Classification

Classification is the process of dividing the datasets into different categories or groups by adding label

* It adds the data point to a particular labelled group on the basis of some condition

Eg: Assigning a given email into spam or non-spam category

Types of classification

* Decision tree
* Random forest
* Naïve bayes
* KNN
* SVM

Decision tree

* Graphical representation of all the possible solutions to a decision
* Decisions are based on some conditions
* Decisions made can be easily explained
* Decision tree builds classification models in the form of a tree structure.
* It breaks down a dataset into smaller and smaller subsets

Decision tree terminology

Root node

It represent the entire population or sample and this further gets divided into two or more homogenous sets.

Branch /sub tree

Formed by splitting the tree/node

Leaf node

Node cannot be further segregated into further nodes.

* Splitting

Splitting is dividing the root node/sub node into different parts on the basis of some condition.

* Pruning

Opposite of splitting basically removing unwanted branches from the tree. Parent/child node

* Parent Node And Child Node

Root node is the parent node and all the other nodes branched from it is known as child node

How does A Tree decide where to split ?

* Information gain
* Gini index
* Reduction in variance
* Chi-square

Random forest

* Build multiple decision trees and merge them together
* More accurate and stable prediction
* Random decision forests correct for decision trees habit of over fitting to

their training set

* Trained with the “bagging” method
* RF is an ensemble classifier made using many decision tree models.
* Ensemble models combine the results from different models.
* Methods for balancing error in unbalanced data sets

What is random Forest?

* Random forest a versatile algorithm capable of performing both
* Regression
* Classification
* It is a type of ensemble learning method
* Commonly used predictive modelling and ML technique

Why we need random forest?

Random forest is suitable for situations when we have a large dataset and

interpretability is not a major concern.

Banking: identification of loan risk applicants by their probability of

defaulting payments.

Medicine: identification of patients and disease trends.

Land use: identification of areas of similar land use.

Marketing : identification of customer

K-Nearest Neighbour(KNN) algorithm

KNN is a simple algorithm that stores all the available cases and classifies

the new data or case based on a similarity measure.

What is K in KNN algorithm?

K= no: of nearest neighbours

How things are predicted using KNN algorithm

KNN algorithm uses feature similarity to predict the values of any new data points.

Naïve bayes

It is a simple but surprisingly powerful algorithm to predictive modelling.

How Naïve Bayes Works?

The Naïve Bayes Classifier works on the principle of conditional probability as given by the Bayes Theorem

Use Case Naïve Bayes

* News Categorization
* Stock Market
* Politics
* Finance
* Spam filtering
* Object and face recognition
* Medical diagnosis
* Weather prediction

Support Vector Machine

Support vector machine(SVM) is a supervised classification method that separates data using hyperplanes.

SVM USES

* Colon cancer classification
* Identifying the classification of genes, patients on the basis of genes and other biological problems