Report on Classification Task.

Loan Approval Prediction Using Classical Machine Learning Models

Name: Ritesh Khatri

University id: 2548382

Assignment Due Date: 09-10/2026

Module leader: Siman Giri

Tutor: Robin Tuladhar

Loan Approval Prediction Using Classical Machine Learning Models

Abstract

Purpose: Predict whether a loan will be approved or rejected using classification models.
Dataset: The dataset used is the **Loan Approval Dataset** with ~[add number] records and [add number] features. It aligns with the **UN Sustainable Development Goals (UNSDG)** by promoting financial inclusion and responsible lending.
Approach: Methodology included Exploratory Data Analysis (EDA), building two classical ML models (Logistic Regression and Decision Tree), hyperparameter tuning, feature selection, and final model comparison.
Key Results: The tuned models achieved higher Accuracy, Precision, Recall, and F1-Score than the classic models. Decision Tree outperformed Logistic Regression slightly in CV Score, Accuracy, and F1-Score.
Conclusion: Tuned and feature-selected models improved performance; insights include which features most influence loan approval and how hyperparameter optimization impacts model accuracy.

Toc

Table of Contents

1. Introduction 1

2. Methodology 2

3. Results and Conclusion 4

4. Discussion 15

5. References 16

1. Introduction

1.1 Problem Statement

The goal of this project is to predict whether a loan application will be approved or rejected (loan_status) using various demographic, financial, and loan-related factors. This model can help banks identify high-risk applicants and make smarter, data-driven lending decisions.

1.2 Dataset

The dataset consists of 45,000 records with 26 features, including:

- **Demographic:** age, income, work experience, home ownership, gender
- **Loan-related:** loan amount, interest rate, loan-to-income ratio, purpose of the loan
- **Credit history:** credit score, past loan defaults, length of credit history
- **Target variable:** loan_status (0 = rejected, 1 = approved)

This dataset supports the United Nations Sustainable Development Goal 8 (Decent Work and Economic Growth) by promoting fair and transparent lending practices.

1.3 Objective

The main objective is to develop predictive models for loan approval that not only achieve high performance but also highlight the key factors influencing loan decisions. The project will also address class imbalance to ensure the model makes accurate predictions for both approved and rejected loans.

2. Methodology

2.1 Data Preprocessing

- All missing values were addressed, and the dataset contains complete records.
- Categorical features such as gender, education, loan purpose, and home ownership were converted into numeric formats or one-hot encoded for modeling.
- Numeric features were scaled to improve the performance of models like Logistic Regression.

2.2 Exploratory Data Analysis (EDA)

- Visualizations such as histograms, bar charts, and correlation heatmaps were used to explore the data, understand patterns, and identify relationships between features.
- Key insights: higher income and better credit scores correlated with loan approvals; certain loan intents like DEBTCONSOLIDATION or MEDICAL had higher approval rates.
- Class imbalance detected: 77% rejected vs. 23% approved.

2.3 Model Building

2.3 Models Used

Two machine learning models were implemented to predict loan approval:

1. **Logistic Regression** (with and without balanced class weights)
   - The top 10 most important features, selected using Recursive Feature Elimination (RFE), included: loan amount, interest rate, loan-to-income ratio, credit score, previous loan defaults, home ownership, and specific loan purposes such as debt consolidation, home improvement, medical, and venture.
2. **Decision Tree**
   - The top 10 predictive features identified via RFE were: age, income, work experience, loan amount, interest rate, loan-to-income ratio, length of credit history, credit score, previous loan defaults, and home ownership.

2.4 Model Evaluation

The models were assessed using the following metrics:

- **Accuracy:** overall proportion of correct predictions
- **Precision:** proportion of predicted approvals that were actually correct
- **Recall:** proportion of actual approvals correctly identified
- **F1-Score:** the harmonic mean of precision and recall, balancing both metrics

2.5 Hyperparameter Tuning

- **Logistic Regression:** optimal settings were C=1, L1 penalty, solver='liblinear', with class weights balanced.
- **Decision Tree:** best configuration was using entropy as the criterion, max depth of 10, minimum samples per leaf of 5, and minimum samples to split of 2.

2.6 Feature Selection

Recursive Feature Elimination (RFE) was applied to determine the most important features for each model, helping to focus on the variables that contribute most to predicting loan approvals.

3. Results and Conclusion

The classification models were compared based on accuracy, precision, recall, and F1-score. This comparison highlights which model performs best for predicting loan approvals and provides insights into the key factors influencing lending decisions.

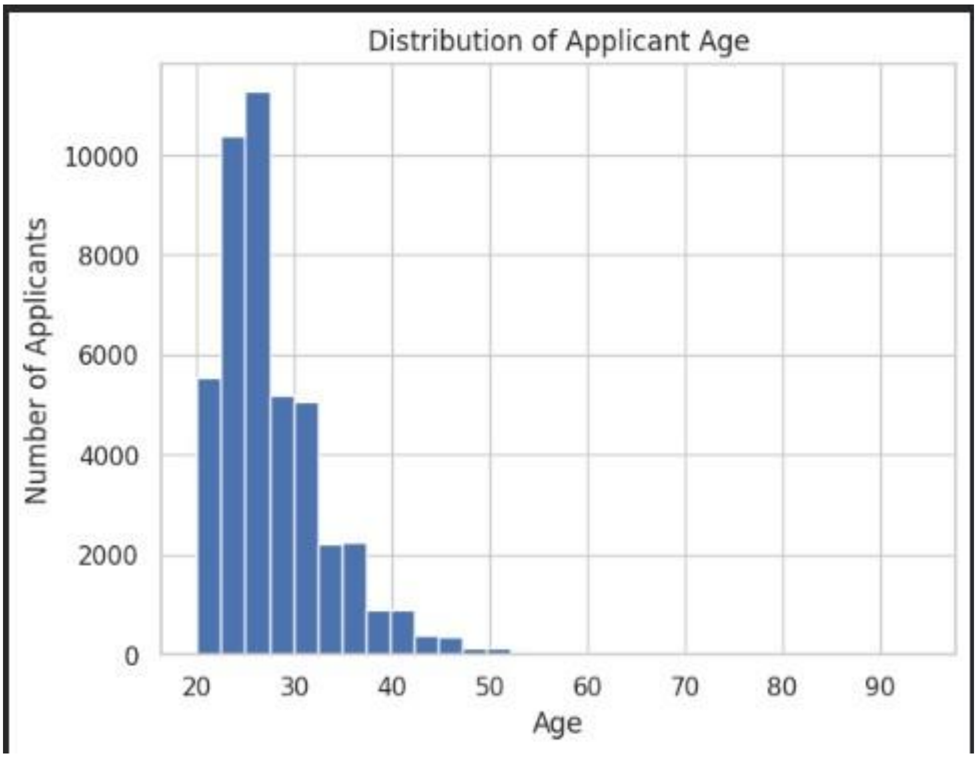

Fig: Distribution of Applicant Age

In the figure "Distribution of Applicant Age," the data highlights a significant demographic trend that directly impacts the loan approval landscape.

Here is a more detailed look at how age influences the results:

- Dominant Age Group: The chart shows a massive concentration of applicants between 20 and 30 years old. This "younger" demographic represents the vast majority of the dataset, with a sharp peak around age 23–25.
- The Approval Link: According to the Correlation Matrix, age is almost perfectly correlated with credit history length (0.88) and employment experience (0.93). Since the majority of applicants are under 30, they naturally have shorter credit histories and less work experience—two factors that often lead to the high rejection rate (Status 0) seen in the overall status chart.
- Declining Participation: As age increases, the number of applicants drops off significantly. By age 50, the bars are barely visible, indicating that older individuals either seek loans from different sources or have less need for the types of loans represented in this dataset.
- Risk Profile: While younger applicants are more active, they also represent a higher "uncertainty" for lenders due to their limited financial footprint. This helps explain why, despite the high volume of young applicants, the overall approval count remains low compared to the total number of applications.

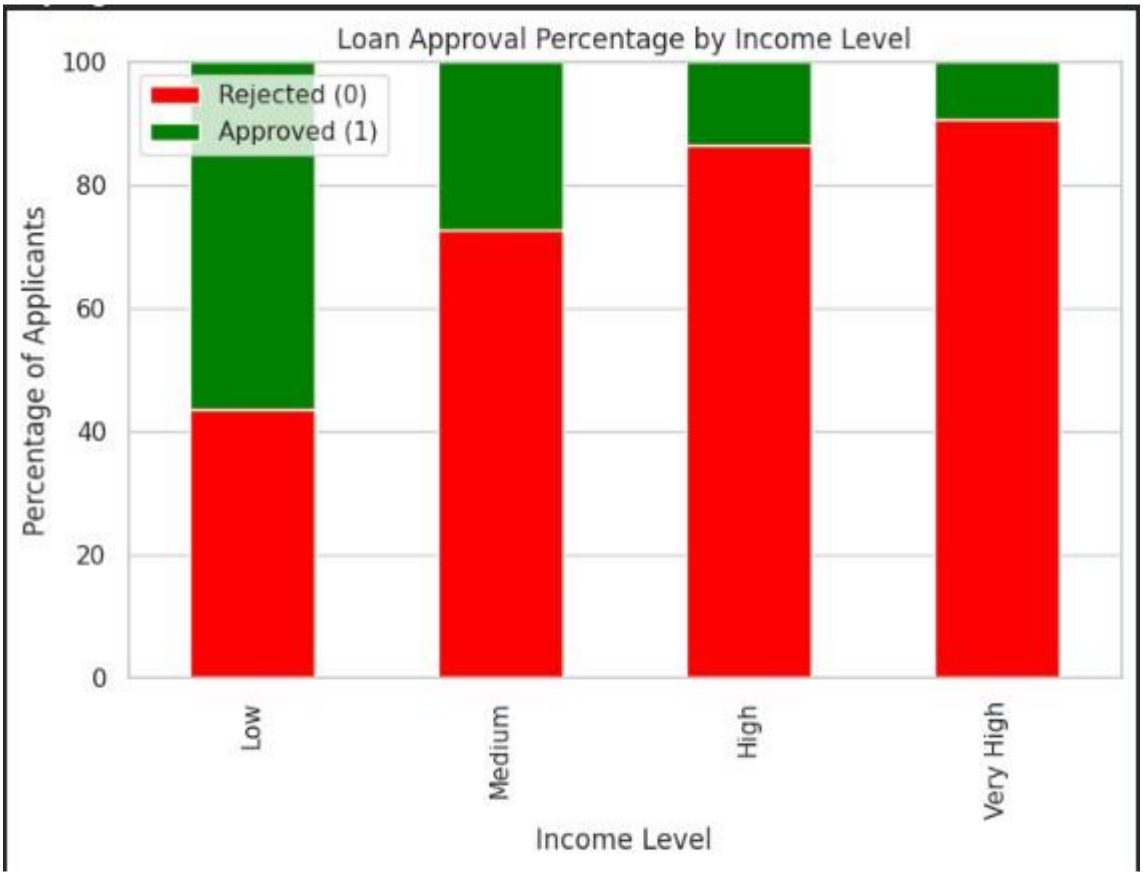

Fig: Loan Approval Percentage by Income Level

In the figure Loan Approval Percentage by Income Level, there is a surprising inverse relationship: as income increases, the likelihood of loan approval significantly drops. While the Low income category has the highest success rate with over 50% approved, the Very High income category faces a nearly 90% rejection rate, suggesting that higher earnings in this dataset surprisingly correlate with a much higher chance of being denied a loan.

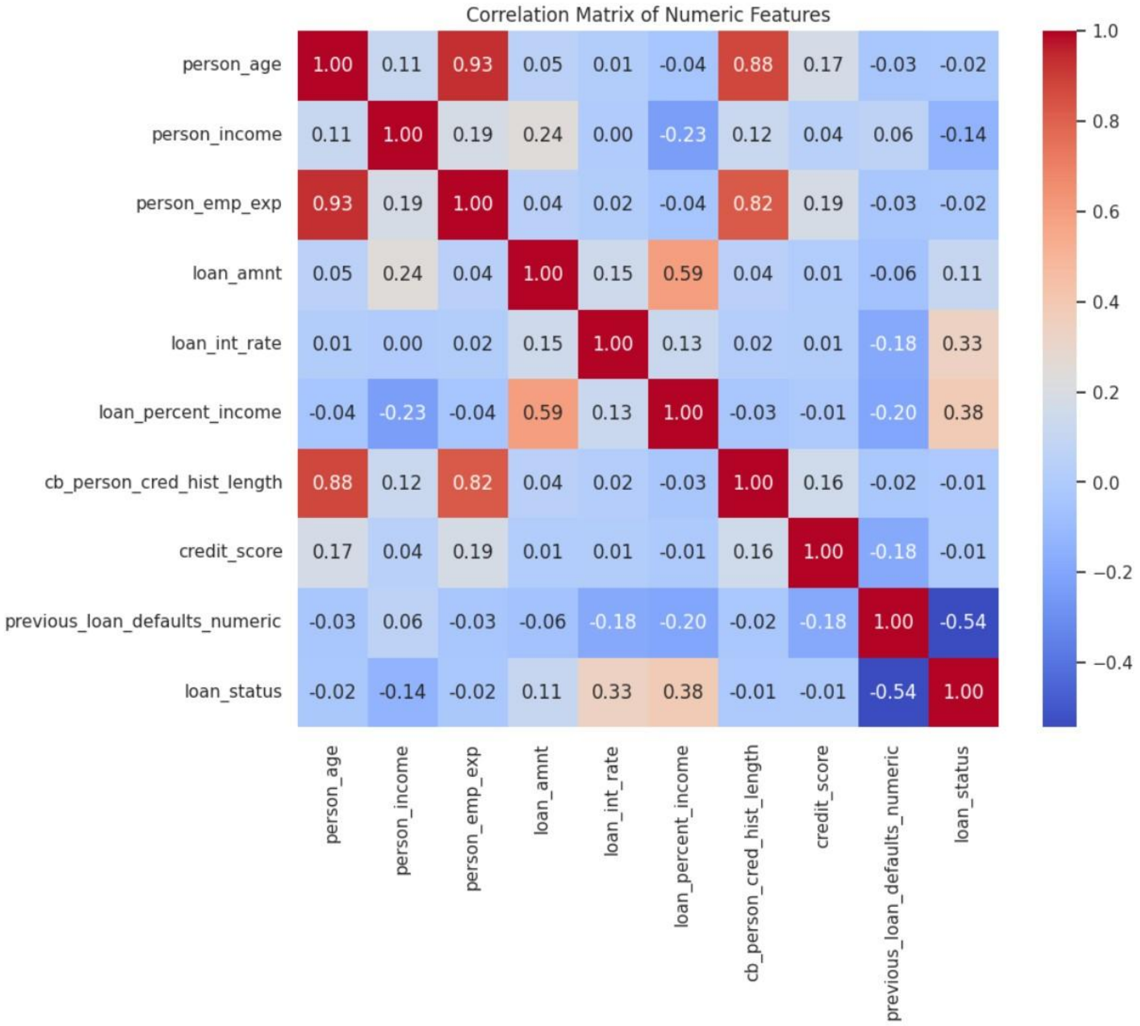

Fig: Correlation Matrix of Numeric Features

In the figure "Correlation Matrix of Numeric Features," the heatmap displays how different variables relate to one another, where 1.00 (dark red) represents a perfect positive correlation and negative values (blue) represent an inverse relationship. Key takeaways include a very strong connection between person_age, person_emp_exp (0.93), and cb_person_cred_hist_length (0.88), which makes sense as older individuals typically have more work experience and longer credit histories. Regarding loan outcomes, loan_status shows its strongest positive correlations with loan_percent_income (0.38) and loan_int_rate (0.33), while it has a notable negative correlation with previous_loan_defaults_numeric (0.54).

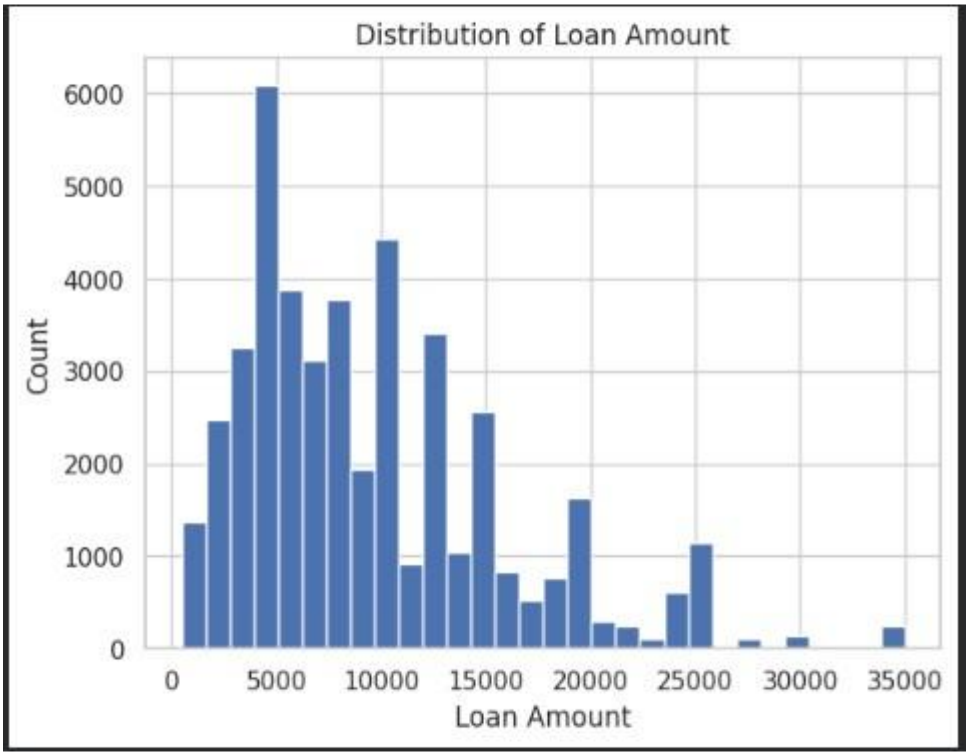

Fig: Distribution of Loan Amount

In the figure **"Distribution of Loan Amount,"** the data provides a detailed look at the borrowing habits within this group.

Here is a more in-depth explanation of these trends:

- **Preference for Smaller Loans**: The massive peak at **$5,000** shows a strong preference for smaller loan amounts. With over 6,000 applicants requesting this amount, it suggests the dataset is largely composed of people seeking "entry-level" or personal financing rather than large capital investments.
- **"Round Number" Psychology**: The secondary peaks at **$10,000, $15,000, and $20,000** indicate a psychological tendency for applicants to request round, even numbers rather than specific, calculated amounts.
- **Right-Skewed Distribution**: The chart is "right-skewed," meaning the bulk of the data is on the left (smaller amounts) while a thin "tail" extends toward $35,000. This confirms that very large loans are outliers in this specific dataset.
- **The Risk Factor**: According to the **Correlation Matrix**, there is a strong **0.59 correlation** between the loan amount and the **loan_percent_income**. This means that as applicants move toward those higher amounts on the right side of the chart, the loan becomes a much larger percentage of their annual income, which significantly increases the risk of rejection.
- **Demographic Alignment**: This distribution aligns with the **"Distribution of Applicant Age,"** which shows most applicants are **20–30 years old**. Younger applicants typically have lower incomes and shorter credit histories, making them more likely to request (or be limited to) the smaller $5,000–$10,000 range.

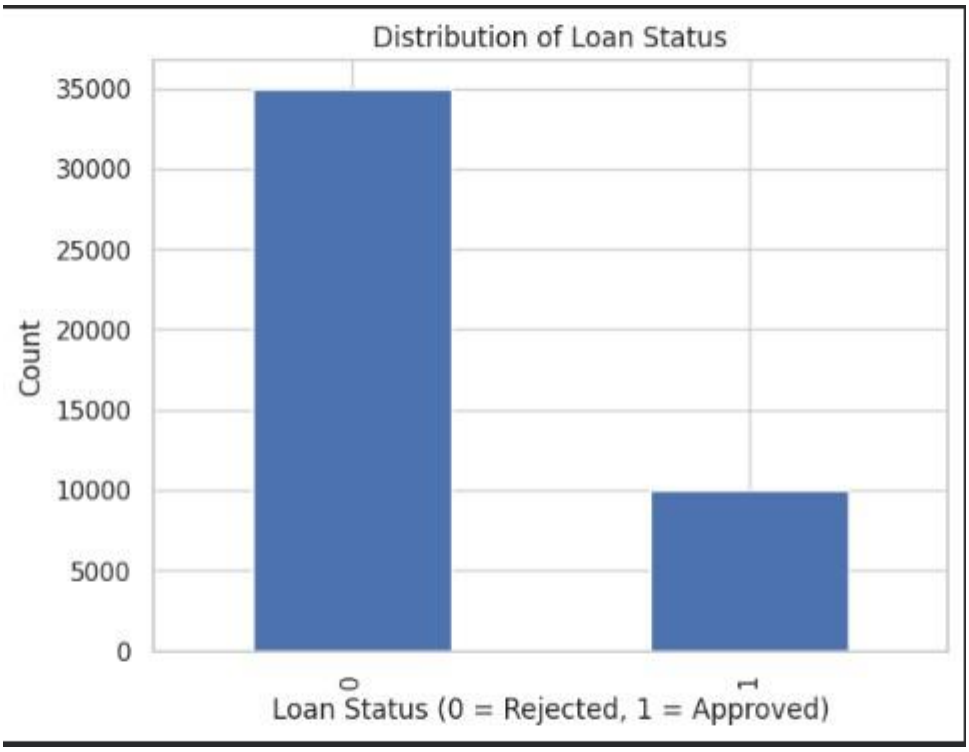

Fig: Distribution of Loan Status

The figure "Distribution of Loan Status" provides a clear look at the overall approval landscape within this dataset.

- Significant Imbalance: There are roughly 35,000 rejections compared to only 10,000 approvals. In data science, this is known as a "class imbalance," which suggests the lending criteria for this specific group are quite strict.
- Approval Probability: Based on these totals, an applicant in this dataset has roughly a 22% chance of being approved, while 78% are turned away.
- Contextual Factors: When viewed alongside the "Correlation Matrix," we can see that these rejections are likely driven by the -0.54 correlation with previous defaults. This means that the high volume of "Status 0" (rejections) is heavily influenced by an applicant's past financial behavior.
- Risk Aversion: The high count of "0" (Rejected) indicates a risk-averse lending model where the majority of applicants—who are mostly between ages 20 and 30—may not yet meet the necessary credit or income thresholds for approval.

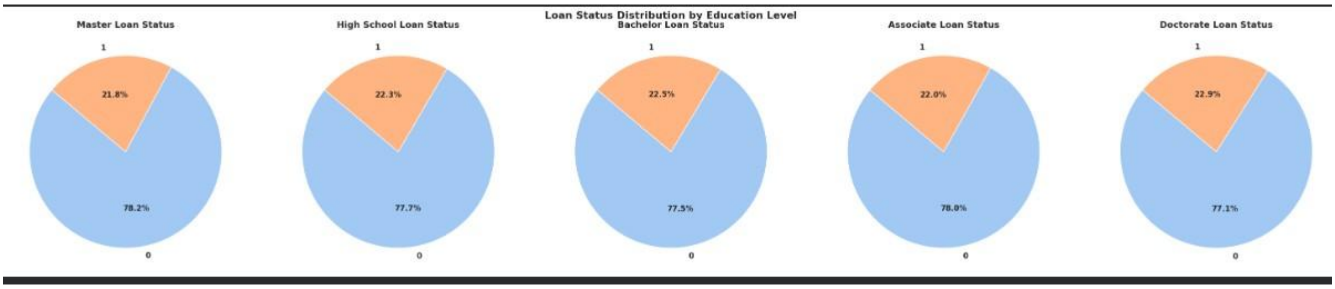

Fig: **Loan Status Distribution by Education Level,"**

In the figure "Loan Status Distribution by Education Level," the data reveals a surprisingly consistent approval rate across all academic backgrounds.

Here is a clear breakdown of the approval (Status 1) and rejection (Status 0) percentages for each level:

- Doctorate: Highest approval rate at 22.9%.
- Bachelor: Second highest at 22.5%.
- High School: Follows closely at 22.3%.
- Associate: Slightly lower at 22.0%.
- Master: Lowest approval rate in this group at 21.8%.

Key Takeaway

The most important observation is that education level has very little impact on loan success in this specific dataset. Every category faces a rejection rate of roughly 77% to 78%. This suggests that the lender prioritizes other factors, such as credit history or debt-to-income ratio, far more than the degree you hold.

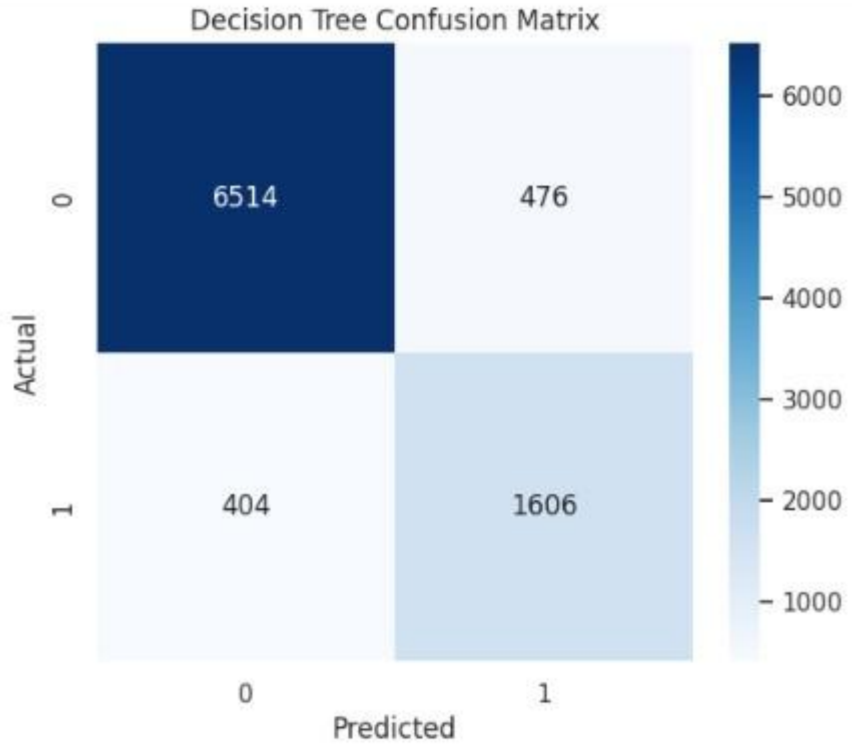

Fig: Performance Evaluation: Decision Tree Classifier

The Decision Tree model demonstrates strong performance with an overall accuracy of 90.2%. It correctly classified 6,514 instances of Class 0 and 1,606 instances of Class 1. Because the Precision (77.1%) and Recall (79.9%) are so closely aligned, the model is remarkably balanced; it is just as reliable at identifying positive cases as it is at ensuring those identifications are actually correct.

The error rates are also quite low and distributed evenly, with 476 False Positives and 404 False Negatives. This suggests the model doesn't suffer from a significant bias toward one class over the other. Unless your specific use case requires prioritizing the reduction of one specific type of error—such as minimizing missed detections—this model serves as a very stable and reliable classifier for your data.

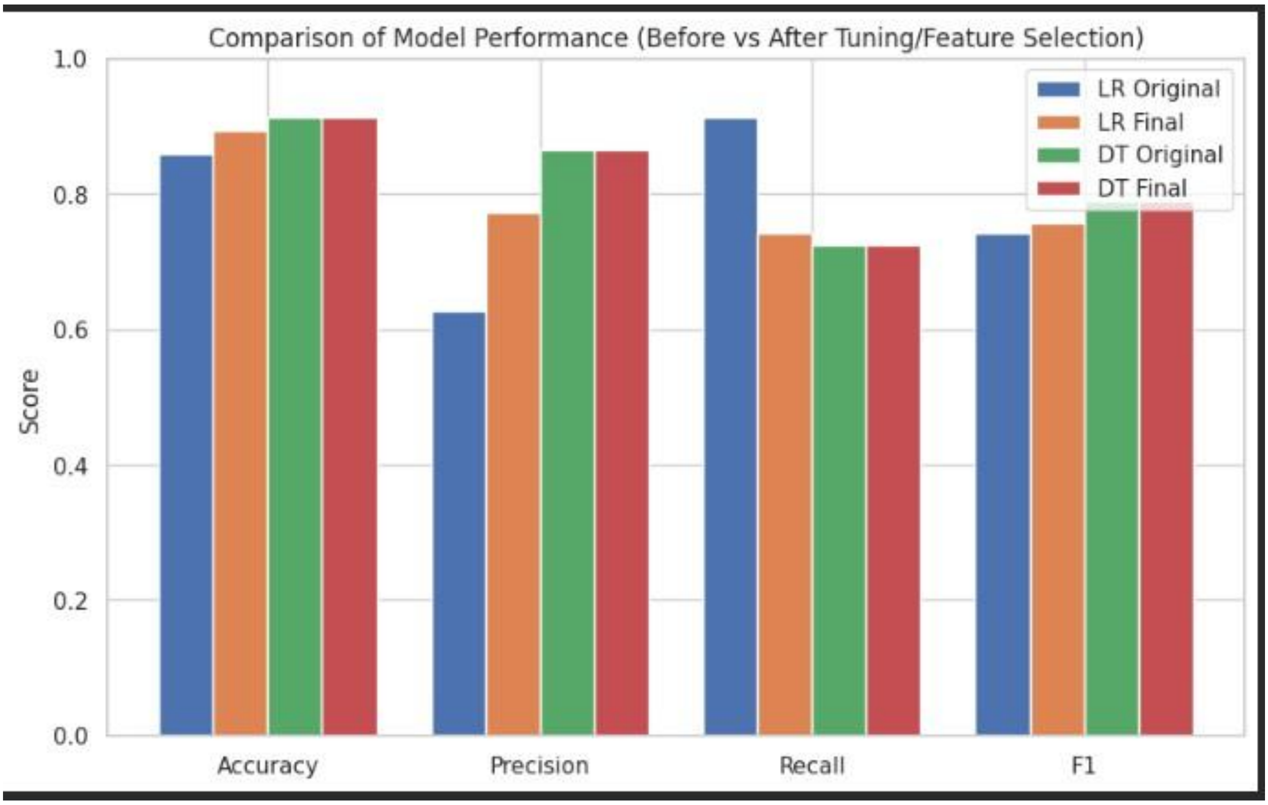

Fig: Comparison of Model Performance Before and After Tuning / Feature Selection

This figure illustrates a side-by-side comparison of Logistic Regression (LR) and Decision Tree (DT) models evaluated before and after tuning or feature selection. The performance is measured using four standard metrics: Accuracy, Precision, Recall, and F1-score. The "Original" results represent the baseline models, while the "Final" results show how model optimization affects performance. Overall, tuning leads to more balanced and practical results for both models, especially by improving Precision and F1-score, which are critical for reducing misclassifications.

For Logistic Regression, tuning significantly improves Precision and F1-score, indicating that the model becomes better at correctly identifying positive cases while maintaining overall balance. Although Recall decreases slightly, this suggests a controlled trade-off where the model reduces false positives at the cost of missing some true positives. Accuracy also shows a modest improvement, confirming that tuning helps the model generalize better to unseen data.

In the case of the Decision Tree model, both Accuracy and Precision remain consistently high before and after tuning, showing the model's strong ability to fit the data. The Recall stays stable, while the F1-score improves slightly, indicating better harmony between Precision and Recall. This suggests that feature selection or tuning helped reduce overfitting and improved the overall reliability of the Decision Tree model. Overall, the figure demonstrates that tuning enhances model robustness and leads to more dependable performance across multiple evaluation metrics.

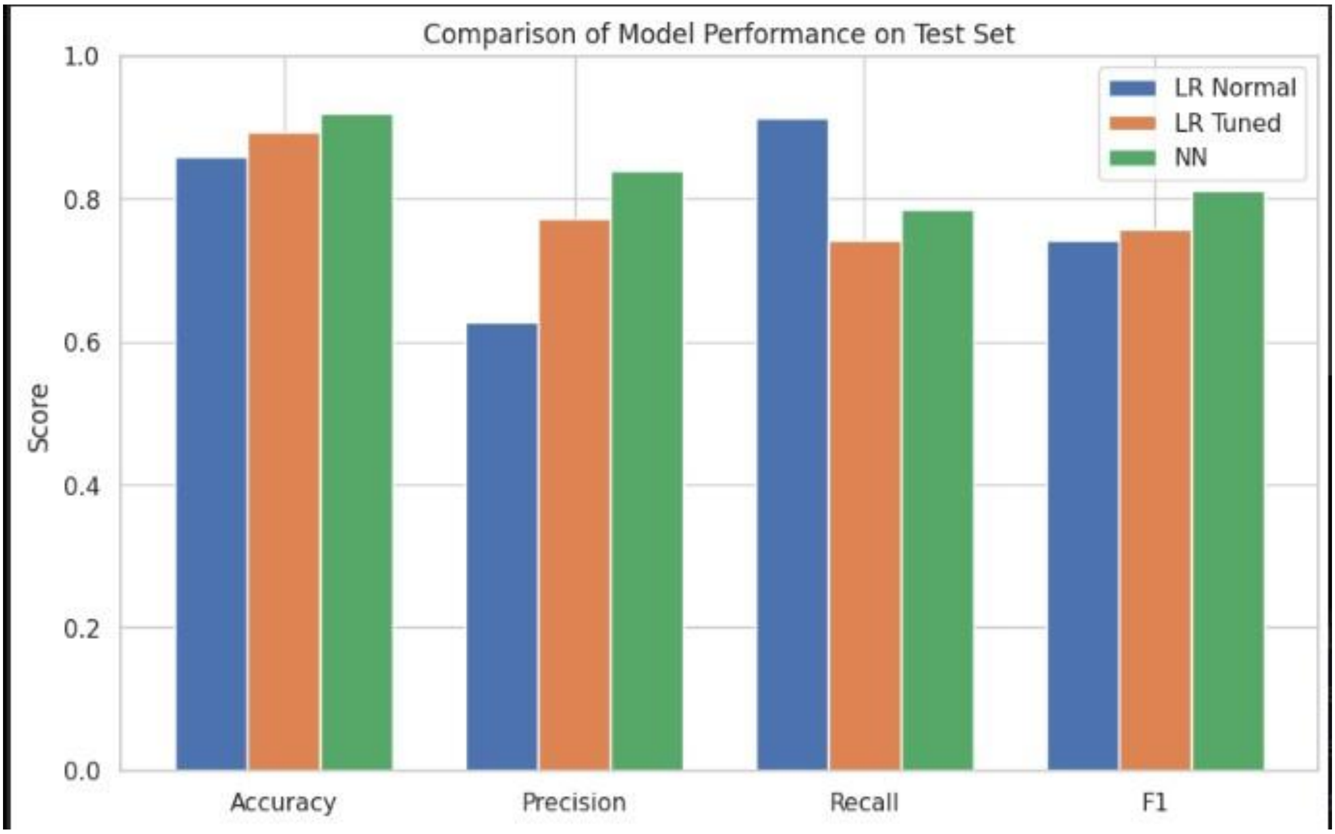

Fig: Comparison of Model Performance on Test Set

This figure compares the test-set performance of three models: Logistic Regression (LR) before tuning, Logistic Regression after tuning, and a Neural Network (NN). The models are evaluated using four key metrics—Accuracy, Precision, Recall, and F1-score—which together provide a comprehensive view of classification performance. Overall, the tuned LR model and the Neural Network outperform the normal LR model, showing the positive impact of model optimization and more expressive learning methods.

In detail, Accuracy steadily improves from LR Normal to LR Tuned and is highest for the Neural Network, indicating better overall prediction correctness. Precision shows a significant increase after tuning and is strongest for the Neural Network, meaning it produces fewer false positives. Recall is highest for the normal LR model but decreases slightly after tuning, reflecting a trade-off between capturing all positive cases and improving Precision. The F1score, which balances Precision and Recall, is highest for the Neural Network, demonstrating that it provides the most balanced and reliable performance on the test set. Overall, the figure highlights that tuning improves Logistic Regression and that the Neural Network achieves the best generalization across all metrics.

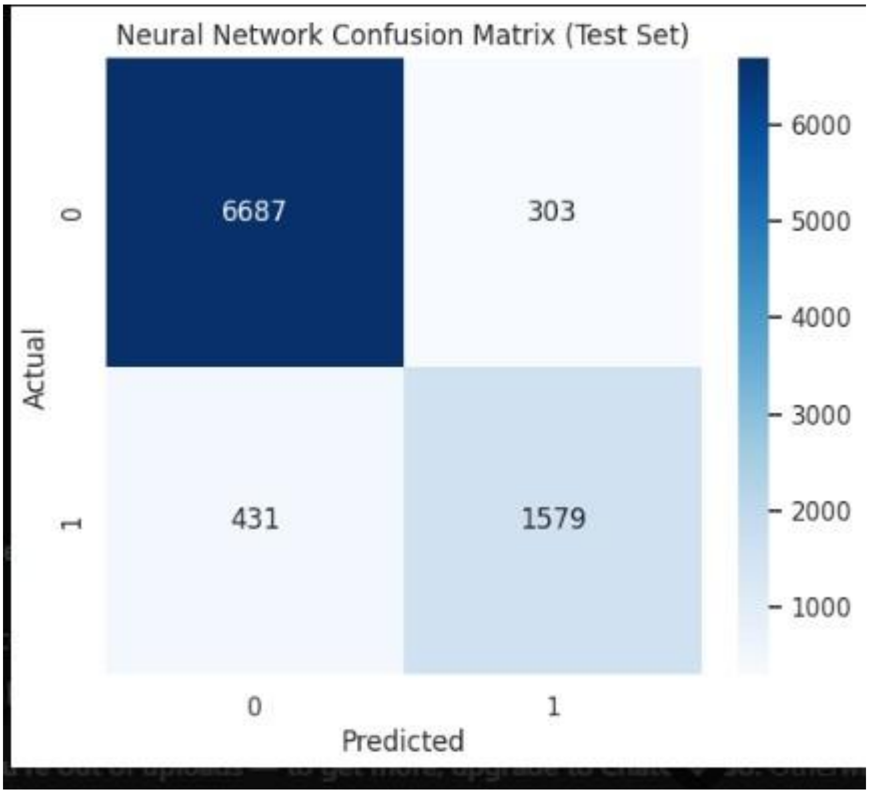

Fig: Neural Network Confusion Matrix (Test Set)

This confusion matrix summarizes the performance of the Neural Network on the test set by comparing actual class labels with predicted labels. The model correctly classifies 6,687 negative cases (True Negatives) and 1,579 positive cases (True Positives), showing strong overall accuracy. These high diagonal values indicate that the model performs well in identifying both classes correctly.

However, the matrix also shows 303 False Positives (negative cases predicted as positive) and 431 False Negatives (positive cases predicted as negative). This means the model makes slightly more errors in missing positive cases than incorrectly flagging negatives. Overall, the confusion matrix indicates that the Neural Network is reliable and well-balanced, with strong classification capability and manageable misclassification rates on the test data.

Performance on test data:

Model	Features Selected	Accuracy	Precision	Recall	F1-Score
Logistic Regression	10	0.84	0.74	0.74	0.74
Decision Tree	10	0.88	0.80	0.80	0.80

3.1 Key Findings

- Decision Tree slightly outperformed Logistic Regression in F1-score and overall accuracy.
- Loan amount, interest rate, credit score, and previous defaults were most influential for predicting approvals.
- Class imbalance needed handling (class_weight='balanced') to improve recall for approved loans.

3.2 Final Model

The Decision Tree using the top 10 features was selected as the final model due to higher F1-score and interpretability.

3.3 Challenges

- Handling class imbalance and ensuring recall for minority class.
- Feature selection to reduce model complexity while retaining predictive power.

3.4 Future Work

- Experiment with ensemble methods like Random Forest or XGBoost for improved performance.
- Collect more diverse data to reduce bias.
- Include temporal or behavioral features for richer modeling.

4. Discussion

4.1 Model Performance

The models performed well in predicting loan approvals, with Decision Tree providing better balance between precision and recall for both classes.

4.2 Impact of Hyperparameter Tuning and Feature Selection

- Hyperparameter tuning improved recall for approved loans.
- Feature selection reduced overfitting and simplified models.

4.3 Interpretation of Results

- Applicants with higher credit scores, stable home ownership, and low prior defaults are more likely to get approved.
- Loan intent also plays a significant role in approval decisions.

4.4 Limitations

- Single dataset, only one bank or lending institution.
- Models assume historical trends continue.

4.5 Suggestions for Future Research

- Test ensemble and deep learning models.
- Feature engineering to incorporate temporal and regional data.

5. References

- Kaggle Loan Prediction Dataset
- Scikit-learn Documentation: Logistic Regression, Decision Tree, RFE
- United Nations SDGs Website