🌳🌳 Random Forest – Introduction & Intuition

1. What is Random Forest?

- A **Random Forest** is an **ensemble of decision trees**.
- Instead of relying on a single decision tree (which might overfit or be unstable), Random Forest builds **many trees** and combines their predictions.
- Think of it as **wisdom of the crowd** → multiple weak opinions combined → stronger and more reliable prediction.

2. Why "Random"?

Two types of randomness make Random Forest powerful:

1. **Random Data (Bootstrap Sampling)**
    1. Each tree gets trained on a random sample of the training data (with replacement).
    2. This ensures trees don't all see the same data → they learn different aspects.
2. **Random Features (Feature Subset Selection)**
    1. At each split in a tree, the model **does not look at all features**, only at a **random subset**.
    2. This makes trees more diverse and less correlated.

3. Intuition / Analogy

- Imagine you ask **100 doctors** for a diagnosis.
- Each doctor sees different **patients** (random data) and checks only some **symptoms** (random features).
- Final decision = majority vote (classification) or average opinion (regression).
- Result: **More accurate and robust** than asking just 1 doctor. ✅

4. How Random Forest Works (Step by Step)

1. Draw multiple bootstrap samples from training data.
2. Train a decision tree on each sample.
    - ❖ But at each split, use only a **random subset of features**.
3. For prediction:
    - ❖ **Classification** → majority vote across all trees.
    - ❖ **Regression** → average the outputs of all trees.

5. Advantages of Random Forest

1. Handles both classification & regression.
2. Reduces overfitting (compared to single tree
3. Works well with high-dimensional data.
4. Handles missing values & outliers well.

6. Disadvantages

❌ Less interpretable than a single tree (hard to visualize 100+ trees).

❌ Can be slower to train and predict if the forest is very large.

❌ May still overfit if trees are too deep and dataset is small.

🎯 Key Takeaway

- Random Forest = **Bagging + Random Features**.
- It's like having a **crowd of decision trees**, each slightly different, and combining them for stable, powerful predictions.