Loan Approval Prediction Report

1. Introduction

Financial institutions rely heavily on data-driven decision-making to evaluate loan applications.

This project focuses on predicting whether a loan application will be approved or rejected using machine learning techniques. Two algorithms were implemented and compared:

- **K-Nearest Neighbors (KNN)**
- **Decision Tree Classifier**

The objective is to identify the better-performing model and provide insights into which applicant and loan features most strongly influence approval decisions.

2. Dataset Overview

The dataset contains multiple financial and applicant-related features, including:

- Annual Income
- Credit Score
- Loan Amount
- Loan Term
- Employment Status
- Education
- Loan Approval Status

  (converted to: 1 = Approved, 0 = Rejected)

Cleaning and Preparation

- Removed rows with missing target labels
- Converted approval labels from text to numeric (0/1)
- Saved cleaned dataset as **loan_approval_cleaned.csv**

3. Preprocessing Workflow

A complete preprocessing pipeline was built using Scikit-learn, ensuring reproducibility and consistency.

3.1 Numeric Features

- Missing values → **Median Imputation**
- Scaling → **StandardScaler**

  (important for KNN performance)

3.2 Categorical Features

- Missing values → **Most Frequent Imputation**
- Encoding → **OneHotEncoder**

  (to convert employment-related (string, object, bool typ) columns into machine-readable format)

3.3 Train–Test Split

- **80% training**, **20% testing**
- Stratified split to maintain class distribution

All transformations were combined using **ColumnTransformer + Pipeline** for clean, modular modeling.

4. Model Development

4.1 K-Nearest Neighbors (KNN)

- Baseline model trained
- Hyperparameter tuning using GridSearchCV:
  - n_neighbors values (3–9)
  - Distance metrics (minkowski, manhattan)
  - Weighting (uniform vs. distance)

4.2 Decision Tree Classifier

- Baseline Decision Tree trained
- GridSearchCV tuning:
  - max_depth
  - min_samples_split
  - min_samples_leaf
  - criterion (gini/entropy)

5. Model Evaluation

5.1 Evaluation metrics:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

5.2 Performance Comparison

Model	Accuracy	Precision	Recall	F1-score
KNN (baseline)	0.892272	0.908752	0.919021	0.913858
Decision Tree (baseline)	0.980094	0.983083	0.984934	0.984008
KNN (tuned)	0.903981	0.913444	0.934087	0.923650
Decision Tree (tuned)	0.981265	0.976516	0.990584	0.985019

5.3 Result Summary

The **tuned Decision Tree** achieved the highest overall F1-score and showed a better balance between correctly approving eligible applicants and rejecting ineligible ones.

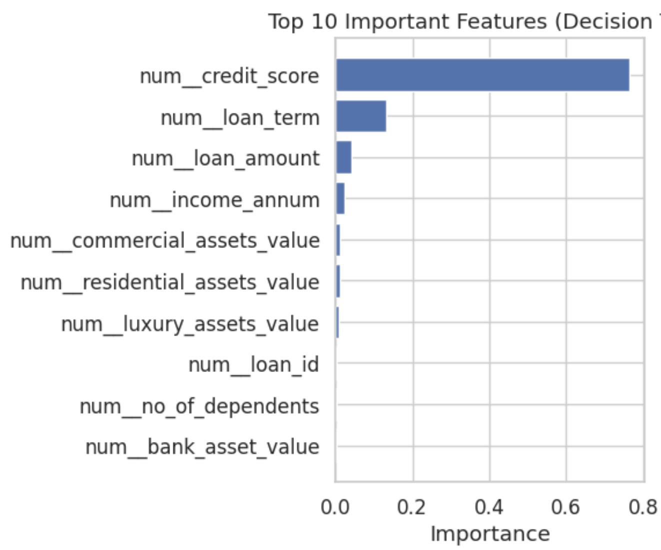
Therefore, it was selected as the **best model** for this task.

6. Feature Importance & Interpretation

Using the tuned Decision Tree model, feature importance values were extracted.

6.1 Top Predictive Features

1. **Credit Score**
2. **Loan Term**
3. **Loan Amount**



6.2 Interpretation

- Higher **credit score** strongly increases the likelihood of loan approval.
- Large **loan amounts** or long **loan terms** indicate higher lender risk.
- These factors heavily influence the approval decision, aligning with real-world lending practices.

7. Business Insights for the Bank

7.1 Creditworthiness is the strongest determinant

The model shows that applicants with poor credit scores are the least likely to be approved.

Recommendation:

- Define clear minimum credit thresholds
- Offer financial improvement programs for borderline applicants

7.2 Loan term affects risk

Longer loan terms correlate with lower approval chances.

Recommendation:

- Provide applicant education on optimizing loan terms
- Adjust risk scoring for extended loan durations

7.3 High loan amounts reduce approval probability

Large requested loan amounts compared to applicant income present increased default risk.

Recommendation:

- Implement loan-to-income ratio guidelines
- Suggest alternative loan sizes or restructuring options

8. Conclusion

This project implemented two machine learning models to predict loan approval status. After training, tuning, and evaluating both models:

- The tuned Decision Tree emerged as the best-performing model.
- It offers:
  - Strong predictive performance
  - High interpretability
  - Clear feature importance insights

The findings help the bank:

- Improve loan decision policies
- Reduce default risk
- Make fairer, data-driven approval decisions

This model can be integrated into the bank's loan screening process to assist human officers and enhance the reliability of approval assessments.