Dr. Vishwanath Karad
MIT-WPU
MIT WORLD PEACE UNIVERSITY | PUNE
।। विश्वशान्तिर्ध्रुवं ध्रुवा ।।
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

WPU School of Science & Environmental Studies
Department of Mathematics and Statistics

Loan Approval Prediction

User Guide -

Prof. Suvarna ranade
Assistant Professor , Program Head (MSc Statistics)
MIT-WPU, Department of Mathematics and Statistics

Introduction

In now days lending institutions wants to automate loan approval process as real-time process based on customer details provided while application for loan. To automate this process they have given a problem to identify customer’s segment that to whom loan will get approved.
We just want to use machine learning algorithm to predict loan approval based on given data.

Problem Statement

Financial institutions currently use manual loan evaluations for their approval process but this approach consumes time and shows human distortions. An automated system needs development to accurately predict loan approvals because it will improve both efficiency and consistency.

The goal is to automate the loan eligibility process in real time using customer details like income, dependents, CIBIL score, and asset values. A binary classification model is built to identify customers eligible for loan approval.

Dataset Overview

Loan Approval dataset is a collection of financial records and associated information used to determine the eligibility of individuals or organizations for obtaining loans from a lending institution. It includes various factors such as cibil score, income, employment status, loan term, asset value, and loan status.
We collect this data from Kaggle. Containing 13 features and 4269 variable in each.

Data Cleaning

We Checked missing value and no missing value were found. It means that data is fully cleaned

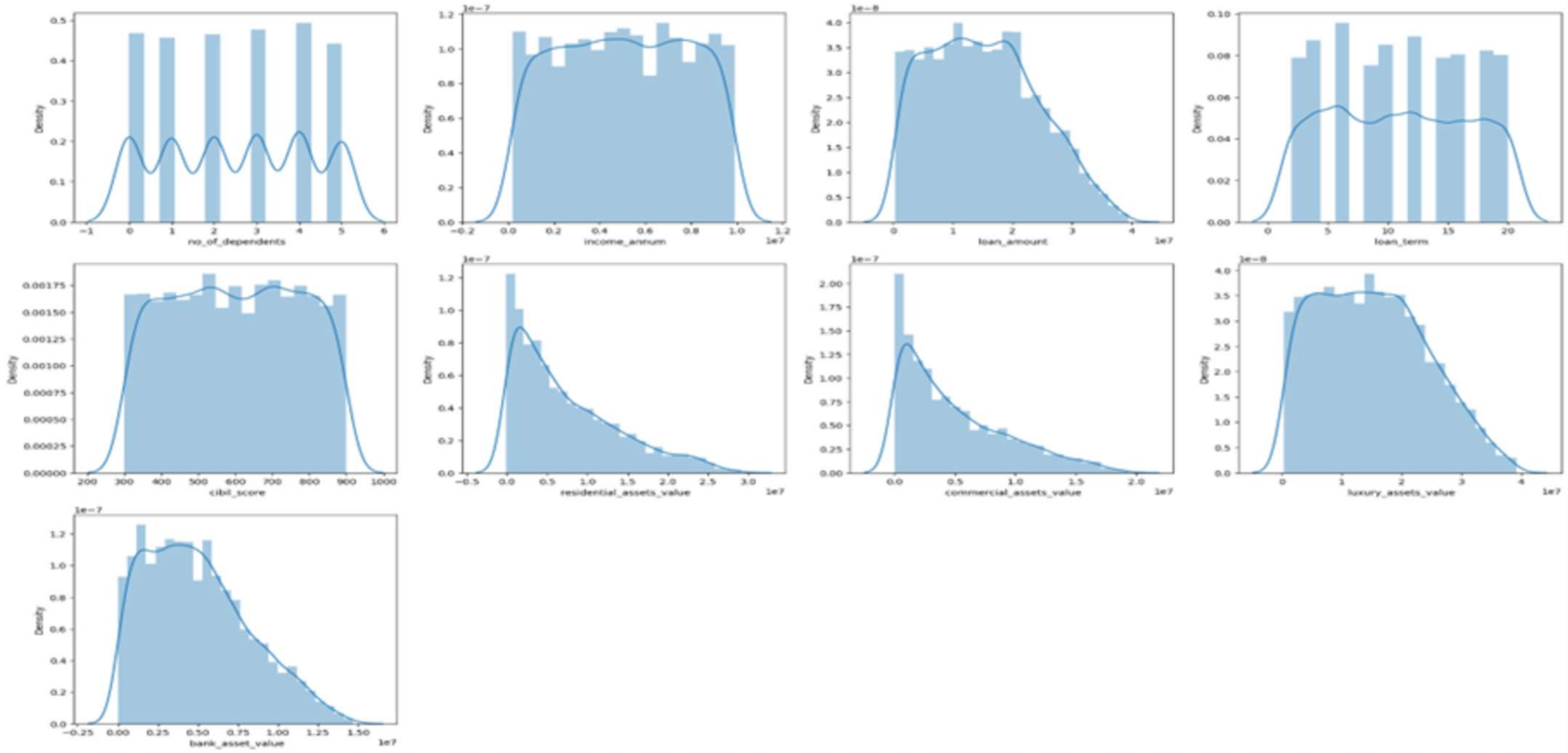
Outlier Treatment

Outliers were detected using the Interquartile Range (IQR) method. Features such as residential_assets_value, commercial_assets_value, and bank_asset_value contained outliers. These outliers were capped using the 1st and 99th percentiles to reduce their impact on the model's performance

EDA- Univariate Analysis:

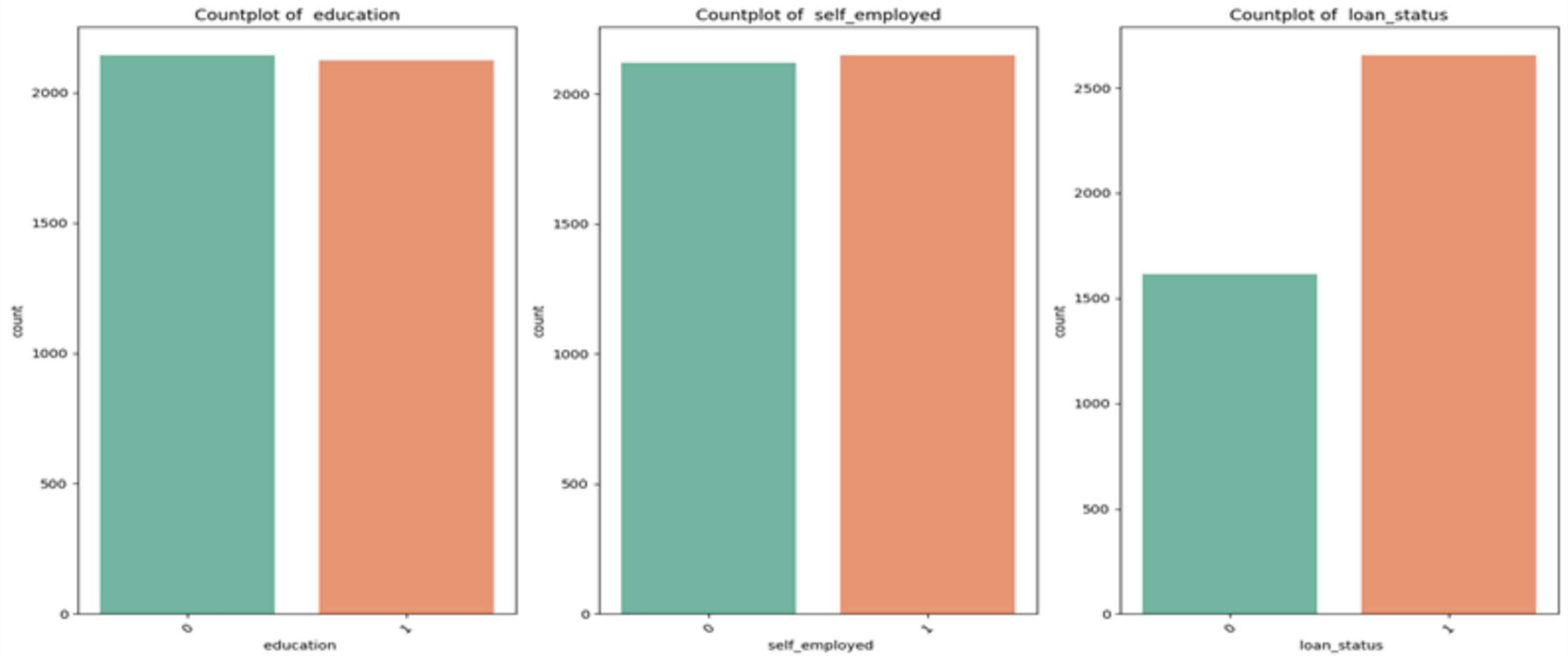
Distribution Plots (Dist-plots for Numerical Features)

- This show how data like income, loan amount, and CIBIL score are spread.
- Most features are right-skewed, especially loan amount and income, indicating a majority of people fall into lower ranges with a few extreme high values.



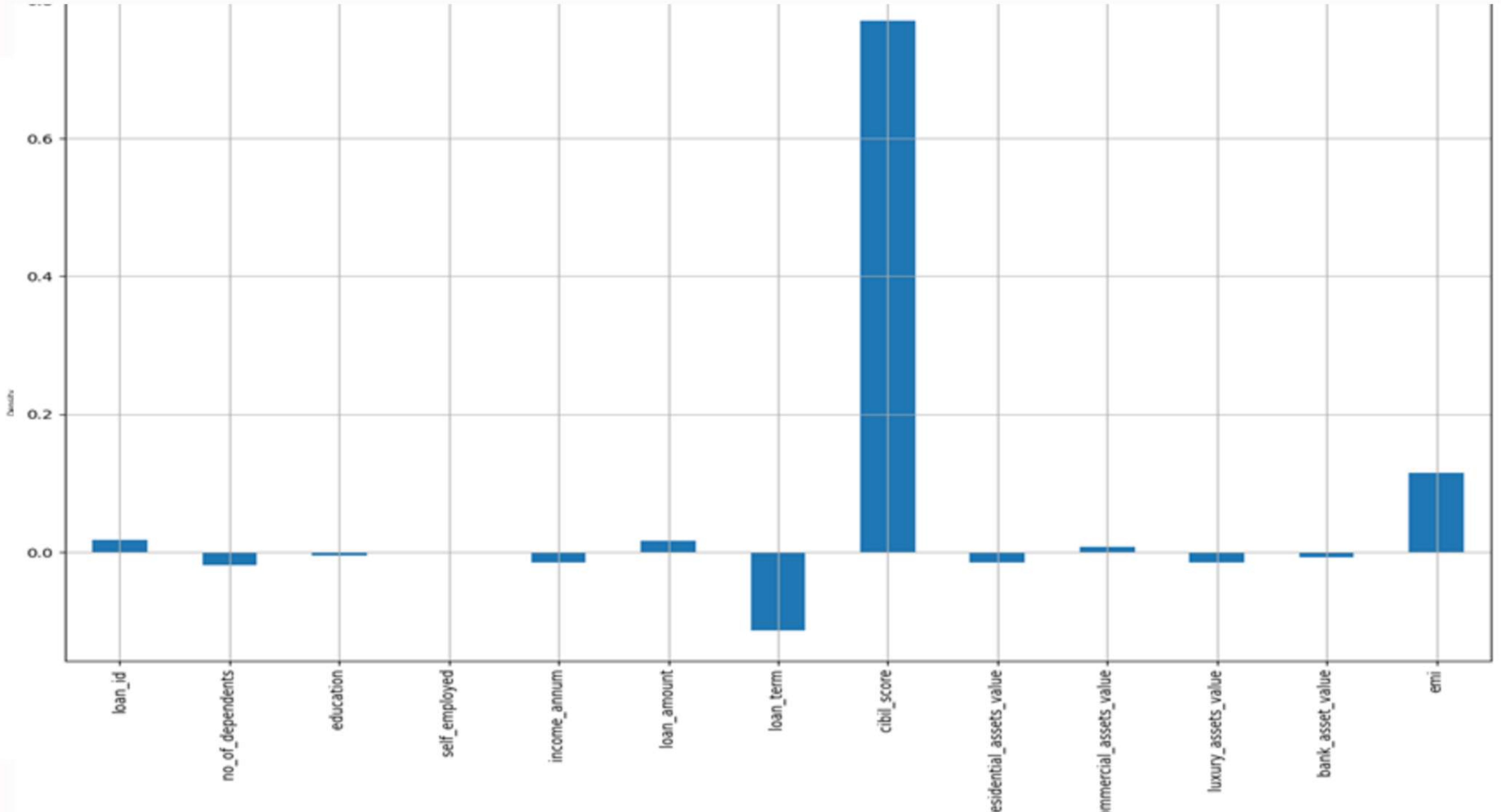
Count Plots (Categorical Features)

- Breakdown of applicants by education, employment type, and loan status.
- Balanced distribution in education and employment.
- More loans are approved than rejected.



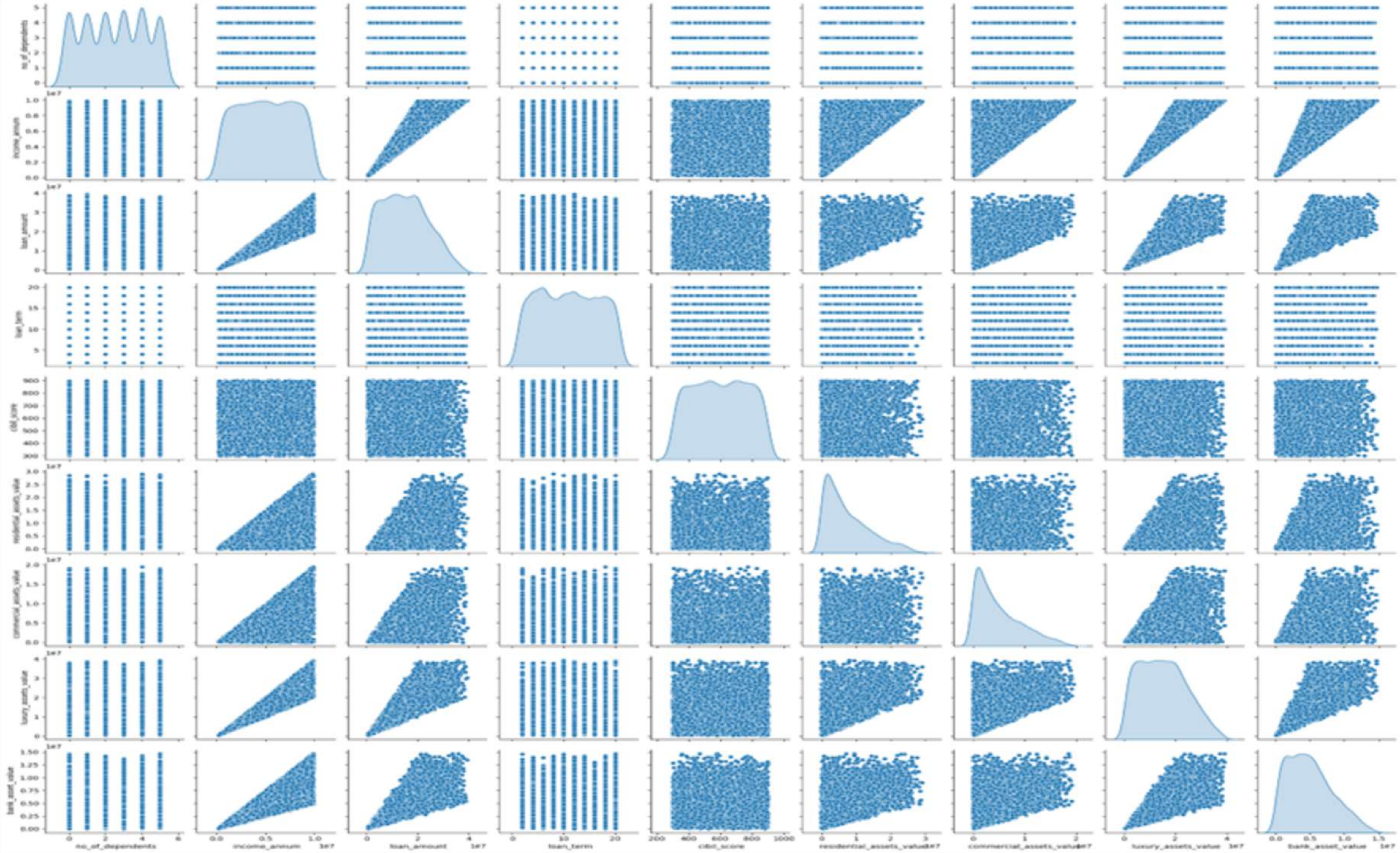
Correlation with Loan Status (Bar Plot)

- Quantifies how each feature correlates with loan approval.
- CIBIL score is the most impactful feature for loan approval.
- Other features have minimal influence individually.



Bivariate Analysis:

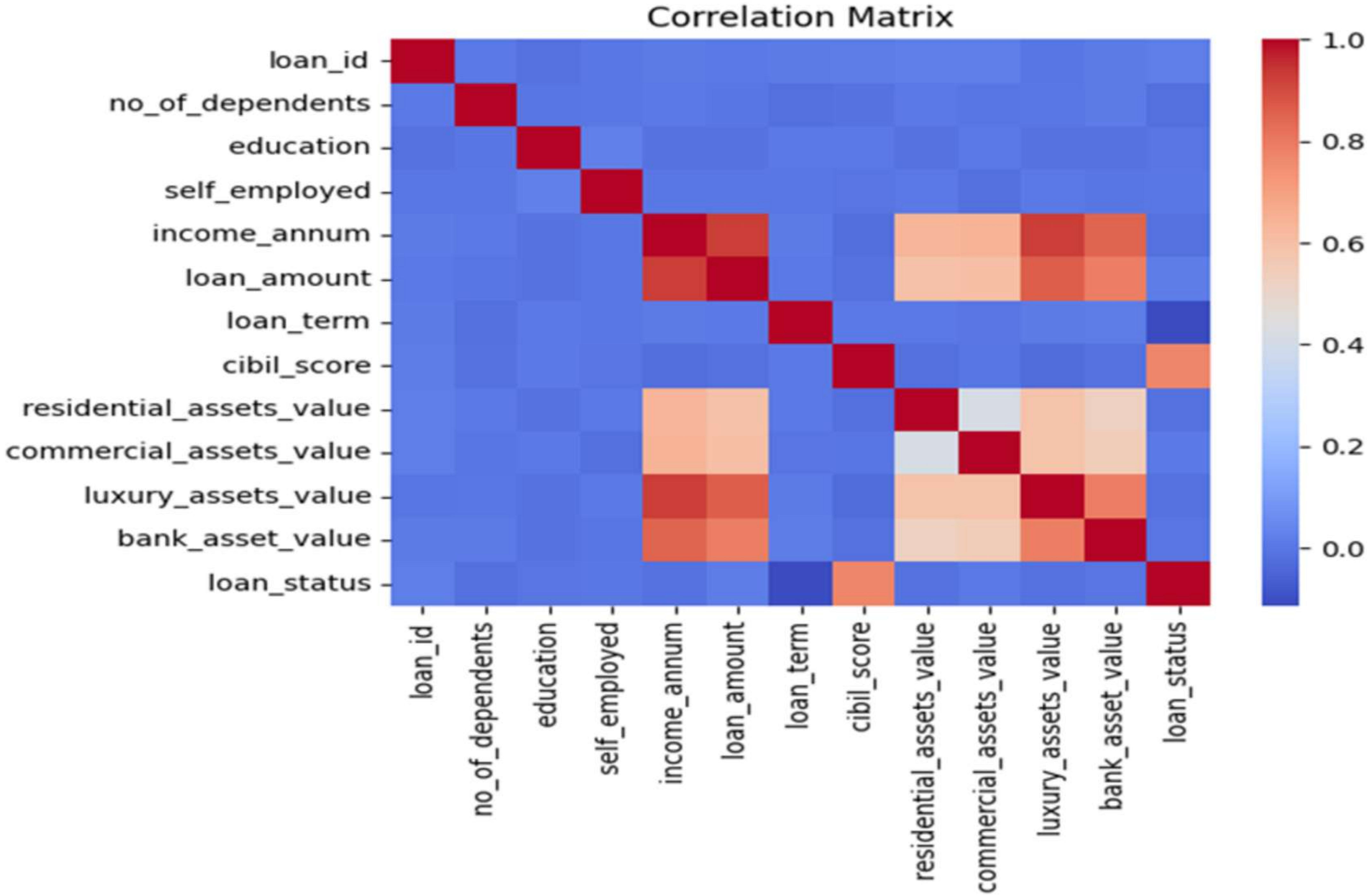
Pairplot



- Visualizes pairwise relationships between numeric features.
- Helps detect correlation and clustering. For instance, CIBIL score and loan amount may show distinct grouping between approved and rejected loans.

Correlation Heatmap

- Displays relationships between numerical features.
- Strong positive correlation between CIBIL score and loan approval; weak or no correlation among many asset-related features.



Feature Engineering

- In feature engineering, EMI feature was created to show the monthly repayment amounts for every loan. The standard amortization formula calculated the EMI feature by combining loan amount and interest rate with the loan term. The new feature enables better prediction of loan approval by assessing how much applicants need to pay each month.

Encoding

- For categorical features such as education, self_employed, and loan_status, encoding was applied. Label Encoding was used to transform these categorical variables into numerical values. Specifically, the education and self_employed features were encoded as binary variables, and the loan_status variable was converted into binary values (1 for 'Approved' and 0 for 'Rejected') to make them compatible with machine learning algorithms

Data Transformation

1. **Yeo–Johnson Transformation**: This transformation was used to handle skewed data by stabilizing the variance and making the distribution more normal (bell-shaped).
2. **Standard Scaling**: After applying the Yeo–Johnson transformation, the features were standardized to have a mean of 0 and a standard deviation of 1. This ensures that all features are on the same scale, preventing any one feature from dominating due to its larger numerical values and improving the performance of machine learning algorithms.

Logistic Regression

What is Logistic Regression?

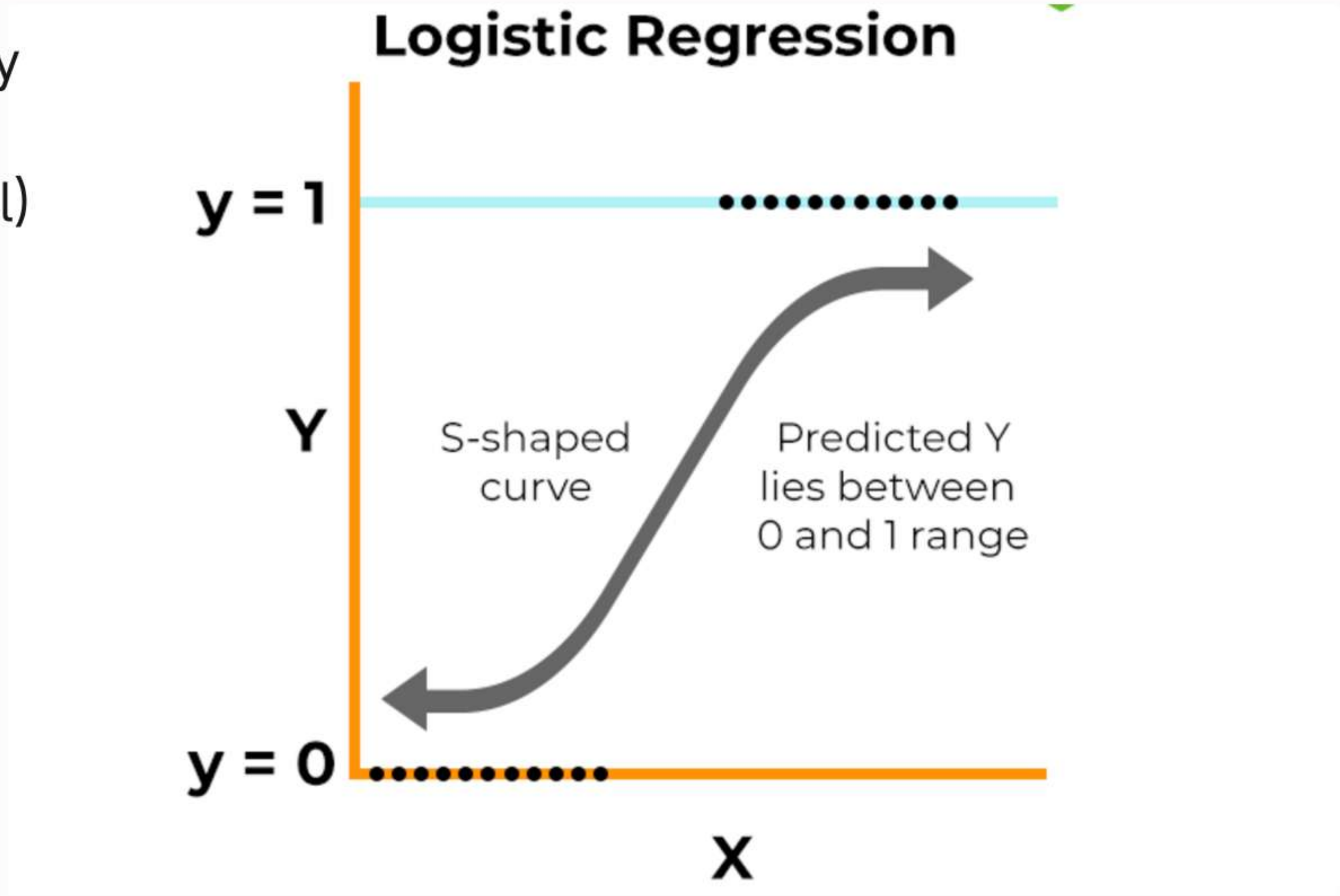
- A statistical model (logit model) used for binary classification (e.g., Approved/Rejected).
- Estimates probability of an event (loan approval) based on input features.

Assumptions:

a. Binary outcome (e.g., Yes/No).
b. Independent variables (no multicollinearity).
c. Linear relationship between features and log-odds.
d. Requires large sample sizes.

Performance:

- ○ Accuracy: 91.33%



Equation:

$$P(Y=1) = \frac{1}{1+e^{-(b_0+b_1X)}}$$

K-Fold Cross-Validation: (for logistic model)

Ensuring Model Generalizability

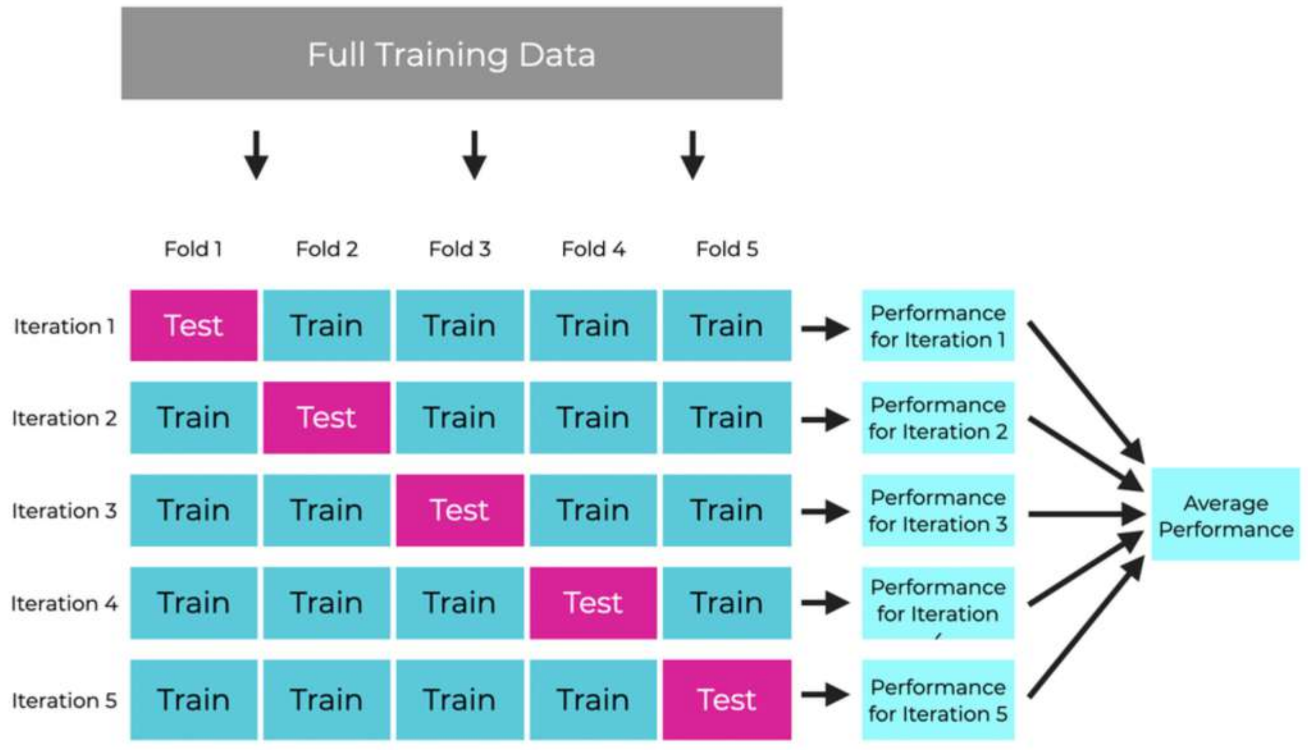
How it works:

- Split data into 5 folds.
- Train on 4 folds, validate on 1 (repeat 5x).
- Prevents overfitting by testing on different subsets.

Results:

- Average Accuracy: ~92.8% (from K-Fold).
- Variation: Scores range from 91.5% to 93.7%.

Performance: Accuracy: 91.33%



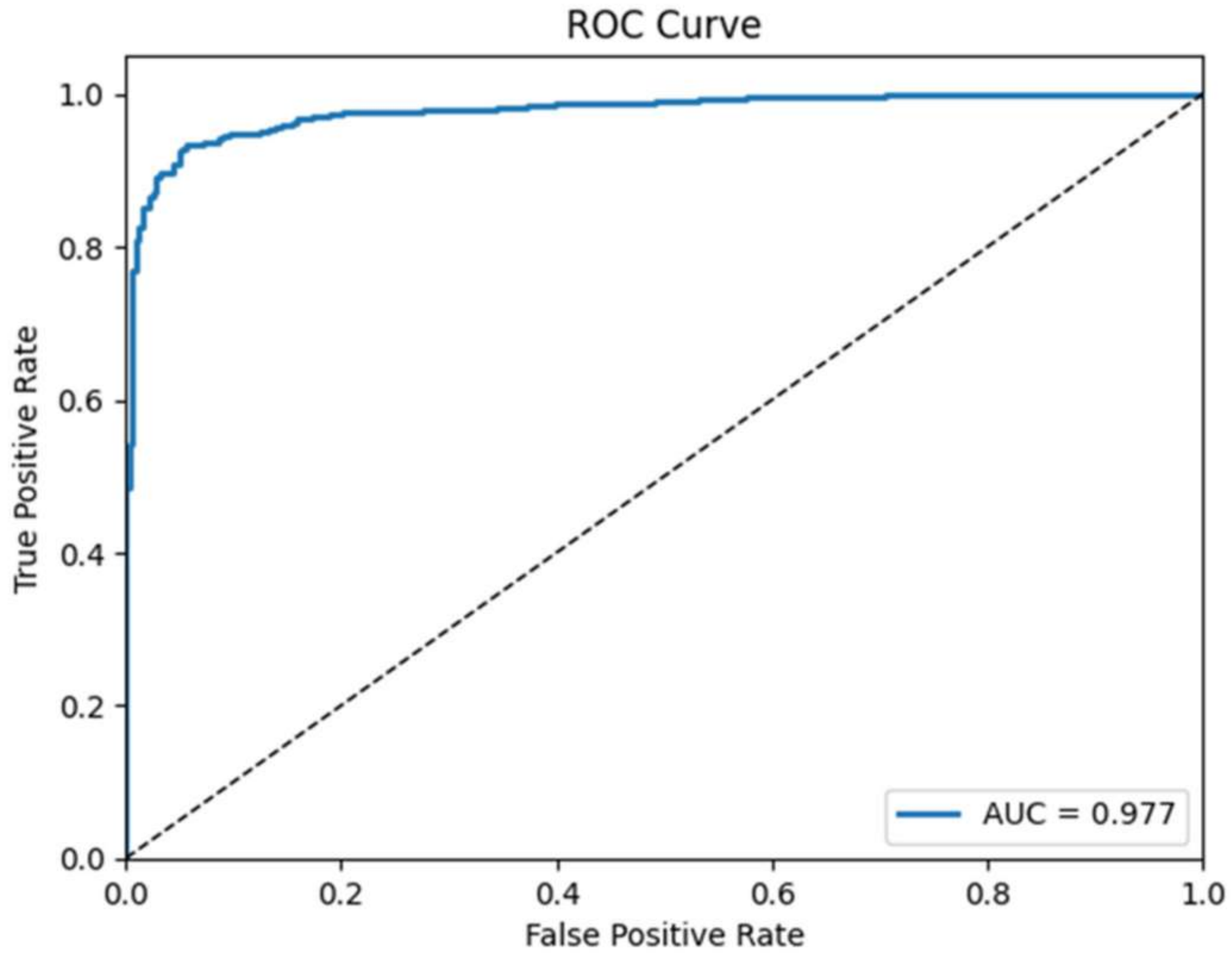
Full Training Data

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	
Iteration 1	Test	Train	Train	Train	Train	Performance for Iteration 1
Iteration 2	Train	Test	Train	Train	Train	Performance for Iteration 2
Iteration 3	Train	Train	Test	Train	Train	Performance for Iteration 3
Iteration 4	Train	Train	Train	Test	Train	Performance for Iteration
Iteration 5	Train	Train	Train	Train	Test	Performance for Iteration 5

Average Performance

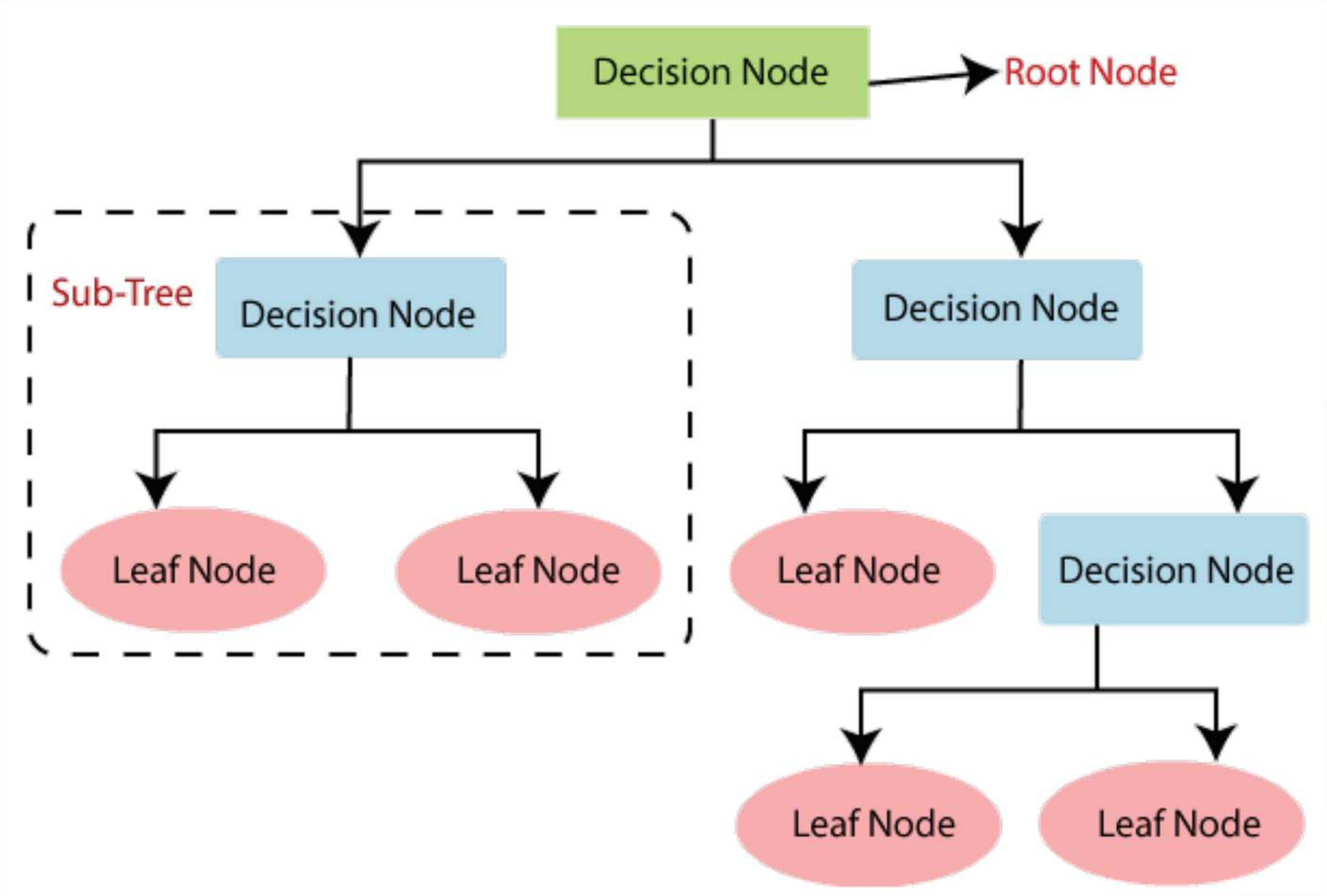
Model Accuracy Across 5 Folds

Fold 1	93.1%
Fold 2	93.5%
Fold 3	92.0%
Fold 4	93.7%
Fold 5	91.5%



Decision Tree

The Decision Tree classifier achieved 98% accuracy but showed signs of overfitting. It’s an interpretable model but can be less stable with minor data changes.



Performance:

- Achieved 98% accuracy on test data
- Shows signs of overfitting (high train accuracy vs. validation)
- Interpretable but sensitive to data changes

K-Fold Cross-Validation (5 folds):

- Consistent high performance across all folds

K-Fold Results Table :

Fold	Accuracy Score
Fold 1	98.6%
Fold 2	97.4%
Fold 3	97.2%
Fold 4	98.0%
Fold 5	97.7%
Avg	97.78%

Random Forest

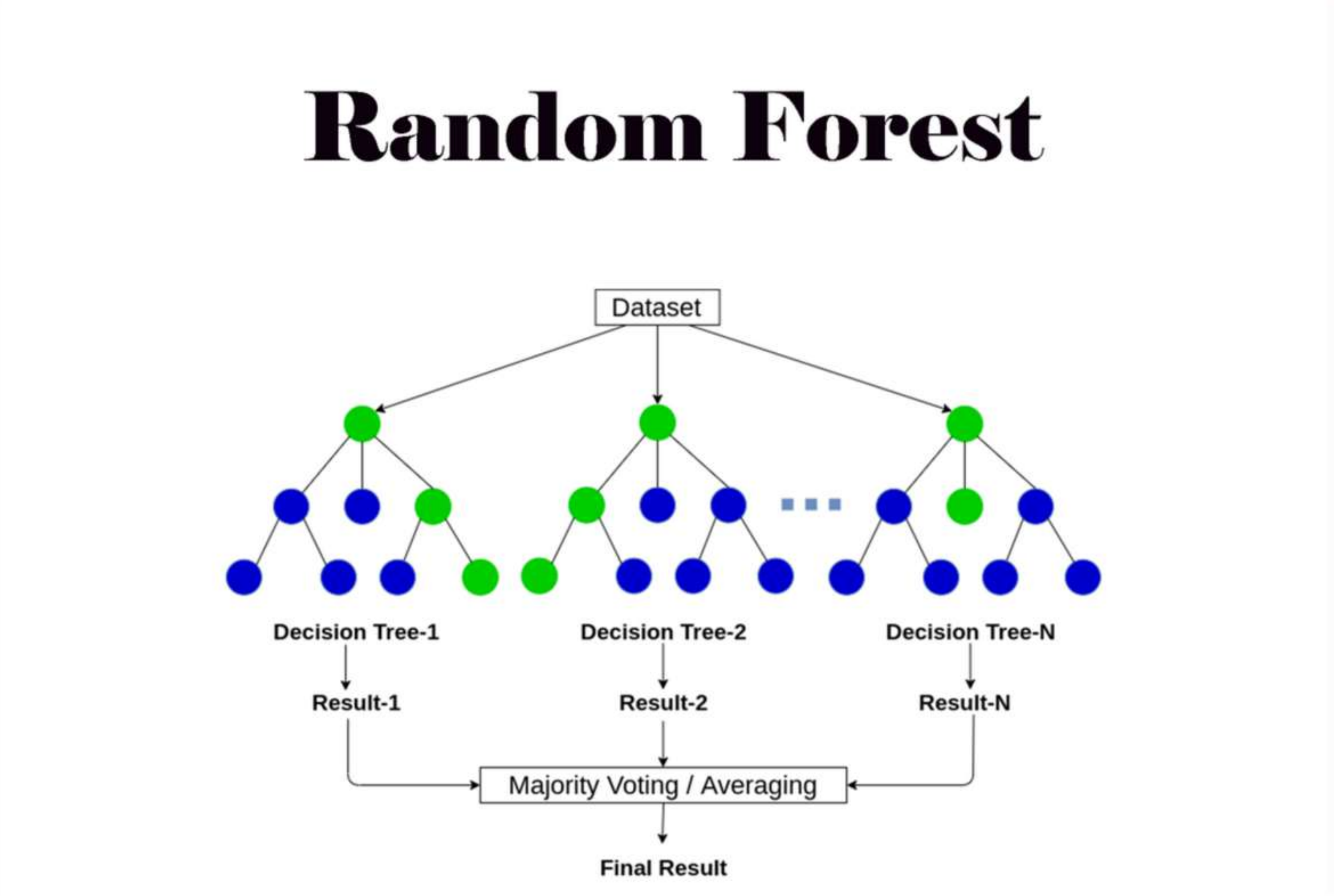
Random Forest was the best performer after hyperparameter tuning (accuracy: 98.8%). It showed high robustness, handled variance well, and generalized better than individual decision trees.

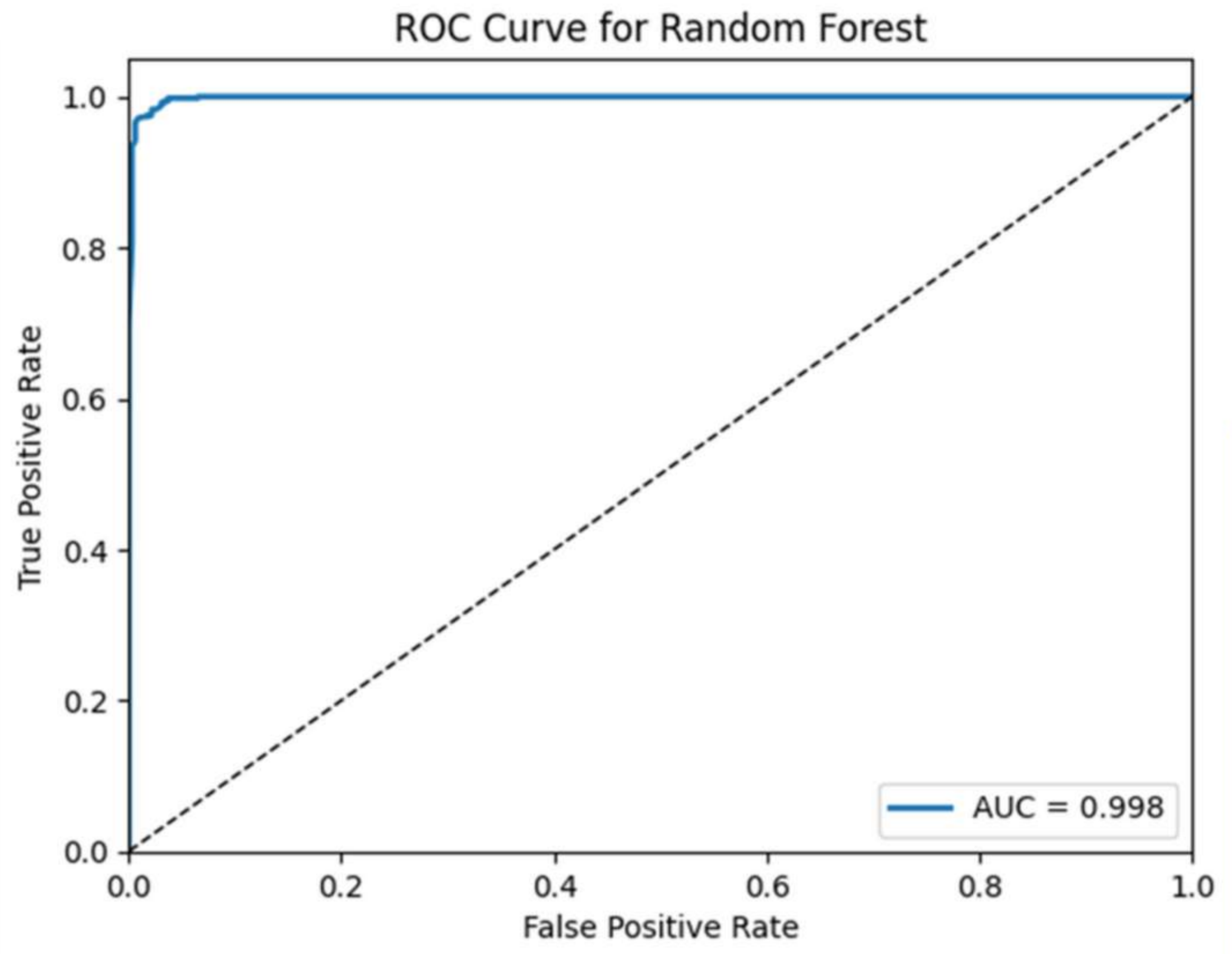
Model Overview

- Ensemble of decision trees (majority voting)
- Advantages: Robust to overfitting, handles noisy data
- Best fold accuracy: 98.8%

K-Fold Validation Scores

Fold	**Accuracy**
1	**98.8%**
2	**98.0%**
3	**98.4%**
4	**98.6%**
5	**98.4%**
Avg	**98.44%**

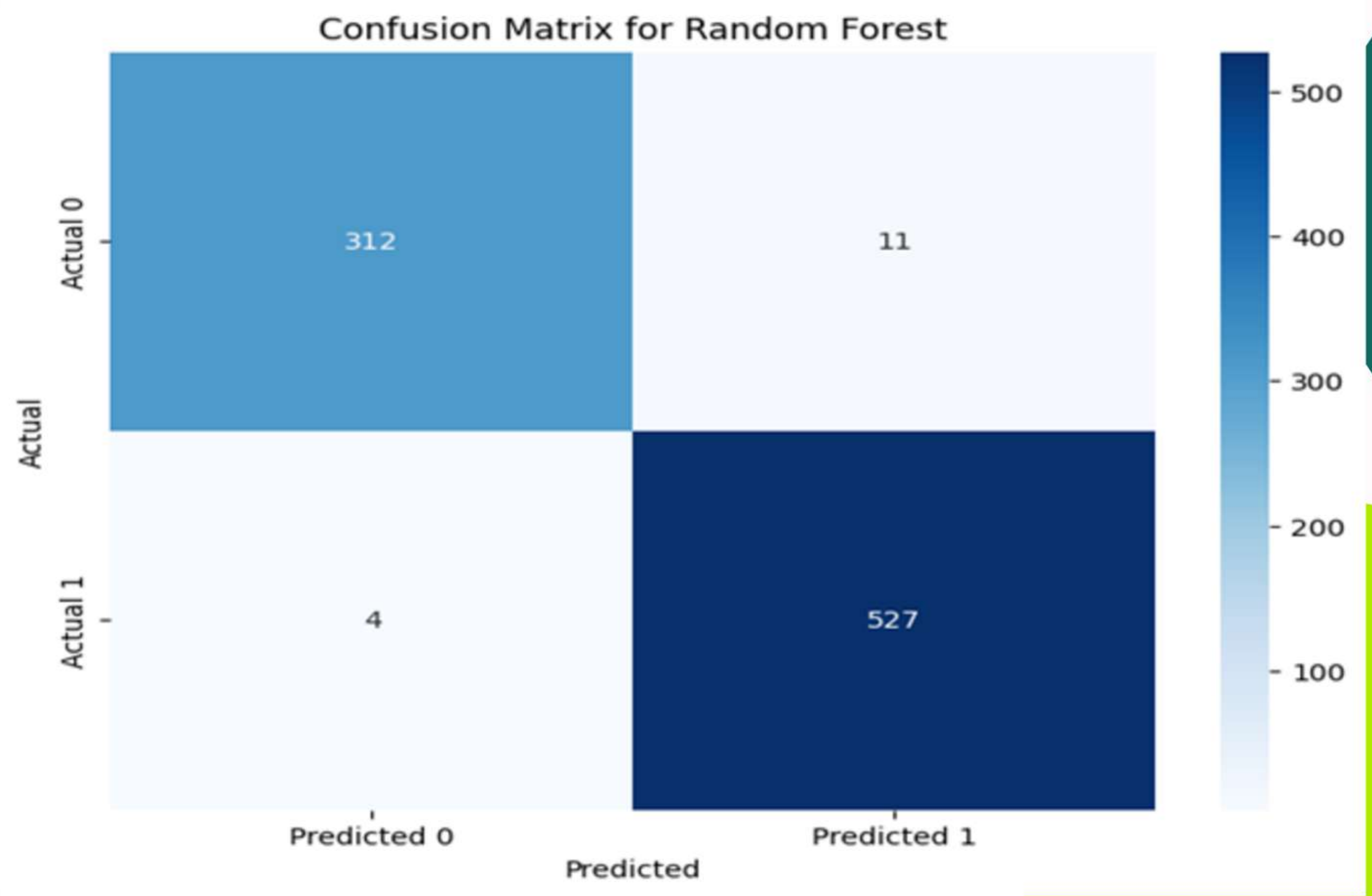




Random Forest

Classification Report Highlights:

- **Precision: 0.99 for class 0 (Rejected), 0.98 for class 1 (Approved)**
- **Recall: 0.97 for Rejected, 0.99 for Approved**
- **F1-Score: 0.98 (Rejected), 0.99 (Approved)**
- **Overall Accuracy: 98%**
- **ROC-AUC Score: 0.9984 (Excellent model performance)**

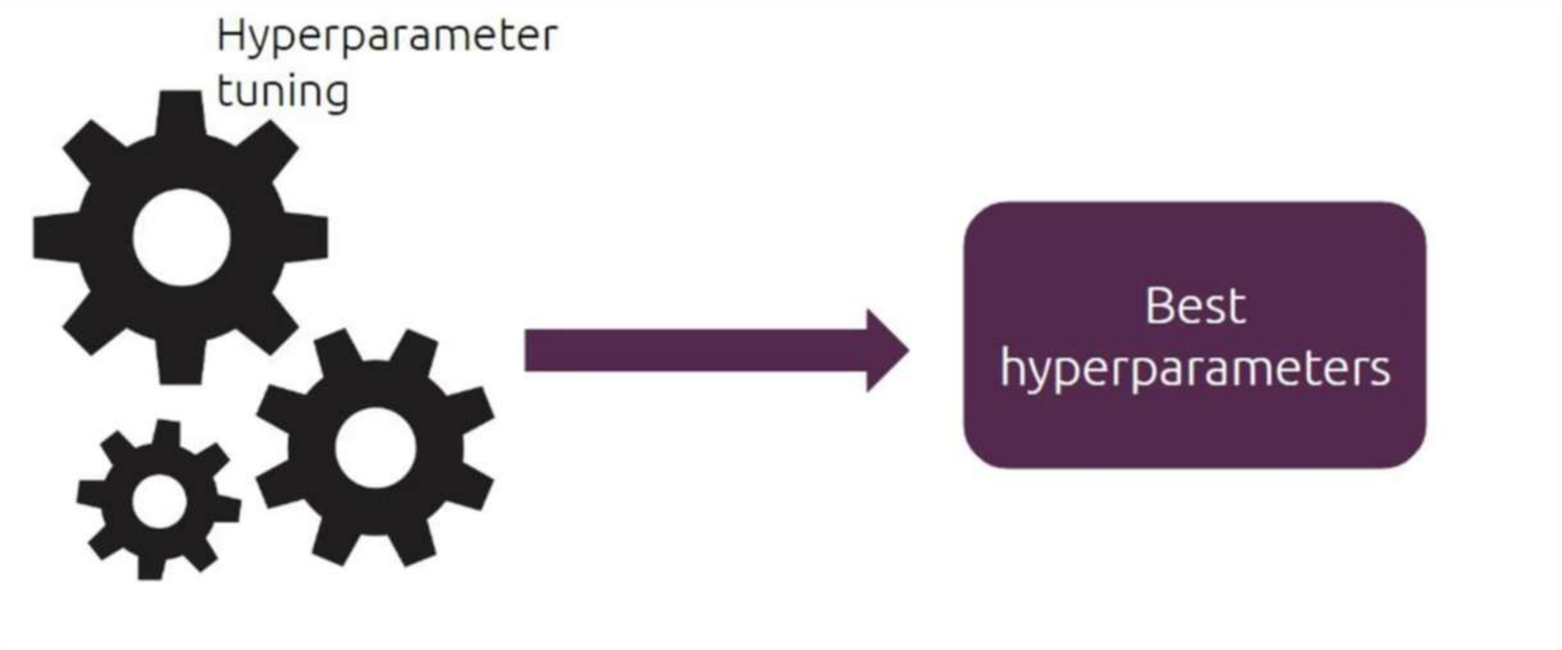


Confusion Matrix Insights:

- True Positives (TP): 527 approved loans correctly predicted.
- True Negatives (TN): 312 rejected loans correctly predicted.
- False Positives (FP): 11 rejected loans incorrectly predicted as approved.
- False Negatives (FN): 4 approved loans incorrectly predicted as rejected.

✅ The Random Forest model achieves high precision, recall, and a near-perfect AUC, indicating excellent discrimination between loan approval and rejection classes.

Hyperparameter Tuning:



- Best Model: Random Forest
- Final Accuracy after Tuning: 99.5%
- Tuning Method: RandomizedSearchCV
- Key Improvements:
  - Optimized the number of trees, depth, and splitting criteria.
  - Improved handling of class imbalance.
  - Increased ROC-AUC score close to 0.998, ensuring better generalization.

✅ Hyperparameter tuning significantly enhanced model performance, making Random Forest the most reliable choice for deployment.

Conclusion

- CIBIL score emerged as the most influential factor in determining loan approval.
- Correlation analysis showed that most variables had low correlation with loan approval, except for the CIBIL score.
- SMOTE was applied to handle class imbalance, improving the model's ability to correctly predict rejected loans.
- Among all models tested, the Random Forest classifier gave the best performance with a ROC-AUC of ~0.996.
- Logistic Regression and Decision Tree also performed well, with ROC-AUC values above 0.97.

The project successfully demonstrates how machine learning and EDA can be combined to build reliable loan approval systems

Future Work

- Deploy model as a web app.
- Include more features (e.g., employment history).

Our Team

Hemanshu S Patil
S.Y.MSc-Statistics

PRN No : 11322131032

Koustubh A Sandbhor
S.Y.Msc-Statistics

PRN No : 11322130142

THANK TOU..!