Ministério da Ciência, Tecnologia e Inovações

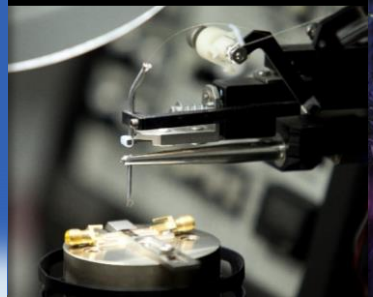
Centro Brasileiro de Pesquisas Físicas

Métodos para Análise de grande volume de dados e Astroinformática

Clécio Roque De Bom – debom@cbpf.br

CBPF

clearnightsrthebest.com






Decision Trees

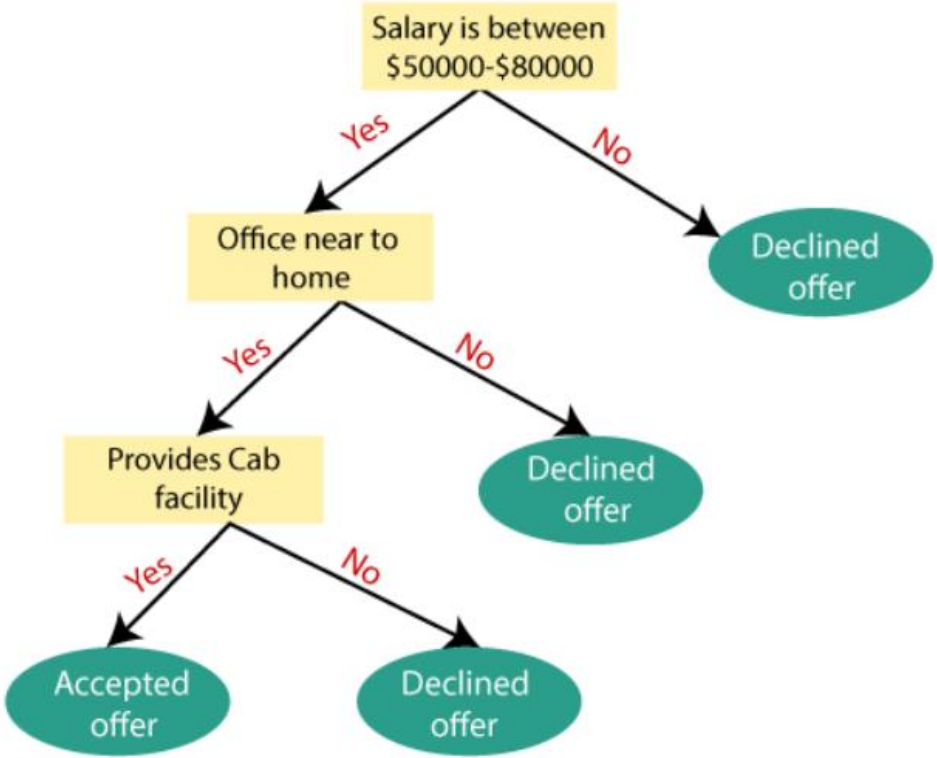
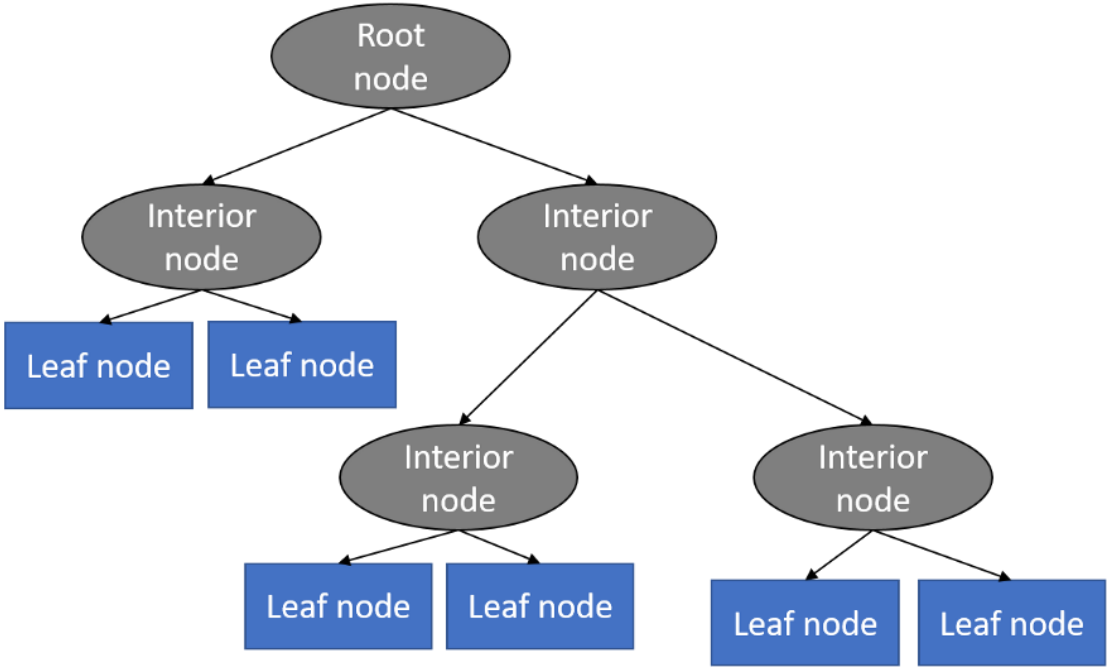
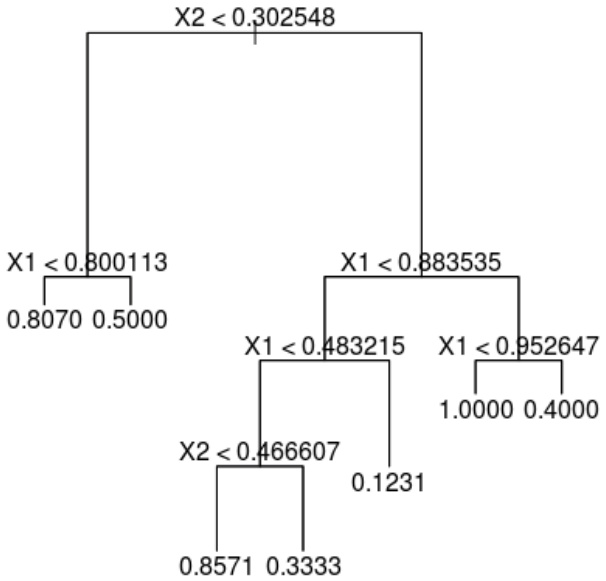


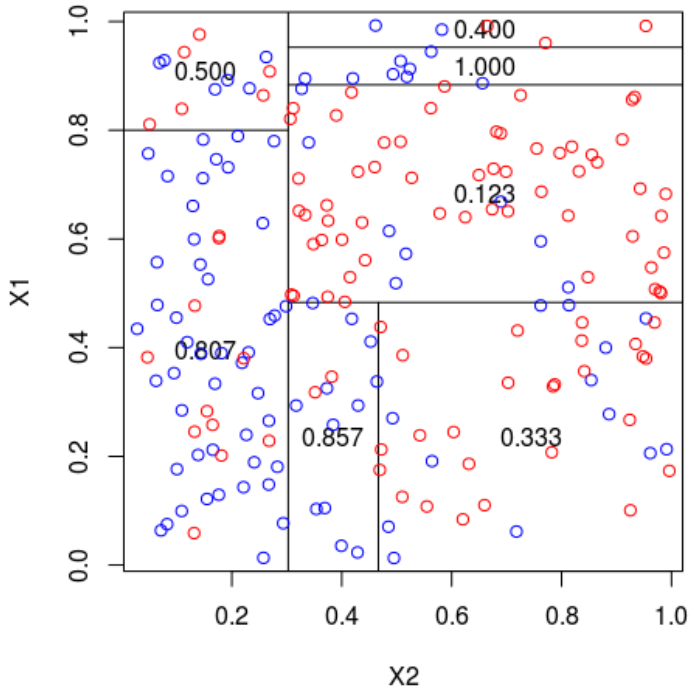
Image Credit: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

Decision Trees



Decision Trees





Decision Trees

	Weight	Heart Disease
Lowest	155	No
	180	Yes
	190	No
	220	Yes
Highest	225	Yes

Example adapted from: https://www.youtube.com/watch?v=7VeUPuFGJHk

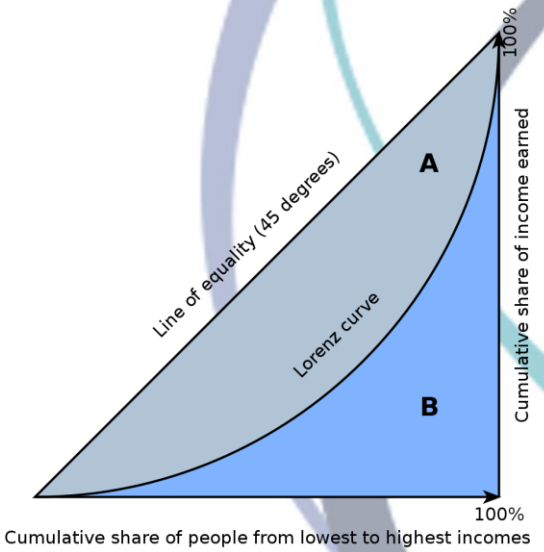
Decision Trees

Weight	Heart Disease
155	No
167.5	→ Gini impurity = 0.3
180	Yes
185	→ Gini impurity = 0.47
190	No
205	→ Gini impurity = 0.27
220	Yes
222.5	→ Gini impurity = 0.4
225	Yes

This is the best value of your metric, This is going to be the Threshold

Gini impurity = 1 – Gini

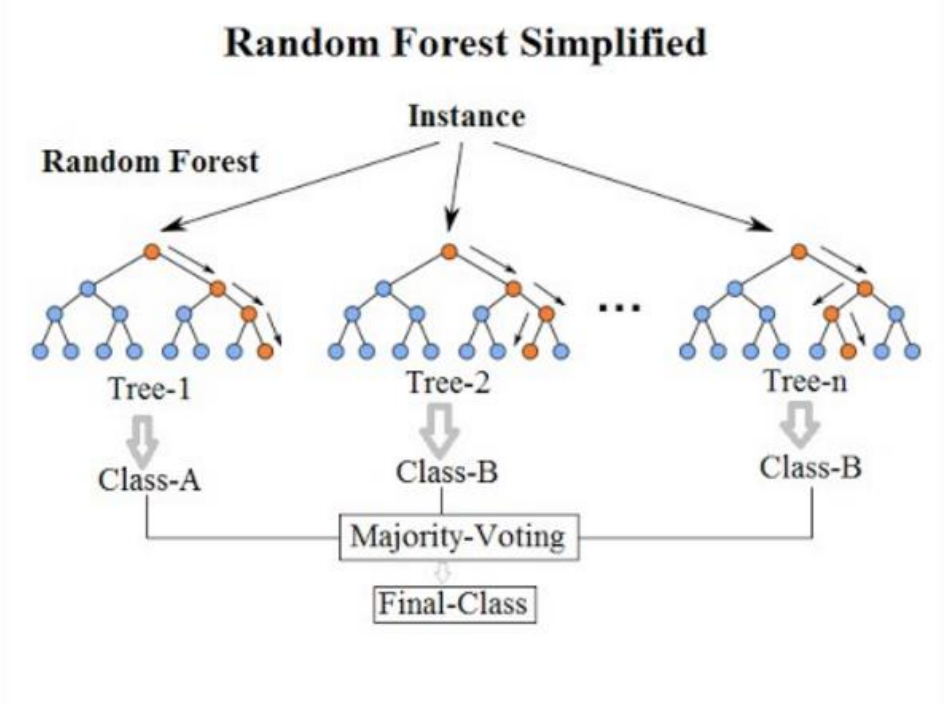
$$G = 1 - \sum_{k=0}^{k=n-1} (X_{k+1} - X_k)(Y_{k+1} + Y_k)$$



Example adapted from: https://www.youtube.com/watch?v=7VeUPuFGJHk

Random Forest

Decision trees overfits... Decision trees are very sensible to the training set... So Random Forest



Now you need a set of (wealy/um) correlated trees.

How you do it?

Bagging the samples.
Bagging the features.

Example adapted from: https://www.youtube.com/watch?v=7VeUPuFGJHk

Ministério da Ciência, Tecnologia e Inovações

Centro Brasileiro de Pesquisas Físicas

Métodos para Análise de grande volume de dados e Astroinformática

Clécio Roque De Bom – debom@cbpf.br

CBPF

clearnightsrthebest.com