RandomForest

Classification and Regression

- Introduction
- How it work?
- Demo with sklearn & h2o

Introduction

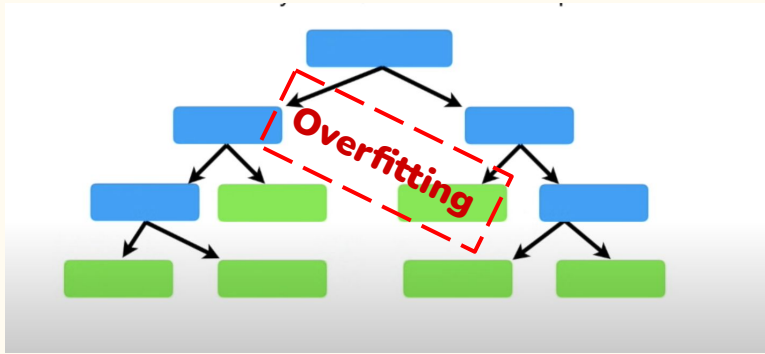
Random Forest is a supervised learning algorithm.

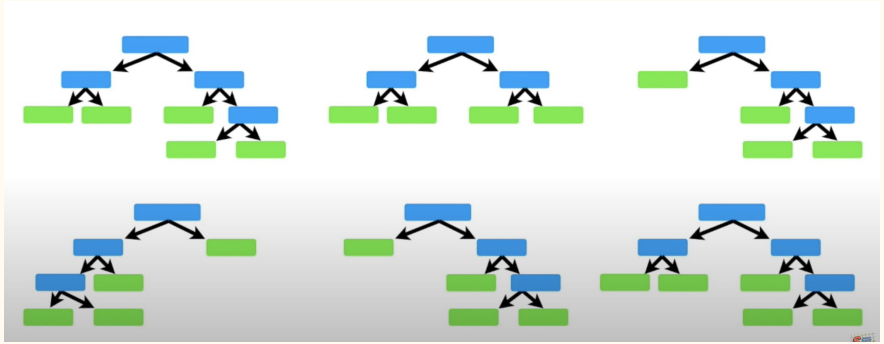
Forest - an ensemble of decision trees, usually trained with the “bagging” method.

Random: Bootstrap dataset, subset of features → wide diversity

General idea: The general idea of the bagging method is that a combination of learning models increases the overall result.

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



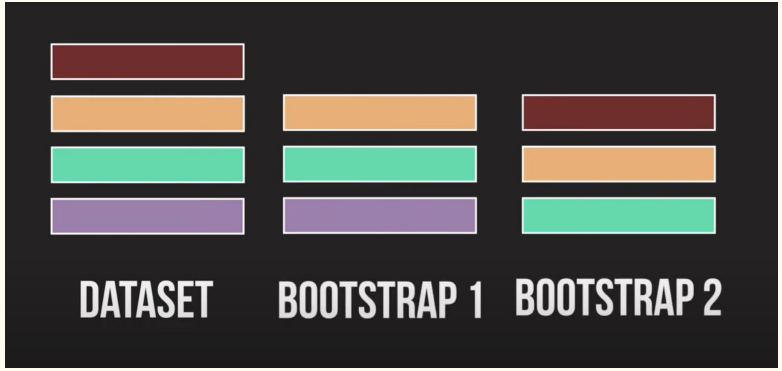


Một cây làm chẳng nên non
n cây chụm lại → Random Forest

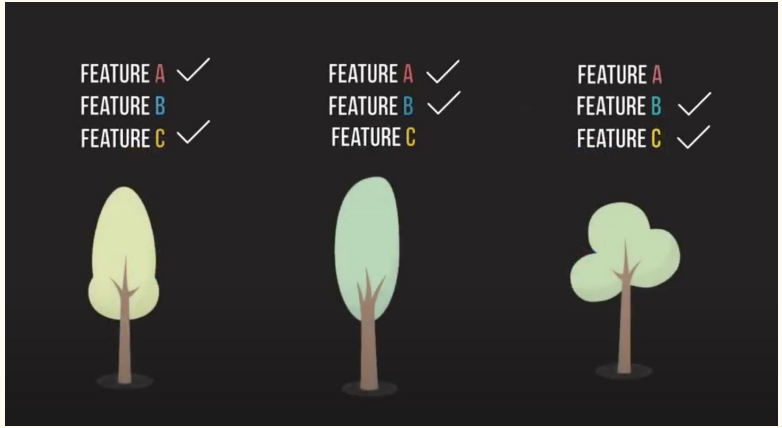
Introduction

How it work?

each Decision Tree is builded with:



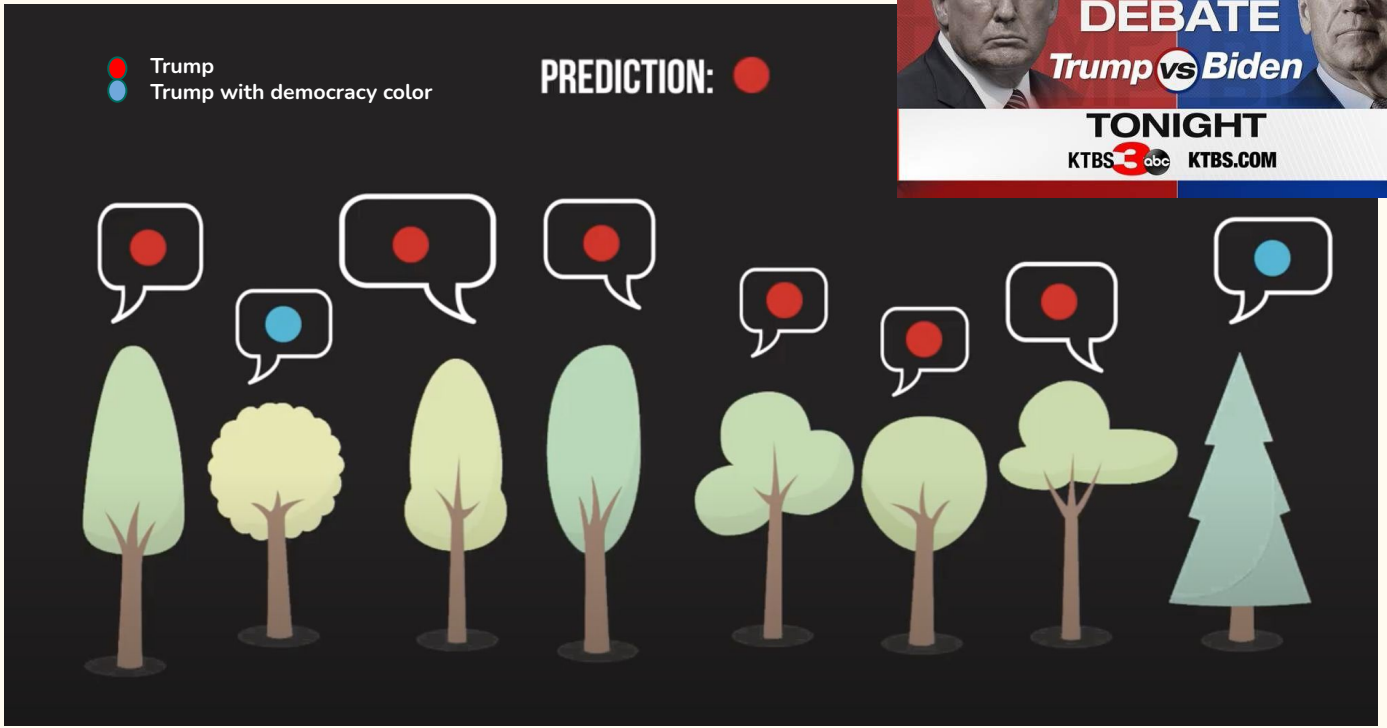
Subset of dataset



Subset of features

How it work?

n Decision Tree
→ Vote



HYPERPARAMETERS of `sklearn.ensemble`.RandomForestClassifier

Share with decision tree:

- criterion: “gini”
- max_depth: None
- min_samples_split
- min_samples_leaf
- min_weight_fraction_leaf
- max_features
- max_leaf_nodes
- min_impurity_decrease
- min_impurity_split

➔ Reduce overfitting

➔ Increasing the predictive power

➔ Increasing the model's speed

Random Forest

- n_estimators
- bootstrap (True)
- max_samples (bootstrap size)
- oob_score (out-of-bag samples)

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Demo with sklearn

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Reference

I. https://www.youtube.com/watch?v=J4Wdy0Wc_xQ&t=69s
II. https://builtin.com/data-science/random-forest-algorithm
III. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
IV. https://www.youtube.com/watch?v=cIbj0WuK41w&t=244s