Loan Status Prediction using Machine Learning:

The project, titled "Loan Status Prediction using Machine Learning," aims to automate the loan approval process for a financial institution. This task involves building a machine learning model to predict whether a loan application should be approved or rejected. By analyzing applicant data, the model provides a data-driven, efficient, and accurate alternative to manual evaluation. The project compares two supervised learning algorithms, Logistic Regression and Decision Tree, to determine the most effective model for this problem.

Dataset Details

The dataset used is `LoanStatus.csv`. It contains detailed information about loan applicants. The dataset includes a mix of categorical and numerical features that are crucial for predicting the target variable, `Loan_Status`.

Key Features:

- **`Gender`**: Gender of the applicant.
- **`Education`**: Educational qualification of the applicant.
- **`ApplicantIncome`**: Applicant's monthly income.
- **`Credit_History`**: A binary variable indicating if the applicant has a credit history.
- **`Property_Area`**: The area where the applicant's property is located.
- **`Loan_Status`**: The target variable, indicating if the loan was approved (Y) or not (N).

The dataset required preprocessing to handle missing values and to convert categorical data into a numerical format suitable for machine learning models.

Steps Followed

1. Data Loading and Preprocessing

- **Data Loading**: The dataset was loaded from the `LoanStatus.csv` file into a pandas DataFrame.
- **Handling Missing Values**: Missing values in all columns were imputed with the most frequent value (mode) of that column.
- **Feature Engineering**: The `Loan_ID` column was dropped as it is a unique identifier and not a useful feature for the model.

- **Categorical Encoding**: All categorical features were converted into numerical format using **one-hot encoding** via `pd.get_dummies()`. This step is crucial for models like Logistic Regression that require numerical input.
- **Data Splitting**: The preprocessed data was split into a training set and a testing set, with 80% of the data used for training and 20% for testing.

2. Modeling

Two machine learning models were chosen and implemented for this binary classification task:

- **Logistic Regression**: A linear model used for binary classification. It is a simple yet powerful model often used as a baseline.
- **Decision Tree Classifier**: A non-linear model that partitions the data into a tree-like structure to make decisions. It's capable of capturing complex relationships in the data.

3. Evaluation and Results

Both models were trained on the preprocessed training data and evaluated on the test set. The primary metric used for evaluation was **accuracy**.

- **Logistic Regression**: Achieved an accuracy of **84.02%**.
- **Decision Tree Classifier**: Achieved an accuracy of **80.49%**.

The performance of the models is visualized below to provide a clear comparison.

4. Visualizations & Insights

- **Model Comparison Bar Chart** : This bar chart visually compares the accuracy of the two models. It clearly shows that the Logistic Regression model performed better than the Decision Tree on this dataset.
- **Data Insights**:
  - The bar chart highlights the superior performance of Logistic Regression. Its simplicity and stability make it a strong choice for this dataset.
  - The Decision Tree's lower accuracy suggests it may be overfitting the training data, capturing noise rather than the underlying patterns. Further tuning of hyperparameters would be necessary to improve its performance.

Bonus Work

The comparison and evaluation of two distinct machine learning models (Logistic Regression and Decision Tree) for the same task serve as a bonus. This step goes beyond a single model implementation, providing a comparative analysis of their effectiveness and offering insights into their respective strengths and weaknesses on the given dataset.

Conclusion & Learning Outcomes

This project successfully built and evaluated a system for predicting loan approval status. The Logistic Regression model emerged as the most suitable choice for this dataset, achieving a solid accuracy of 84.02%. This project reinforced key concepts in the machine learning workflow, including data preprocessing (handling missing values and one-hot encoding), model selection, and performance evaluation. The comparative analysis between the two models demonstrated that a seemingly simple model can sometimes outperform a more complex one, emphasizing the importance of a data-driven approach to model selection.