Random forest

What is Random Forest?

Random Forest is an **ensemble learning** algorithm that combines **multiple decision trees** to improve prediction accuracy and control overfitting. It's used for both **classification** and **regression** tasks.

Why is it called “Random Forest”?

- **Forest** = A group of decision trees.
- **Random** = It uses random subsets of data and features to grow each tree.

How Random Forest Works (Step-by-step):

1. Bootstrap Sampling (Bagging):

- Randomly select **samples** (rows) **with replacement** from the dataset to create different training subsets for each tree.
- This is called **bootstrapping**.

2. Grow Many Decision Trees:

- Each tree is trained on a different **bootstrapped dataset**.
- At every split in a tree, it selects the **best feature from a random subset** of features (not all features). This adds **feature randomness**.

3. Voting (for Classification):

- Once all trees are built, a new instance is classified by **majority vote** — i.e., most trees say "Class A", so the forest says "Class A".

4. Averaging (for Regression):

- For regression, it returns the **average** prediction of all the trees.

Why Random Forest is Powerful:

Advantage	Description
✅ **High Accuracy**	By averaging multiple trees, errors and overfitting are reduced.
✅ **Robust to Outliers**	Less sensitive than a single tree.
✅ **Feature Importance**	Can rank features by importance, useful for feature selection.
✅ **Non-linear Capable**	Works well with complex and non-linear relationships in data.

Some Limitations:

Limitation	Explanation
🦕 **Slower**	Training and predicting can be slower due to many trees.
🖤 **Less Interpretability**	Harder to interpret than a single decision tree.
🧠 **Memory Intensive**	Needs more memory due to many trees being stored.