🌲 Random Forest Hyperparameters (Classification & Regression)

Random Forest has many hyperparameters, but only a few matter the most. They control **tree size, randomness, number of trees, and splits**.

---

1. Number of Trees → n_estimators

- **What it does:** Number of decision trees in the forest.
- Larger = more stable, less variance, but slower.
- **Trade-off:** Too few trees → high variance. Too many trees → more training time but not much accuracy gain after a point.
- Rule of Thumb: Start with **100–500**, increase if needed.

---

2. Tree Depth → max_depth

- Limits how deep each tree can grow.
- Deep trees → more complex, risk of **overfitting**.
- Shallow trees → risk of **underfitting**.
- Tune depending on dataset complexity.

---

3. Minimum Samples to Split → min_samples_split

- Minimum number of samples required to **split a node**.
- Small value → more splits → complex trees (risk overfitting).
- Large value → fewer splits → simpler trees (risk underfitting).

---

4. Minimum Samples per Leaf → min_samples_leaf

- Minimum samples required to be at a **leaf node** (end node).
- Helps avoid leaves with only 1–2 samples (which overfit).
- Larger datasets → keep higher values (e.g., 5, 10).

---

5. Number of Features → max_features

- Number of features considered at each split.
    1. **Classification**: default = $\sqrt{}$(n_features).
    2. **Regression**: default = all features.
- More features → less randomness, stronger trees but more correlation.

- Fewer features → more randomness, trees less correlated, improves ensemble diversity.

---

6. Bootstrap Sampling → bootstrap

- Whether to use bootstrap samples (sampling with replacement).
- Default = True → standard Random Forest.
- If False → trees trained on full dataset (less randomness).

---

7. Random State → random_state

- Ensures reproducibility. Same random state → same results.

---

8. Others (Less common but useful)

- max_samples: Fraction/number of samples drawn for each tree.
- class_weight: Helps with **imbalanced datasets** (important in classification).
- oob_score: Out-of-bag score → gives internal validation accuracy without test set.

⚖️ Bias–Variance Intuition

- **More trees (n_estimators)** → lower variance, stabler model.
- **Deeper trees (max_depth)** → lower bias but higher variance.
- **Higher min_samples_split or min_samples_leaf** → higher bias but lower variance.
- **Lower max_features** → more randomness, less variance, but slightly higher bias.

➢ Summary (Easy to Remember)

- **n_estimators**: How many trees?
- **max_depth**: How deep can trees grow?
- **min_samples_split / min_samples_leaf**: Prevent overfitting by controlling splits.
- **max_features**: Adds randomness, reduces correlation among trees.
- **bootstrap**: Whether to use bootstrapping.

👉 For **Classification & Regression**, the hyperparameters are the same — the only difference is the **default max_features setting** (√ for classification, all for regression).
