Unsupervised machine learning is a machine learning method that allows us to cluster data in order to find hidden or unknown patterns.

Preparing the data comes in 3 steps: Data Selection, Data Pre-processing and Data Transformation.

There are several types of unsupervised machine learning models such as Customer Segregation, K-means and Principal Component analysis.

PCA machine learning which enhances and optimizes machine learning algorithms by applying PCA.

PCA is a statistical technique used to speed up machine learning algorithms and works by reducing the number of input features or dimensions. As you can see from the diagram the PCA reduced the number of features from 29 to 13.

The TSNE algorithm also known as Stochastic Neighbourhood Embedding is a clustering and visualization method and is different from PCA as it is able to account for non-linear relationships.

This algorithm models the probability distribution of neighbours around each point.

As the image on the slide shows there are 2 points with the value 1 shown by the blue circle and shows that the TSNE algorithm was unable to distinguish the classes.

K means is an unsupervised learning algorithm used to identify clusters and solve clustering issues.

K represents the number of clusters.

The K-means algorithm groups the data into clusters where each piece of data is assigned to a cluster based on similiarity or the distance measured to a centroid.

A centroid represents a data point that is the arithmetic mean position of all the points in a cluster. This process is repeated until the data is separated into distinct groups.

The elbow curve represents the number for k and the number of clusters as it is the inflection point where the slope takes a sharp turn and flattens out.

As we can see in the graph, the elbow curve has 2 bends and is therefore said to have 2 clusters. The inertia value is 11267 and as we read it off the y-axis it can be seen to have 2 clusters on the x-axis.

We also tried to carry out hierarchy clustering and tried running the cluster, however due to the size of the data and the RAM available, we had difficulty obtaining results.