**Iris Flower Species Prediction**

**1. Introduction**

**Problem Statement**

The goal of this project is to build machine learning models that predict the species of iris flowers based on their sepal and petal measurements.

**2. Data Overview**

**Data Source**

The dataset used in this project is in a CSV file named Iris.csv. It includes data on iris flower measurements and their corresponding species.

**Data Description**

The dataset contains the following columns:

* **Id**: Identifier for each record.
* **SepalLengthCm**: Sepal length in centimeters.
* **SepalWidthCm**: Sepal width in centimeters.
* **PetalLengthCm**: Petal length in centimeters.
* **PetalWidthCm**: Petal width in centimeters.
* **Species**: Species of the iris flower (Iris-setosa, Iris-versicolor, Iris-virginica).

**3. Data Preprocessing**

1. **Loading the Data**: The dataset is loaded into a DataFrame using pandas.
2. **Drop Unnecessary Columns**: Remove the Id column as it does not contribute to the analysis.
3. **Initial Examination**: Inspect the first few rows and overall structure of the dataset.
4. **Check for Missing Values**: Identify and handle any missing values.

**4. Exploratory Data Analysis (EDA)**

**Visualizations**

1. **Histograms**:
   * Distribution of Sepal Length
   * Distribution of Sepal Width
   * Distribution of Petal Length
   * Distribution of Petal Width
2. **Scatter Plots**:
   * Sepal Length vs Sepal Width (colored by Species)
   * Petal Length vs Petal Width (colored by Species)
3. **Pair Plot**: Explore the pairwise relationships between features to understand their interactions.

**Correlation Matrix**

* **Label Encoding**: Convert categorical values (Species) to numeric values using LabelEncoder.
* **Heatmap**: Visualize the correlation matrix to understand the strength of relationships between different features.

**5. Model Development**

**Feature Selection**

The features used in the model include:

* SepalLengthCm
* SepalWidthCm
* PetalLengthCm
* PetalWidthCm

The target variable is:

* Species

**Data Splitting**

The dataset is divided into training and testing sets to evaluate the performance of the model. A common split ratio is 75% for training and 25% for testing.

**6. Model Training and Evaluation**

**Logistic Regression**

1. **Training**: Fit a Logistic Regression model to the training data.
2. **Evaluation**: Assess the model’s performance on the testing data using accuracy metrics and a confusion matrix.

**Random Forest Classifier**

1. **Training**: Fit a Random Forest Classifier model to the training data. This model uses an ensemble of decision trees to make predictions.
2. **Evaluation**: Evaluate the model using accuracy metrics and a confusion matrix.

**7. Conclusions**

The project involved building and evaluating two machine learning models—Logistic Regression and Random Forest Classifier—to predict the species of iris flowers based on sepal and petal measurements. The models were assessed using accuracy and confusion matrices, providing insights into their predictive accuracy. Future work could include trying more advanced algorithms or fine-tuning the existing models for better performance.