PROJECT: LOAN APPROVAL PREDICTION

The goal of this project is to develop and evaluate various machine learning models to predict the loan status of people based on their financial records and key information related to loan applications. By analyzing and comparing different models, the aim is to identify the most effective model for predicting whether a loan will be approved or not. #

Steps Involved

1. Data Exploration

- **Load the Dataset**: Import the dataset using Pandas from a CSV file.
- **Initial Inspection**: Display the first few rows, last few rows, and a random sample of the data to understand its structure.
- **Data Overview**: Check the shape of the dataset and review the column names and data types.
- **Missing Values and Duplicates**: Identify any missing values and duplicate records in the dataset and handle them accordingly.
- **Descriptive Statistics**: Generate summary statistics for numerical and categorical features to understand data distributions and uni

2. Data Cleaning

- **Handling Missing Values**: Handle the missing values in the dataset.
- **Handling Outliers**: Check the outliers in the dataset using IQR or Z-score then use necessary technique to clean them.
- **Handling Duplicated**: Removing duplicates from the dataset.
- **Ready for EDA Step**: Ensuring the dataset is in best condition for EDA.

3. EDA Visualization

- **Numerical Features**: Create histograms to visualize the distribution of numerical features.
- **Correlation Heatmap**: Generate a heatmap to visualize the correlation between numerical features.
- **Loan Status Distribution**: Use countplots to visualize the distribution of loan status.
- **Categorical Features**: Visualize the relationship between categorical features and loan status using countplots.
- **Pairplot**: Create pairplots to visualize relationships between numerical features and 4oan status.

4. Data Preprocessing¶

- **Encoding Categorical Variables**: Convert categorical features into numeric using One Hot Encoder.

- **Scaling Numerical Features**: Standardize numerical features using StandardScaler to ensure they are on same scale.

5. Model Building and Evaluation

- **Splitting the Dataset**: Divide the dataset into training and testing sets.
- **Model Selection**: Choose a variety of machine learning models, including:
    - Random Forest
    - Gradient Boosting
    - Logistic Regression
    - Support Vector Classifier (SVC)
    - Decision Tree
    - Bagging Classifier
    - LightGBM
- **Model Training and Prediction**: Train each model on the training data and make predictions on the testing data.
- **Evaluation Metrics**: Calculate and compare performance metrics for each model, including accuracy, pre6ision, recall, and F1 score.

6. Model Comparison and Visualization

- **Metric Comparison**: Compare the performance of each model using the calculated metrics.
- **Visualization**: Create bar plots to visualize the performance metrics for each model.
- **Best Model Selection**: Identify the best-performing model based on metrics such as accuracy7 precision, recall, and F1 score.

7. Final Analysis

- **Best Model for Imbalanced Data**: Discuss the model that performs best based on the F1 score, which is a good metric for imbalanced datasets.
- **Model Training on Full Dataset**: Optionally, train the best-performing model on full dataset to optimize its performance.