Lecture inf620 2025 - Random Forest (Bag) and Boost

Depto de Informática - UFV

Ricardo Ferreira
ricardo@ufv.br

2025

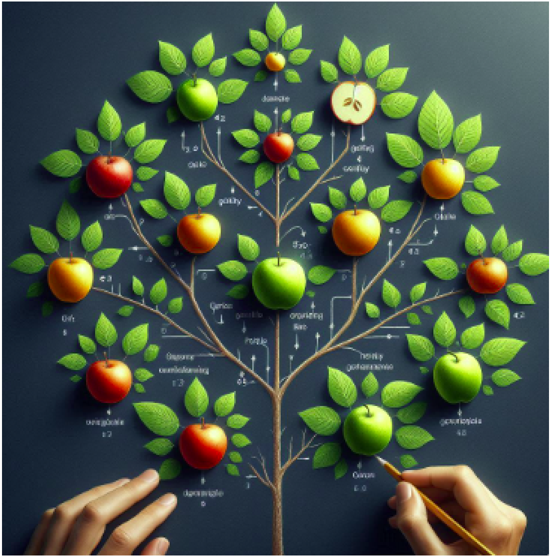
Introduction

- Class Material (click here for the Colab)
- **Review**: Supervised Learning with Decision Trees
- **Problems**: Classification and Regression
- TODAY's class: Ensemble Techniques
    - Random Forest and Bagging

Review of Decision Trees

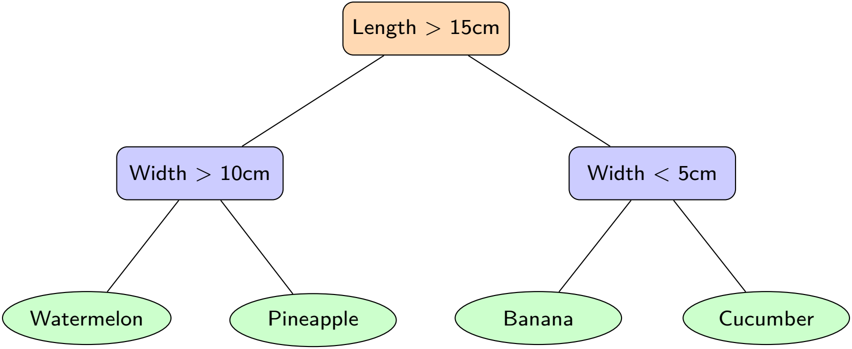
- **Definition:** Classification model that uses a tree structure for decision making
- **Main characteristics:**
  - Splits data into subsets based on attribute conditions
  - Hierarchical structure with decision nodes and class leaves
  - Easily interpretable and visually intuitive
  - Can lead to overfitting

Tree



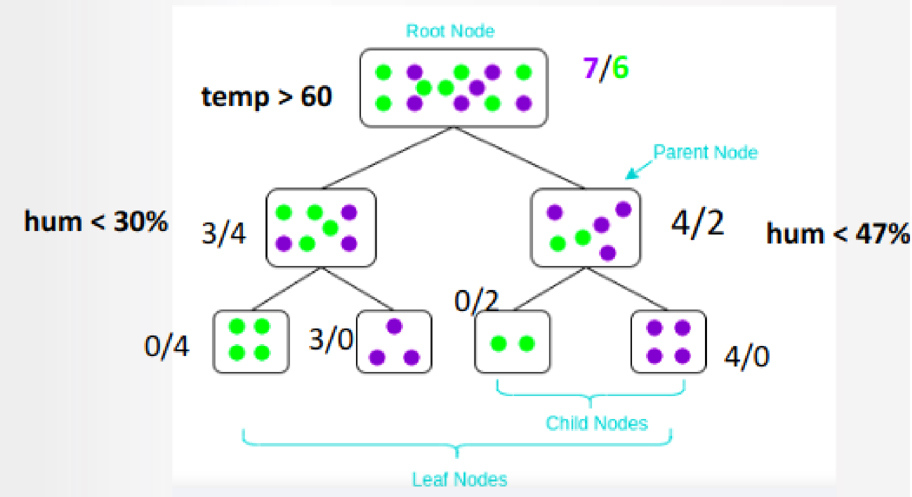
Tree in Computing

Review: Example of a Decision Tree



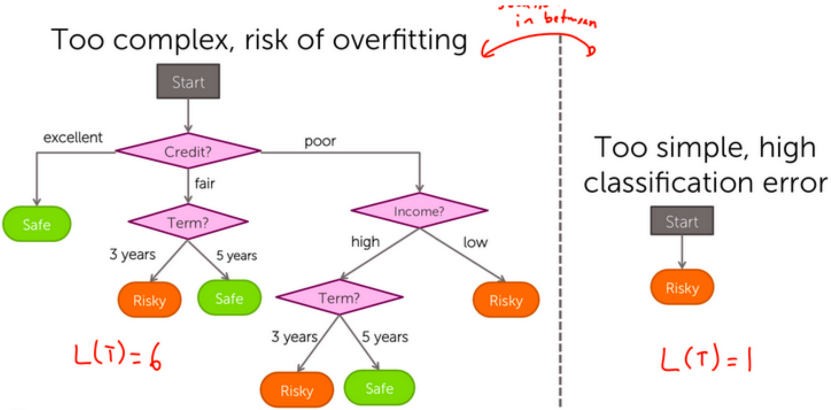
- Root: First decision based on length
- Decisions: Based on fruit width
- Fruits: Final classification

Example of a Decision Tree

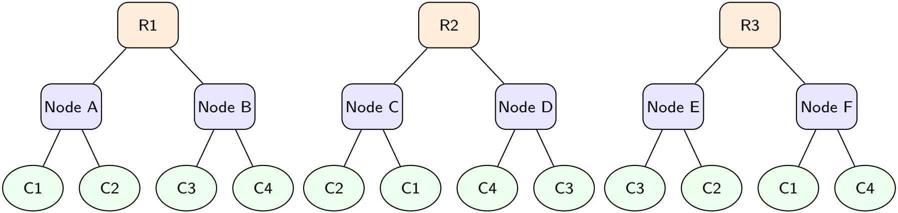


How to classify?

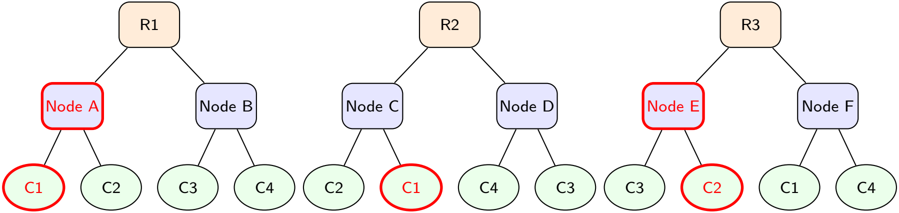
Overfitting



Review: Example of a Random Forest (3 Trees)



Review: Example of a Random Forest (3 Trees)



Random Forest - Basic Concepts

- Ensemble method based on multiple decision trees
- Combines two main concepts:
    - Bagging (Bootstrap Aggregating)
    - Random Attribute Selection
- Final result by voting or averaging

Random Forest - Characteristics

Bagging

- Samples with replacement
- Independent training
- Reduces variance

Attribute Selection

- Random subset
- Lower correlation
- Greater diversity

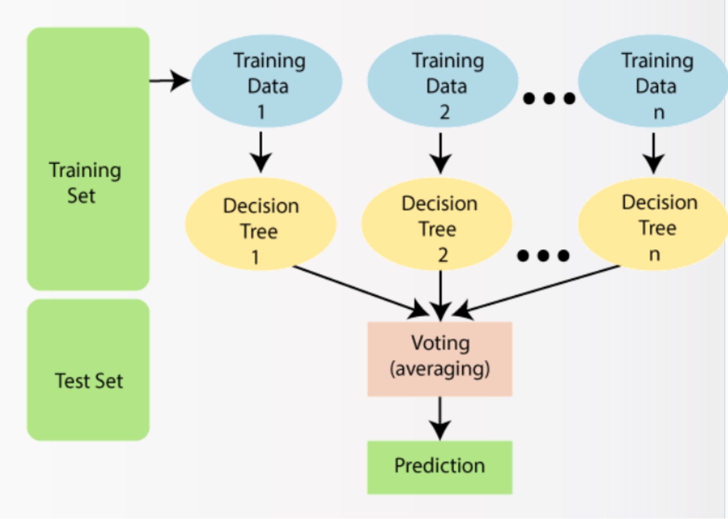
Advantages of Random Forest

- **Overfitting Reduction**
    - Combination of multiple models
    - Better generalization
- **Robustness**
    - Less sensitive to noise
    - Tolerant to outliers
- **Features**
    - Natural variable importance
    - Facilitates interpretation

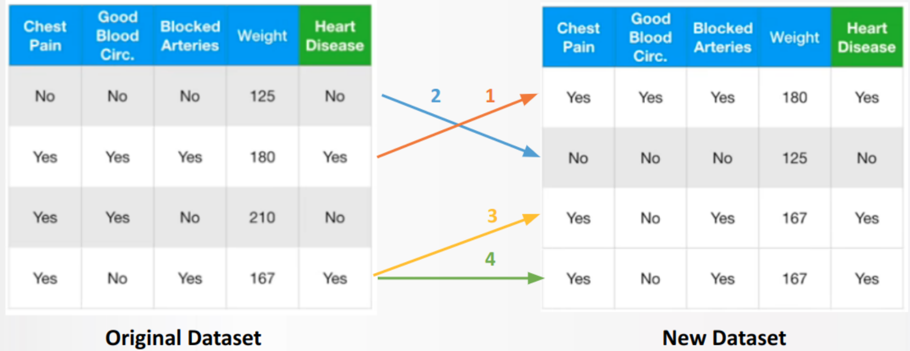
Best Practices

- **Main Hyperparameters**
    - Number of trees
    - Maximum depth
    - Learning rate
- **Validation**
    - Cross-validation
    - Early stopping
    - Continuous monitoring

Random Forest - Bagging



Random Forest - Bagging/Bootstrap



Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	No	Yes	167	Yes

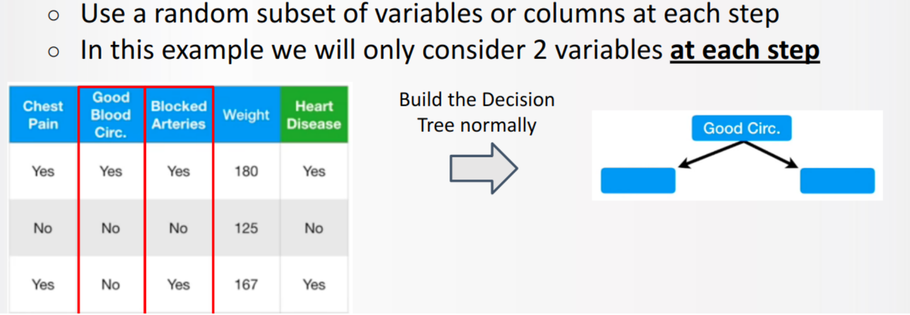
Original Dataset

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes
Yes	No	Yes	167	Yes

New Dataset

Click here for statquest

Random Forest - Bagging/Bootstrap



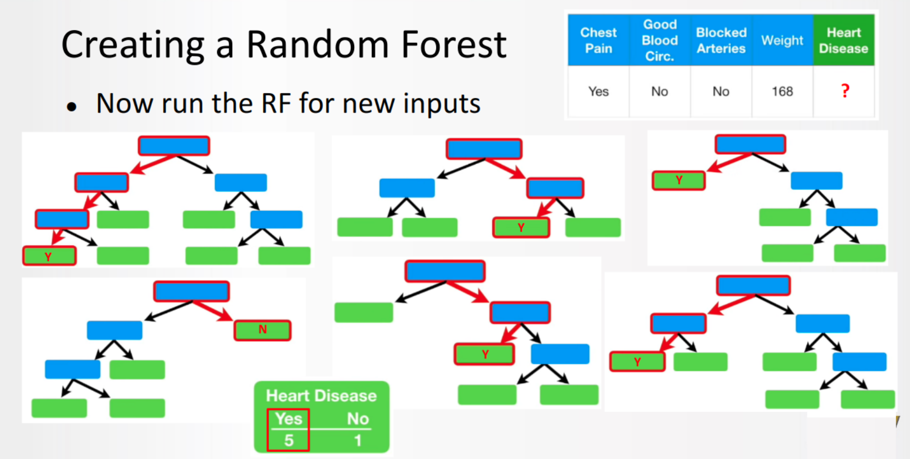
- Use a random subset of variables or columns at each step
- In this example we will only consider 2 variables **at each step**

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	Yes	Yes	180	Yes
No	No	No	125	No
Yes	No	Yes	167	Yes

Build the Decision Tree normally

Click here for statquest

Random Forest - Bagging/Bootstrap



Creating a Random Forest

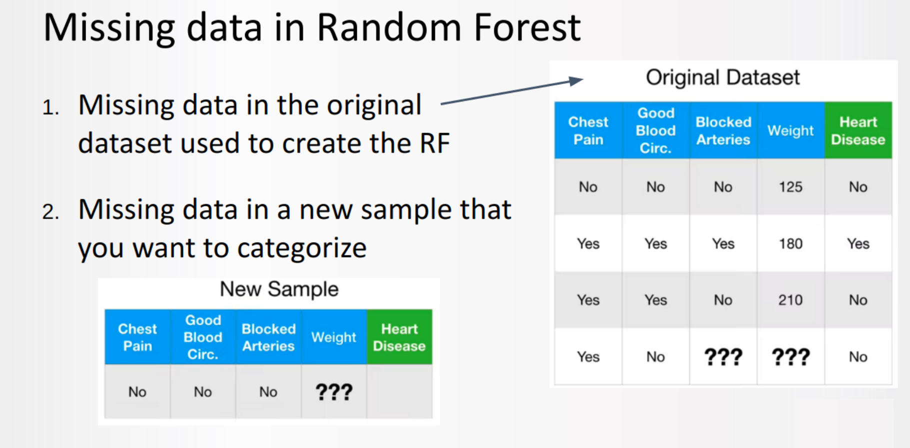
- Now run the RF for new inputs

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	No	168	?

Heart Disease	
Yes	No
5	1

Click here for statquest

Missing Data



Missing data in Random Forest

1. Missing data in the original dataset used to create the RF
2. Missing data in a new sample that you want to categorize

Original Dataset

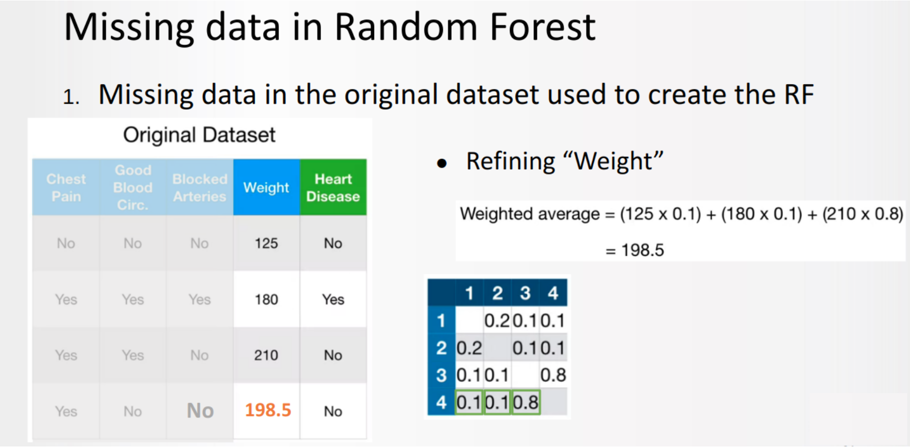
Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	No	???	???	No

New Sample

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	???	

Click here statquest

Missing Data



Missing data in Random Forest

1. Missing data in the original dataset used to create the RF

Original Dataset

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	No	No	198.5	No

- Refining "Weight"

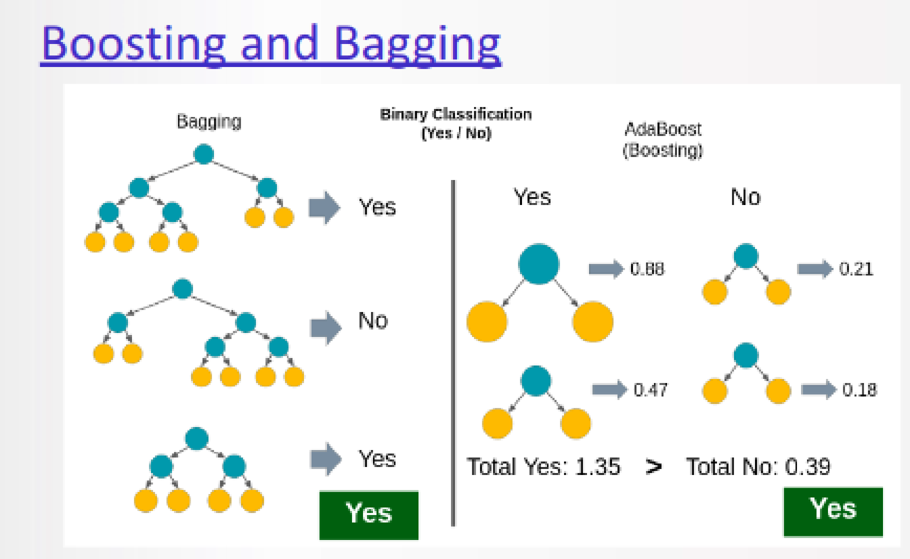
Weighted average = (125 x 0.1) + (180 x 0.1) + (210 x 0.8)

= 198.5

	1	2	3	4
1		0.2	0.1	0.1
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

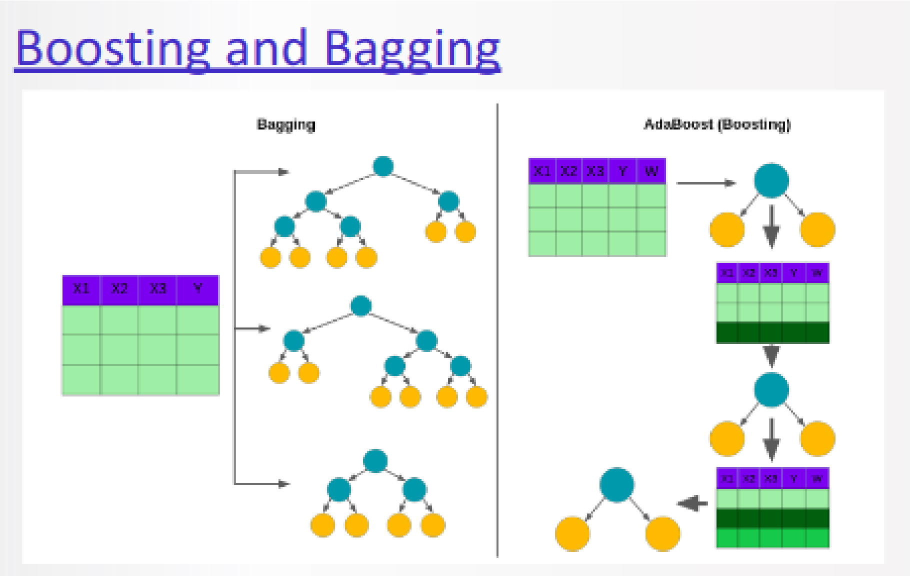
Click here statquest

Bag or Boost



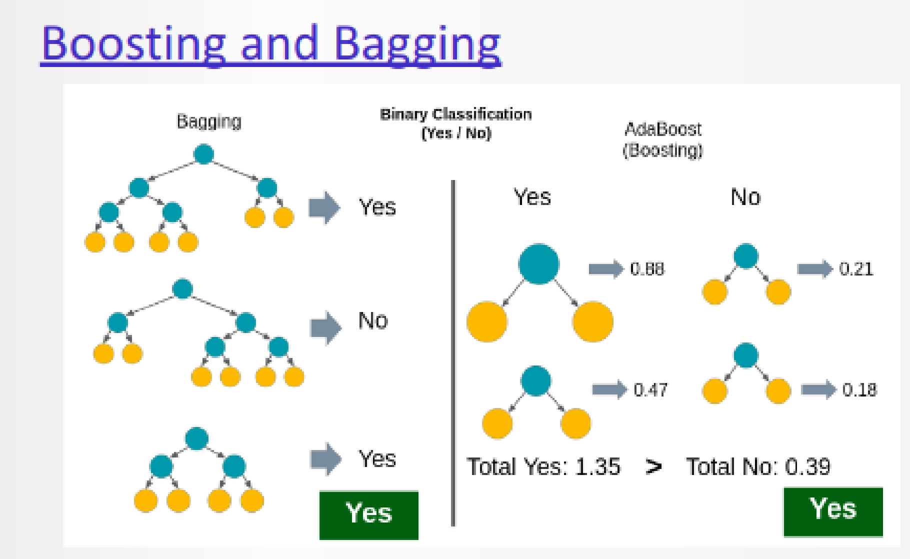
Boosting and Bagging

Bag or Boost



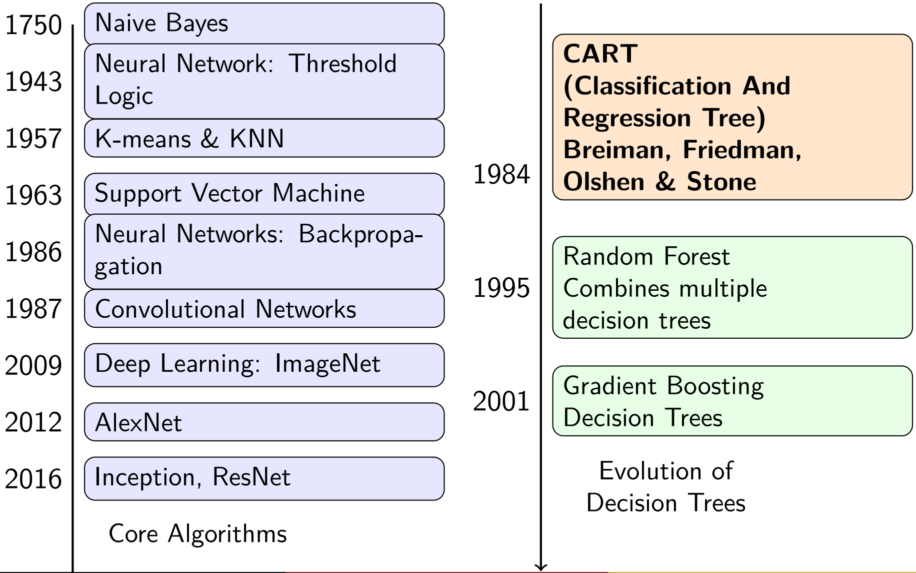
Boosting and Bagging

XGBoost



Boosting and Bagging

Evolution of Machine Learning Algorithms



1750	Naive Bayes
1943	Neural Network: Threshold Logic
1957	K-means & KNN
1963	Support Vector Machine
1986	Neural Networks: Backpropagation
1987	Convolutional Networks
2009	Deep Learning: ImageNet
2012	AlexNet
2016	Inception, ResNet

Core Algorithms

1984	**CART (Classification And Regression Tree) Breiman, Friedman, Olshen & Stone**
1995	Random Forest Combines multiple decision trees
2001	Gradient Boosting Decision Trees

Evolution of Decision Trees