TSSM
Page No. :

ASSIGNMENT - 4

Title :-

Implement K-Nearest Neighbours algorithm on diabetes.csv dataset. Compute confusion matrix, accuracy, error rate, precision & recall on the given dataset.

Objective :-

Students should be able to preprocess dataset & identify outliers, to check correlation & implement KNN algorithm & random forest classification models.

Theory :-

K-Nearest Neighbour (KNN) is a supervised machine learning model.

Supervised learning is when a model learns from data that is already labeled.

A supervised learning models takes in a set of input objects and output values.

The model then trains on that data to learn how to map the inputs to the desired output so it can learn to make predictions on unseen data.

KNN model works by taking a data point and looking at the 'k' closest labeled data points.

The data point is the assigned the label of the majority of K closest points.

For KNN model.

First step is to read in the data we will use as input.

Next we will split dataset into inputs and target.

Next we will split dataset into training data and testing data. Training data is the data that the model will learn from. Testing data is the data we will use to see how well model performs. on unseen data.

Next we will build the model.

Once we have trained the model we will test our model and check accuracy.

K-Fold Cross Validation.

Cross-fold validation is when the dataset is randomly split up into 'k' groups. One of the groups is used as the test set & the rest are used as training set. The model is trained on training set and scored on the test set. Then the process is repeated until each unique group has been used as test.

Hypertuning model parameters using Gridsearchcv.

Hypertuning parameters is when you go through a process to find the optimal parameters for your model to improve accuracy.

Grid search CV works by training model multiple times on a range of parameters that we specify.
That way, we can ~~trus~~ test our model with each parameter & figure out the optimal values to get the best accuracy results.

Conclusion:-

In this way we build a neural network based classifier that can determine whether they will leave or not in the next 6 months.

TSSM
Page No. :

ASSIGNMENT - 5.

Title:-

Implement K-means clustering / hierarchical clustering on sales_data_sample.csv dataset. Determine the number of clusters using the elbow method.

Objective:-

Students should be able to understand how to use unsupervised learning to segment different - different clusters or groups & used to them to train your model to predict future things.

Theory:-

Clustering:-

Clustering algorithms try to find natural clusters in data, the various aspects of how the algorithms to cluster data can be tuned & modified.

Clustering is based on the principle that items within the same cluster must be similar to each other.

The data is grouped in such a way that related elements are close to each other.

Applications:-

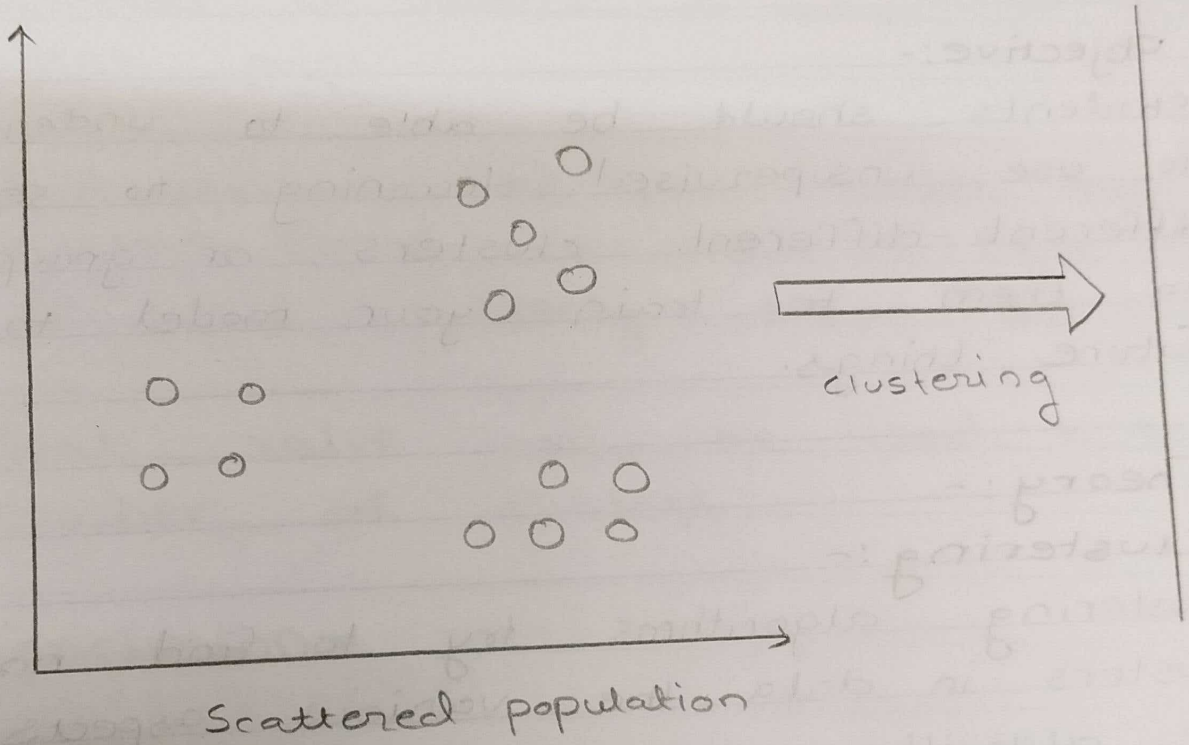
Marketing.

Real Estate.

Libraries & Bookstores.

Document Analysis.

Unsupervised learning - C.



TSSM
Page No. :

K means Clustering.

- K-Means clustering is an unsupervised machine learning algorithms that divides the given data into the given number of cluster.
Here, K is the given number of predefined clusters, that need to be created.
- It is a centroid based algorithm in which each cluster is associated with a centroid.
The main idea is to reduce the distance between the datapoints & their respective cluster centroid.

- The algorithm takes raw unlabelled data as an input and divides the dataset into clusters & the process is repeated until the best cluster found.

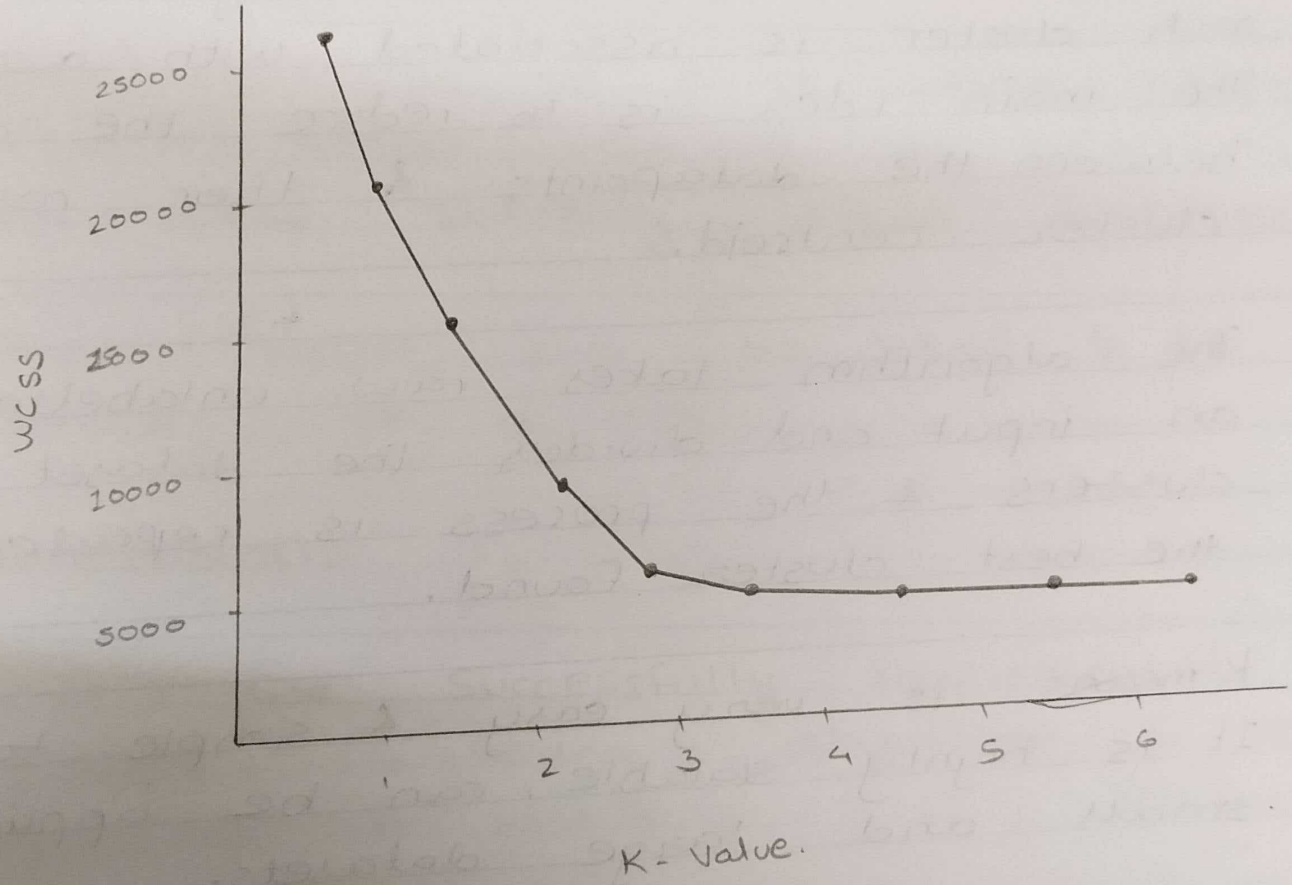
- K-means is very easy & simple to implement It is highly scalable, can be applied to both small and large datasets.

- There is however, a problem with choosing the number of clusters or K.

- Also, with the increase in dimensions, stability decreases.

- But, overall K means is a simple & robust algorithm that makes clustering very easy.

Elbow plot



within Cluster Sum of Squared Errors (WSS)

This ~~num~~ value of K gives us the best number of clusters to make the raw data.

WCSS has elbow graph, the x-axis being the ~~nob~~ number of clusters, the number of clusters is taken at the elbow joint point.

This point is the point where making clusters is most relevant as here the value of WCSS suddenly stops decreasing.

That value can be used to be the number of clusters

· Conclusion:-

Hence, we successfully implemented K-means clustering algorithm & determined the no. of clusters using elbow method.