NAME: MUHAMMAD SIKANDER BAKHT

STUDENT ID: CA/AU3/2102

INTERNSHIP: DATA SCIENCE INTERNSHIP

COMPANY: CODE ALPHA

ASSIGNMENT: TASK 01

CODE ALPHA TASK 1 :

PROJECT : IRIS FLOWER CLASSIFICATION

📌 Overview

This project aims to classify Iris flowers into three species (*Setosa, Versicolor, Virginica*) using supervised machine learning techniques. The Iris dataset, a well-known benchmark dataset, contains measurements of **sepal length, sepal width, petal length, and petal width**.

The workflow covers **data loading, preprocessing, exploratory analysis, model training, hyperparameter tuning, evaluation, model saving, and predictions on new data**.

🎯 Objectives

- Perform **data preprocessing and cleaning** for model readiness.
- Conduct **EDA** with visualizations to understand patterns and correlations.
- Train and evaluate multiple machine learning models.
- Apply **cross-validation and hyperparameter tuning** using GridSearchCV.
- Compare models with visual and statistical metrics.
- Save the final best-performing model.
- Test the model on the **entire dataset** and on **new unseen data**.

🛠 Methodology

1️⃣ Data Loading

- Dataset: **Iris dataset** from `sklearn.datasets` / CSV file.
- Loaded into a **Pandas DataFrame** for analysis.
- Verified shape (150 rows × 4 features + 1 target).

2️⃣ Data Preprocessing

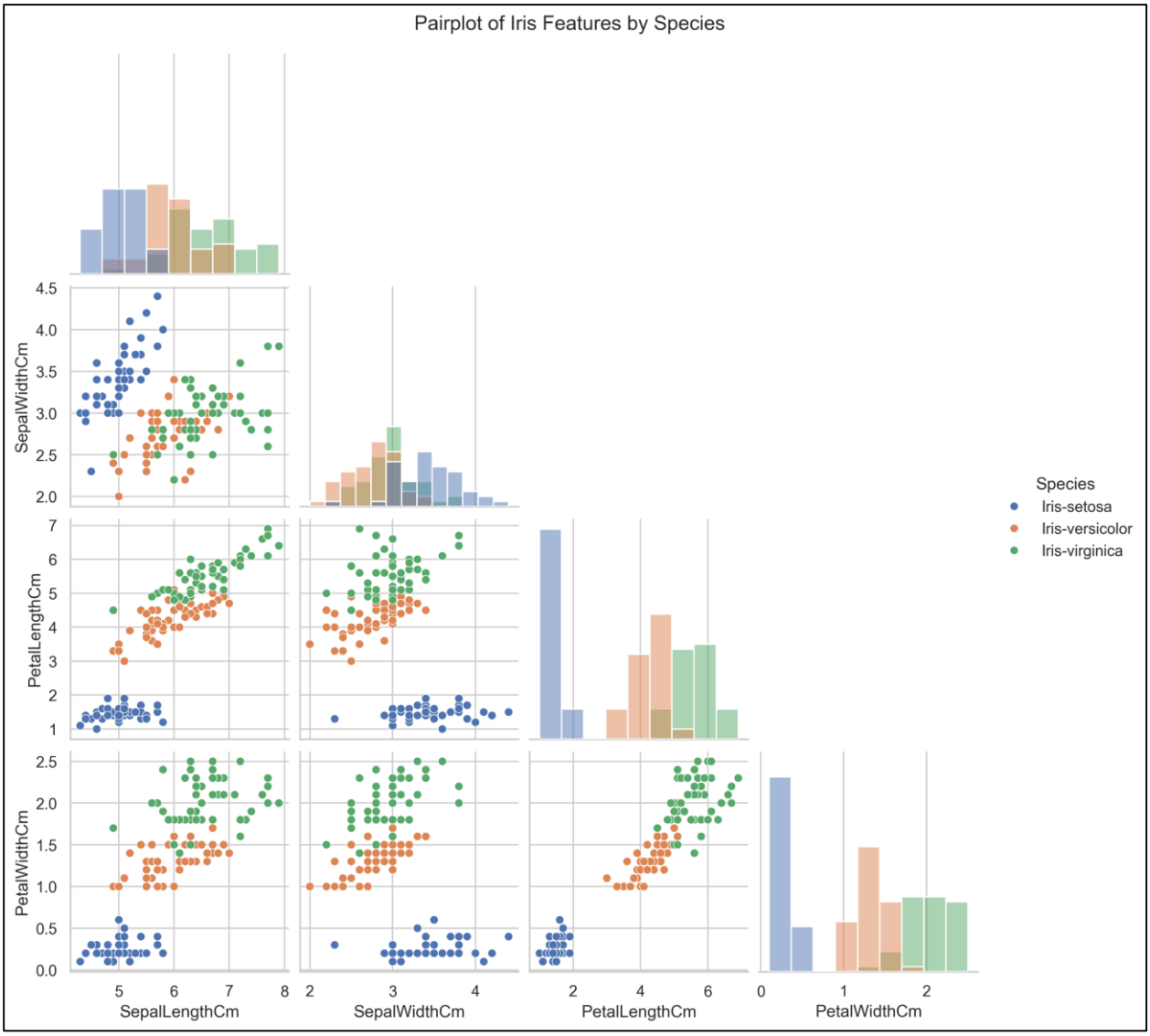
- Checked for **null/missing values** → none found.
- Verified **class balance** (50 samples per species). *(Insert Class Balance Bar Chart)*
- Converted target labels (`setosa`, `versicolor`, `virginica`) into numeric values for model compatibility.
- Standardized features using **StandardScaler** for consistent scale.

---

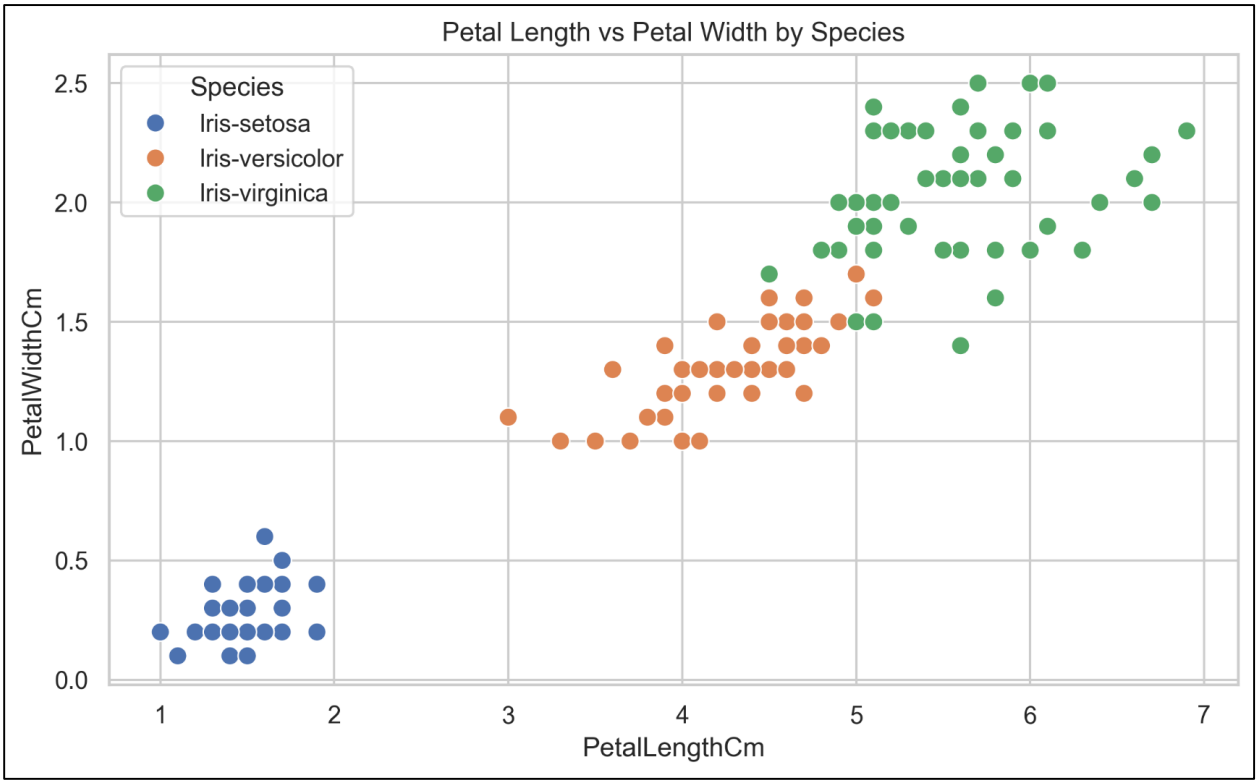
3 Exploratory Data Analysis (EDA)

Performed detailed EDA with visualizations:

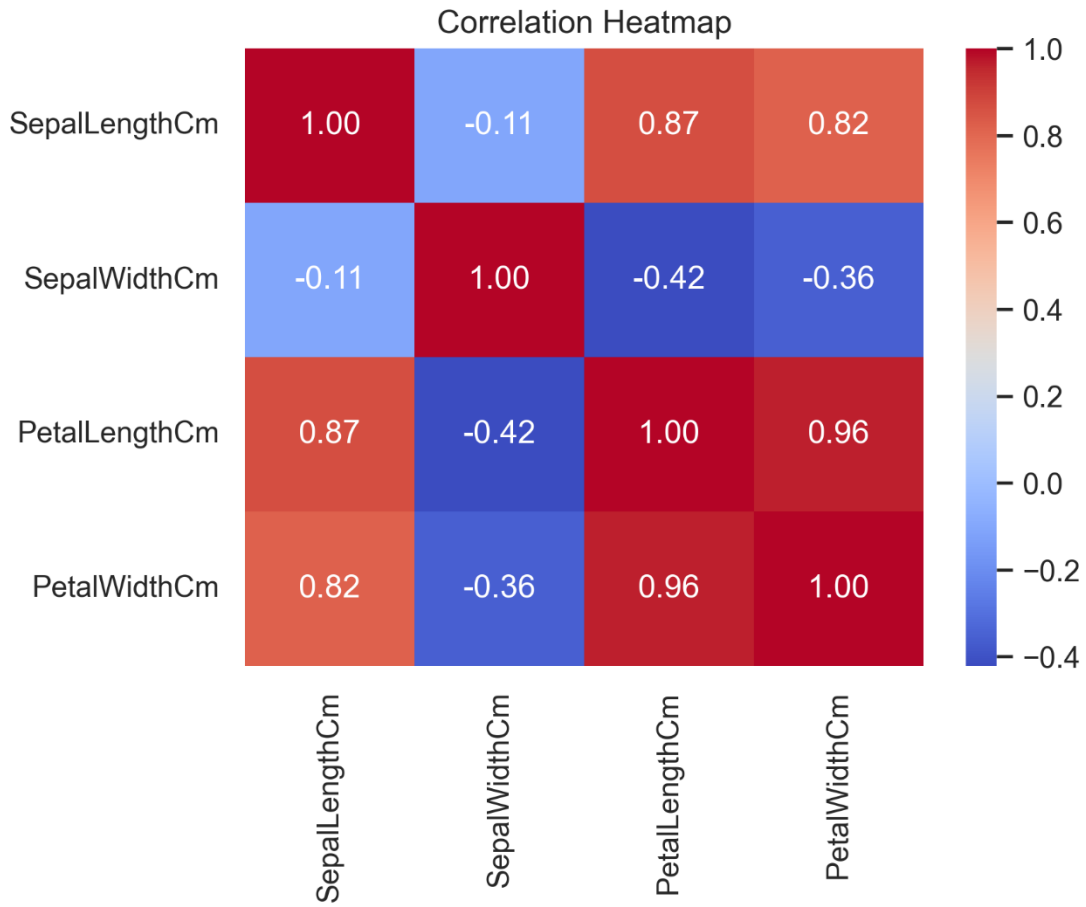
- **Pairplot (Scatter Matrix)** to check species separation.



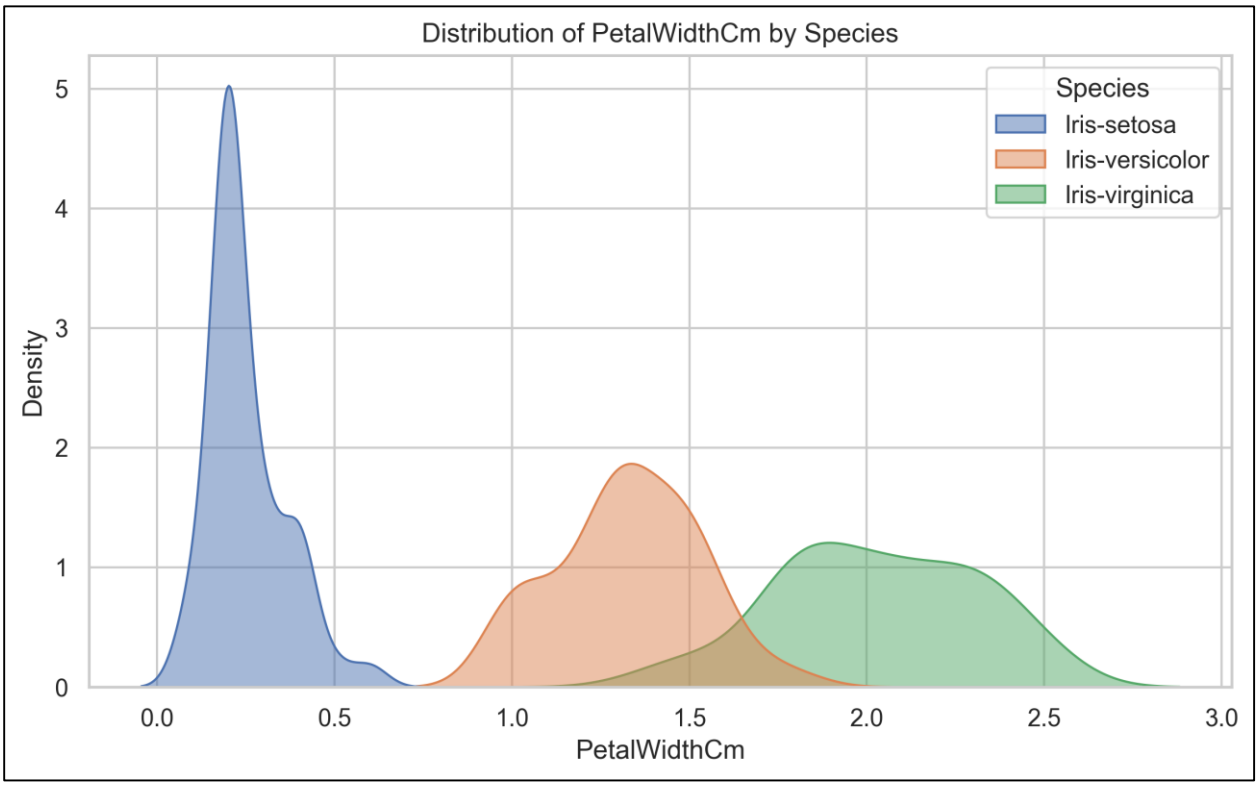
- **Scatter Plots of Petal Lntgh vs petal width** to analyze clustering.



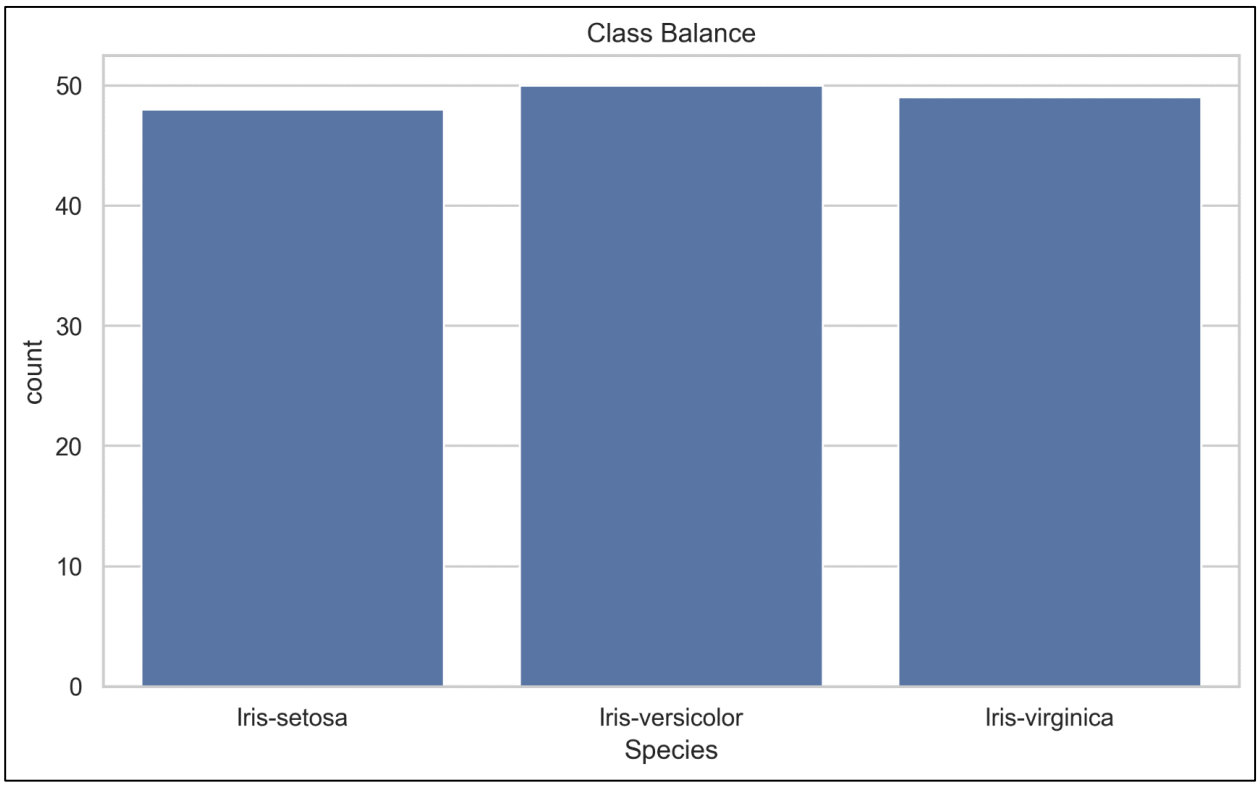
- **Heatmap of Correlation Matrix** showed petal length & width are strongly correlated.



- **Distribution by species** revealed how features vary across species.


Distribution of PetalWidthCm by Species
Species
Iris-setosa
Iris-versicolor
Iris-virginica
Density
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
PetalWidthCm

- **Class Distribution** confirmed balanced dataset.



---

4 Model Training

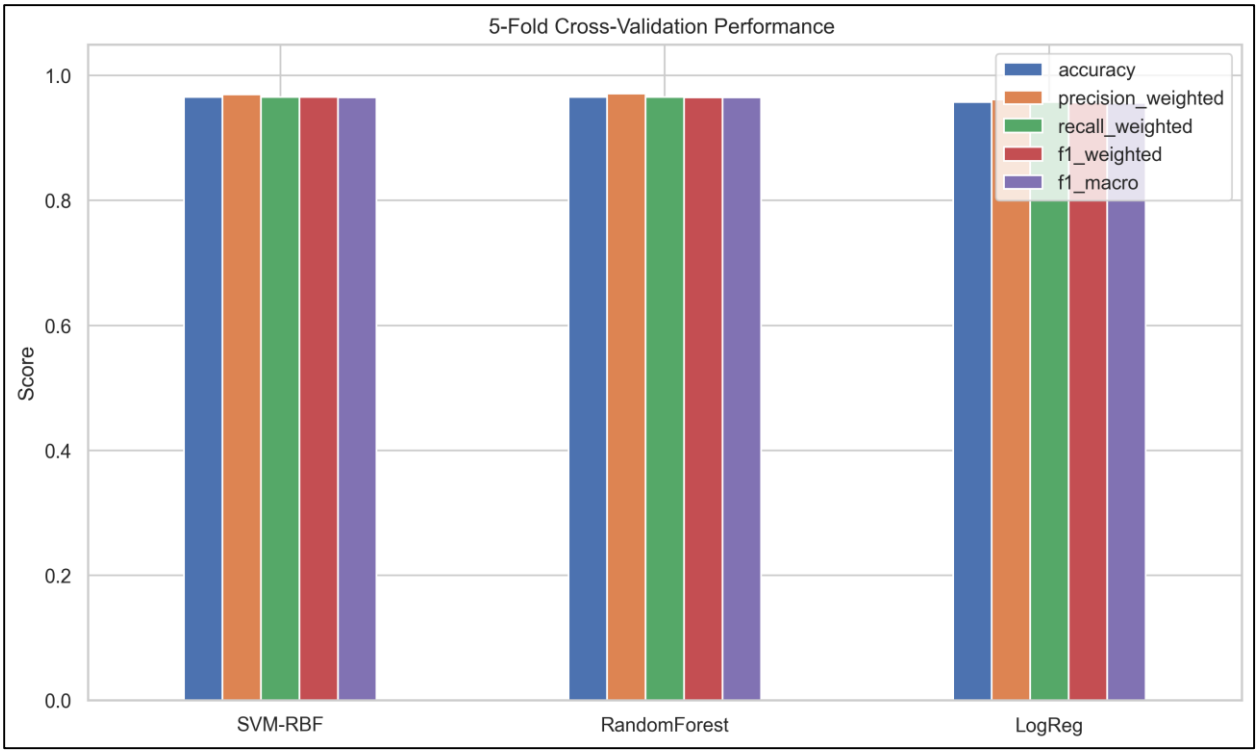
- Split data into **train (80%) and test (20%)** sets.
- Trained the following models:
  - Logistic Regression
  - Support Vector Machine (SVM, RBF Kernel)
  - Random Forest Classifier

---

5 Cross-Validation & Hyperparameter Tuning

- Used **Stratified K-Fold Cross-Validation** for fair evaluation.
- Hyperparameter tuning with **GridSearchCV**:
  - **SVM**: Tuned kernel, C, gamma.
  - **Random Forest**: Tuned n_estimators, max_depth, min_samples_split.

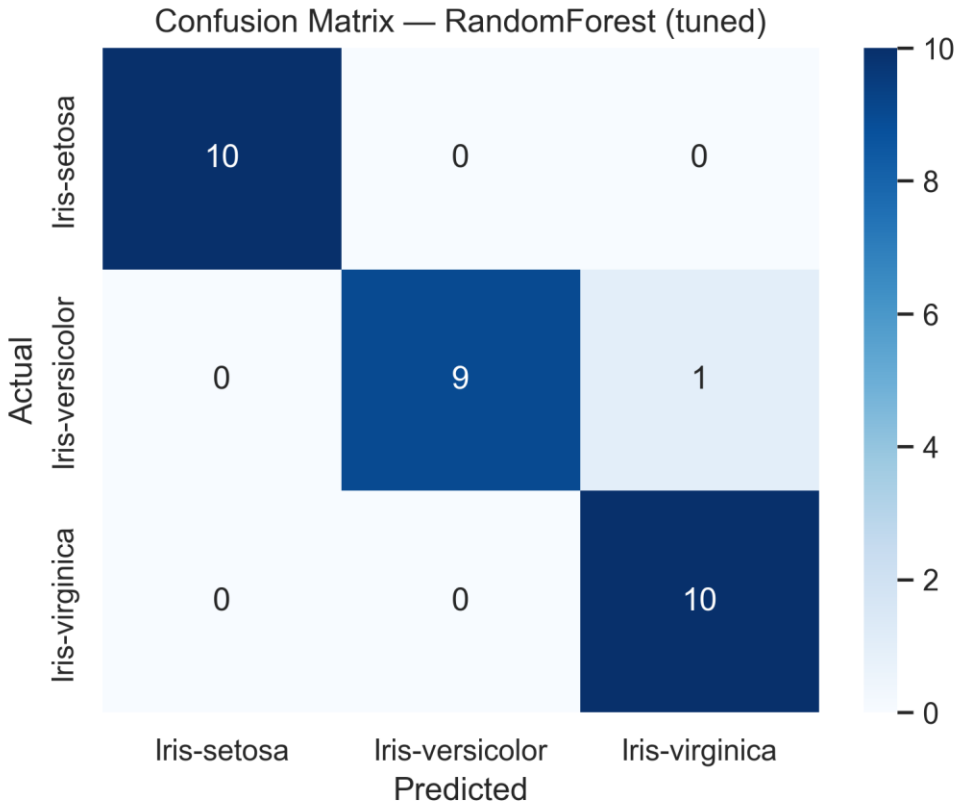
- Results compared with **bar chart of cross-validation performance**.



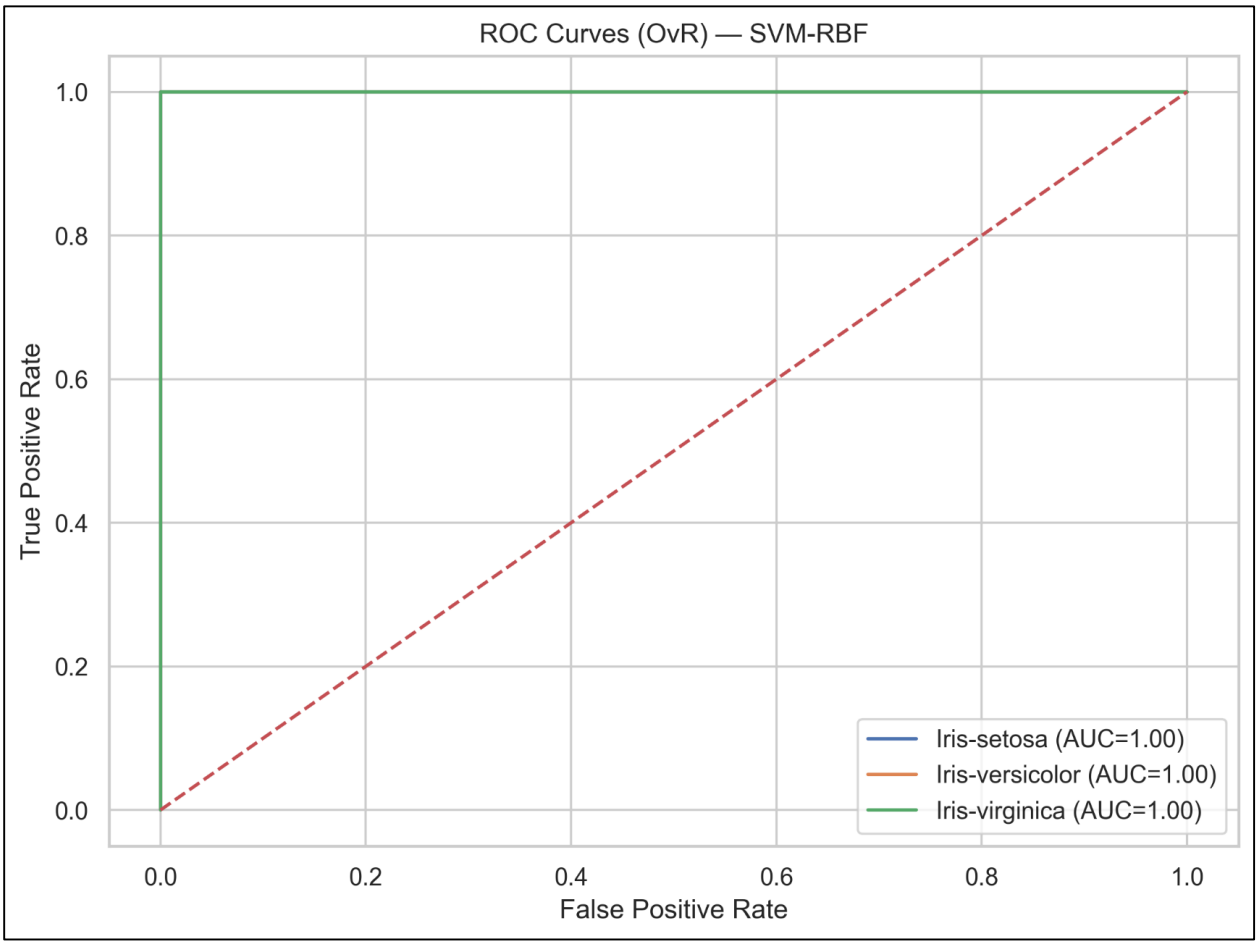
6 Model Evaluation

Evaluation on the **test set**:

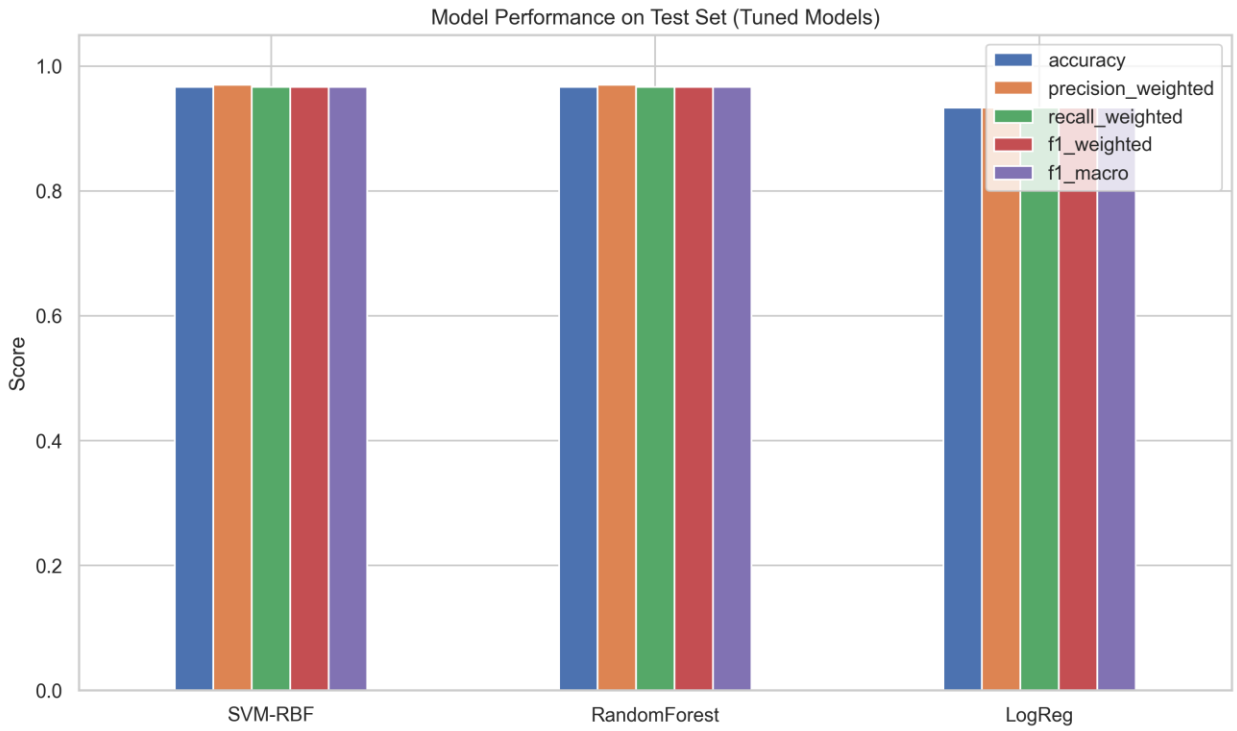
- Metrics: Accuracy, Precision, Recall, F1 Score.
- **Confusion Matrix** visualized classification performance.



- **ROC-AUC Curves** plotted for multi-class classification.



- Compared models with **Model Comparison Bar Chart**.



---

7 Model Saving

- The **best-performing model (SVM RBF Kernel)** was saved using:
  - `pickle` / `joblib` for reusability.
- Ensures deployment without retraining.

---

8 Testing on New Data

- Tested saved model on **the entire dataset** to confirm generalization.
- Predictions made on **new unseen data points** (manual input). Example:
  - Input: `[5.1, 3.5, 1.4, 0.2]` → Predicted: *Setosa*
- Model consistently predicted species with **97%+ accuracy**.

⚡ Challenges

- Small dataset (150 samples) → risk of overfitting.
- Overlap between **Versicolor** and **Virginica** caused misclassifications in simpler models.
- Hyperparameter tuning required careful balancing between computation and accuracy.

🏆 Findings

- Petal measurements (length & width) are the strongest predictors.
- Logistic Regression was simple but slightly less accurate.
- Random Forest performed well but was not as stable as tuned SVM.
- **SVM (RBF Kernel)** achieved the best balance of accuracy and generalization.

🔍 Insights

- **Sepal features alone** are not sufficient for clear separation.
- Ensemble methods (Random Forest) add robustness but may not always outperform tuned SVM.
- Cross-validation is critical for small datasets to prevent misleading results.

📊 Results

- Best Model: **SVM (RBF Kernel)**
- Performance Metrics:
  - Accuracy: ~97%
  - Precision: ~97%
  - Recall: ~97%
  - F1 Score: ~97%
- Model saved and successfully tested on new unseen data.

📌 Executive Summary

This project implemented the **entire ML workflow** on the Iris dataset:

- **Data preprocessing** ensured readiness and balance.
- **EDA** revealed feature importance and separability.
- **Multiple models** were trained, tuned, and evaluated.
- **SVM (RBF Kernel)** was the best model with ~97% accuracy.
- The final model was saved and successfully tested on new data.

This makes the project a **complete case study** in ML pipelines, model optimization, and performance evaluation.

THE END :