# Random Forest – Study Notes

## 🌲 What is Random Forest?

Random Forest is an ensemble machine learning algorithm that builds multiple decision trees and combines their predictions for better accuracy and generalization.  
  
- Used for classification and regression  
- Part of the Bagging (Bootstrap Aggregating) family

## ⚙️ How It Works (Step-by-Step)

1. Bootstrapping (Sampling with Replacement):  
 - Create multiple random datasets (samples) from the original training data.  
 - Each sample may have repeated rows.  
  
2. Random Feature Selection at Splits:  
 - At each node of every tree, select a random subset of features to find the best split.  
 - Introduces diversity among trees.  
  
3. Train Multiple Decision Trees:  
 - Each tree is trained on its own bootstrapped data and random features.  
 - Trees are built independently.  
  
4. Aggregate Predictions:  
 - Classification: Majority vote  
 - Regression: Average of all tree outputs

## 📌 Key Characteristics

Property | Random Forest  
---------------------- | ----------------  
Type | Ensemble (Bagging)  
Learners | Multiple Decision Trees  
Combines Trees By | Voting (classification), Averaging (regression)  
Handles Overfitting? | Yes, better than single trees  
Handles Noise & Outliers? | Yes, robust  
Interpretability | Lower than single decision trees

## 🔍 Important Hyperparameters

Hyperparameter | What It Does  
---------------------|--------------  
n\_estimators | Number of trees  
max\_features | Number of features considered per split  
max\_depth | Maximum depth of a tree  
min\_samples\_split | Min samples to split a node  
bootstrap | Whether bootstrapping is used  
random\_state | For reproducibility

## ✅ Pros

- High accuracy on many datasets  
- Reduces overfitting compared to single decision trees  
- Handles missing values & outliers well  
- Feature importance insight  
- Works well on both small and large datasets

## ❌ Cons

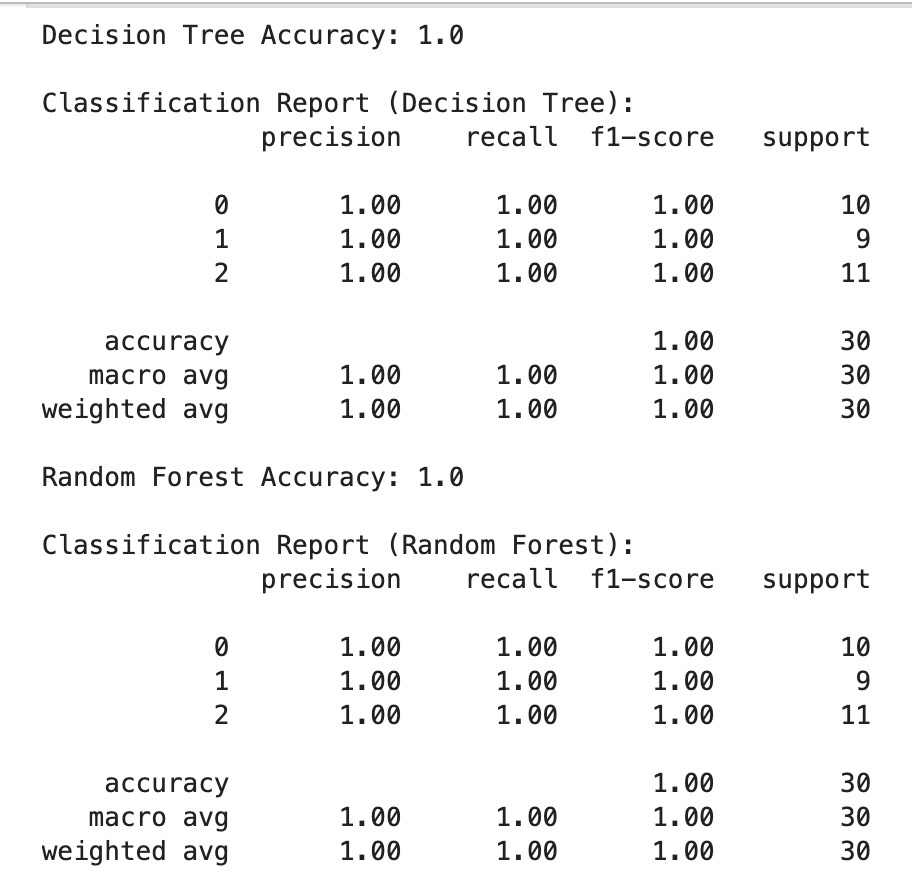
- Slower to train and predict (many trees)  
- Less interpretable ("black box" model)  
- May require more memory and computation  
- Can still overfit on very noisy data if not tuned

## 🧠 Good to Know

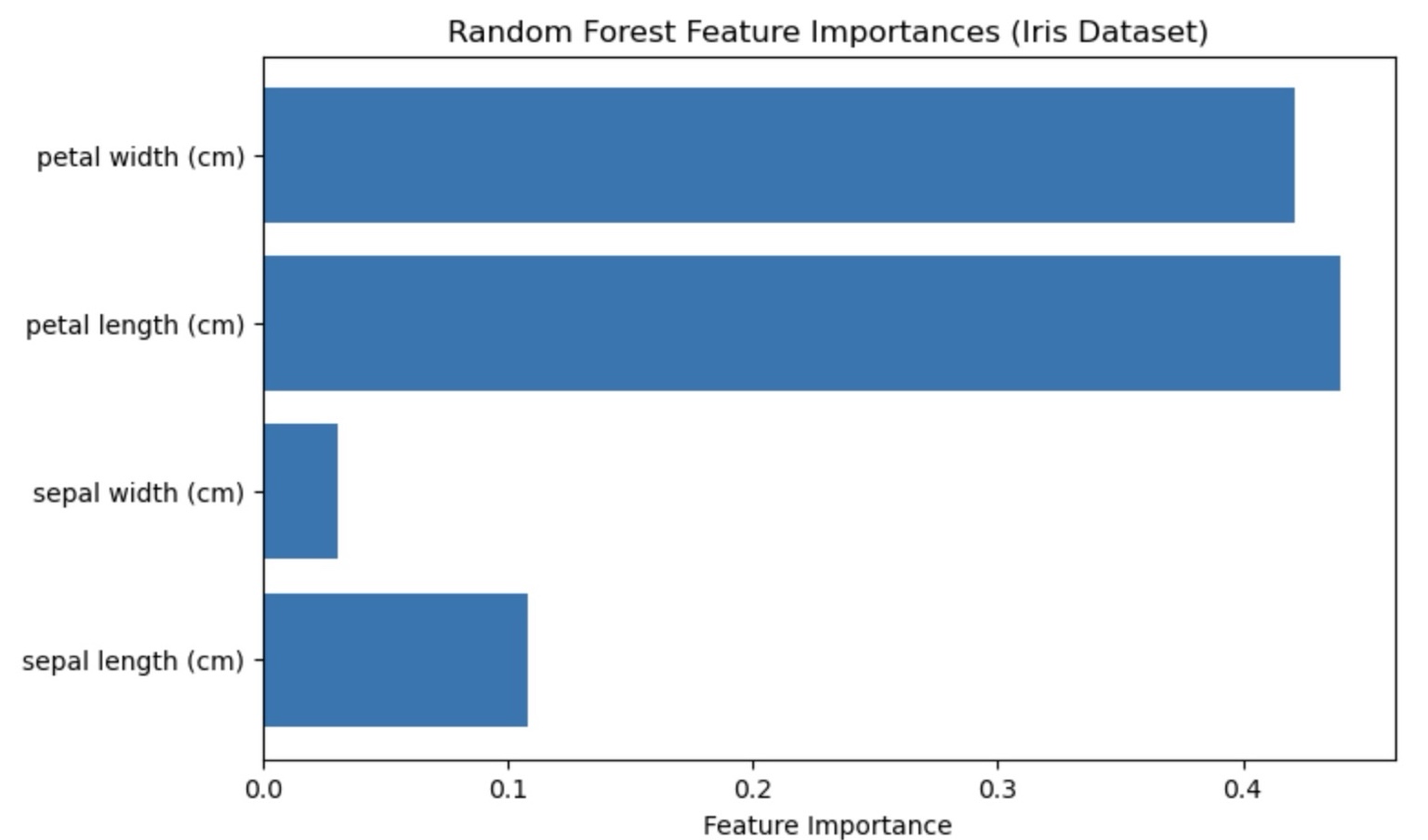
- Random Forest is less prone to high variance than a single decision tree.  
- Adding more trees usually improves stability but increases computational cost.  
- Feature Importance is one of its biggest interpretability advantages.

## 📊 Model Outputs & Interpretation

The classification reports for both Decision Tree and Random Forest models showed 100% accuracy. This means that all test samples were correctly predicted. However, since the Iris dataset is very clean and small, this perfect accuracy is common and might not generalize to more complex, noisy data.



📌 This image shows the precision, recall, and F1-scores for each class. All metrics are 1.00, indicating perfect classification.



📊 This bar chart displays feature importances as determined by the Random Forest model. Petal width and petal length are the most influential features in predicting the flower species. Sepal features contribute much less to the decision-making process.