Decision Trees and Random Forests: Theory Explained

1. Decision Trees

A **decision tree** is a supervised machine learning algorithm used for both **classification** and **regression**. It works by recursively splitting the dataset into subsets based on the most significant feature at each step.

Key Concepts:

- **Root Node:** The topmost decision node (starting point).
- **Internal Nodes:** Decision nodes that split the data.
- **Leaf Nodes:** Terminal nodes that give the final prediction.
- **Splitting Criteria:**
    - **Classification:** Gini Impurity or Entropy (Information Gain).
    - **Regression:** Mean Squared Error (MSE) or Variance Reduction.

How it Works:

1. **Select the Best Feature:** Choose the feature that best splits the data (maximizes information gain or minimizes impurity).
2. **Split the Data:** Divide the dataset into subsets based on the selected feature.
3. **Repeat:** Continue splitting until a stopping condition is met (max depth, minimum samples per leaf, etc.).
4. **Prediction:** New data points traverse the tree from root to leaf, where the majority class (classification) or average value (regression) is predicted.

Advantages:

- Easy to interpret.
- Handles both numerical and categorical data.
- No need for feature scaling.

Disadvantages:

- Prone to overfitting (high variance).
- Sensitive to small changes in data (unstable).

---

2. Random Forest

A **Random Forest** is an **ensemble learning** method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting.

Key Concepts:

- **Ensemble Method:** Uses **bagging (Bootstrap Aggregating)** to train multiple trees on different subsets of data.
- **Feature Randomness:** Each tree considers a random subset of features at each split (reduces correlation between trees).
- **Majority Voting (Classification) / Averaging (Regression):** Final prediction is based on the consensus of all trees.

How it Works:

1. **Bootstrap Sampling:** Randomly select subsets of data (with replacement) to train each tree.
2. **Random Feature Selection:** At each split, only a random subset of features is considered.
3. **Build Multiple Trees:** Each tree grows independently to maximum depth (no pruning).
4. **Aggregate Predictions:**
   - For **classification**, the majority vote wins.
   - For **regression**, the average of all tree predictions is taken.

Advantages:

- Reduces overfitting compared to a single decision tree.

- Handles high-dimensional data well.
- Robust to noise and outliers.
- Provides feature importance scores.

Disadvantages:

- Less interpretable than a single decision tree.
- Slower training and prediction time due to multiple trees.

Key Differences:

Feature	Decision Tree	Random Forest
Model Type	Single tree	Ensemble of trees
Overfitting	High risk	Reduced (due to averaging)
Stability	Sensitive to data changes	More stable
Interpretability	High	Lower (black-box)
Performance	Lower accuracy	Higher accuracy

When to Use?

- **Decision Tree:** When interpretability is crucial and dataset is small.
- **Random Forest:** When higher accuracy is needed and computational cost is acceptable.