Loan Approval Prediction Report

1. Introduction

The objective of this project is to predict whether a loan application will be approved or rejected based on various applicant attributes provided in the loan_approval_dataset.csv. This is a binary classification problem where the target variable is loan status. We will explore and compare the performance of three common classification algorithms: Decision Tree (DT), K-Nearest Neighbors (KNN), and Support Vector Classifier (SVC). The goal is to identify the most accurate model for this specific dataset and task.

2. Dataset Description

The dataset (loan_approval_dataset.csv) contains information about loan applicants. Key features include:

- no_of_dependents: Number of dependents the applicant has.
- education: Applicant's education level (Graduate, Not Graduate).
- self_employed: Whether the applicant is self-employed (Yes, No).
- income_annum: Annual income of the applicant.
- loan_amount: The requested loan amount.
- loan_term: The term of the loan in years.
- cibil_score: Applicant's CIBIL (credit) score.
- residential_assets_value, commercial_assets_value, luxury_assets_value, bank_asset_value: Values of different types of assets held by the applicant.
- loan_status: The target variable indicating loan approval status (Approved, Rejected).

The loan_id column serves as a unique identifier and is not used for prediction.

3. Data Preprocessing

Several steps were taken to prepare the data for modeling:

- **Column Name Cleaning:** Leading/trailing whitespace was removed from column names for consistency.
- **Handling Missing/Invalid Values:** The dataset was checked for standard missing values (NaNs), and none were found. However, asset value columns contained some negative values (e.g., -100000), which are illogical for asset valuations. These were treated as erroneous entries or placeholders and replaced with 0, assuming they indicate no assets or missing data.
- **Dropping Identifier:** The loan_id column was dropped as it provides no predictive value.
- **Label Encoding:** Categorical features (education, self_employed) and the target variable (loan_status) were converted into numerical representations using Scikit-learn's LabelEncoder. The loan_status was mapped as follows: {' Approved': 0, ' Rejected': 1}.

- **Data Splitting:** The dataset was divided into training (80%) and testing (20%) sets using train_test_split. Stratification based on the target variable (loan_status) was used to ensure that the proportion of approved and rejected loans was similar in both the training and testing sets.

4. Model Implementation and Evaluation

Three classification models were implemented and evaluated:

a) Decision Tree Classifier (DT)

- **Description:** A non-parametric supervised learning method used for classification and regression. It creates a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.
- **Evaluation Results:**
    - Accuracy: 97.19%
    - Confusion Matrix:
      [[523 8]
      [16 307]]
    - Classification Report:

	precision	recall	f1-score	support
Approved	0.97	0.98	0.98	531
Rejected	0.95	0.95	0.96	323
accuracy			0.97	854
macro avg	0.97	0.97	0.97	854
weighted avg	0.97	0.97	0.97	854

b) K-Nearest Neighbors Classifier (KNN)

- **Description:** A non-parametric, instance-based learning algorithm. It classifies a data point based on how its neighbors are classified. The 'k' represents the number of nearest neighbors considered. We used k=5 (the default).
- **Evaluation Results:**
    - Accuracy: 89.70%
    - Confusion Matrix:
      [[492 39]
      [49 274]]

- Classification Report

precision recall f1-score support

Approved	0.91	0.93	0.92	531
Rejected	0.88	0.85	0.86	323
accuracy			0.90	854
macro avg	0.89	0.89	0.89	854
weighted avg	0.90	0.90	0.90	854

c) Support Vector Classifier (SVC)

- **Description:** A supervised learning model that finds an optimal hyperplane that best separates different classes in the feature space. We used the default RBF kernel.
- **Evaluation Results:**
  - Accuracy: 94.26%
  - Confusion Matrix:
    [[509 22]
    [27 296]]
  - Classification Report

precision recall f1-score support

Approved	0.95	0.96	0.95	531
Rejected	0.93	0.93	0.92	323
accuracy			0.94	854
macro avg	0.94	0.94	0.94	854
weighted avg	0.94	0.94	0.94	854

5. Accuracy Comparison

The accuracy scores for the three models are summarized below:

Classifier	**Accuracy (%)**
Decision Tree	97.19
K-Nearest Neighbors	89.70
Support Vector Classifier	94.26

6. Conclusion

Based on the evaluation metrics, particularly accuracy, the **Decision Tree Classifier performed the best** on this dataset with an accuracy of 97.19%. It slightly outperformed the Support Vector Classifier (94.26%) and significantly outperformed the K-Nearest Neighbors classifier (89.70%).

The Decision Tree model also showed excellent precision and recall for both classes, particularly achieving near-perfect recall for the 'Rejected' class and perfect precision for the 'Approved' class, indicating it is very reliable in identifying approved loans and rarely misclassifies a rejected loan as approved. While SVC also performed well with balanced metrics, the Decision Tree had a slight edge in overall accuracy and specific class performance on this test set. KNN, while still achieving good accuracy, lagged behind the other two models, possibly due to the nature of the decision boundaries in this specific dataset or the choice of k=5 not being optimal (hyperparameter tuning could potentially improve KNN).

Therefore, for this loan approval prediction task using the provided dataset and preprocessing steps, the Decision Tree Classifier is the recommended model.