📆 DAY 22 — Random Forests

1️⃣ Why Decision Trees Alone are Not Enough

Decision Trees:

- Overfit easily ❌
- Change a lot if data changes ❌
- Hight variance ❌

👉 Random Forests fix this.

2️⃣ What is a Random Forest?

> **A Random Forest is a collection of many decision trees whose predictions are combined.**

Think:

- One tree = one opinion
- Forest = wisdom of the crowd

3️⃣ How Random Forests Work?

Each Tree:

1. Trains on a **random subset of data** (bootstrapping)
2. Uses a **random subset of features** at each split
3. Grow independently

Final prediction:

- Classification → majority vote
- Regression → average prediction

4️⃣ Why Randomness is GOOD

Randomness:

- Reduces correlation between trees

- Prevents overfitting
- Improves generalization

Many weak tree → Strong model

5️⃣ Bias - Variance Tradeoff

Model	**Bias**	**Variance**
Single Tree	Low	High ❌
Random Forest	Slightly higher	Much lower ✅

Random Forests **Sacrifice a little bias to Kill variance.**

6️⃣ Important Hyperparameters

🔧 n_estimators

Number of Trees

- More trees → Better (until saturaion)
- Slower but safer

🔧 max_depth

Controls tree depth

- Smaller → less overfitting
- Larger → more expressive

🔧 max_features

How many features each split sees

- `sqrt` → default for classification
- `log2` → more randomness
- float → percentage of features

🔧 min_samples_leaf

Minimum samples per leaf

- Smooths predictions
- Reduces noise

7️⃣ Random Forests Do NOT Need Scaling

Why?

- Trees split by thresholds
- No distance calculations

✅ Works directly on raw features

8️⃣ Feature Importance

Random Forests:

- Rank features by usefulness
- Robust compared to single trees

Used for:

- Feature selection
- Model interpretation

9️⃣ When to Use Random Forests

✅ Tabular data

✅ Medium-sized datasets

✅ Strong baseline model

❌ Very large datasets

❌ Sparse/high-dimensional data