Introduction to Machine Learning

17th

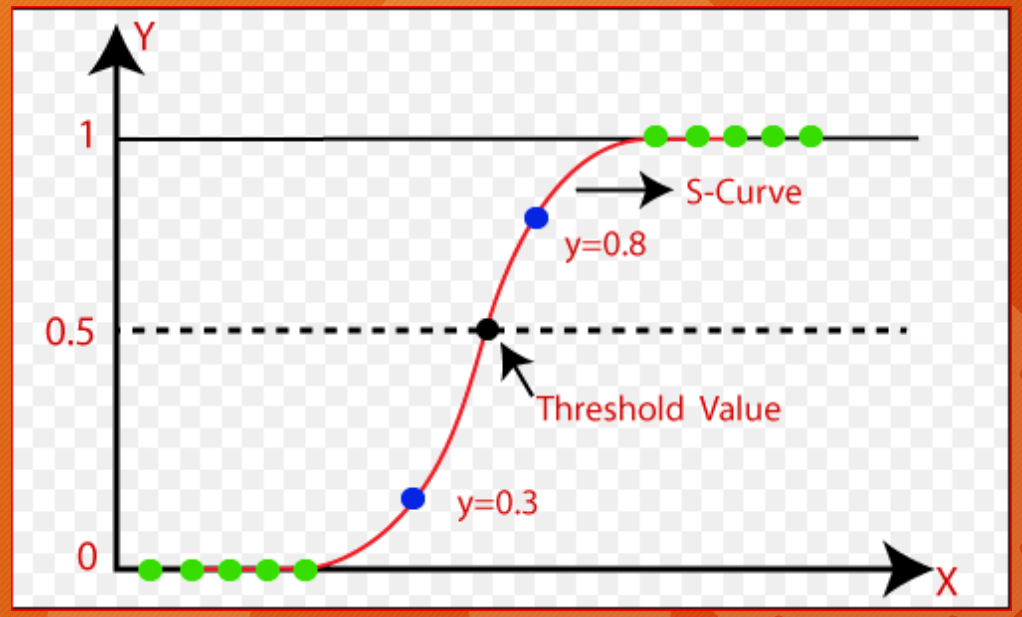
By Ayub Odhiambo

http://www.a4ayub.me/

Objectives

- Key Statistical Concepts
- Review of Previous Class
- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machines
- SVM Kernel Functions
- Naïve Bayes
- Decision Trees Classification
- Random Forest Classification
- Theory and Practical for 7 Classifiers

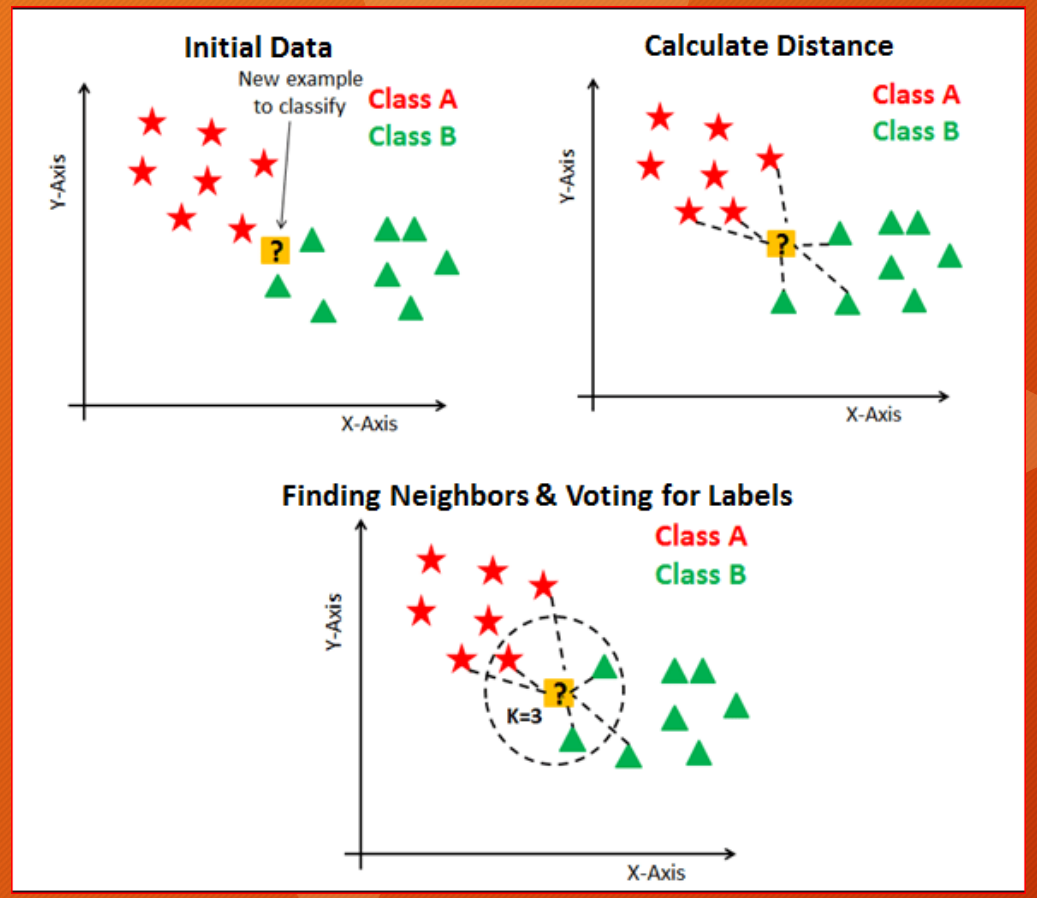
Logistic Regression



$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

- Dataset is labelled
- Used for solving classification problems
- Output variables i.e. target variable fall between 0 and 1
- Target variable (output variable is categorical in nature)
- This is used where the probabilities between two classes are required but there are instances of multi-class problems
- We pass weighted sum of inputs into an activation function that will map these values into values between 0 and 1
- Maximum likelihood estimation method is used for estimating accuracy
- It is not required to have a linear relationship between the dependent and the independent variables

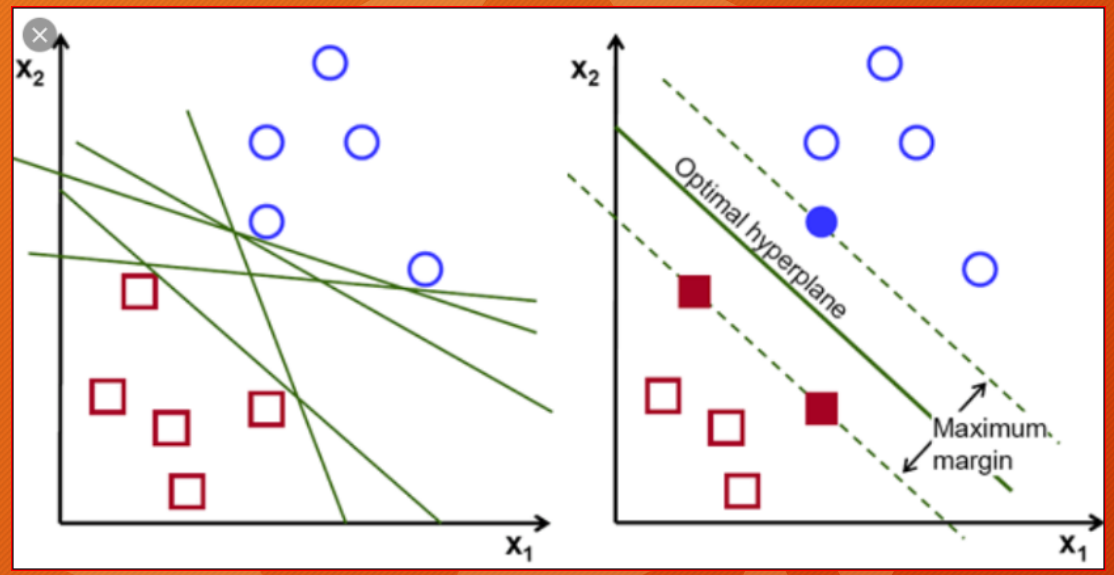
K-Nearest Neighbors k-NN

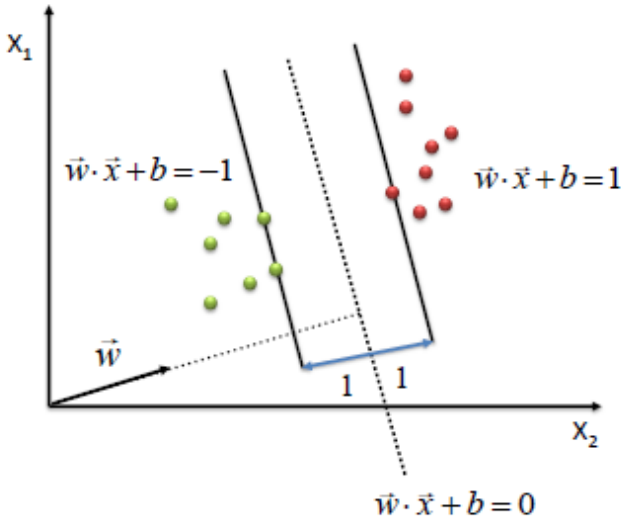


- It is **supervised**
- Assumes the similarity between the new case/data and available cases and put the new case into the category that it is most similar to the available categories
- It is non-parametric i.e. does not make any assumptions on the data
- **A lazy learner** i.e. it does not learn from the training set but instead stores the dataset and at the time of the classification, it performs and action on the data
- Examples of possible use cases are recommender systems
- Other examples of measuring the distances are eucledan squared, Cityblock, Chebyshev but the most commonly used is Eucledian Distance

$$d(p,q) = d(q,p)$$
$$= \sqrt{(q_1 - p_1)^2 + (q_3 - p_3)^2 + \cdots + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Support Vector Machines - SVMs





$$\max \frac{2}{\|w\|}$$

s.t.

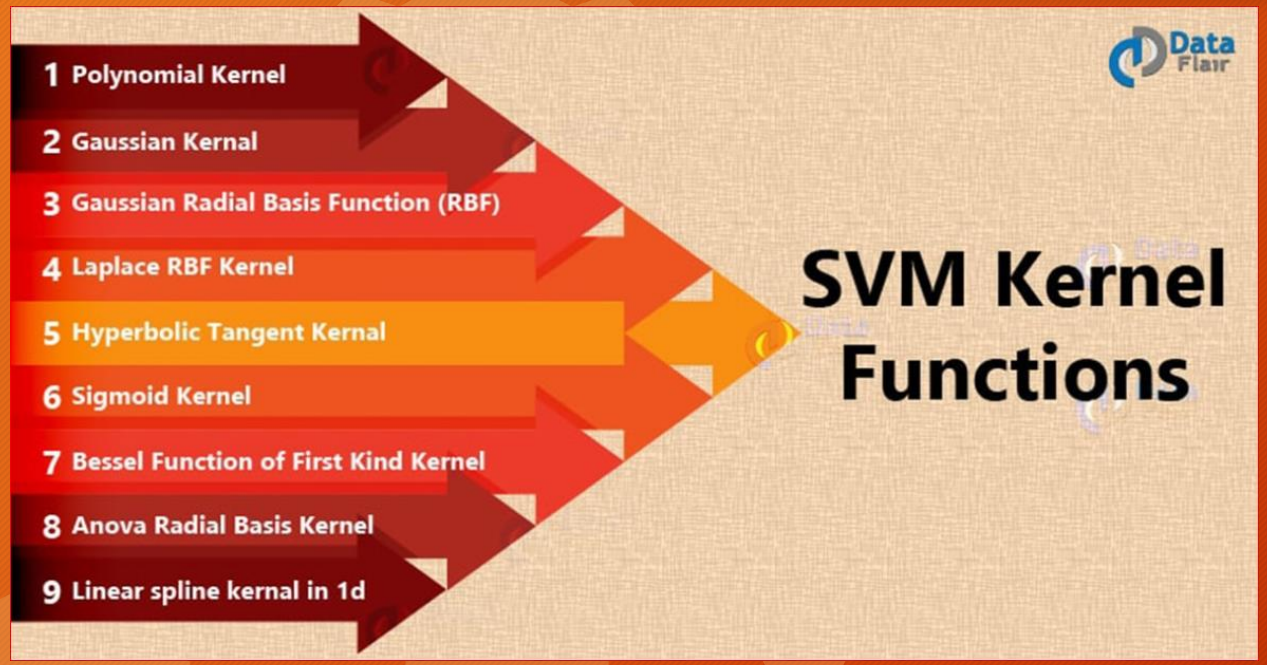
$(w\cdot x+b)\geq 1, \forall x$ of class 1

$(w\cdot x+b)\leq -1, \forall x$ of class 2

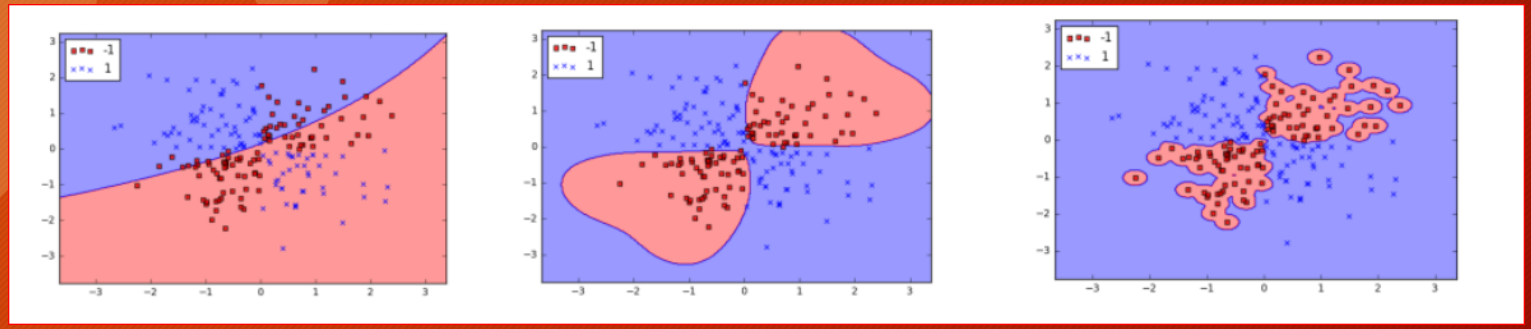
$$\min \frac{1}{2}\|w\|^2$$

$$s.t.\ y_i(w\cdot x_i+b)\geq 1,\ \ \forall x_i$$

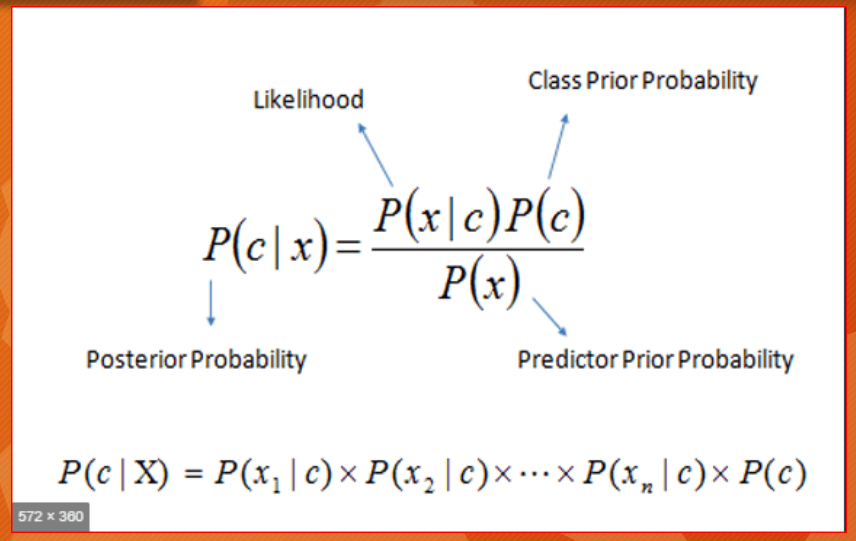
SVM Kernel Functions



- Kernel methods are a class of algorithms for pattern analysis i.e. any linear model can be transformed into a non-linear model by applying the kernel trick to that model.
- These are just mathematical functions that take data as input and transforms it into the required form.
- There are kernel functions for data, graphs, text, images as well as vectors with the most used type of kernel function being the RBF
- A kernel trick is a way of using linear classifier to solve none-linear problems



Naïve Bayes Theorem



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times P(B \mid A)}{P(B)}$$

where:

$P(A)$ = the prior probability of A occurring

$P(A \mid B)$ = the conditional probability of A given that B occurs

$P(B \mid A)$ = the conditional probability of B given that A occurs

$P(B)$ = the probability of B occurring

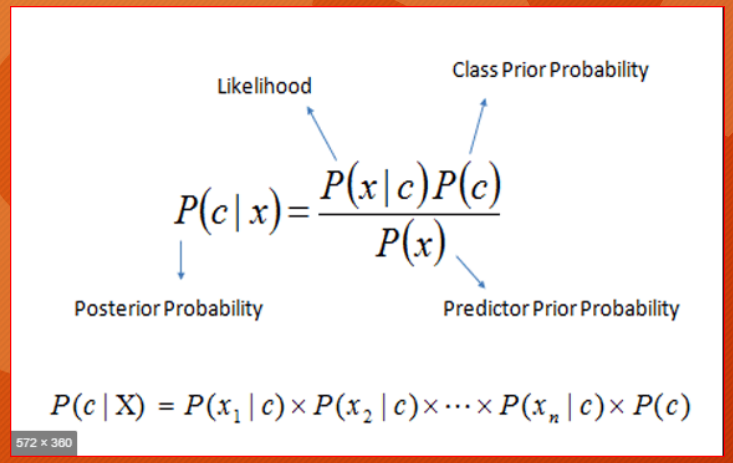
- This is a mathematical formula for determining conditional probability.
- This is based on the Bayes theorem for calculating probabilities and conditional probabilities.
- The theorem provides a way of revising existing predictions or theories given new or additional evidence.
- Considered to be fast relative to other classification algorithms
- It is based on the assumption that it's predictor variables are independent, basically meaning the presence of a particular feature is unrelated to the presence of any other feature.
- Works well with large datasets and is known to outperform sophisticated classification methods.
- **Posterior Probability** : This is the updated probability of an event occurring after taking into consideration new information
- **Prior Probability** : This is the probability of an event occurring before new data is collected. The best rational assessment of the probability of an outcome based on the knowledge before an experiment is done.
- **Likelihood** : This is the probability of the predictor given the class

Naïve Bayes Theorem : Class Problem

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		



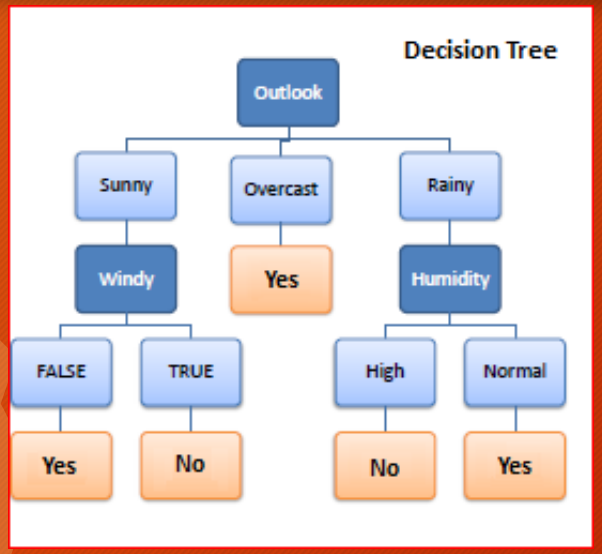
Problem Statement : Players will play if weather is sunny. Is this a correct statement?

Steps:

- **Posterior Probability** : This is what we need to calculate usine naïve bayes – P(Play | Sunny)
- **Likelihood** : P (Sunny | Yes)
- **Class Prior Probability** : P (Yes)
- **Predictor Prior Probability** : P (Sunny)

Decision Trees Classification

Predictors				Target
Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



Decision Tree breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

- **Decision Nodes** : This e.g. Outlook has two or more branches e.g. (Sunny, Overcast, Rainy), each representing values for the attribute tested.
- **Leaf Nodes** : e.g. Yes or No , the output is a categorical target
- **Root Node** : This is the top-most decision node in a tree which corresponds to the best predictor.
- **Impurity** : This is when we have one class division into the other.

Important Terms:

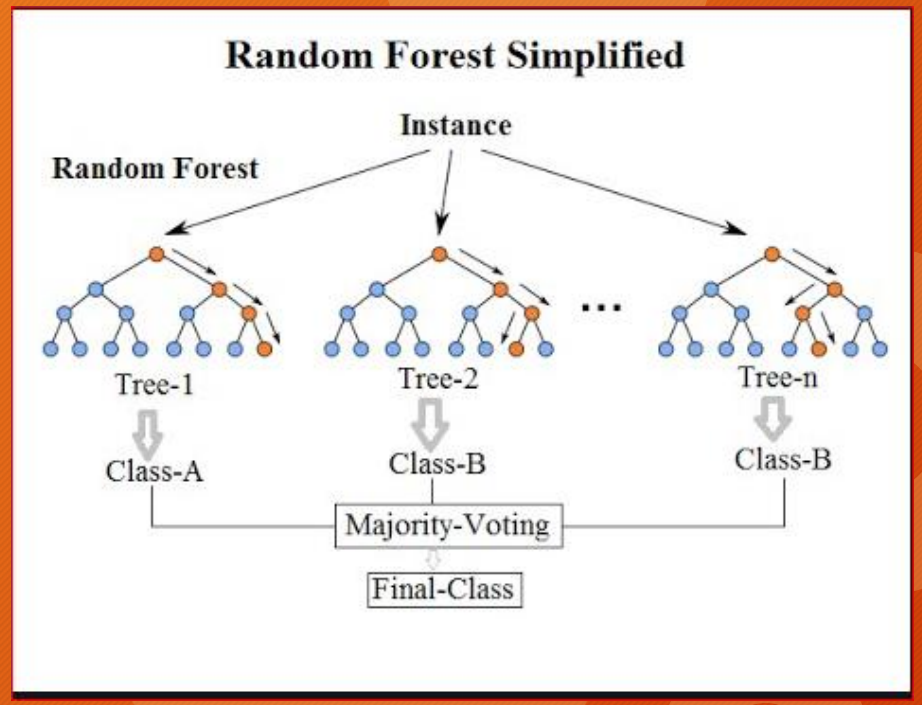
- **Entropy** : This is the degree of randomness of elements or in other words it is a measure of impurity.
- **Information Gain** : This is the decrease in entropy after a dataset is split on an attribute. The objective is to find the attribute that returns the highest information gain

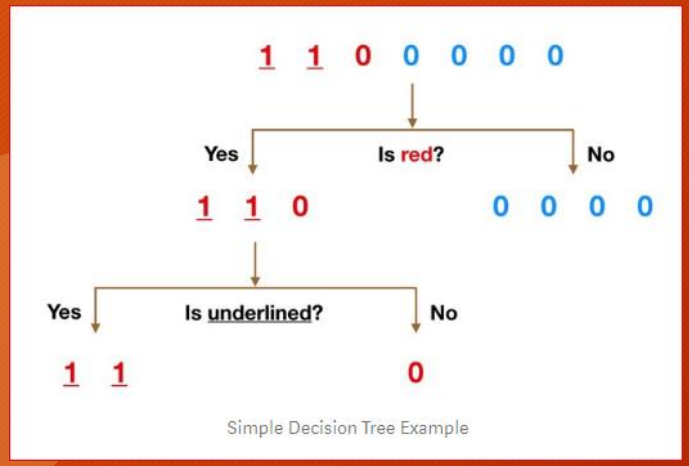
$$H = -\sum p(x) \log p(x)$$

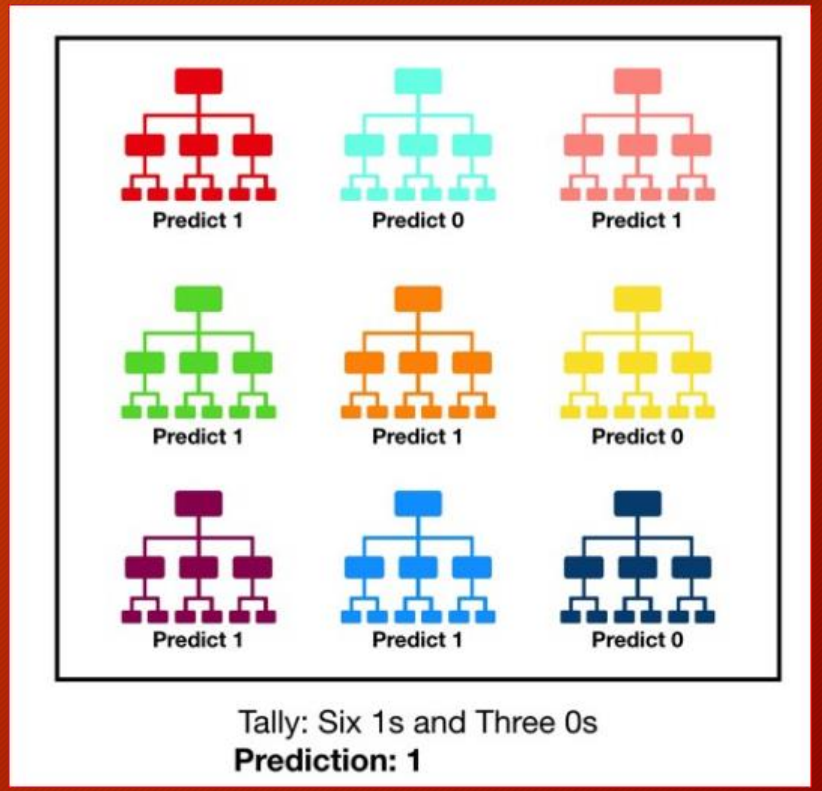
H = Entropy

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

Random Forest Classification

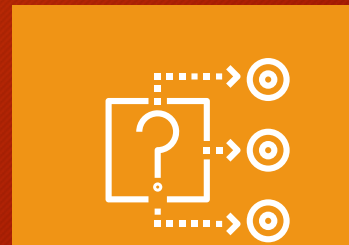






Random forest is a **Supervised Learning algorithm** which uses ensemble learning method for **classification and regression.**

Regression Jupyter Notebooks



Hands-On practical sessions on 6 Regressors.