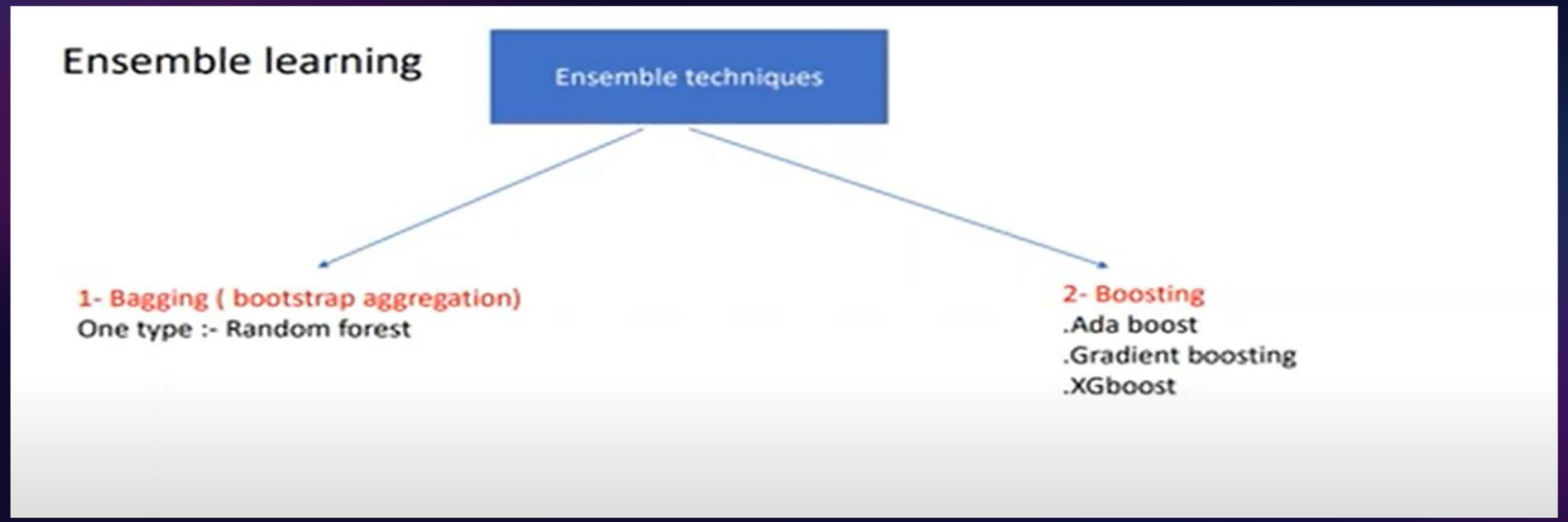
Random Forest Algorithm

- Mohamed Mahmoud Ali
- Ali Mohamed Hussein
- Anas Emad AbdElmoaty
- Hazem Sayed Abdallah

- DR. Cherry Ahmed

A Powerful Machine Learning Ensemble Method

The Random Forest algorithm is an ensemble learning method that combines multiple decision trees to make predictions. Each tree in the forest is built independently, using a random subset of the training data and a random subset of the features. This randomness helps to reduce overfitting and improve the generalization of the model.



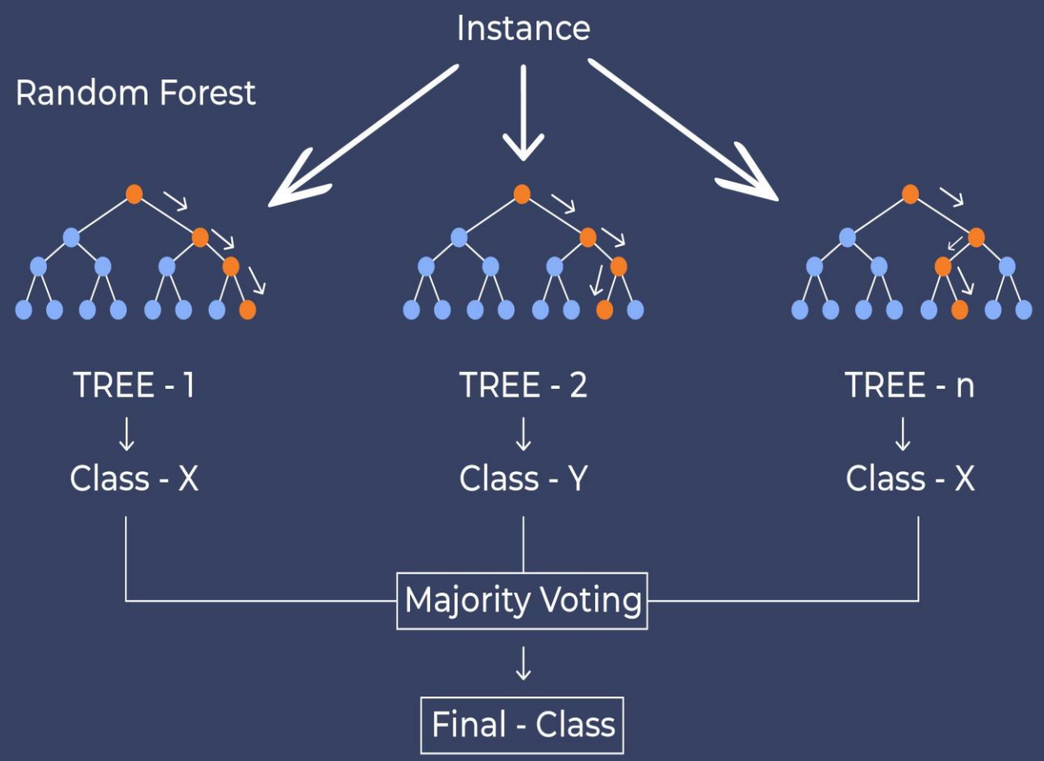
Bootstrap Aggregation (Bagging)

Randomly sampling data (with replacement) to create diverse decision trees, each trained on a unique subset of the training data

Learnbay

RANDOM FOREST

CLASSIFICATION



learnbay.co

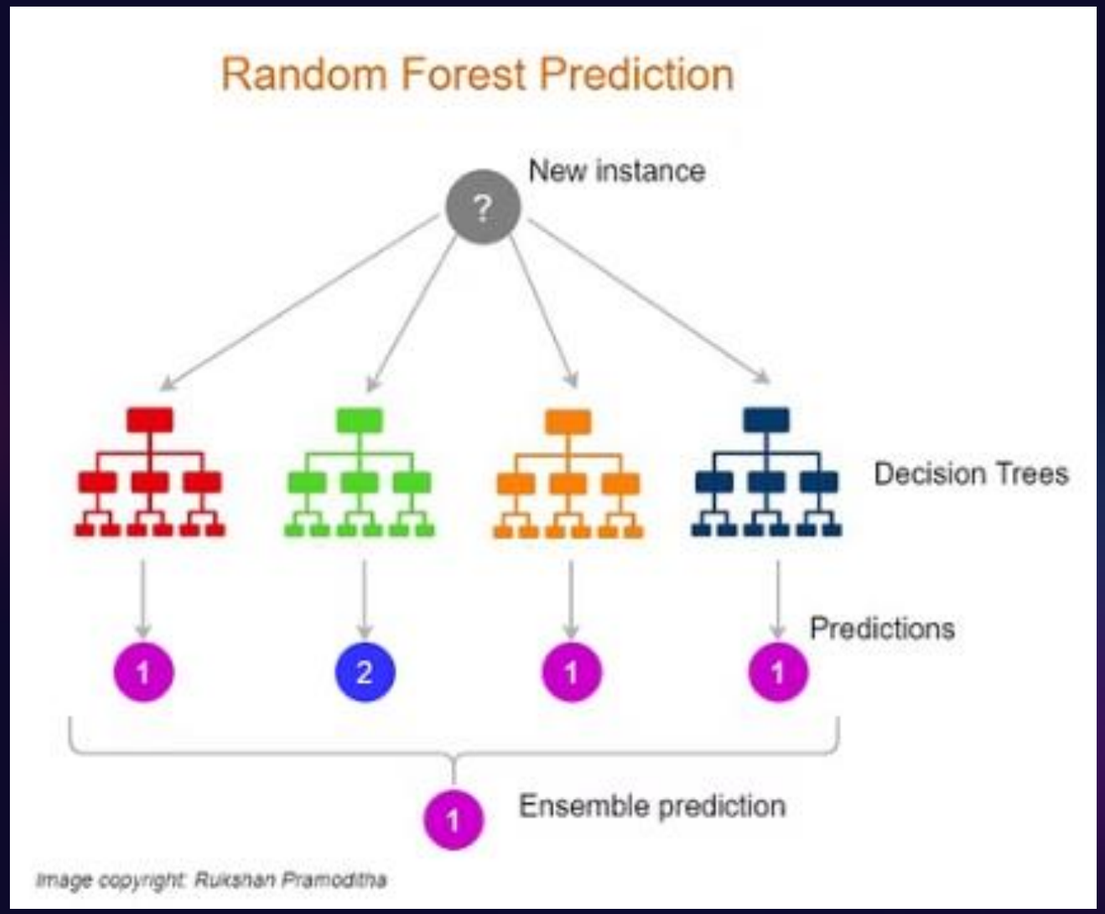
What is Random forest

- **Random forest is a supervised learning algorithm.**
- **The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method.**
- **The general idea of the bagging method is that a combination of learning models increases the overall result**

How it works

Put simply: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems.



Disadvantages of Random Forest Algorithm

Lack of Interpretability

Random Forest models are often considered as "black-box" models, making it challenging to interpret how individual decisions are made.

Resource Intensive

Training a Random Forest model can be computationally expensive, especially with a large number of trees and features, due to the need for bootstrapping and feature selection.

Overfitting

Although Random Forest mitigates overfitting compared to individual decision trees, it can still occur, especially when using a large number of trees or if the dataset is noisy.

Less Effective with Imbalanced Data

Random Forest may struggle to effectively handle imbalanced datasets, where one class significantly outnumbers the other, leading to biased predictions.

Differences between Random Forest and Decision Trees

Basic Structure

Decision Trees: Constructs a tree-like structure where each internal node represents a feature, each branch represents a decision based on that feature, and each leaf node represents the outcome.

Random Forest: Ensemble method that consists of multiple decision trees, where each tree is trained independently on a random subset of the data and features.

Bias-Variance Tradeoff

Decision Trees: Prone to high variance and overfitting due to capturing noise in the data.

Random Forest: Helps mitigate overfitting by averaging predictions from multiple trees, reducing variance at the expense of introducing some bias.

Differences between Random Forest and Decision Trees cont.

Interpretability

Decision Trees: Generally more interpretable as the decision-making process can be visualized in a tree structure.

Random Forest: Less interpretable due to the combination of multiple decision trees and the averaging of predictions.

Training Speed

Decision Trees: Faster to train compared to Random Forest, as it involves constructing a single tree.

Random Forest: Slower to train due to the training of multiple decision trees and the need for bootstrapping and feature selection.

Handling Missing Values

Decision Trees: Can handle missing values by selecting an appropriate split during tree construction.

Random Forest: Can handle missing values as well, but the method used may vary depending on the implementation and parameter settings.

Why Should We Use Random Forest ?

- Random forest algorithm is suitable for both classifications and regression task.
- It gives a higher accuracy through cross validation.
- Random forest classifier can handle the missing values and maintain the accuracy of a large proportion of data.
- If there are more trees, it doesn’t allow over-fitting trees in the model.
- It has the ability to work upon a large data set with higher dimensionality.

Random Forest Applications

1 Banking Industry

- **Credit Card Fraud Detection**
- **Customer Segmentation**

2 Healthcare and Medicine

- **Cardiovascular Disease Prediction**
- **Diabetes Prediction**
- **Breast Cancer Prediction**

3 Stock Market

- **Stock Market Prediction**

4 E-Commerce

- **Product Recommendation**
- **Price Optimization**

Application of Boosted Random Forest in Sepsis Prediction

(Applying Machine Learning to Blood Count Data Predicts Sepsis with ICU Admission 2022)

Study Overview:

Aim: To predict sepsis with ICU admission using Complete Blood Count (CBC) diagnostics.

Data Source: Non-ICU patient data from a German tertiary care center (January 2014 to December 2021).

Model Description:

Algorithm: Boosted Random Forest.

Features: Patient demographics and CBC parameters (e.g., hemoglobin, platelets, mean corpuscular volume, white and red blood cells).

Findings:

The model demonstrated significant predictive performance, indicating its potential in early sepsis prediction in non-ICU settings.

External validations highlighted the robustness across different datasets.

Conclusion:

Highlighting the potential of leveraging routine CBC results with Machine Learning models for clinical decision support systems to enhance patient safety.

THANKS