**Machine Learning Algorithm**

K-Means is an example of unsupervised learning because we use it as a form of estimation to try and conclude how many clusters we have in the data we are analysing. We can use it to make predictions/inferences from the datasets without the use of known labels. So it enables us to determine groups in the data and which data point belongs to a certain cluster because the data within the cluster boundary means that it is only part of one cluster and not multiple clusters. From this we can then assume that the data points in a given cluster display certain similarities as they have been grouped together. K-Means finds the nearest cluster centre for each data point and assigns that data point to the cluster closest to it. For this algorithm we need to define a target number, called k, which denotes the number of centroids (centre of the cluster) we need in the dataset and the number of clusters to be formed. The K-Means algorithm identifies k number of centroids by randomly selecting k data points from the data set as the initial cluster centroids. For each data point the algorithm computes the distance between the data point and the cluster centroid to then allocate every data point to the nearest cluster, whilst at the same time keeping the centroids as small as possible by iterative calculations to optimize the positions of the centroids. The algorithm does this for each cluster by calculating the new mean based on the data points in the cluster. The algorithm stops when the centroids have stabilized as there is no change in their mean values because the clustering has been completed or if the user defined number of iterations has been accomplished. The K-Means algorithm is most useful for cluster analysis in data mining to solve clustering problems. An application where we can utilise K-Means is document classification. This involves cluster documents in multiple categories based on tags, topics, and the content of the documents. K-Means is frequently used in this application as this is a very standard classification problem and k-means is a highly suitable algorithm for this purpose. The initial processing of the documents is needed to represent each document as a vector and uses term frequency to identify commonly used terms that help classify the document. The document vectors are then clustered to help identify similarities in document groups.