Iris Flower Classification Project 🚀

In this project, I developed a machine learning model to classify Iris flower species based on their features, utilising the well-known Iris dataset. Here's a detailed breakdown of the steps I undertook, the methodologies applied, and the results obtained.

1. Data Loading 📥

I began by importing the necessary libraries, including pandas for data manipulation. I loaded the dataset from a specified path, ensuring I had a clear understanding of the data's structure. The initial data inspection revealed five features: SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm, and Species.

2. Data Exploration 🔍

To gain insights into the dataset, I performed exploratory data analysis (EDA) using seaborn and matplotlib. This involved:

- Checking for missing values to ensure data integrity.
- Generating summary statistics to understand the distribution and central tendencies of each feature.
- Creating pair plots to visualize relationships between features, highlighting distinctions between the three species: Iris-setosa, Iris-versicolor, and Iris-virginica.

3. Data Preprocessing 🛠️

Preprocessing was crucial for preparing the data for model training. I executed the following steps:

- Dropped unnecessary columns, if any.
- Separated the features (X) from the target variable (y).
- Encoded categorical labels into numerical format for the model.
- Split the dataset into training (70%) and testing (30%) sets using stratified sampling.
- Standardised the features with StandardScaler, which is vital for algorithms sensitive to feature scaling.

I also created box plots to visually assess feature distributions across species, identifying potential outliers and trends.

4. Model Building 🏗️

For the classification task, I chose the RandomForestClassifier, a robust ensemble method known for its high accuracy and interpretability. I initialised the model with 100 estimators, laying the groundwork for training.

5. Model Training 📊

To optimise the model's performance, I implemented HalvingGridSearchCV, allowing for hyperparameter tuning across various configurations. I defined a grid of parameters such as the number of estimators, maximum depth, and minimum samples split.

Upon fitting the model, I found the optimal parameters:

- **Max Depth:** None
- **Min Samples Leaf:** 2
- **Min Samples Split:** 2
- **N Estimators:** 100

The best cross-validation score achieved was **93.33%**, demonstrating strong predictive capabilities.

6. Model Evaluation 🧪

I evaluated the model's performance using several metrics:

- **Confusion Matrix:** It highlighted how well the model performed across the different species.
- **Classification Report:** Provided precision, recall, and F1-scores, indicating that the model was particularly accurate in identifying Iris-setosa (100% accuracy).
- **Accuracy Score:** Achieved an impressive **88.89%** on the test set, with cross-validation accuracy reaching **96.67%**.

These metrics confirmed the model's reliability and its ability to generalise well to unseen data.

7. Visualization 📈

To enhance understanding of the model's performance and feature significance, I created several visualisations:

- Histograms of feature distributions showcased how the features varied across different species.
- A heatmap of the confusion matrix provided a clear overview of model predictions versus actual values.
- A bar plot illustrating feature importance highlighted that PetalLengthCm and PetalWidthCm were the most significant predictors in classifying Iris species.

Final Results 🌟

The model demonstrated exceptional accuracy, with cross-validation confirming its robustness. This project deepened my understanding of machine learning methodologies and solidified my skills in data analysis, model evaluation, and visualisation techniques. Notably, I learned the importance of data preprocessing, feature selection, and hyperparameter tuning in enhancing model performance.

Skills Acquired:

Data Analysis, Machine Learning, Data Visualisation, Model Evaluation, Feature Engineering

🚀 **Top Hashtags:**

#DataScience #MachineLearning #IrisFlowerClassification #RandomForest #DataAnalysis #ModelEvaluation #HyperparameterTuning #DataVisualization #ExploratoryDataAnalysis #FeatureImportance #Python #Pandas #Seaborn #Matplotlib #ScikitLearn #AI #Statistics #DeepLearning #DataPreprocessing #Kaggle