# **IRIS CLASSIFICATION USING RANDOM FOREST CLASSIFIER**

## **1. Introduction**

This report presents an analysis of the Iris dataset using a **Random Forest Classifier**, a powerful ensemble learning method that aggregates multiple decision trees to enhance prediction accuracy and prevent overfitting. The model was trained and evaluated using a stratified subset of the dataset to ensure balanced representation of all three classes: **Setosa, Versicolor, and Virginica**.

## **2. Dataset Overview**

The **Iris dataset** is a well-known dataset in machine learning, consisting of **150 instances** with **four numerical features**:

* **Sepal length (cm)**
* **Sepal width (cm)**
* **Petal length (cm)**
* **Petal width (cm)**

Each instance is labeled into one of three species:

1. **Setosa**
2. **Versicolor**
3. **Virginica**

To prevent overfitting and improve generalization, a **smaller, stratified subset** of 20 samples per class was taken, leading to a dataset of **60 instances**.

## **3. Model Selection & Training**

### **3.1 Random Forest Classifier**

Random Forest is an ensemble learning method that builds multiple decision trees and merges them for better accuracy and stability.

#### **Advantages of Random Forest:**

✔ **Reduces Overfitting**: By averaging multiple trees, it prevents overfitting compared to a single decision tree.  
 ✔ **Handles Missing Values**: Works well with missing or noisy data.  
 ✔ **Feature Importance**: Can identify the most relevant features.

### **3.2 Hyperparameters Used**

The model was configured with:

* **n\_estimators = 50** → Uses 50 decision trees.
* **max\_depth = 5** → Limits depth to prevent overfitting.
* **random\_state = 42** → Ensures reproducibility.
* **train-test split = 70%-30%** for training and testing.

## **4. Model Performance**

### **4.1 Cross-Validation Accuracy**

Cross-validation was performed with **5 folds** to assess model generalization:

* **Cross-Validation Accuracy ≈ 98.00%**

### **4.2 Test Accuracy**

The model was tested on **18 test samples**, achieving:

* **Test Accuracy ≈ 94.44%**

This accuracy is **below 100%**, indicating some generalization but still quite high, suggesting possible minor overfitting.

## **5. Confusion Matrix Analysis**

* **Setosa was classified perfectly (6/6 correct).**
* **Versicolor was classified perfectly (6/6 correct).**
* **1 Virginica sample was misclassified as Versicolor.**

The slight misclassification suggests **overfitting is not extreme**, but the model may still be biased towards some features.

## **6. Feature Importance Analysis**

The feature importance plot was generated, showing which attributes contributed most to classification.

### **Key Findings:**

* **Petal length and petal width were the most influential features.**
* **Sepal width contributed the least.**

These findings align with prior research on the Iris dataset, reinforcing that petal dimensions hold the highest discriminatory power between species.

## **7. Recommendations for Improvement**

Although the model performs well, improvements could be made:

✅ **Reduce Overfitting:**

* Use **fewer estimators (e.g., 30 trees)** to lower model complexity.
* Increase **max\_depth limitation (e.g., reduce to 3 levels)**.

✅ **Enhance Generalization:**

* Perform **more aggressive data augmentation** (e.g., adding noise).
* Consider **other classifiers (e.g., SVM, KNN)** for comparison.

✅ **Collect More Diverse Data:**

* If deployed in real-world settings, data from different environments should be included to ensure better generalization.

## **8. Conclusion**

This project successfully classified the **Iris dataset** using a **Random Forest Classifier** with an accuracy of **94.44%**. The results indicate strong performance but suggest mild overfitting. Further optimization, such as tuning hyperparameters and exploring additional classifiers, can enhance robustness.