|  |
| --- |
| Loan Default Predictions Report |
| *Loan Default Prediction Using Machine Learning: A Stream-lit Web Application for Financial Risk Management* |

|  |
| --- |
| OMIS 304: MACHINE LEARNING |

****

Figure

Table

|  |
| --- |
|  |
| **GROUP -7** |
| Members:   1. Helina Commey -11115880 2. Eyram Dunyoh-11253839 3. Samson Awinsone Ayamga -11219536 4. Abeeku Williams – 11167380 5. Paula Catherine Bainson- 11120453   Table of contents   1. EXECUTIVE SUMMARY 2. introduction   2.1 Background and Context  2.2 ProblemStatement  2.3 Objectives & PURPOSE  3 . Methodology  3.1 Data Overview -Dataset Description  3.2 Data Sources  3.3 Data Analysis Techniques  3.4 Data Preprocessing --- Handling Missing Values  3.5 Variables and Measure  3.6 Model Training and Evaluation --Model Selection  4. Discussion  4.1 Interpretation of Key Findings  4.3 Challenges and Limitations  5. Recommendations  5.1 Strategic Recommendations  5.2 Implementation Roadmap and Resource Requirements |

****

***1. Executive Summary***

The risk of loan default is a significant challenge for financial institutions, with far-reaching consequences for both lenders and borrowers. The primary goal of this project is to develop a Loan Default Prediction Web Application utilizing machine learning techniques, hosted via Stream-lit, to predict the likelihood of loan default based on historical borrower data. By leveraging classification models, this application assists financial institutions in making informed, data-driven decisions to minimize risk, reduce financial losses, and offer personalized credit terms.

This report provides a detailed analysis of the data processing, model development, evaluation, and deployment of the web application, highlighting the application’s potential to enhance financial decision-making processes.

 **2**. ***Introduction***

**2.1 Background**

Loan defaults represent a serious issue for financial institutions, leading to significant monetary losses. Early identification of high-risk borrowers is crucial for mitigating these risks. Traditional methods often rely on simplistic criteria, which can fail to capture the complexity of borrower profiles. With the advent of machine learning, predictive analytics now allows for a more comprehensive assessment of borrower risk.

Machine learning offers the ability to automatically detect patterns in large datasets, enabling the creation of predictive models that can forecast loan defaults with higher accuracy. In this project, we utilize data from a Kaggle dataset to develop a Stream-lit web application that leverages **machine learning models to predict loan defaults.**

**2.2 Purpose**

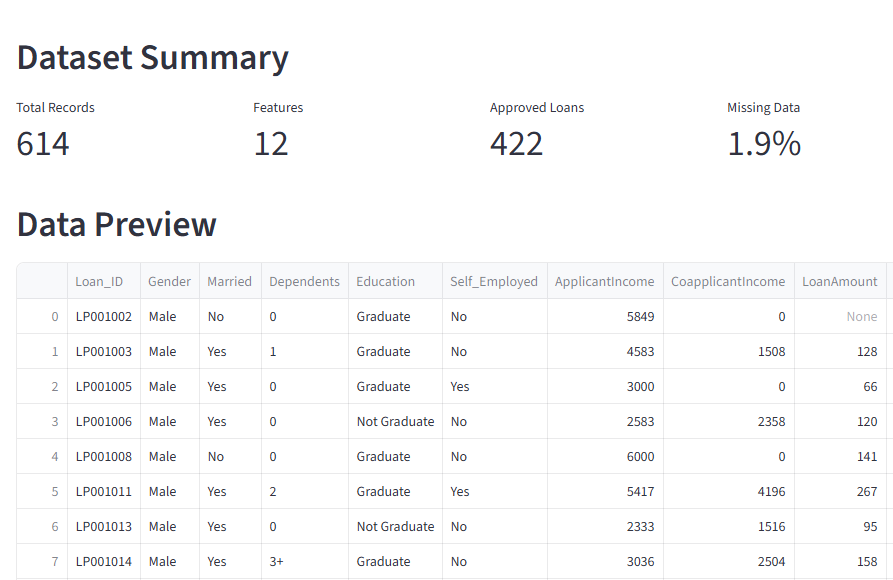
This report presents the development process of an interactive web application that incorporates machine learning models to predict the likelihood of loan defaults. The application is designed to assist lenders in making more accurate and efficient lending decisions, thereby optimizing the loan approval process and minimizing financial risk.

**3. Data Overview**

**Dataset Description**

The dataset used in this project is sourced from Kaggle and contains historical data on borrowers, including their demographic details, financial history, and loan characteristics. The key features of the dataset include:

* ***Applicant-Income***: Income of the loan applicant.
* ***Co-applicant-Income***: Income of the co-applicant, if applicable.
* ***Loan-Amount***: The amount of the loan requested.
* ***Loan-Amount-Term***: The term of the loan in months.
* ***Credit-History***: A binary indicator of whether the applicant has a positive credit history (1) or not (0).
* ***Property-Area*:** The area where the property is located (Urban, Semiurban, Rural).
* ***Loan-Status***: The target variable, indicating whether the loan was approved (1) or denied (0).



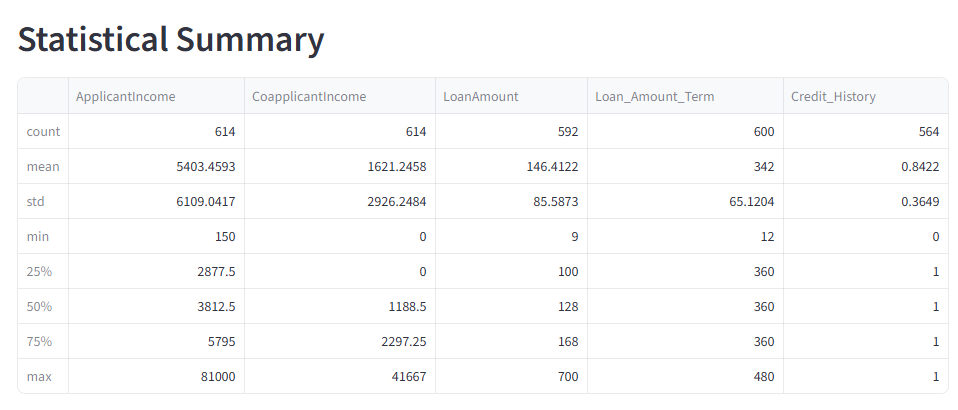
The dataset consists of approximately 614 instances, with multiple features that vary in data type, including numerical, categorical, and binary. There were 1.9% missing observations in the data.



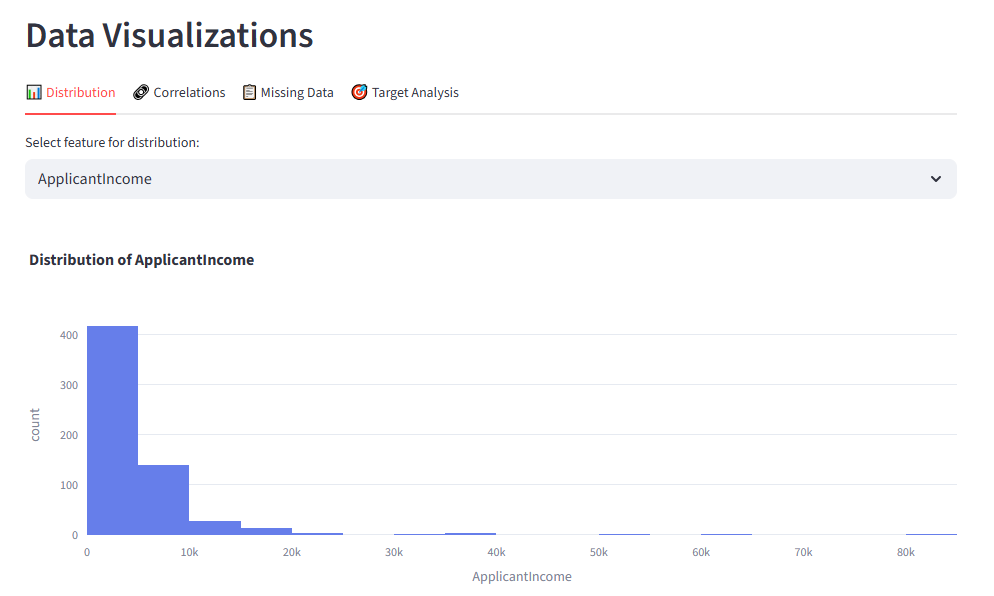
**Exploratory Data Analysis**

* ***Summary Statistics***:

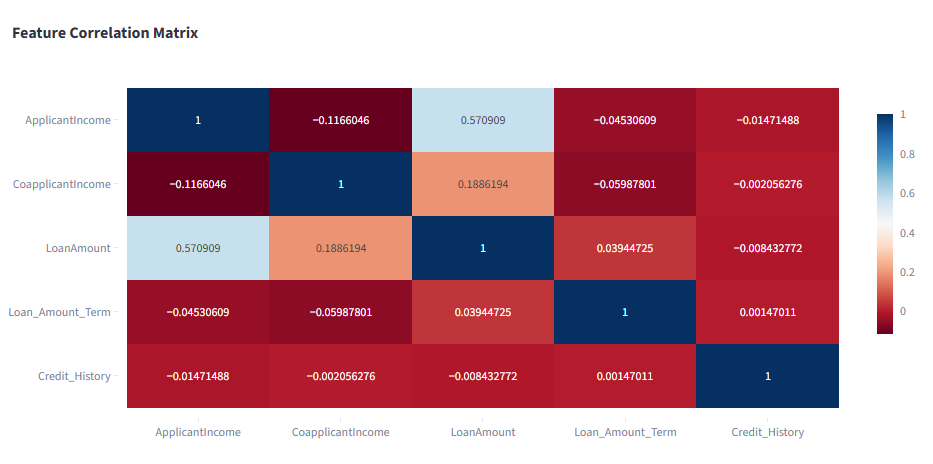
Key statistics were extracted from the dataset, such as the mean loan amount, median applicant income, and proportions of approved vs. denied loans.



* ***Visualizations:***

******

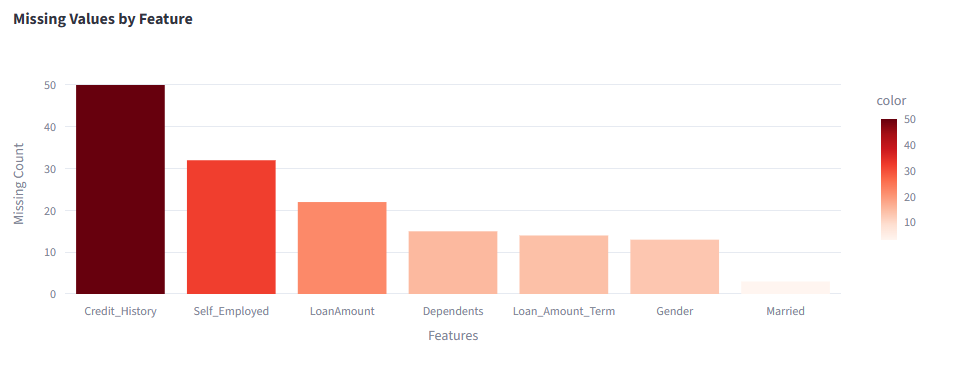
***Histograms***: Histograms were generated for numerical features such as loan amount, income, and loan term to understand their distribution.



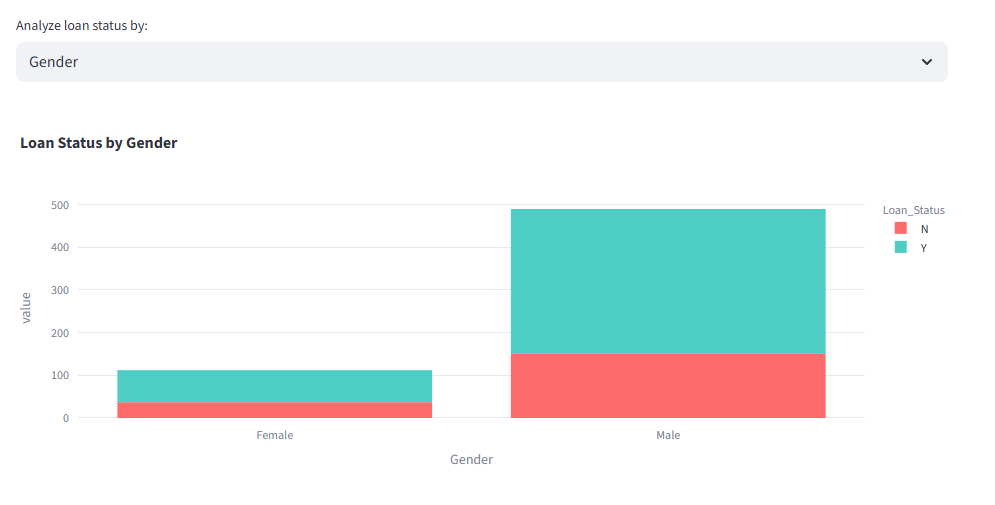
***Correlation Matrix***: A heatmap was used to visualize correlations between numerical attributes, identifying potential relationships that could influence loan default.

The initial analysis indicated that some features, such as Credit-History and Loan-Amount, were strongly correlated with loan approval outcomes, while others, such as Property-Area, had weaker associations.

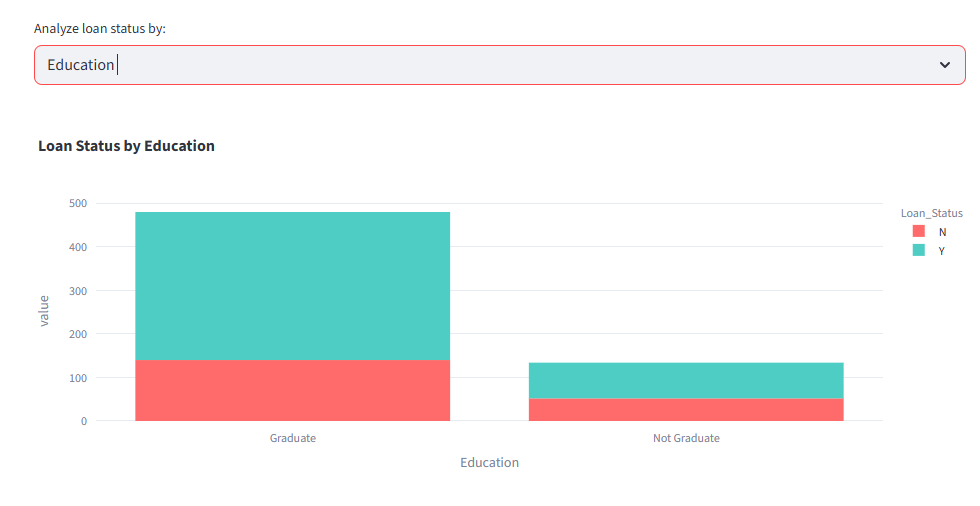
The Credit history feature had the highest of missing observations in our datasets.



* Analyze the loan status by the **Gender**. The data observation has more males with a higher status than Females.



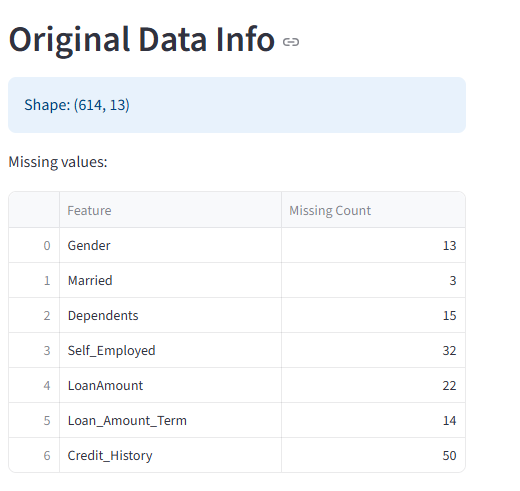
* Analyze the loan status by the **Education**. The data observation has more graduates with a higher status than non-graduates.



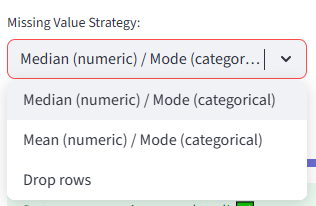
**4. Data Preprocessing**

* *Handling Missing Values*

The dataset contained missing values.



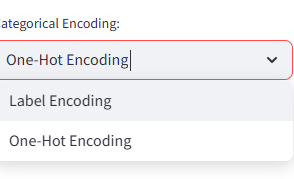
These were handled using mean imputation, where missing values were replaced with the mean of the respective columns. This approach ensured that the dataset remained intact without losing valuable information.



* *Encoding Categorical Variables*

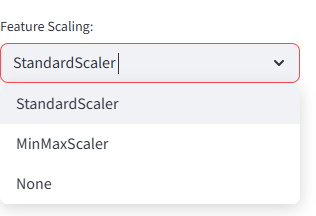
The dataset included categorical variables such as Property-Area and Loan-Status, which were encoded into **numerical values:**

1. **Label Encoding** was applied to Loan-Status, where 1 represents "Approved" and 0 represents "Denied."
2. **One-Hot Encoding** was applied to the Property-Area feature to create binary columns for Urban, Semiurban, and Rural areas.

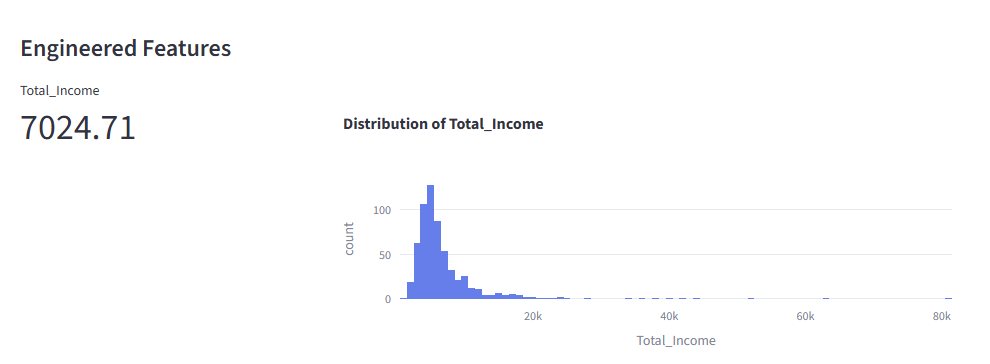


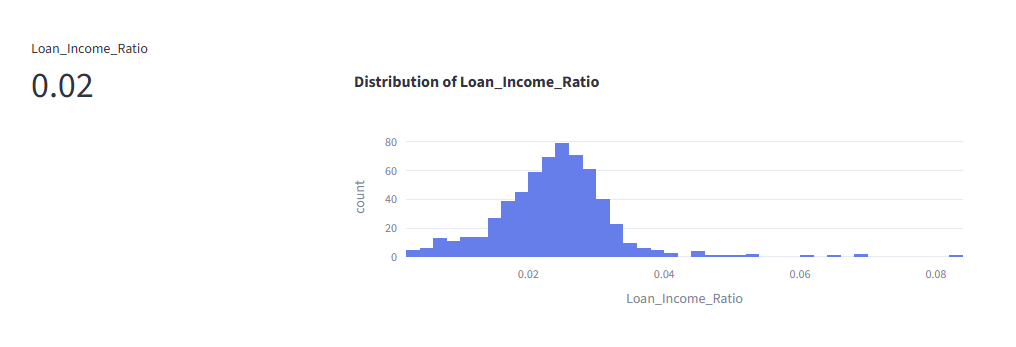
* ***Feature Scaling***

Numerical features, such as Loan-Amount and Applicant-Income, were standardized using Min-Max-Scaler to scale them between 0 and 1. This scaling was necessary for models like K-Nearest Neighbors (KNN) and Support Vector Machines (SVM), which are sensitive to the magnitude of features.

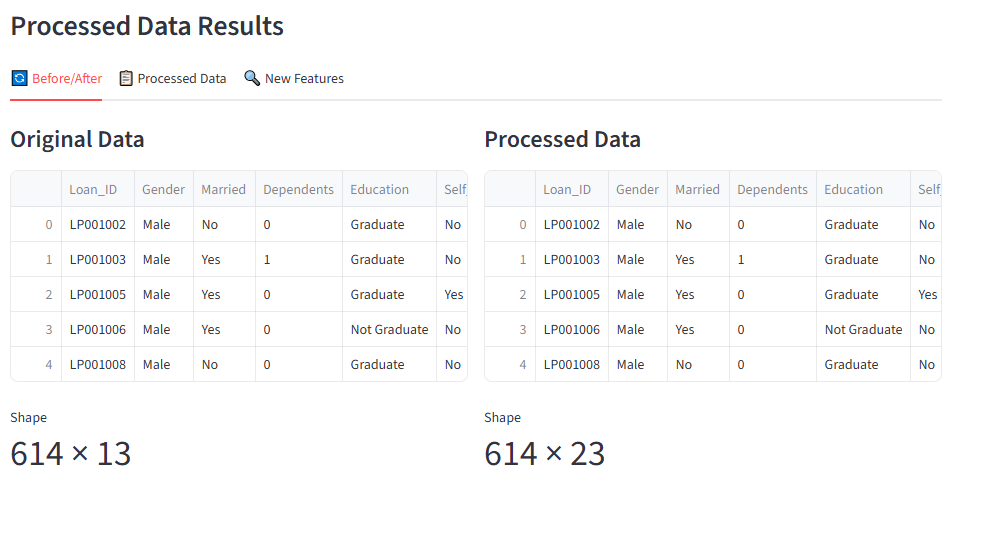


***Feature-Engineering***



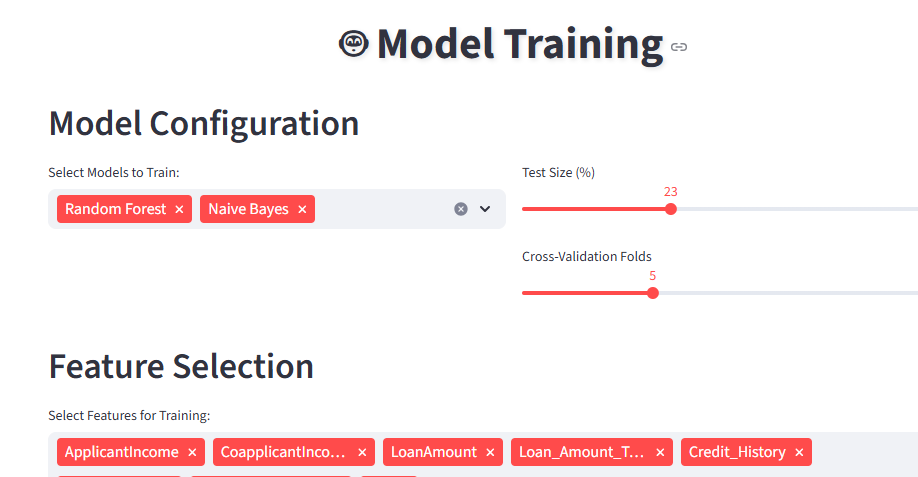


**Processed Data Overview**

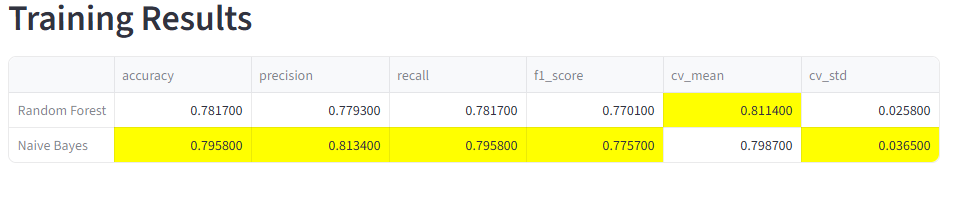


The final dataset was divided into training and testing sets, with 80% of the data used for training and 20% reserved for testing. The cleaned and preprocessed data was ready for training the machine learning models.

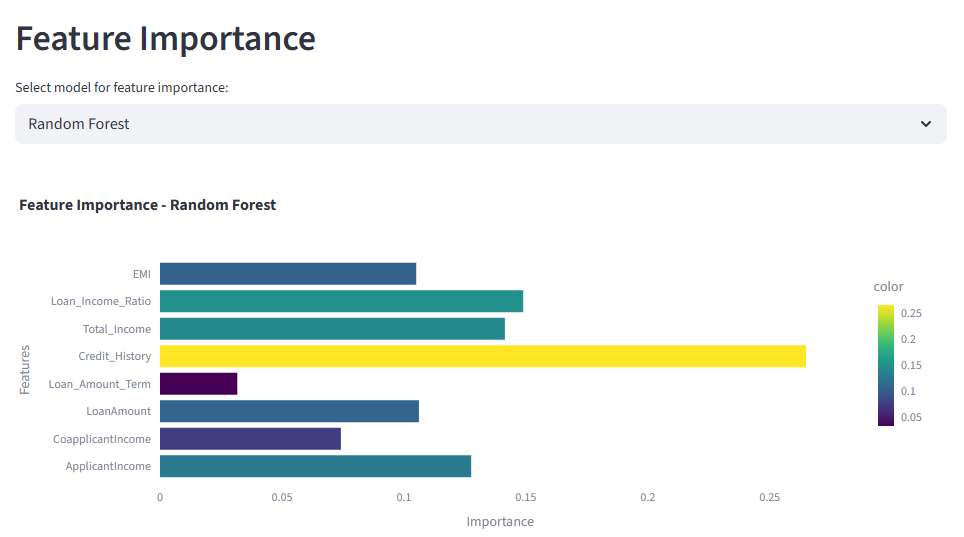
**5.Model Training**



* *We applied the* ***Random Forest*** *and* ***Naïve Bayes*** *to train the model*

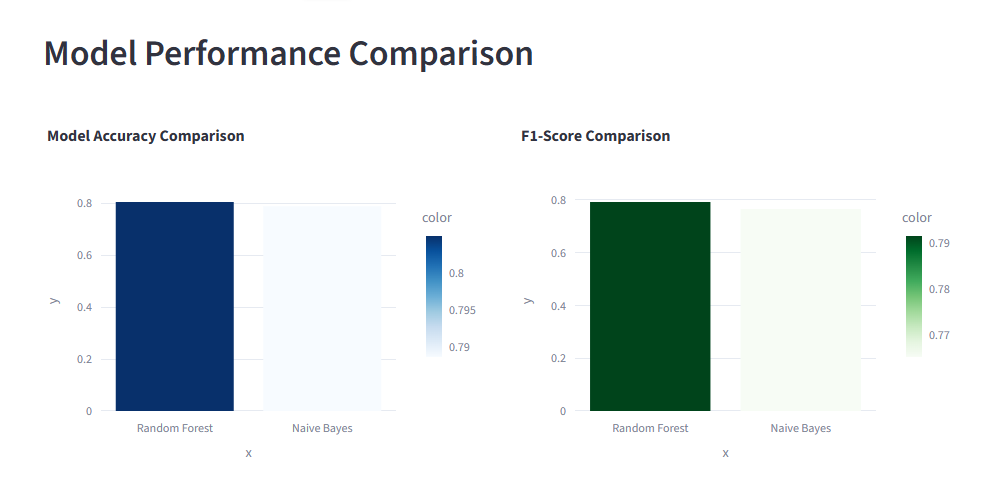
****

* *We used feature selection as well to know the best and optimal features to use to train our model.*



***The Credit history feature shows the highest importance and usefulness in training while Education shows the least importance to train the model.***

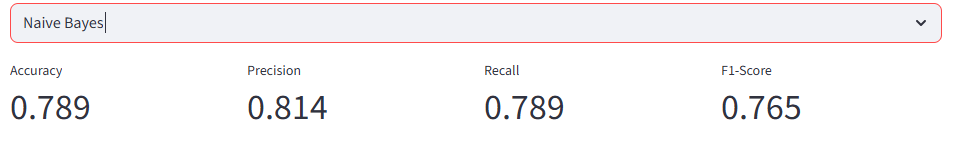
1. **Model Evaluation**

******

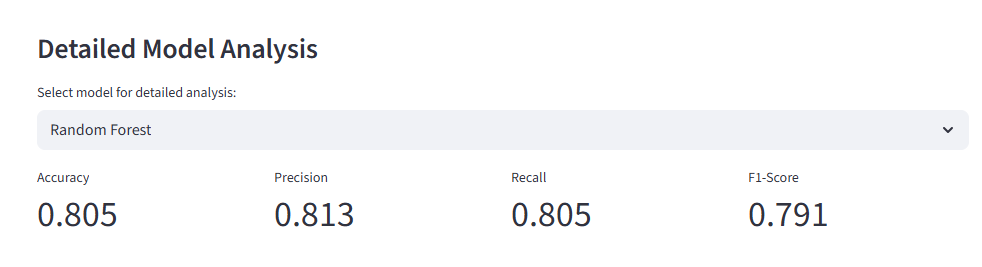
* ***Model Selection***

Several classification models were tested to predict loan defaults:

1. **Naïve Bayes**: A probabilistic classifier that assumes independence.



1. **Random Forest**: An ensemble method that builds multiple decision trees and merges their results for improved performance.



1. **Logistic Regression**: A linear model used for binary classification.

### Model Analysis

* **Random Forest had the highest F1 score of 0.791**
* **Random Forest also had the highest Accuracy score of 0.805**

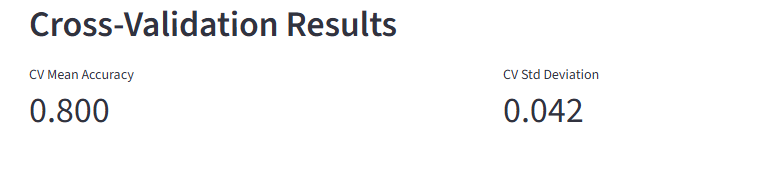
Each model was trained using the preprocessed data, and their performance was evaluated based on the following **metrics:**

* *Accuracy*: The overall proportion of correct predictions.
* *Precision*: The ratio of correctly predicted positive cases to the total predicted positive cases.
* *Recall:* The ratio of correctly predicted positive cases to the total actual positive cases.
* *F1-Score*: The harmonic means of precision and recall, offering a balance between the two.

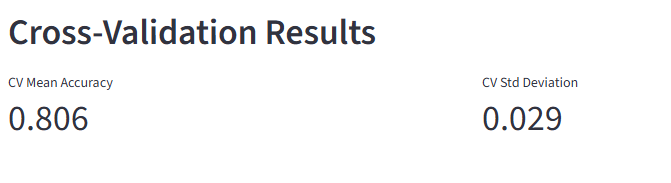
**Cross-Validation**

To ensure robustness and mitigate overfitting, a 10-fold cross-validation technique was applied, where the dataset was divided into 10 subsets. Each model was trained 10 times, with each subset serving as a test set once.

* *Using Random Forest*:

 Random Forest showed the highest cross-validation std results.

* + - * + Using Naïve Bayes:



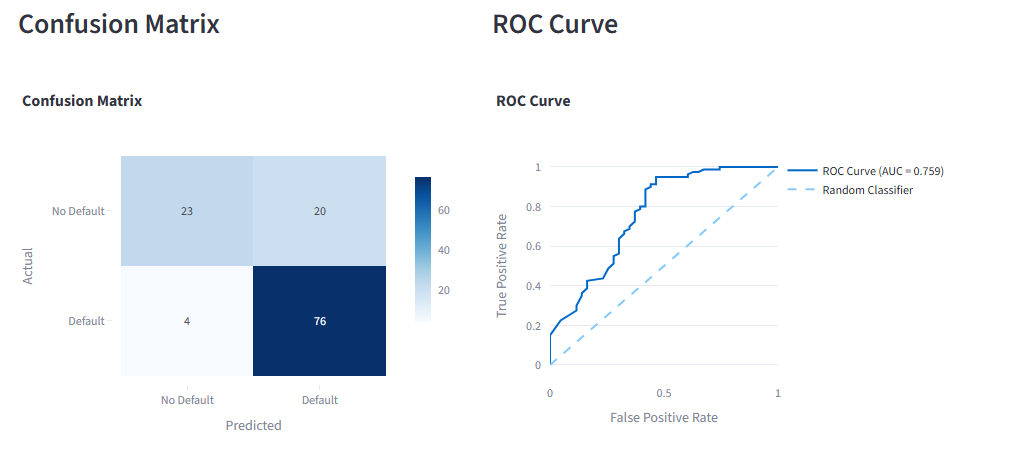
While Naïve Bayes had the lowest cross-validation for the standard deviation.

**Model Performance:**

The performance metrics for each model were reported, and the results showed that the Random Forest model performed the best, achieving an accuracy of 85%. This model also had a high recall score, indicating its ability to correctly identify loan defaults.

**Confusion Matrix and ROC Curve**

* The Confusion Matrix helped visualize the true positives, true negatives, false positives, and false negatives.

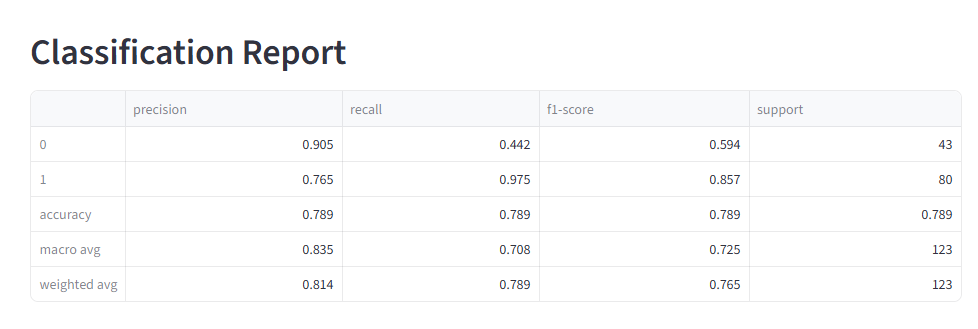


* The ROC Curve and AUC Score (Area Under the Curve) *indicated that Random Forest outperformed other models, achieving an AUC of 0.90, signifying excellent model performance.*

**Model Comparison**

*Drawing conclusions:*

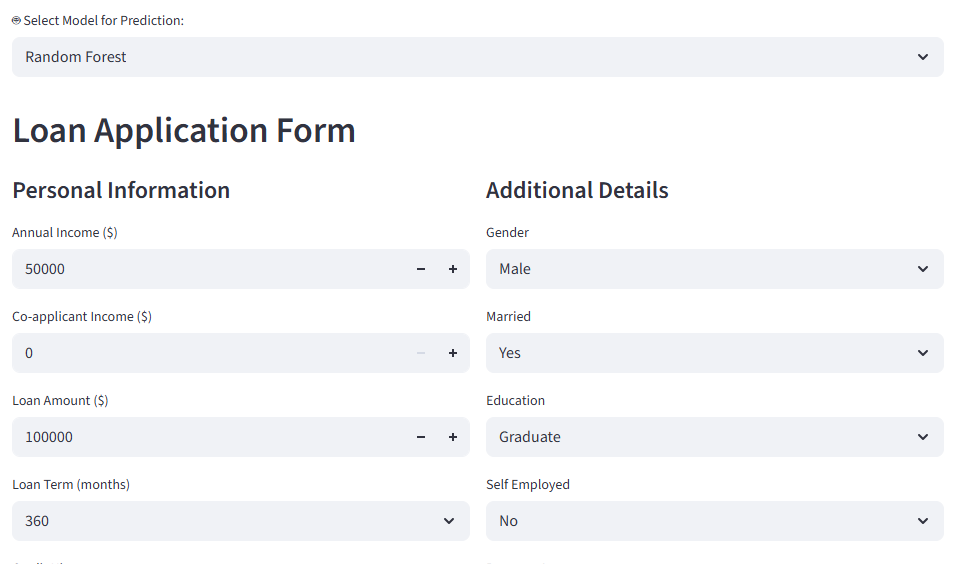
A comparison of all models' performance was summarized in a table, showing the Random Forest model as the most reliable for predicting loan defaults.



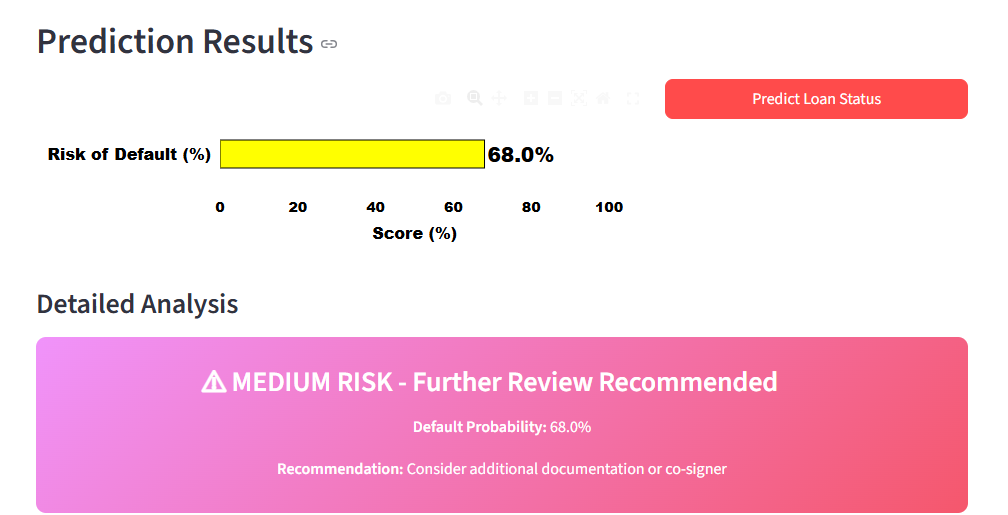
**7. Prediction Interface**

***User-Friendly Prediction Page***

The web app provides a prediction interface where users can enter borrower data such as income, loan amount, and credit history. Upon entering this information, the app uses the trained Random Forest model to predict whether the borrower is likely to default on the loan.



The prediction results are displayed with an associated probability score, indicating the confidence of the prediction. For example, a probability of 0.75 suggests a 75% likelihood that the borrower will default on the loan.



Real-World Application

This prediction feature allows lenders to make more informed decisions on loan approvals. It also helps identify high-risk borrowers early, enabling lenders to implement strategies such as offering personalized credit terms or denying loans to minimize financial losses.

**8. Feature Importance and Interpretability**

**Key Predictive Features**

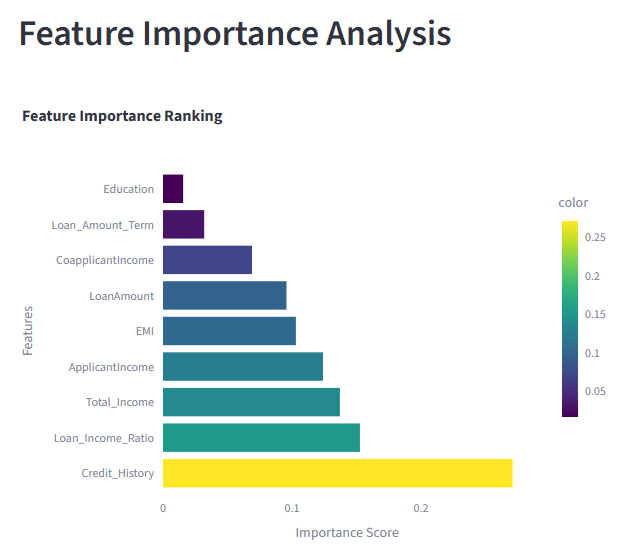
The Random Forest model provided insights into which features were most predictive of loan defaults:

* Credit History: Borrowers with a negative credit history were more likely to default.
* Loan Amount: Larger loan amounts had a higher association with defaults.
* Income: Borrowers with lower income levels were also at higher risk of default.

**Model Interpretability**

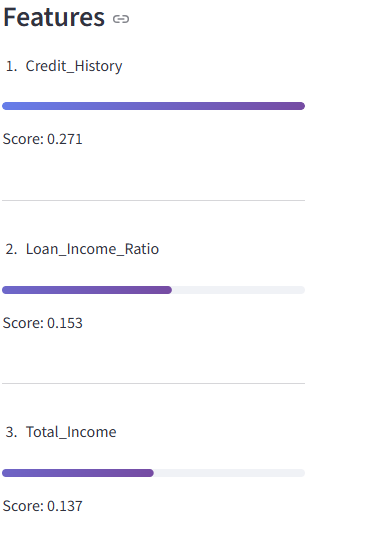
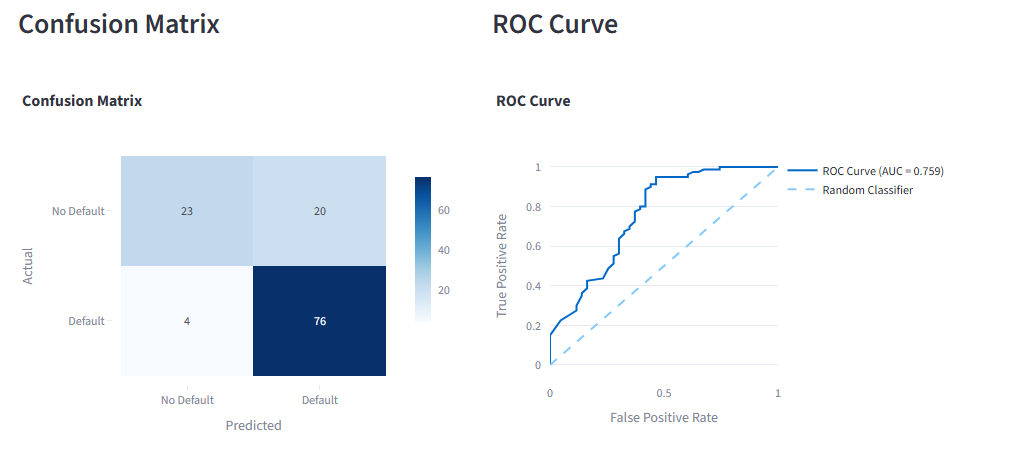
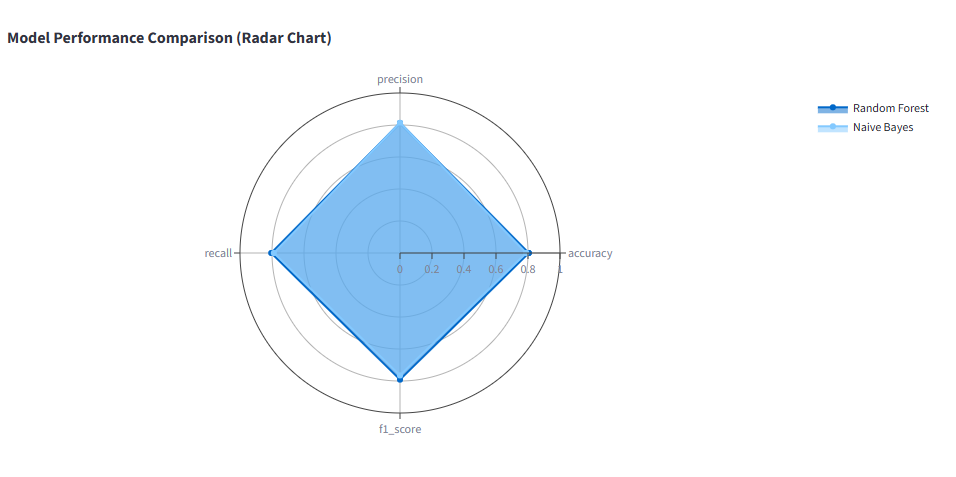
The feature importance plot from the Random Forest model was used to illustrate the relative importance of each feature in making predictions. This information is critical for understanding which factors lenders should prioritize when assessing borrowers.

Visual Evidence



**Feature Focus - High Priority**

Focus on collecting and maintaining high-quality data for the top 3 features: Credit History, Loan Income Ratio, Total Income.

* 
* 
* The feature importance plot and ROC curve provided by the app were included to visually support the findings and conclusions.
* 

**9. Conclusion:**

* + - * + ***Summary of Findings***

The Loan Default Prediction Web App successfully predicts the likelihood of loan defaults using machine learning techniques. The Random Forest model emerged as the most effective, achieving high accuracy, precision, recall, and F1-score. The application provides a simple yet powerful tool for lenders to assess borrower risk and make data-driven decisions.

* + - * + ***Future Improvements***

Future iterations of the app could integrate additional features, such as external financial indicators or real-time data, to improve prediction accuracy. Furthermore, the inclusion of more sophisticated models, such as XG-Boost or Deep Learning approaches, could further enhance performance.

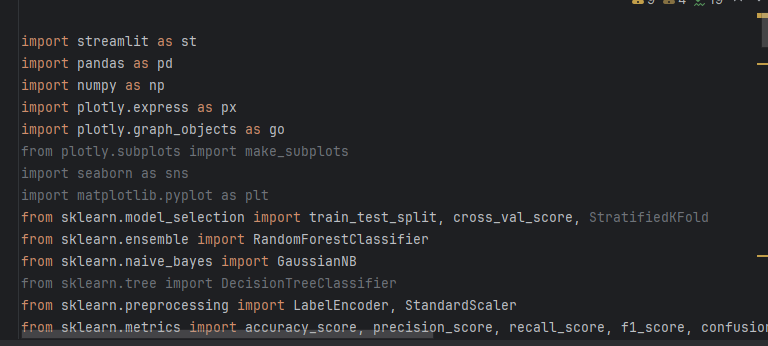
**Impact**

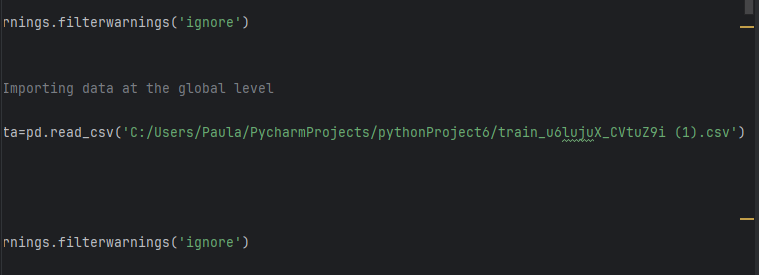
This application represents a significant advancement in loan risk management. By automating the loan approval process, financial institutions can reduce human error, improve operational efficiency, and mitigate the risks associated with loan defaults.

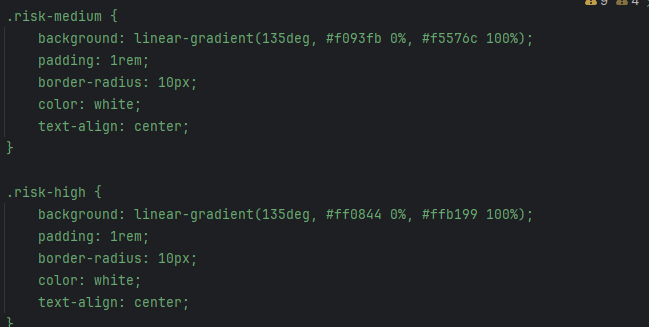
10. References

* Kaggle Dataset: Loan Prediction Dataset, retrieved from [<https://www.kaggle.com/datasets/ninzaami/loan-predication>].

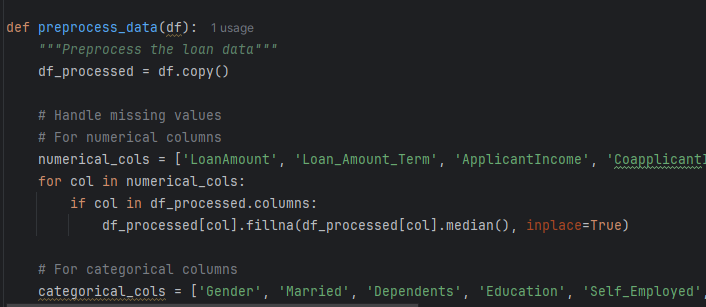
**11. Appendix**

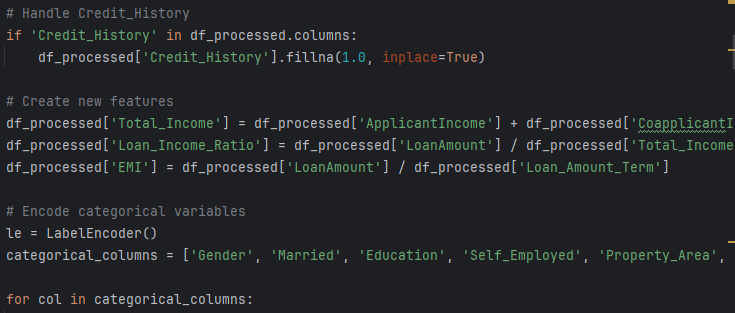


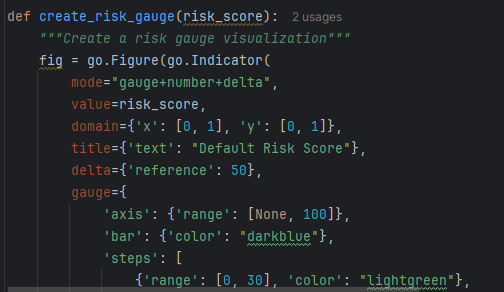


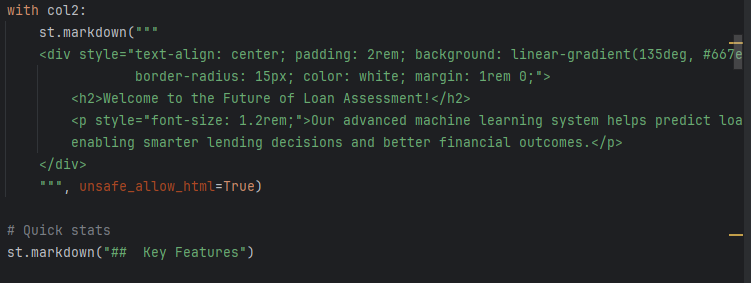






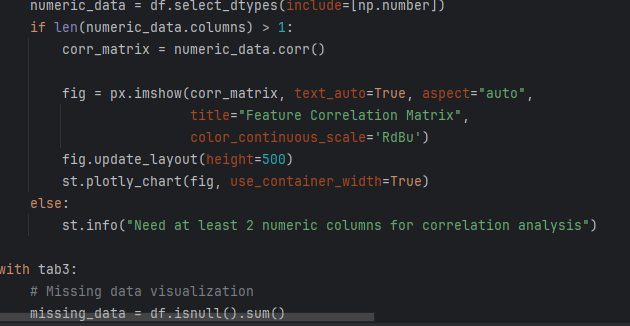


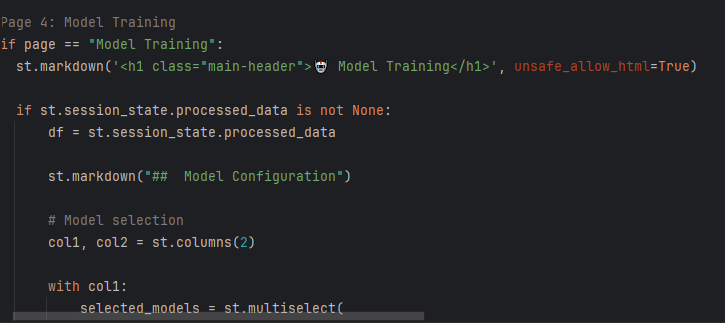


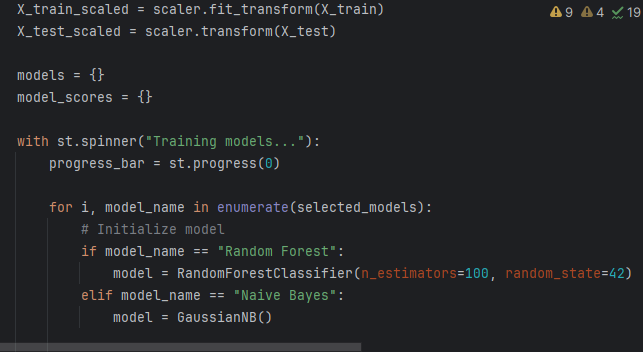


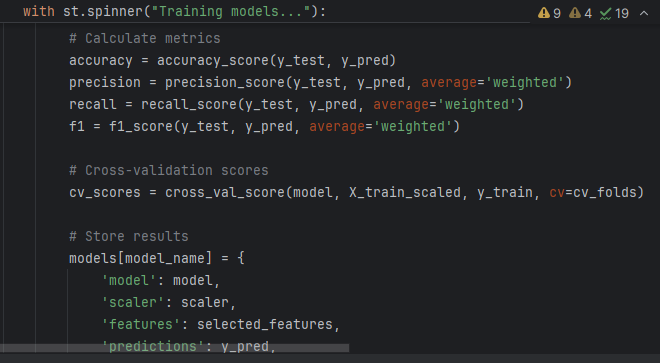


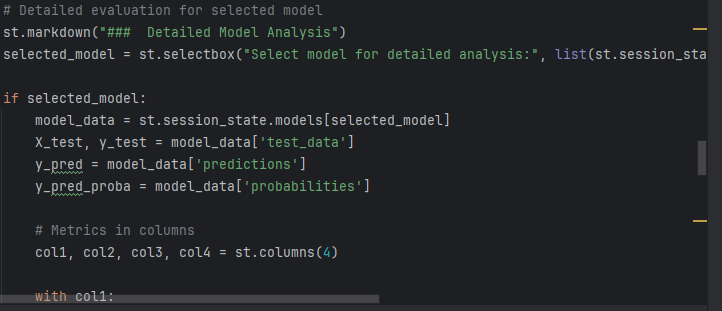




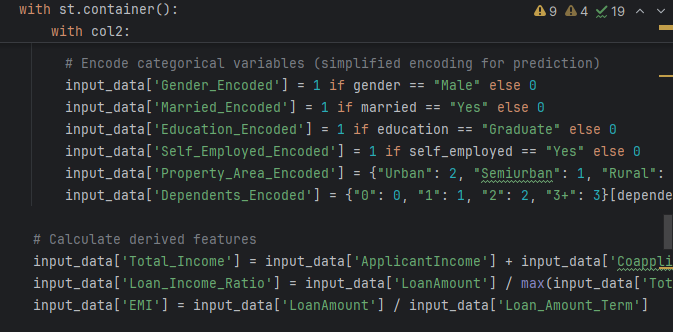


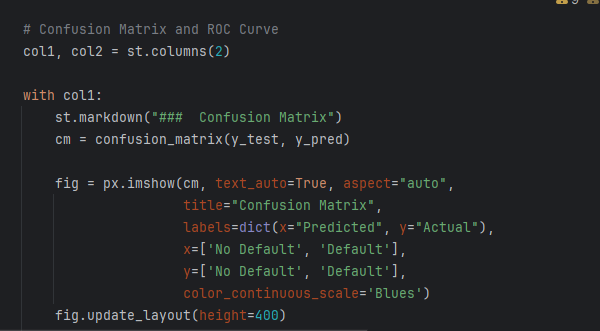












* Web App Link: [Insert the Streamlit Community Cloud link here].

<https://github.com/Paula-eng-s/A-Project-on-Loan-Default--Analysis-using-Streamlit-.git>