**Problem Statement** – Loan Repayment Predication system

**Team Members:**

1. Kaarthikeyan AV – 2133021
2. Vishwa R – 2133050

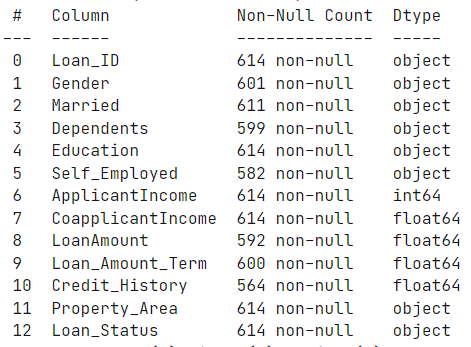
**Abstract:**

In the domain of financial lending, accurately predicting loan repayment is a critical task for financial institutions. This project aims to develop a loan repayment system that utilizes machine learning algorithms to predict the risk of loan default and manage repayment processes effectively. The system will analyse various borrower data points, such as demographics, financial history, and loan details, to identify potential risk factors and predict loan repayment behaviour. This information will be presented to lenders and borrowers, enabling informed decision-making and proactive interventions to prevent defaults. This project aims to develop predictive models for loan repayment using six distinct techniques: Simple Linear Regression, Multiple Linear Regression, Support Vector Machines (SVM), Random Forest, K-Nearest Neighbours (KNN), and Artificial Neural Networks (ANN). These models leverage various borrower characteristics, historical data, and financial features to make informed predictions.

**Key Components:**

1. Data Collection: Comprehensive data related to past loan transactions, borrower information, credit scores, and financial statements are collected from reliable sources, ensuring a diverse and representative dataset.
2. Data Preprocessing: The dataset undergoes thorough preprocessing, involving cleaning, handling missing values, and transforming data into a format suitable for model training. Features like credit scores, income, employment history, and debt-to-income ratios are prepared for analysis.
3. Feature Engineering: Relevant features are selected and engineered to create meaningful inputs for the predictive models. This includes creating new variables, handling categorical data, and considering historical repayment patterns.
4. Model Selection:
5. Simple Linear Regression
6. Support Vector Machines (SVM)
7. Random Forest
8. Logistic Regression
9. Decision Tree
10. K-Nearest Neighbours (KNN)
11. Artificial Neural Networks (ANN)

**Exploratory Data Analysis:**



This dataset contains information on 614 loan applications, including applicant demographics, financial details, and loan characteristics.

**Key features:**

Loan\_ID: Unique identifier for each loan application.

Applicant information: Gender, marital status, dependents, education, self-employed status.

Financial information: Applicant income, co-applicant income (if applicable).

Loan details: Loan amount, loan term, credit history, property area.

Target variable: Loan\_Status (it’s a binary value).

The average applicant has an income of 5403.46 and a loan amount of 14641.22. The average loan term is 342 months and the average credit history is 0.84. The 25th percentile for applicant income is 2877.50, the 50th percentile is 3812.50, and the 75th percentile is 5795.00. The 25th percentile for loan amount is 128.00, the 50th percentile is 168.00, and the 75th percentile is 261.00. The 25th percentile for loan term is 360.00, the 50th percentile is 360.00, and the 75th percentile is 360.00. The minimum credit history is 0.00 and the maximum is 1.00.

**Data completeness:**

Most features have a high percentage of non-null values, except for:

Gender: 601 non-null values (98%)

Married: 611 non-null values (99%)

Dependents: 599 non-null values (97%)

Self\_Employed: 582 non-null values (95%)

Credit\_History: 564 non-null values (92%)

**Data types:**

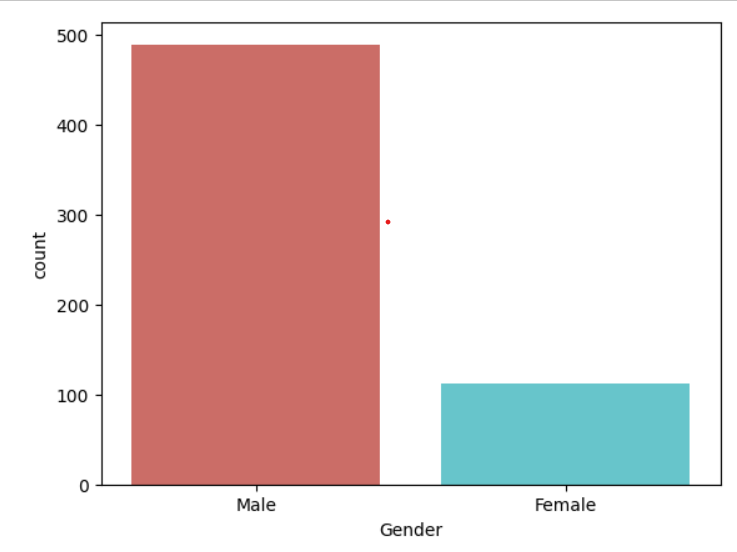
The dataset contains a mix of numerical and categorical features.

Numerical features include: ApplicantIncome, CoapplicantIncome, LoanAmount, Loan\_Amount\_Term, Credit\_History.

Categorical features include: Loan\_ID, Gender, Married, Dependents, Education, Self\_Employed, Property\_Area, Loan\_Status.

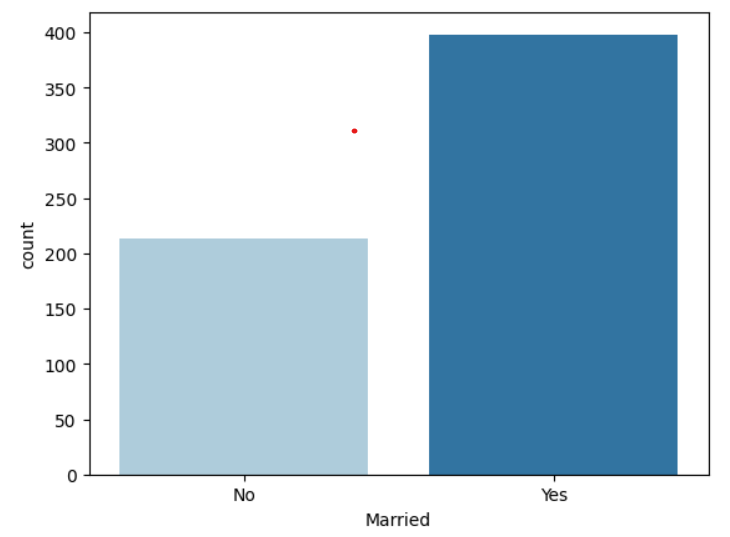
This dataset provides valuable insights into loan repayment patterns and can be used to develop machine learning models for predicting loan defaults. This can help lenders make informed decisions, manage risk, and improve loan portfolio performance.

**Gender:**



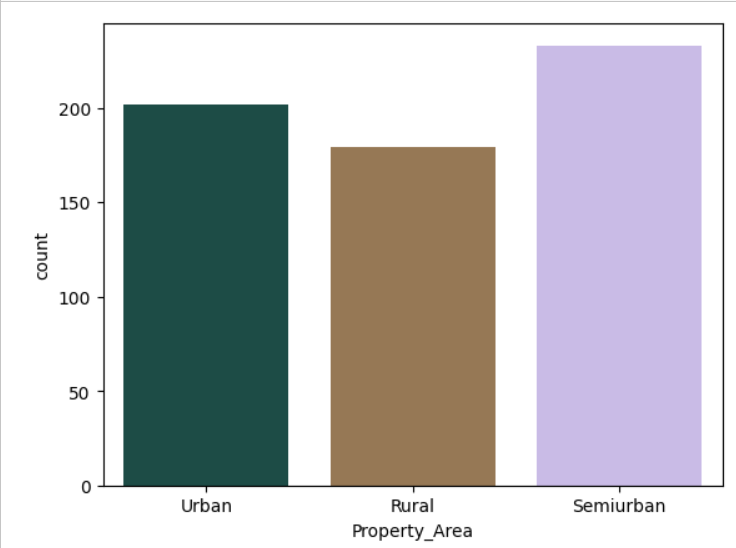
The total of 614 in which there were count of male around 489(79.7%) and female around 112, rest of them were null values. The data is significantly skewed towards the Male category, with nearly 80% of the individuals being male.

**Married:**



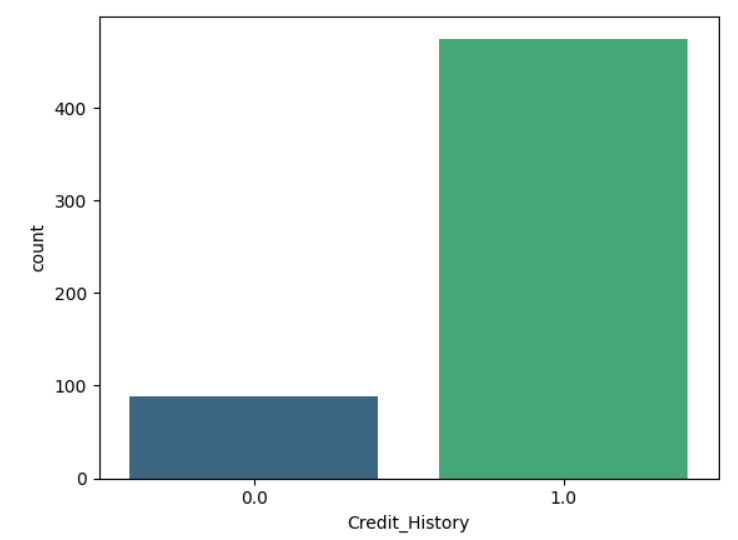
The above graph mentions about the count of married in the given dataset with 398 married and 213 are not married and rest of them were null values.

**Property Area:**



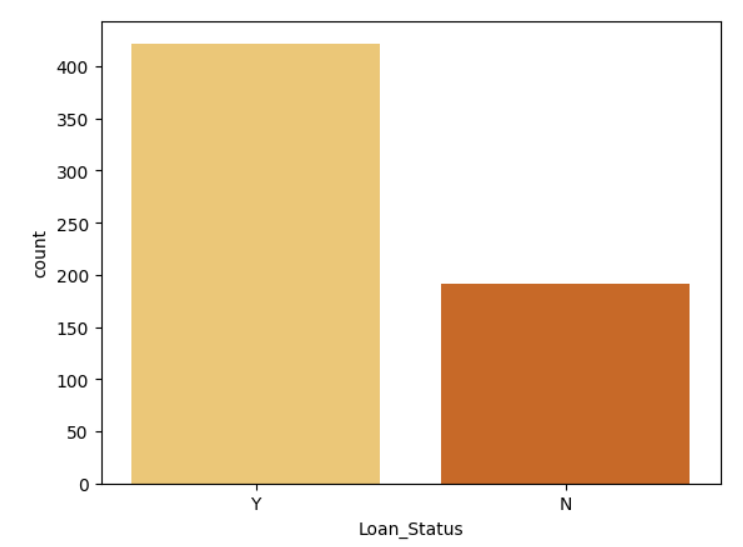
The above bar plots the represents the property area where the customer holds in are divided into three different sections are Urban (202), Semiurban (233), Rural (179).

**Credit History:**



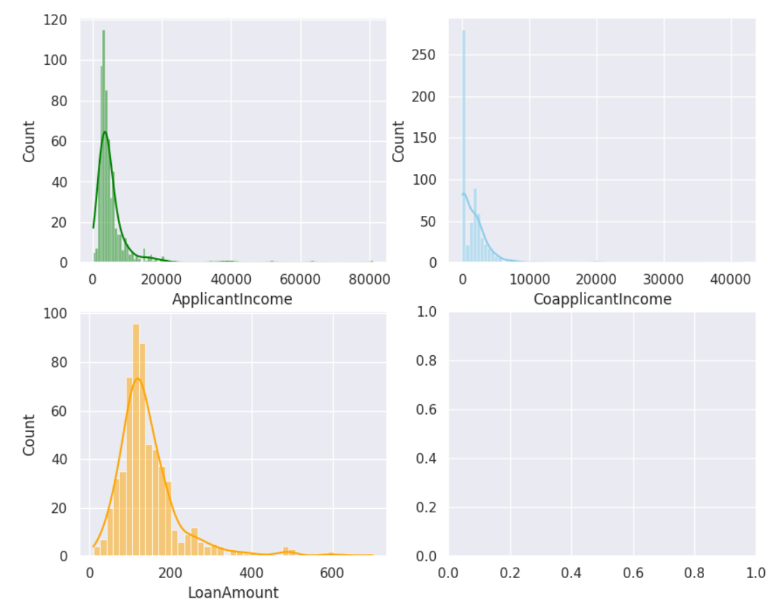
The above bar plot explains about the credit history good credit history (1.0): 475 (77.5%), Bad credit history (0.0): 89 (14.5%), Missing values: 50 (8.1%). The majority of individuals (77.5%) have good credit history. A significant minority (14.5%) have bad credit history.

**Loan Status:**



The above bar chart explains about the loan status that has been change of loan repayment (Y): 422 (68.7%), (N): 192 (31.3%). The majority of loans (68.7%) were repaid. A significant minority of loans (31.3%) defaulted.

**Income:**



The average applicant has an income of 5403.46, a co-applicant income of 1621.25, and a loan amount of 14641.22. There is a significant range in income and loan amount values, with minimums of 150 and 9 respectively, and maximums of 81000 and 700. The 25th percentile for applicant income is 2877.50, the 50th percentile is 3812.50, and the 75th percentile is 5795.00. The 25th percentile for co-applicant income is 0.00, the 50th percentile is 1188.50, and the 75th percentile is 2297.25. The 25th percentile for loan amount is 128.00, the 50th percentile is 168.00, and the 75th percentile is 261.00.

**Machine Learning Techniques:**

1. Linear Regression:

R-squared: 0. 377, this indicates that the model explains 37.7% of the variance in the target variable. This value suggests a moderate fit of the model to the data.

Mean Squared Error: 0.156, this represents the average squared difference between predicted and actual values. A lower MSE indicates better model performance.

Accuracy: 0.844, this reflects the percentage of correct predictions made by the model. An accuracy of 84.4% suggests that the model performs well in classifying data points.

Overall:

The model achieves a moderate R-squared and high accuracy, indicating its ability to explain a portion of the data and make accurate predictions.

However, the relatively high mean squared error suggests room for improvement in the model's ability to closely predict actual values.

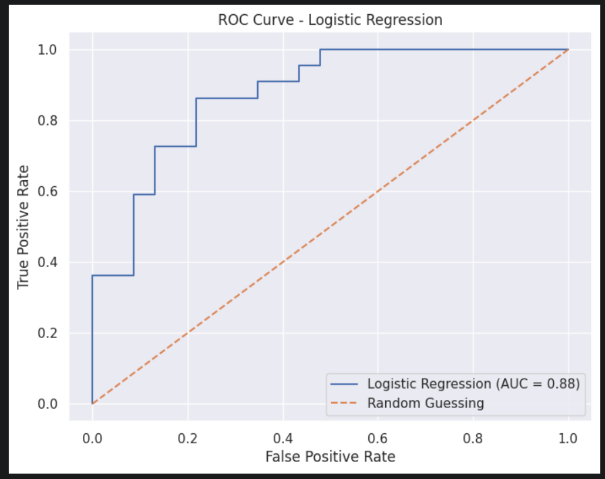
1. Logistic regression:

Class Precision Recall F1-Score Support

0 0.82 0.78 0.80 23

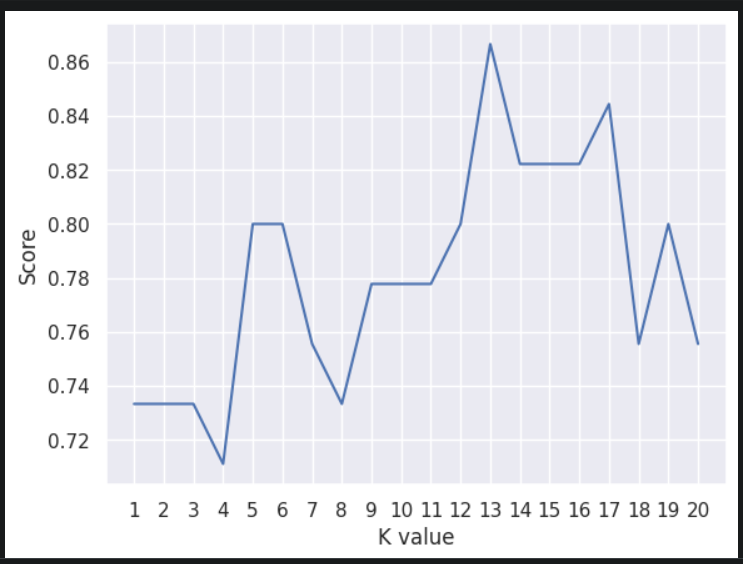
1 0.78 0.82 0.80 22

This Logistic Regression model achieves a relatively high accuracy of 80% on the classification task. Precision, recall and F1-Score are balanced across the two classes, showcasing the model's ability to correctly identify both positive and negative cases. The confusion matrix reveals a small number of misclassified samples, suggesting potential areas for improvement. Further investigation into these misclassified samples and exploration of different model configurations could lead to further refinement and enhanced performance.



1. K-Nearest Neighbour (KNN):

The KNN model achieved a best accuracy of 86.67%. This indicates that the model correctly predicted the target variable for 86.67% of the data points. This is a very good performance, suggesting that the model is well-suited for this classification task.



1. Support Vector Machine (SVM)

Class Precision Recall F1-Score Support

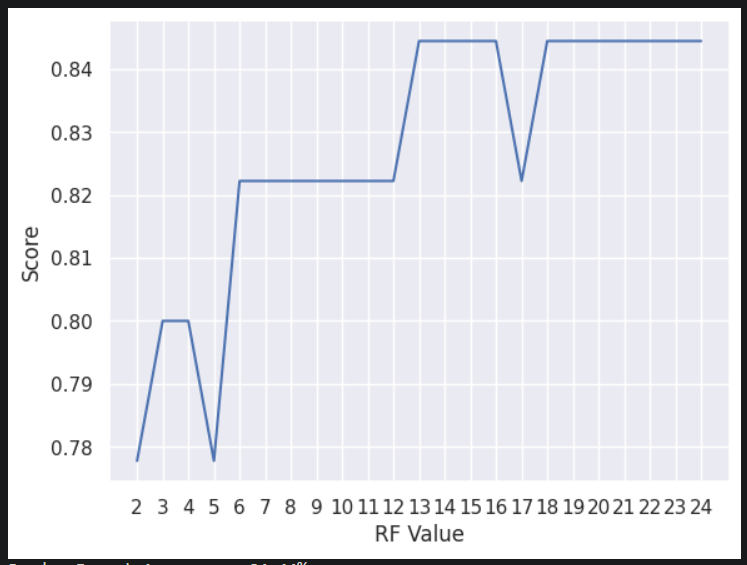
0 0.90 0.78 0.84 23

1 0.80 0.91 0.85 22

This Support Vector Machine (SVM) model demonstrates good performance in classifying data points, achieving a high accuracy of 84.44%. It exhibits balanced F1-scores across both classes, showcasing its ability to accurately identify both positive and negative cases. The confusion matrix reveals a small number of misclassified samples, indicating potential areas for further improvement. Investigating these misclassified samples and exploring different hyperparameter configurations could lead to enhanced model performance.

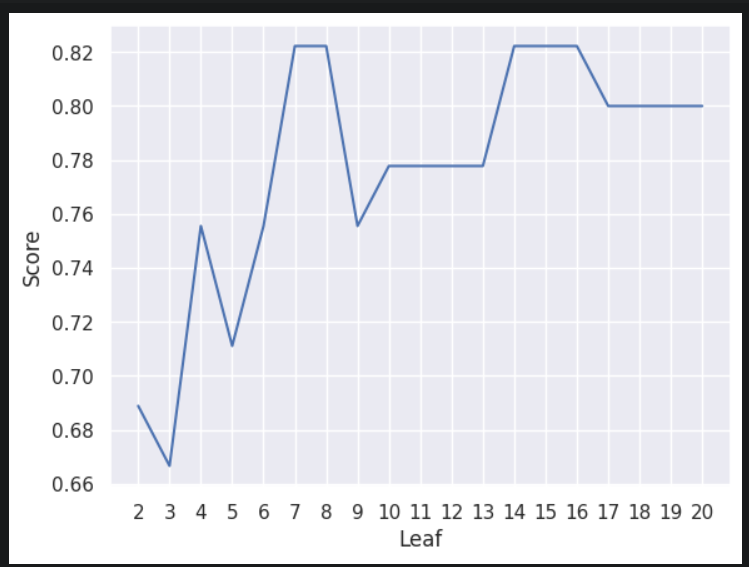
1. Random Forest

The Random Forest model achieved an accuracy of 86.67% on the classification task. This suggests that the model correctly predicted the target variable for 86.67% of the data points. This is a very good performance, indicating that the model is well-suited for this classification task.



1. Decision Tree

The Decision Tree model achieved an accuracy of 82.22% on the classification task. This indicates that the model correctly predicted the target variable for 82.22% of the data points. This is a good performance, suggesting that the model is capable of learning the underlying patterns in the data.



1. **Artificial Neural Network (ANN):**

The neural network trained for 10 epochs. Loss and accuracy steadily improved throughout training. After 10 epochs, the training accuracy reached 71.33% and the validation accuracy reached 66.67%.

On the test set, the model achieved an accuracy of 80.00%.

Key

Takeaways:

The training process converged within 10 epochs.

The model achieved good performance on both the training and test sets, indicating successful generalization.

Further training with more epochs might potentially improve the model's performance even further.

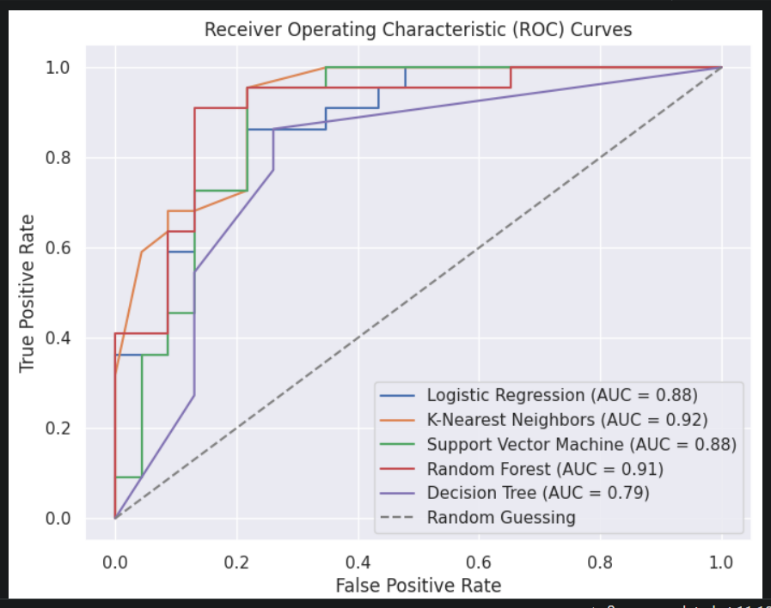
**Comparative Analysis of Machine Learning Models:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Name** | **Accuracy (%)** | **Precision (%)** | **Confusion Matrix** |
| Logistic Regression | 77.78 | 78.77 | [[16, 7], [3, 19]] |
| K-Nearest Neighbour | 86.67 | 75.58 | [[18, 5], [6, 16]] |
| Support Vector Machine | 84.44 | 85.11 | [[18, 5], [2, 20]] |
| Random Forest | 84.44 | 84.49 | [[20, 3], [4, 18]] |
| Decision Tree | 82.22 | 80.6 | [[17, 6], [3, 19]] |

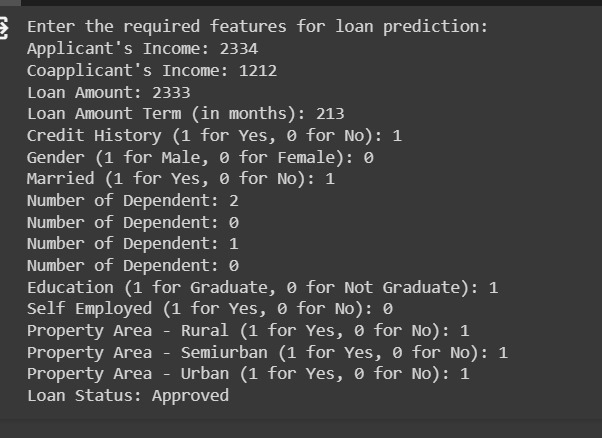
**Key Findings:**

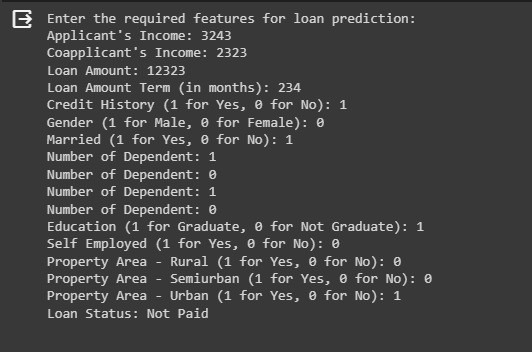
KNN achieved the highest accuracy (86.67%) among all models, suggesting its effectiveness in capturing relationships between data points. Logistic Regression and Support Vector Machine follow closely with accuracies of 77.78% and 84.44% respectively. Random Forest and Decision Tree have slightly lower accuracies (84.44% and 82.22%) but exhibit balanced precision scores. The confusion matrices reveal that all models have some misclassified samples, indicating potential areas for improvement.

**Recommendations:**

KNN seems to be the best performing model based on accuracy. However, further analysis is needed to assess its generalizability on unseen data. Tuning hyperparameters for each model may lead to improved performance. Investigating the misclassified samples can provide insights into the models' weaknesses and guide further improvements. Comparing the models on additional metrics like recall and F1-score can offer a more comprehensive evaluation. Considering the specific task and data characteristics, the most suitable model can be chosen based on its strengths and limitations.

**Test Set:**





**Conclusion:**

In overall the test sample that are obtained from the user is able to classify the loan repayment as paid and not paid, by using the KNN classifier algorithm.