Random Forest

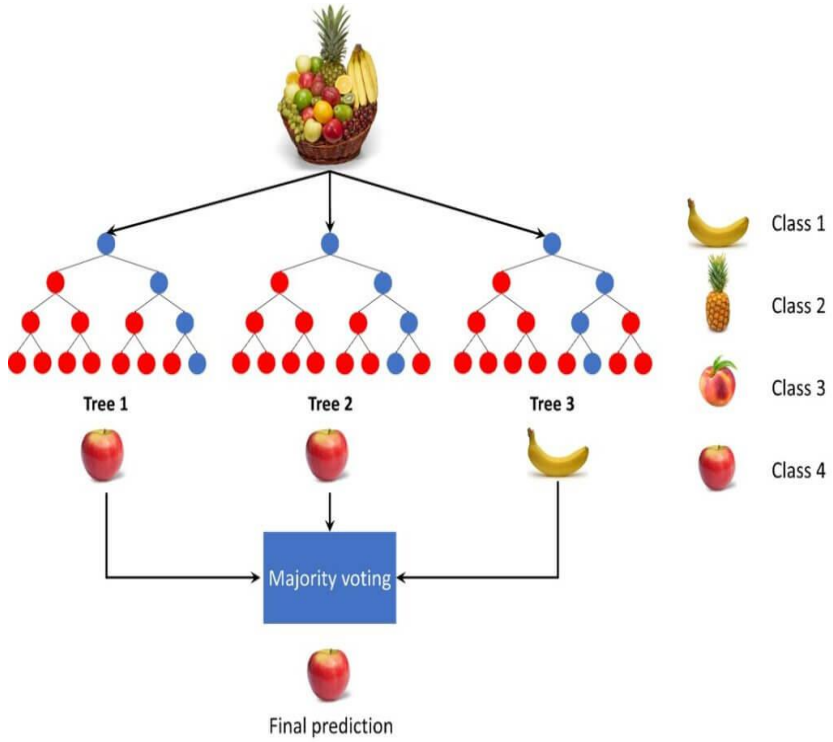
Presented by : Muhammad Zaqeem

Overview:

- **Random Forest**
- **How it Works**
- **Assumption of Random Forest**
- **Hyperparameters in Random Forest**
- **OOB (Out Of Bag)**

Random Forest:

- Random forest is developed by Leo Breiman and Adele Cutler in 2001
- It is a powerful machine learning algorithm that belongs to the **ensemble learning** family.
- Ensemble methods combine multiple models to improve prediction accuracy.
- In the case of Random Forest, it's a combination of **multiple decision trees.**
- it handles both classification and regression problems.



How it Works:

1: Create Multiple Bootstrap Samples: Random Forest starts by creating multiple **bootstrap samples**. These are random samples taken from the original dataset **with replacement**.

2: Build Decision Trees on Each Sample For each bootstrap sample, we build a **decision tree** with a few key differences from a regular decision tree:

- **Random Feature Selection**: At each split in the tree, only a random subset of the features is considered, not all features.

3: Make Predictions with Each Tree: Once all the trees are built, they can be used to make predictions. For a new data point, each tree makes its own prediction.

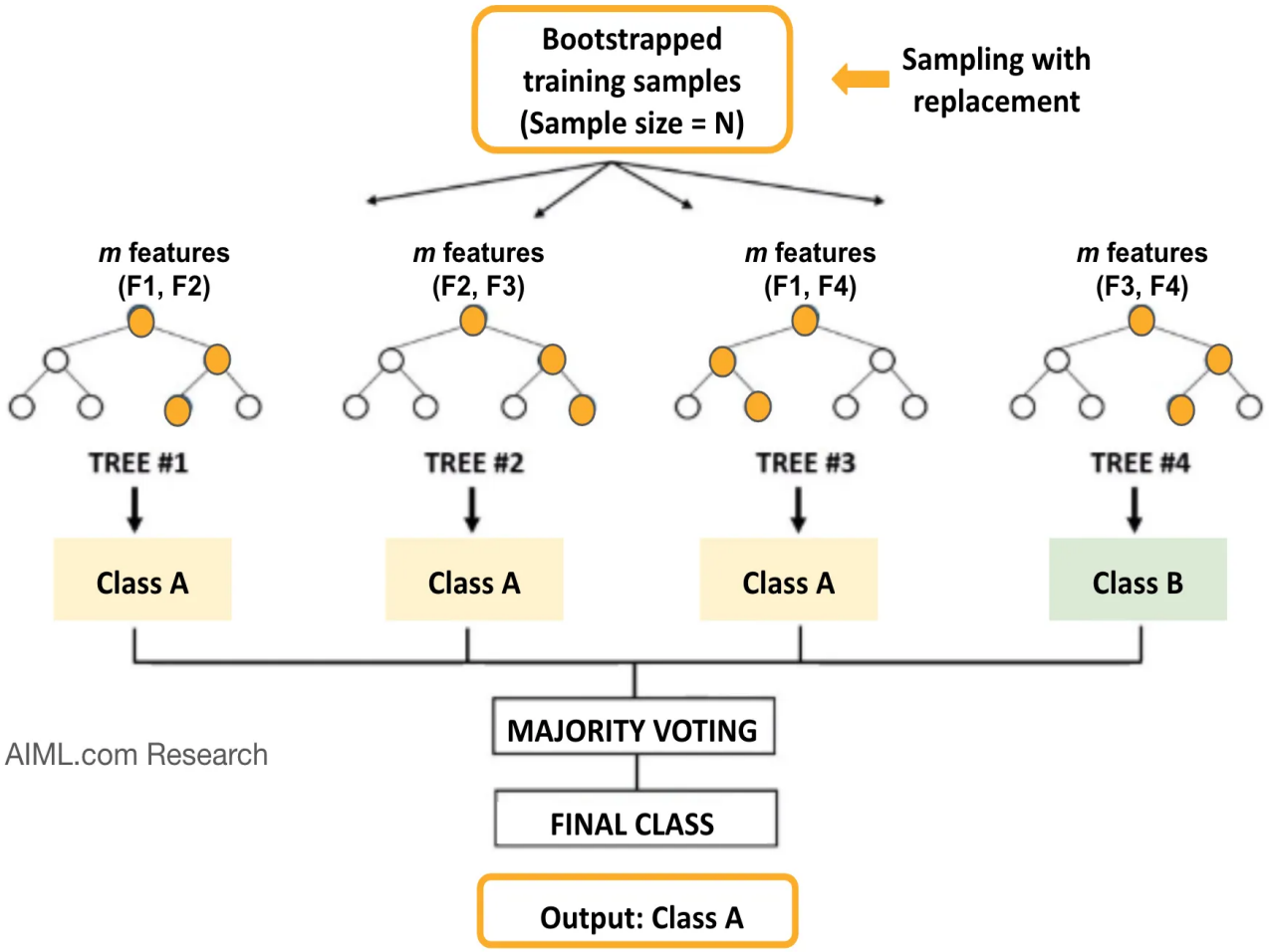
4: Using Majority Voting (for Classification)

- To make a final prediction, Random Forest uses a **majority voting** method
- For **regression problems**, instead of voting, Random Forest averages the predictions from all trees.

Random Forest Classifier

Training Data
(Sample size, N= 6, No. of features, F = 4)

F1	F2	F3	F4	Y
2.1	0	400	-9	A
3.0	1	890	-42	B
2.2	1	929	0	B
4.0	0	324	-23	A
3.5	1	333	-15	A
6.0	0	215	-9	A



Key parameters of Random Forest Model are: (a) Number of trees , (b) Maximum depth of the trees (c) Size of the random subset of features
In this example, No. of trees = 4, Depth = 2, and Feature subset size, m = 2 (no. of features/2)

Assumptions of Random Forest :

- To effectively use Random Forest, it is important to understand the underlying assumptions of the algorithm:

Independence of Trees: The decision trees in the forest should be independent of each other. This is achieved through bootstrap sampling and feature randomness.

Sufficient Data: Random Forest requires a large amount of data to build diverse trees and achieve optimal performance.

Balanced Trees: The algorithm assumes that the individual trees are grown sufficiently deep to capture the underlying patterns in the data.

Data Handling: Random Forest can handle noisy data, but it assumes that the noise is randomly distributed and not systematic

Hyperparameters in Random Forest:

- Hyperparameters are used in random forests to either enhance the performance and predictive power of models or to make the model faster.

n_estimators: Number of trees the algorithm builds before averaging the predictions.

max_features: Maximum number of features random forest considers splitting a node.

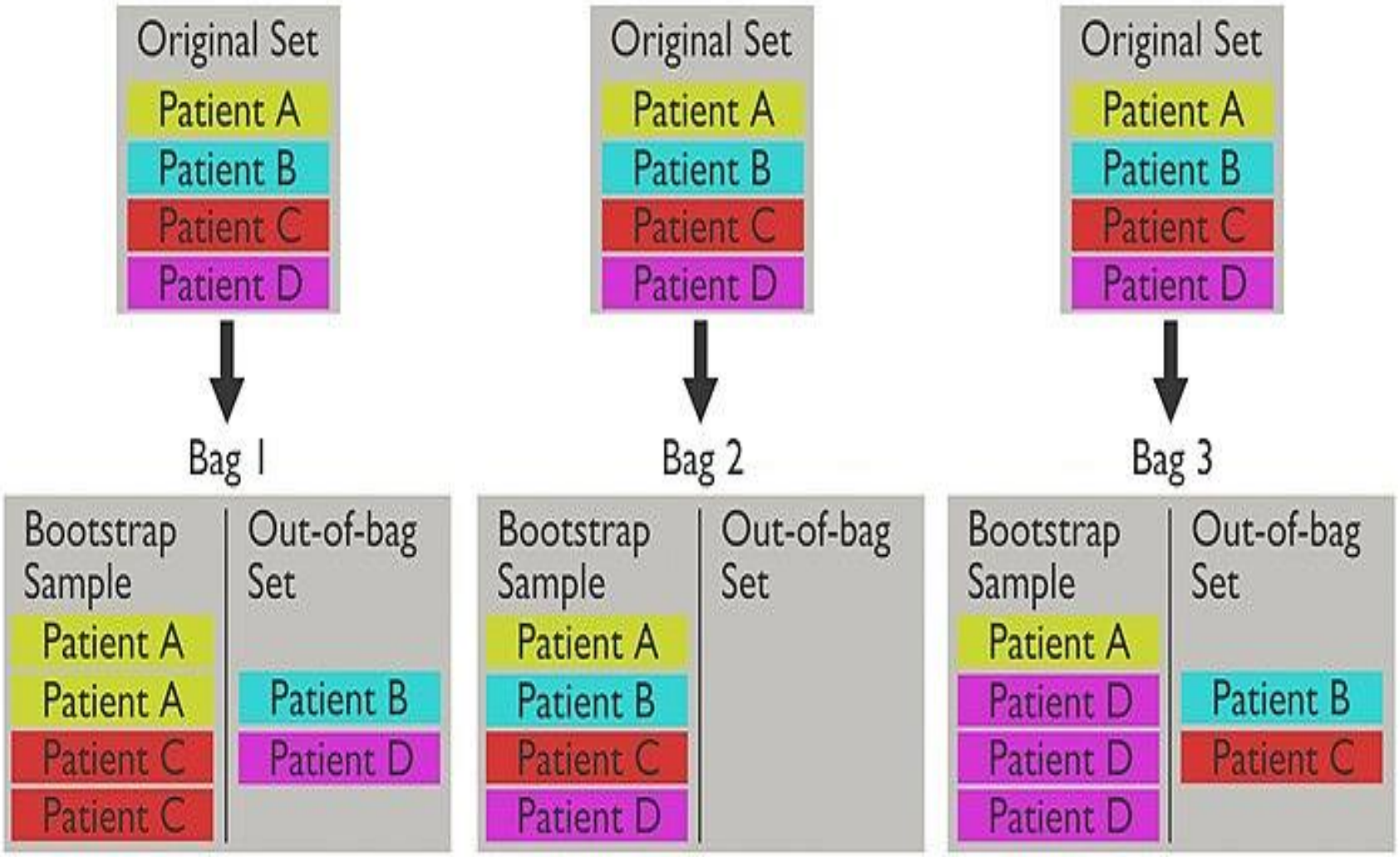
min_samples_leaf: The minimum number of samples that a leaf node (end node) can have.

max_depth: Limits the maximum depth (levels) of each tree. Deeper trees capture more details, but may lead to overfitting.

min_samples_split: This is the minimum number of samples required to split an internal node. Higher values prevent the model from learning too specific patterns, helping avoid overfitting.

OOB (out of bag):

- When a Random Forest builds each tree, it uses a method called "bootstrapping." This means it creates each tree using a random sample of the data
- Not all data points are used in every tree.
- Usually, about **63% of the data points are sampled**, and **37% are left out** for that particular tree.
- The data points **not used in a tree** are called **Out-of-Bag (OOB) samples** for that tree.
- Once all the trees are built, the Random Forest can use these OOB samples to test how well the trees are performing.
- For each data point, the Random Forest looks at all the trees that didn't use that data point (those for which it's an OOB sample) and predicts the target label using only those trees


Original Set
Patient A
Patient B
Patient C
Patient D
Bag 1
Bootstrap Sample
Patient A
Patient A
Patient C
Patient C
Out-of-bag Set
Patient B
Patient D
Original Set
Patient A
Patient B
Patient C
Patient D
Bag 2
Bootstrap Sample
Patient A
Patient B
Patient C
Patient D
Out-of-bag Set
Original Set
Patient A
Patient B
Patient C
Patient D
Bag 3
Bootstrap Sample
Patient A
Patient D
Patient D
Patient D
Out-of-bag Set
Patient B
Patient C

Thank You