Loan Application Prediction

Name : Kunal Sapkal.

I've approached the above problem with the following steps:

- Preprocessing Steps:
  - Libraries are imported (e.g., Pandas, NumPy, Scikit-learn).
  - The dataset is loaded from CSV files (Assignment_Train.csv and Assignment_Test.csv).
  - Training and test datasets are concatenated for preprocessing.

- Handling Missing Values:
  - Code is provided to identify and handle missing values, particularly in numeric columns.

- Encoding Categorical features:
  - Encoded categorical features using Label-Encoders.

- Standardization :
  - Normalized the features to get better results.

- Feature Extraction :
  - Included only relevant features for training the model.

- Training :
  - Two models are trained parallely.
  - The Naïve Bayes and Random Forest Classifier.

- Evaluation :
  - Both models are evaluated using confusion metrics.

- Test File prediction :
  - Predicted the values for test file provided.

Insights through data :

- I have plotted 3 important graphs according to which the the age factor and the CIBIL score factor influences the output in a major way.
- Other than that the basic details like name , presence on the applications , etc does not influence the data.

- The heatmap shows the detailed description of the influencing features , those features are only considered for training the model.

Two machine learning algorithms were tested:

- Naive Bayes: Known for simplicity and speed, particularly effective on small datasets.
- Random Forest: An ensemble method, often outperforming simpler models by reducing overfitting.

Metrics Used:

- Accuracy: To measure the percentage of correctly predicted instances.
- Confusion Matrix: To visualize true vs. false predictions.