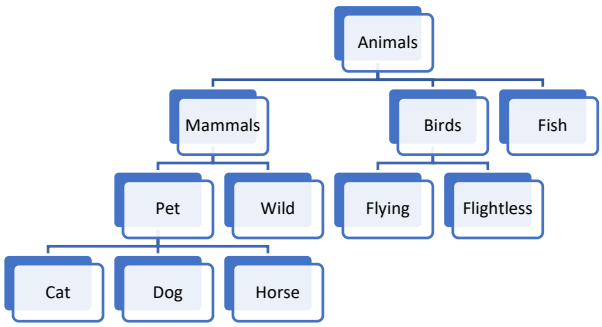
Types of Clustering

Clustering may be classified broadly into two categories – Flat and Hierarchical. K- Mean Clustering is a kind of Flat clustering that only requires the number of clusters to be considered. As the name suggests, in hierarchical clusters there is an hierarchy in the clusters. For example



The hierarchical clustering can further be divided into – Agglomerative (Bottom-up) or Divisive (Bottom down). In Agglomerative clustering, we start with each individual item as a cluster and then start clubbing them together. In the diagram look at the leaves and start associating with a parent till you reach a single parent (the root). The clustering is depicted by a dendrogram. It shows all the possible linkages between the clusters, gives a better understanding of the data and there is no need to preset the number of clusters though it may get messy if we start with too many leaves. But you can create heatmaps to understand the clustering in a better way.

Naïve Bayes Algorithm

This is a classification technique that works on the Bayes theorem that states

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

For example consider the example of getting a spam mail is 30%. Assume that 20% of the mails that you receive contain the word "win" and 50% of these are spam mails. Then P(Spam) = 0.3, P(Win) = 0.2 and P(Win|Spam) = 0.5. Therefore, if a mail contains the letter win, the probability of it being a spam will be given by

$$P(Spam|Win) = \frac{P(Win|Spam)P(Spam)}{P(Win)} = 7.5\%$$

The Naïve-Bayes algorithm assumes that each feature is independent. In the previous example, consider now another word prize in the mails. Naïve-Bayes algorithm assumes that the occurrence of these two words in a mail are independent of each other. Assume that the word prize is there in 80% of the mails and 75% in spam mails. Then,

$$P(Spam|Prize) = \frac{P(Prize|Spam)P(Spam)}{P(Prize)} = \frac{0.75 \times 0.5}{0.8} = 47.875\%$$

On the basis of the occurrence of these two words – win and prize – the probability

$$P(Spam|Win) \times P(Spam|Prize)$$

is calculated. If this is greater than a threshold value it is classified as a spam else not a spam (ham).

Read more about it at https://scikit-learn.org/stable/modules/naive_bayes.html

Note: Naïve-Bayes is not a single algorithm but a collection of different implementation