secure and protect user data from unauthorized access.
* The web page must be easy to use and navigate.

# 5. SOFTWARE DESIGN

## 4.1 UML DIAGRAMS

The Device Architecture Manual describes the application requirements, operating state, application and subsystem functionality, documents and repository setup, input locations, yield types, human-machine interfaces, management reasoning, and external interfaces. The Unified Modeling Language (UML) assists software developers in expressing an analysis model through documents that contain a plethora of syntactic and semantic instructions. A UML context is defined as five distinct viewpoints that present the system from a particularly different point of view.

The components are similar to modules that can be combined in a variety of ways to create a complete UML diagram. As a result, comprehension of the various diagrams is essential for utilizing the knowledge in real-world systems. The best method to understand any complex system is to draw diagrams or images of it. These designs have a bigger influence on our understanding. Looking around, we can see that info-graphics are not a new concept, but they are frequently utilized in a variety of businesses in various ways.

**User Model View**

The perspective refers to the system from the clients' point of view. The exam's depiction depicts a situation of utilization from the perspective of end-clients. The user view provides a window into the system from the perspective of the user, with the system's operation defined in light of the user and what the user wants from it.

**Structural model view**

This layout represents the details and functionality of the device. This software design maps out the static structures. This view includes activity diagrams, sequence diagrams and state machine diagrams

**Behavioral Model View**

It refers to the social dynamics as framework components, delineating the assortment cooperation between various auxiliary components depicted in the client model and basic model view. UML Behavioral Diagrams illustrate time-dependent aspects of a system and communicate the system's dynamics and how they interact. Behavioral diagrams include interaction diagrams, use case diagrams, activity diagrams and state–chart diagrams.

**Implementation Model View**

The essential and actions as frame pieces are discussed in this when they are to be

manufactured. This is also referred to as the implementation view. It uses the UML Component diagram to describe system components. One of the UML diagrams used to illustrate the development view is the Package diagram.

**Environmental Model View**

The systemic and functional component of the world where the program is to be introduced was expressed within this. The diagram in the environmental view explains the software model's after-deployment behavior. This diagram typically explains user interactions and the effects of software on the system. The following diagrams are included in the environmental model: Diagram of deployment.

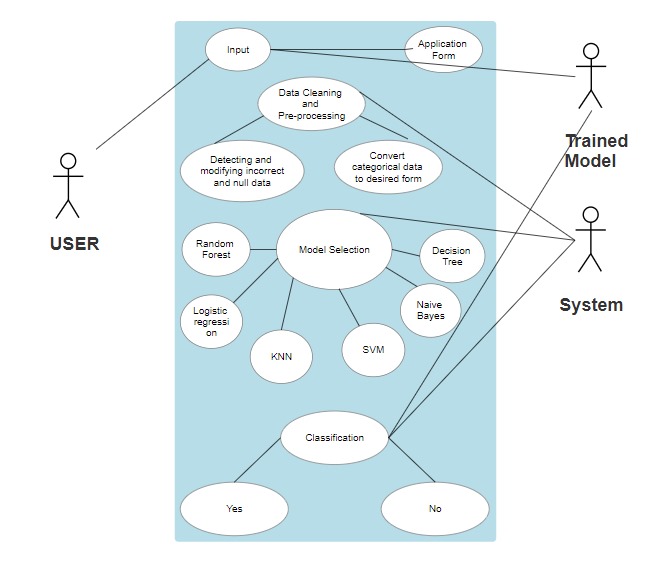
The UML model is made up of two separate domains:

* Demonstration of UML Analysis, with a focus on the client model and auxiliary model perspectives on the framework.
* UML configuration presenting, which focuses on demonstrations, usage, and natural model perspectives.

### USE CASE DIAGRAM

A use case diagram is a kind of behavioral diagram that is used in the Unified Modeling Language (UML). This type of diagram is defined by and developed from use case research. Its purpose is to provide a graphical representation of a system's functionality in terms of its actors, the goals of the actors that they want to achieve (which are stated as use cases), and any relationships that exist between those use cases. The primary objective of a use case diagram is to specify which system functions are carried out for particular actor.

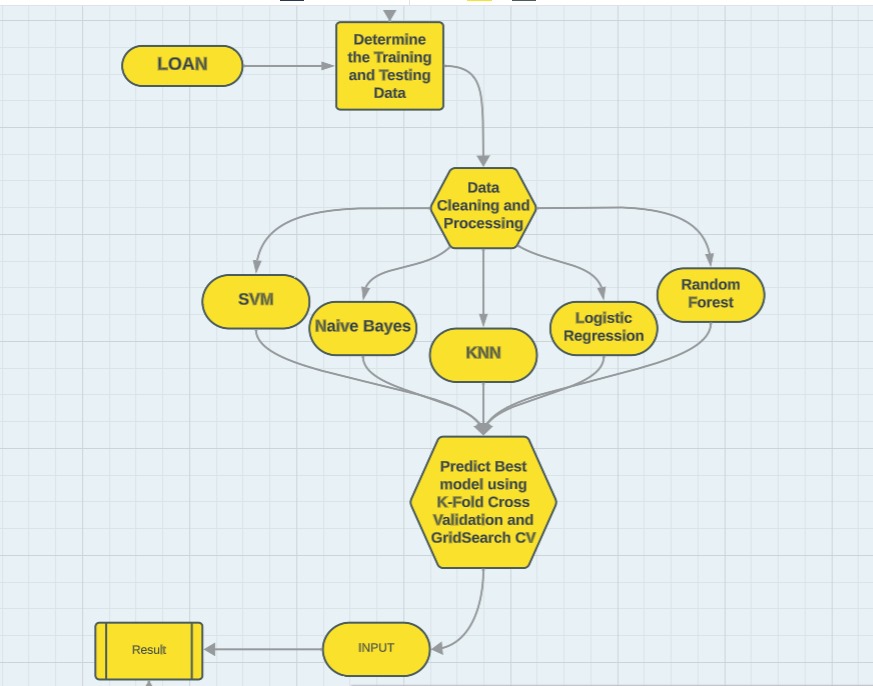
It is possible to demonstrate the roles that each player plays inside the system.



### 

### ACTIVITY DIAGRAM

The activity diagram presents a representation of the system's process flows. A state diagram is similar to an activity diagram in that it consists of activities, actions, transitions, beginning and end states, and guard conditions.



# 6. PROPOSED SYSTEM

To address the challenges associated with loan processing, we have implemented an automatic loan prediction system leveraging advanced machine learning techniques. Departing from traditional methods, our approach involves comprehensive data gathering on loan applicants, encompassing personal details such as average income, marital status, number of dependents, educational background, properties owned, past loan records, and even criminal records if any. Subsequently, we employ a series of machine learning model trained on a substantial dataset to enable the system to analyze and comprehend the intricacies of the loan approval process. The machine, through its learning, evaluates applicants to determine eligibility, providing us with efficient and accurate results. The most accurate model is then selected for further predictions by the user. The user enters the required details like income status, educational qualification, employment status and various property details into a sample webpage which is powered by a machine learning This automated system boasts several advantages, notably a significant reduction in the time required for loan sanctioning. By automating the entire process, the system minimizes the likelihood of human errors, ensuring a streamlined and error-free assessment. Eligible applicants benefit from swift loan sanctioning without unnecessary delays, contributing to a more seamless and expedited lending process. The implementation of this system also guarantees higher accuracy in decision-making, enhancing the overall reliability and efficiency of the loan approval process.

## 

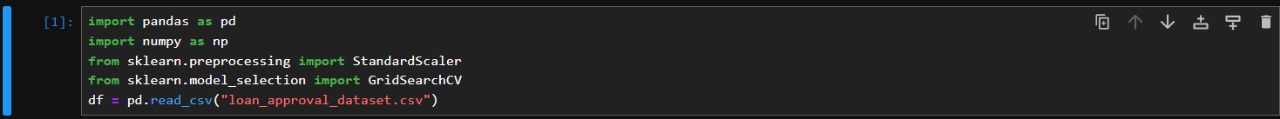
## MODULES

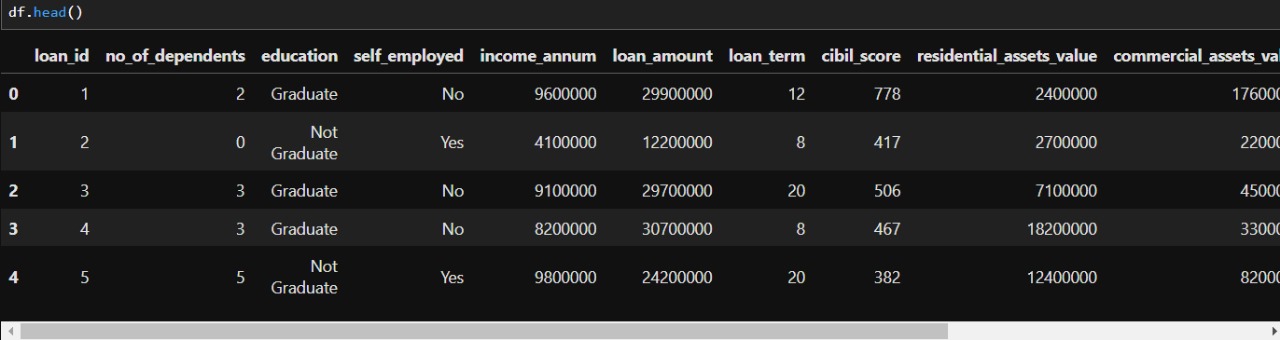
* **Data collection**: This module is responsible for collecting historical data on loan lending, including income status, bank assets, luxury assets, residential assets and commercial assets among many more details of users. This data can be obtained from various sources such as Kaggle.
* **Data preprocessing**: This module is responsible for cleaning and preparing the collected data for machine learning. This might involve removing irrelevant or incomplete data, transforming the data into a consistent format, and normalizing the data.
* **Model training**: This module is responsible for training the selected machine learning model on the preprocessed data. The training process involves splitting the data into training and validation sets, fitting the model on the training set, and evaluating the model on the validation set. Machine learning models like Logistic Regression, Support Vector Machine classifier, KNN, Naïve-Bayes, Decision Tree Classifier and Random Forest Classifier are trained on the prepared dataset.
* **Model evaluation**: This module is responsible for evaluating the trained machine learning model on a held-out test set to assess its performance on unseen data. Libraries such as GridSearchCv are used for parameter tuning to obtain high accuracy. This helps in determining how well the model is likely to generalize to new data. Evaluation metrics, including RMSE, MSE, MAE, r2, and adjusted r2, are employed to measure the accuracy and effectiveness of the models.
* **Model Selection**: Based on the evaluation results, the Model Selection module chooses the best-performing prediction model. Factors considered include accuracy, interpretability, and computational efficiency to ensure the selected model aligns with the system's goals and constraints.
* **Model deployment**: This module is responsible for deploying the trained and evaluated machine learning model to production so that it can be used to make predictions on new data. This involves saving the model to a file, deploying it to a cloud platform, or embedding it in a software application.
* **User Interface**: Creating an easy-to-use interface is crucial for user interaction. This module focuses on designing a user-friendly interface that allows users to enter the required details which are then fed into the model for prediction. The simple UI is created using HTML and CSS.
* **Visualization:** This module is responsible for creating data visualizations to help you understand the data and evaluate the performance of the machine learning model

# 7. CODING AND IMPLEMENTATION

**Importing and Loading Dataset:**

Dataset is collected from Kaggle website and below are the required libraries that we need to import.





# Information about the dataset:

# In this section, we are obtaining information about the dataset. The information includes the dimensions of the dataset, mean, standard deviation, range, etc. This is helpful for us to understand more about the dataset.

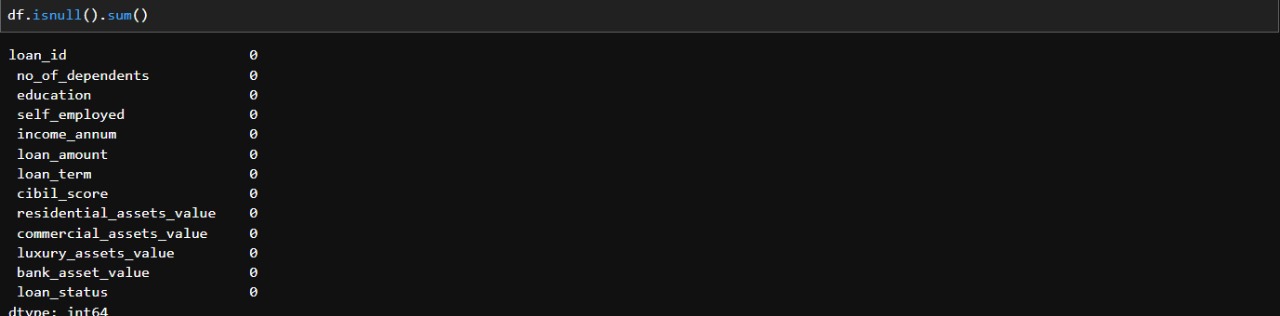
# 

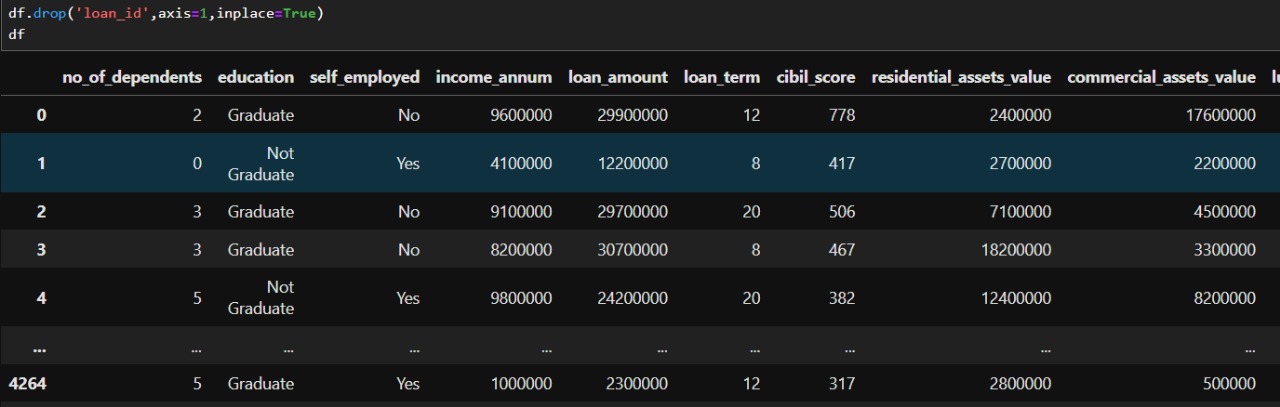
# 

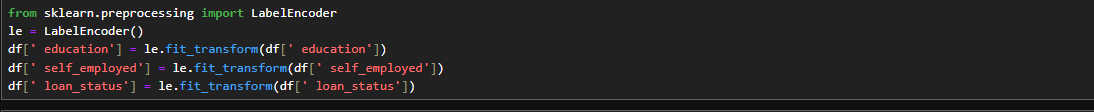
# 

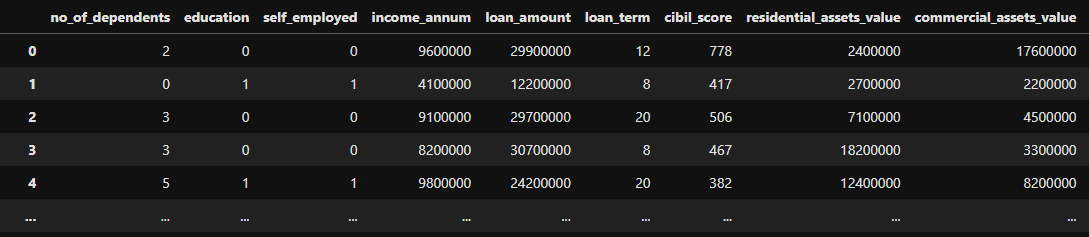
# Data Processing:

In this section we checked if there are any null values in different feature columns in order to replace with some standard measure(like mean etc.). Also, the columns which are not required for training (like ‘loan\_id’) are removed. The columns which contain categorical values are identified and replaced with numerical values, using the LabelEncoder class in the scikit-learn library.



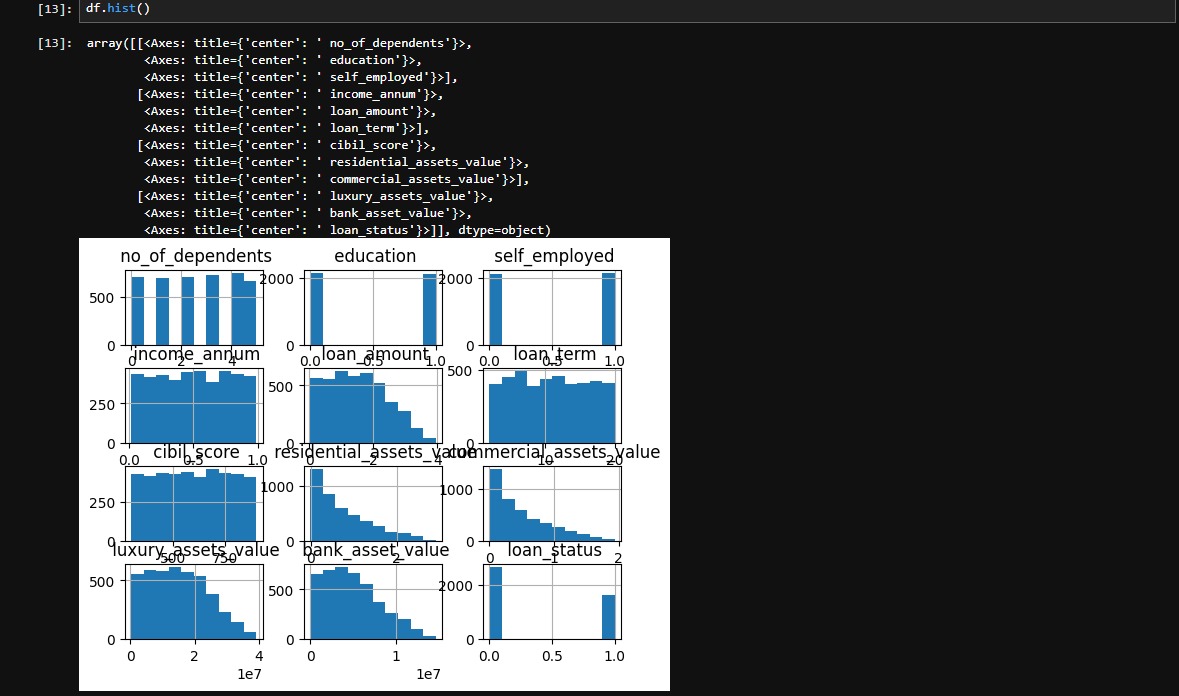






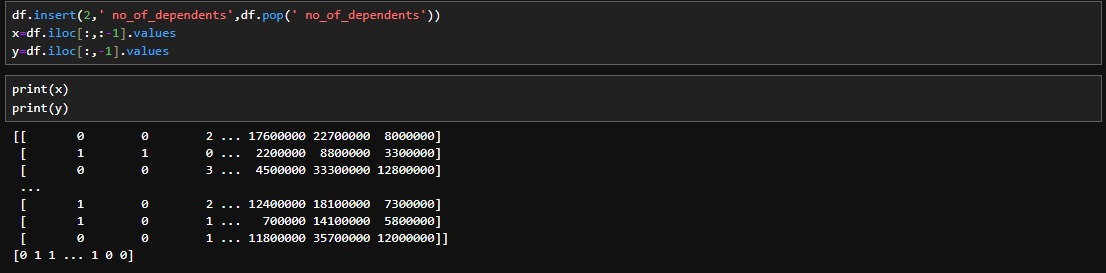
# Visualization the data:

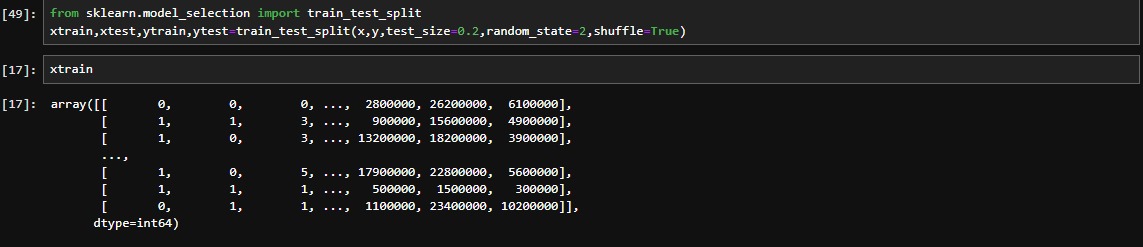
Exploring the 12 columns of the dataset and understanding how the data varies.



# Splitting into training and test sets:

In below section we are splitting the dataset into training and test sets. 80% of the dataset is used for training and remaining for testing. The train\_test\_split class of scikitlearn library is used.

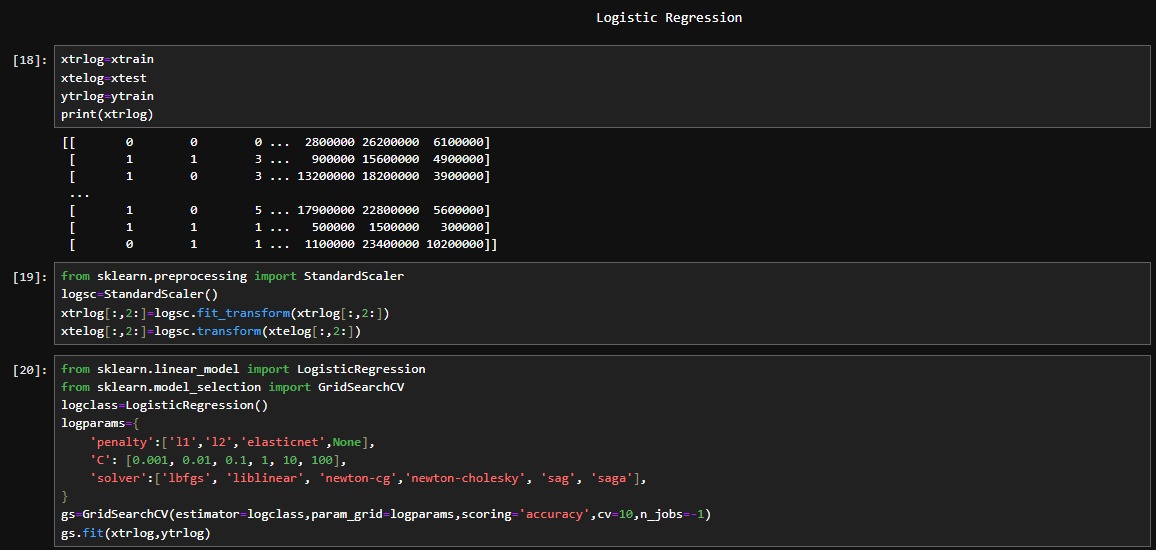




# Model Implementation:

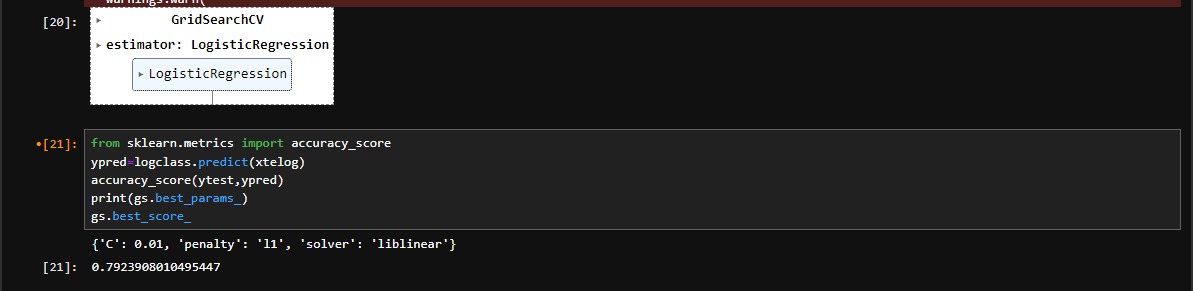
# Logistic Regression:

Logistic regression is a statistical method in machine learning that models the probability of a binary outcome by applying a logistic function to a linear combination of input features.



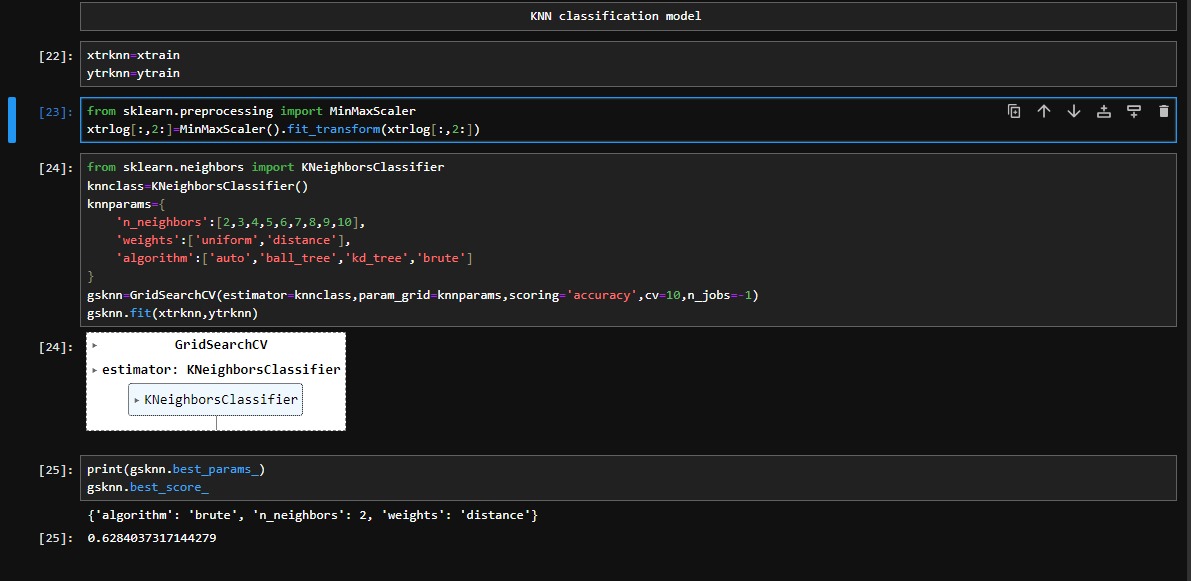
The training and testing sets are first scaled using Standard Scaler available in the scikit learn library.

The required classes such as GridSarchCV are imported. A list of parameters are given for the model to train on and the best parameters which give the highest accuracy are displayed. We can obtain an accuracy of around 79.2% for the best parameters in logistic regression.



# KNN classification:

K-Nearest Neighbors (KNN) classification is a supervised machine learning algorithm that assigns a data point to the majority class of its k nearest neighbors in the feature space, based on a chosen distance metric. It is effective for both binary and multiclass classification tasks.



The training and testing sets are first normalized using MinMax Scaler available in the scikit learn library.

The required classes such as GridSarchCV are imported. A list of parameters are given for the model to train on and the best parameters which give the highest accuracy are displayed. We can obtain an accuracy of around 62.8% for the best parameters in KNN classification.

**SVM classification:**

Support Vector Machine (SVM) classification is a supervised machine learning algorithm that finds the hyperplane in a high-dimensional space to separate classes and maximize the margin between them. It is effective for linear and non-linear classification tasks.



The training and testing sets are first scaled using Standard Scaler available in the scikit learn library.

A list of parameters are given for the model to train on and the best parameters which give the highest accuracy are displayed. We can obtain an accuracy of around 63.5% for the best parameters in SVM regression.

**Naïve-Bayes Classifier:**

Naive-Bayes classification is a probabilistic machine learning algorithm based on Bayes' theorem. It assumes independence among features and calculates the probability of each class given a set of features, making it efficient for text classification and simple yet effective in various applications.

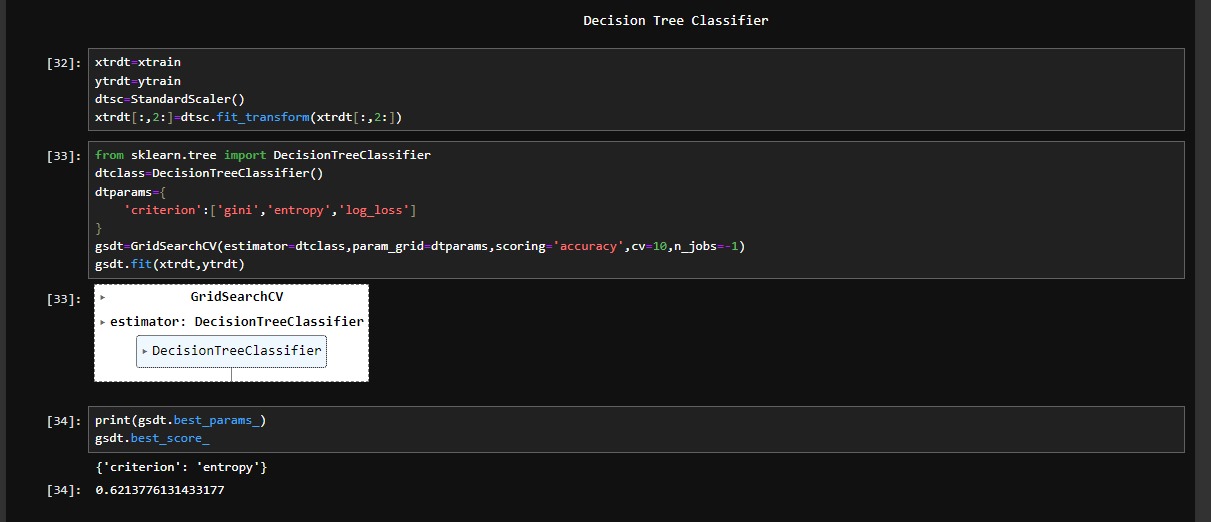


The training and testing sets are first scaled using Standard Scaler available in the scikit learn library.

A list of parameters are given for the model to train on and the best parameters which give the highest accuracy are displayed. We can obtain an accuracy of around 58.6% for the best parameters in Naïve-Bayes classifier.

**Decision Tree Classifier:**

Decision Tree classification is a machine learning algorithm that recursively splits the dataset into subsets based on the most significant features, creating a tree-like structure for decision-making. It is intuitive, interpretable, and widely used for both classification and regression tasks.

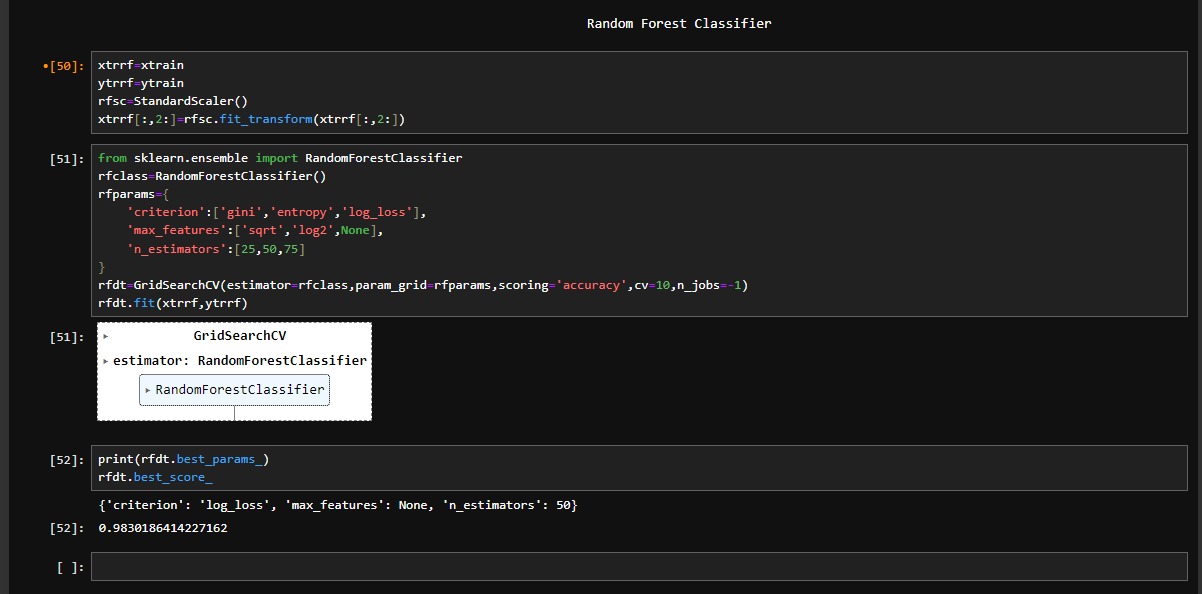


The training set is first scaled using Standard Scaler available in the scikit learn library.

A list of parameters are given for the model to train on and the best parameters which give the highest accuracy are displayed. We can obtain an accuracy of around 62.1% for the best parameters in Decision Tree Classifier.

**Random Forest Classifier:**

Random Forest is an ensemble learning method in machine learning that constructs multiple decision trees during training and outputs the mode of the classes for classification. It enhances accuracy and reduces overfitting by combining the predictions of individual trees.



The training set is first scaled using Standard Scaler available in the scikit learn library.

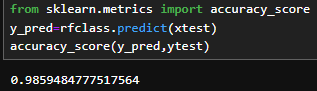
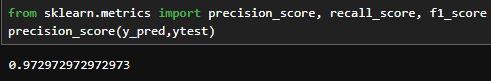
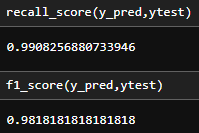
A list of parameters are given for the model to train on and the best parameters which give the highest accuracy are displayed. We can obtain an accuracy of around 98.3% for the best parameters in Random Forest Classifier. This model has the highest accuracy.

# 8. RESULTS

Accuracy score, Precision score, Recall Score, and f1 score were utilized as the assessment parameters.

1. **Accuracy Score:** Measures the overall correctness of your model by calculating the ratio of correctly predicted instances to the total instances. It's a good general metric but may be insufficient if there is a class imbalance (unequal distribution of loan approval and rejection), as high accuracy can be achieved by simply predicting the majority class.
2. **Precision Score:** Indicates the accuracy of positive predictions. In the context of loan eligibility, precision assesses the proportion of correctly predicted approved loans among all instances predicted as approved. High precision is crucial when false positives (incorrectly approving a loan) have significant consequences.
3. **Recall Score:** Evaluates the model's ability to capture all positive instances. In the loan approval scenario, recall measures the proportion of correctly predicted approved loans among all actual approved instances. High recall is essential when false negatives (incorrectly rejecting a legitimate loan) are a concern.
4. **F1 Score:** Represents the harmonic mean of precision and recall. It provides a balance between precision and recall, making it a useful metric when both false positives and false negatives need to be minimized. F1 score is especially valuable in situations where there is an uneven cost associated with misclassifying different classes.

The results obtained are as follows:

A sample prediction has also been made with practical values being fed into the model for estimating the outcome.

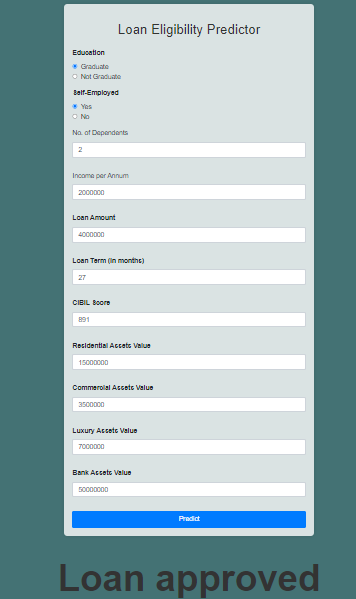


As the label encoder assigned the value 0 to ‘Approved’ and 1 to ‘Rejected’, the outcome here means that the loan is approved.

**Web Integration:**

A web application has been developed using HTML and CSS so that the user can enter his details in the web portal and get an instant result. The backend has been developed using Flask, which is a library in python. Flask is a lightweight and extensible web framework for Python, designed to make it easy to build web applications quickly. It follows the WSGI (Web Server Gateway Interface) standard and is known for its simplicity and flexibility. First, the trained model is stored as a pickle file and then it is loaded into the main application. Then, the user inputs are converted into a numpy array and stored in a variable. This variable is then fed into the model to predict the outcome, and thus the outcome is rendered onto the webpage.





# 9. CONCLUSION AND FUTURE SCOPE

# In conclusion, the loan eligibility prediction project has demonstrated its significance in automating and enhancing the decision-making process for financial institutions. By leveraging machine learning algorithms and predictive modeling, the project contributes to more efficient and objective evaluations of loan applications. The incorporation of assessment metrics such as accuracy, precision, recall, and F1 score ensures a robust evaluation of the model's performance, considering both false positives and false negatives, which are crucial factors in the context of loan approvals. As a result, financial institutions can benefit from reduced manual workload, faster processing times, and improved risk management, ultimately leading to better-informed lending decisions.

# Looking ahead, the future scope of the loan eligibility prediction project involves continuous refinement and optimization of the machine learning model. The incorporation of more diverse and real-time data sources can enhance the model's accuracy and adaptability to changing economic conditions. Additionally, exploring advanced techniques such as deep learning or ensemble methods could provide further improvements. Moreover, the project's scalability could be expanded to handle a larger volume of data and accommodate more complex decision-making scenarios. Collaborations with regulatory bodies and industry stakeholders can also play a role in establishing standardized practices and ensuring compliance with evolving financial regulations. Overall, the ongoing evolution of the project should focus on staying abreast of technological advancements, refining model interpretability, and addressing emerging challenges in the financial landscape.

# REFERENCES

1. <https://www.researchgate.net/publication/355061139_Prediction_of_IPL_Match_Outcome_Using_Machine_Learning_Techniques>
2. <https://www.researchgate.net/publication/261058699_CricAI_A_classification_based_tool_to_predict_the_outcome_in_ODI_cricket>
3. <https://www.researchgate.net/publication/323611656_Predicting_Players'_Performance_in_One_Day_International_Cricket_Matches_Using_Machine_Learning>
4. <https://www.semanticscholar.org/paper/Dynamic-Winner-Prediction-in-Twenty20-Cricket%3A-on-Viswanadha-Sivalenka/9dfe2fe41fa40304b1826a3b93818ed68b4acb57>
5. <https://ictactjournals.in/paper/IJSC_Vol_11_Iss_1_Paper_2_2199_2204.pdf>
6. <https://www.analyticsvidhya.com/blog/2022/05/ipl-team-win-prediction-project-using-machine-learning/>
7. <https://www.researchgate.net/publication/339681163_Prediction_of_Indian_Premier_LeagueIPL_2020_using_Data_Mining_Algorithms#:~:text=Priyanka%20et%20al.,of%2082.73%25.%20...&text=India's%20most%20popular%20sport%20is,T20%2C%20ODI%2C%20and%20Test>.
8. <https://www.ijraset.com/research-paper/super-predictor-of-indian-premier-league-using-various-ml-techniques>
9. <https://www.researchgate.net/publication/346348756_Application_of_Machine_Learning_in_Cricket_and_Predictive_Analytics_of_IPL_2020>