DECISION TREE AND RANDOM FOREST

DECISION TREE

A Decision tree is a graphical representation of all the possible solutions to a decision based on certain conditions.

Terminologies

Nodes: Split for the value of a certain attribute.

Edges: Edges are the outcome of a split to next node.

Root: Root node is the base node of a tree. It is the node that performs the first split.

Leaf node: Node that cannot be further segregated into further nodes. These are the terminal nodes that predict the outcome.

Intuition Behind Splits

Entropy and Information gain are the mathematical methods of choosing the best split.

Entropy: It is the measure of impurity. It defines randomness in the data.

Information Gain: Measures the reduction in entropy. It decides which attribute should be selected as the decision node.

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Where, *S* is the sample space.

RANDOM FOREST

Random forest is an ensemble learning method for classification, regression and other tasks. It is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

DECISION TREE AND RANDOM FOREST WITH PYTHON

Project Background

Data Source: https://www.kaggle.com/c/titanic

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

Variables

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

pclass: A proxy for socio-economic status (SES)
1st = Upper
2nd = Middle
3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations,
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations,
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Some children travelled only with a nanny, therefore parch=0 for them.