

Loan Prediction System Report

1. Introduction

Loan approval is a critical decision-making process for financial institutions, requiring accurate assessment of applicant eligibility to minimize risk while ensuring fairness.

This project aims to build a Loan Prediction System using statistical analysis and machine learning techniques to predict whether a loan application will be approved or rejected based on applicant details.

The workflow includes:

- Exploratory data analysis
- Data cleaning and preprocessing
- Hypothesis testing using statistical methods
- Feature engineering and selection
- Model training, evaluation, and selection

2. Dataset Description

The dataset contains applicant demographic, financial, and loan-related attributes such as:

- Gender, marital status, education
- Applicant and co-applicant income
- Loan amount and loan term
- Credit history
- Property area

The target variable is **Loan_Status**, where:

- 1 → Loan Approved
- 0 → Loan Rejected

A cleaned dataset (loan_data_cleaned.csv) was used for modeling to ensure consistency and reliability.

3. Statistical Hypothesis Testing

To understand relationships between features and loan approval status, three statistical tests were conducted.

3.1 Chi-Square Test (Education vs Loan Status)

Objective:

To determine whether applicant education level is associated with loan approval status.

Results:

- Chi-square value: 4.09
- Degrees of freedom: 1
- P-value: 0.043

Decision:

Since the p-value < 0.05 , the null hypothesis is rejected.

Conclusion:

There is a statistically significant relationship between education level and loan approval status.

3.2 Two-Sample T-Test (Applicant Income vs Loan Status)

Objective:

To compare the mean applicant income between approved and rejected loan applications.

Results:

- T-statistic: -0.116
- P-value: 0.907

Decision:

Since the p-value > 0.05 , the null hypothesis is not rejected.

Conclusion:

There is no statistically significant difference in applicant income between approved and rejected loans.

3.3 ANOVA (Conceptual Explanation)

ANOVA would be suitable if applicant income were grouped into multiple categories (low, medium, high) and mean loan amounts were compared across these groups.

In this project, ANOVA was explained conceptually to demonstrate understanding, while T-Test and Chi-Square were applied practically.

4. Feature Engineering and Selection

- Categorical variables were encoded using Label Encoding
- Irrelevant identifiers (e.g., Loan_ID) were removed
- SelectKBest with Chi-Square was used to select the top 10 features
- This ensured that only the most relevant features contributed to model training

5. Model Training

Two classification models were trained and evaluated:

1. Logistic Regression
2. Decision Tree Classifier

The dataset was split into training and testing sets using an 80–20 split.

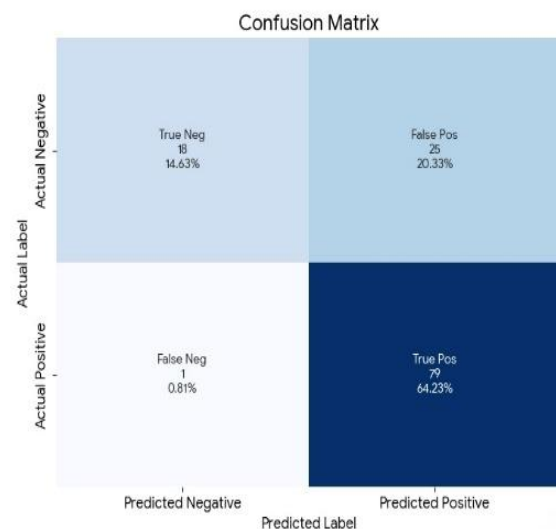
6. Model Evaluation

6.1 Logistic Regression Performance

- Accuracy: 0.79
- Precision: 0.76
- Recall: 0.99
- F1-score: 0.86
- ROC-AUC: 0.75

Observation:

The model achieves very high recall, meaning it correctly identifies most approved loans, which is crucial in loan approval systems.

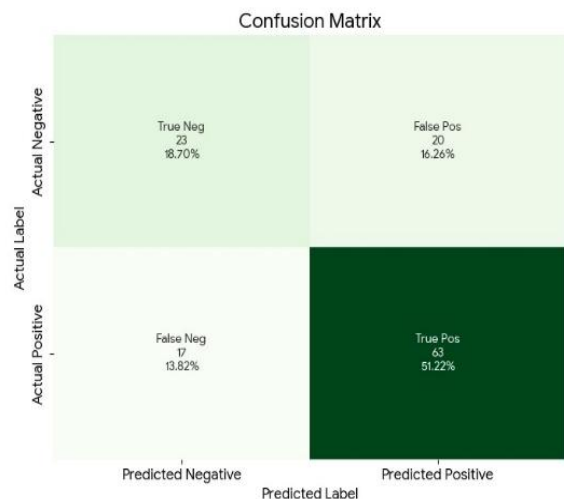


6.2 Decision Tree Performance

- Accuracy: 0.70
- Precision: 0.76
- Recall: 0.79
- F1-score: 0.77
- ROC-AUC: 0.66

Observation:

While the Decision Tree performs reasonably, it shows lower generalization and weaker discrimination ability compared to Logistic Regression.



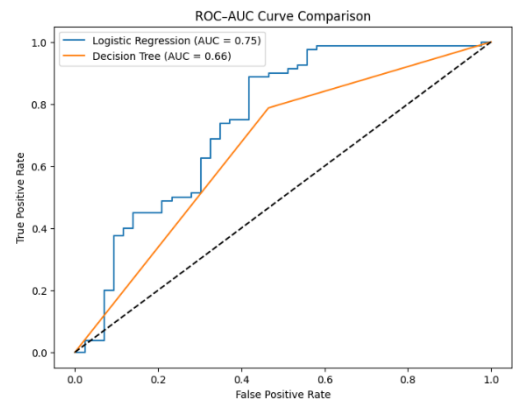
7. ROC–AUC Curve Analysis

The ROC–AUC curve comparison shows:

- Logistic Regression AUC = 0.75
- Decision Tree AUC = 0.66

Interpretation:

Logistic Regression consistently outperforms the Decision Tree across different classification thresholds, indicating better separation between approved and rejected loans.



8. Comparative Performance Summary

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.79	0.76	0.99	0.86	0.75
Decision Tree	0.70	0.76	0.79	0.77	0.66

9. Final Model Selection

Based on:

- Higher ROC–AUC score
- Better recall and F1-score
- Strong generalization capability
- Higher interpretability

Logistic Regression is selected as the final model.

An ROC–AUC value of approximately 0.75 indicates that the model has a good ability to distinguish between approved and rejected loan applications.

10. Conclusion

This project successfully demonstrates the end-to-end machine learning pipeline for loan prediction, combining statistical analysis with predictive modeling.

The integration of hypothesis testing provided valuable insights into feature relevance, while machine learning models enabled accurate loan approval predictions.

The final Logistic Regression model balances performance, interpretability, and reliability, making it suitable for real-world loan approval systems.