**ELG5255 Applied Machine Learning Group Assignment 3**

**Group 9**

1. Wael Mofreh Ismail Ali Elsharkawy – 300273116

2. Sameh Mohamed Ismail Deabes - 300273126

**Goal:**

Dimensionality Reduction and Feature Selection methods along with the unsupervised methods.

**Dataset:**

Our dataset that we use is the pokemon dataset is divided to 2 csv file Pokemon\_train.csv and Pokemon\_test.csv used to train and test model. The provided dataset is preprocessed and prepared to use, we upload the data to GitHub and read it from GitHub.

Application

Description automatically generated with medium confidence

# **Load dataset:**

We load the data and divide it to train\_feature and train\_label and test\_feature ,test\_label .we read the data set and represent it

Text

Description automatically generated

Graphical user interface

Description automatically generated

# **Apply Gaussian Naïve Bayes classifier (GNB) and (SVM)**

**Apply the GNB on the pokemon .**

* train model on dataset
* test the accuracy of the model by test data
* get accuracy 51.437%

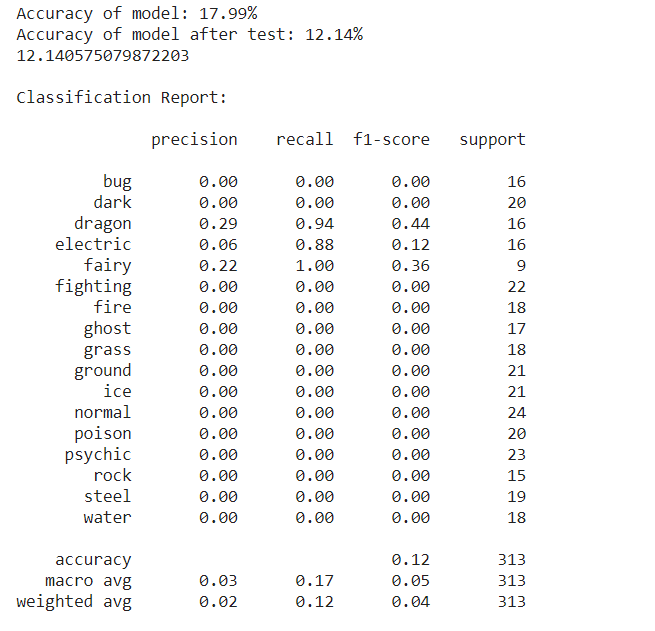
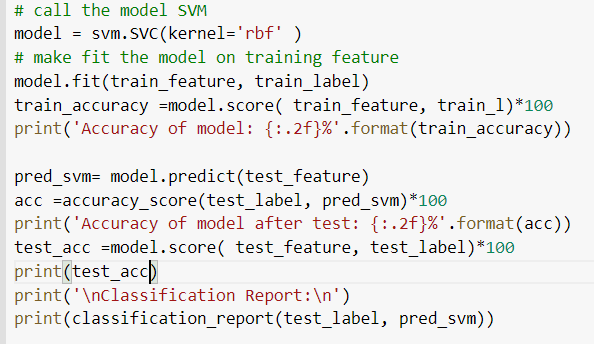
Text

Description automatically generatedTable

Description automatically generated

**Apply the SVM on the pokemon** .

* train model on dataset
* test the accuracy of the model by test data
* get accuracy 17.99%



**Apply TSNE(n\_components=2, random\_state=0) to training and test set and visualize**

* **TSNE on train set**

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

* **TSNE on test set**

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

# **Choose the best number of clusters for k-means cluster**

* **Using the elbow rule, plot the distortion score (a.k.a inertia) vs the number of clusters**

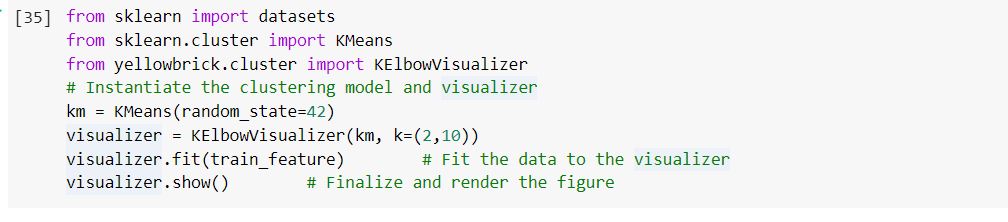
Text

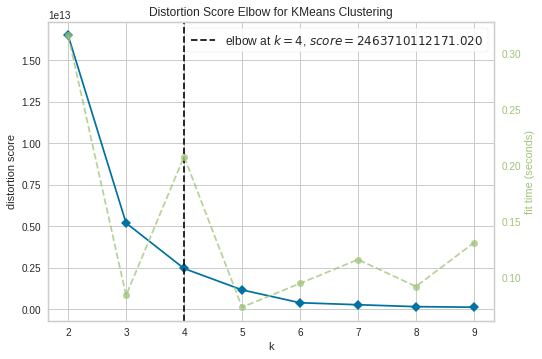
Description automatically generatedChart, line chart

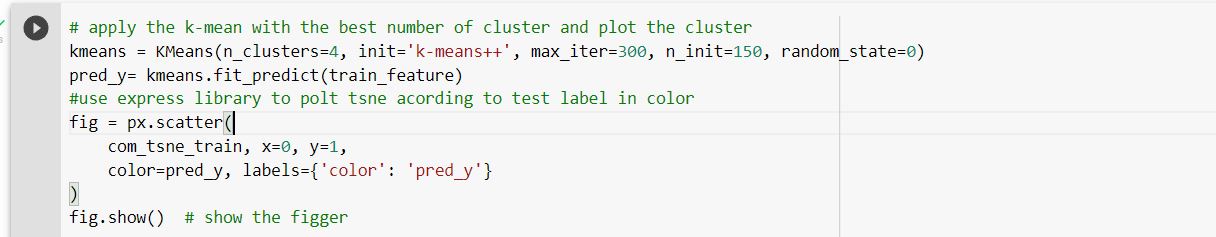
Description automatically generated

* **Determine the optimal number of clusters for k-means:**

As shown the optimal number of clusters in k-means is **4** clusters

****

****

* **Plot clustered data with optimum number of cluster**

Graphical user interface, chart, scatter chart

Description automatically generated

# **Dimensionality Reduction**

We apply dimensionality reduction on GNB and SVM classifiers as Followed:

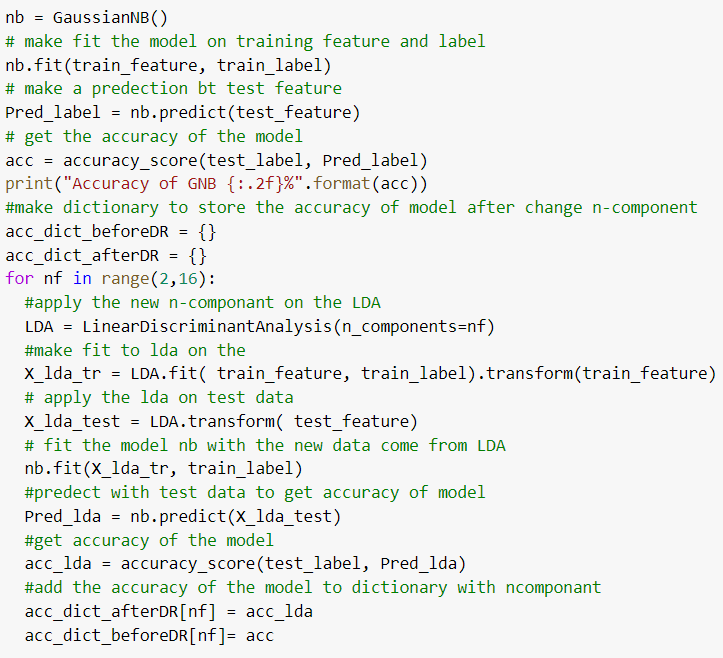
**First applying dimensionality reduction on GNB model:**

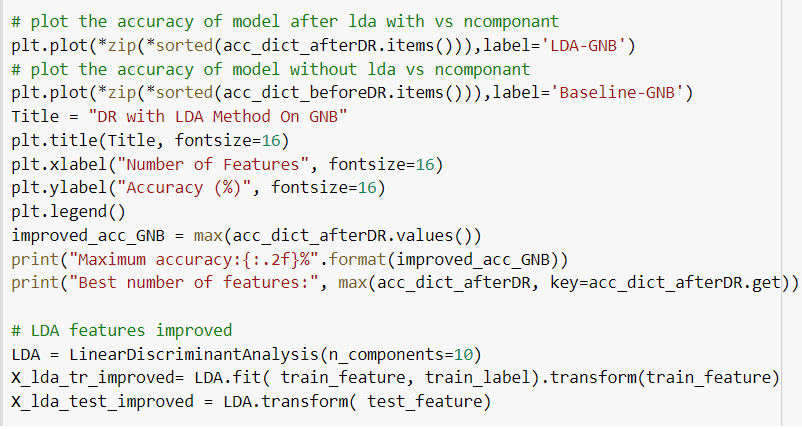
We apply GMB on training set, and test set with LDA and PCA and find accuracies as followed:

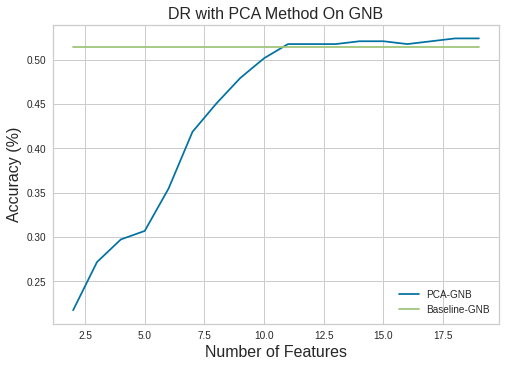
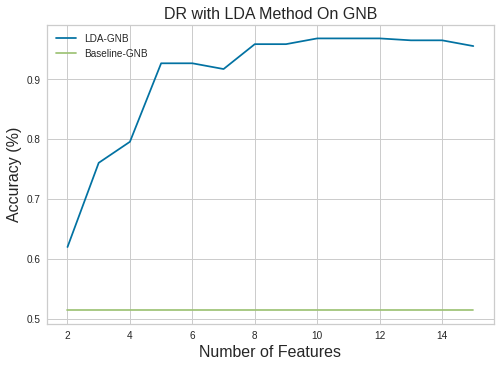
* + With LDA(n\_components=n, random\_state=0) Accuracy changed from ( 0.51% to 0.97%)
  + With PCA(n\_components=n, random\_state=0) Accuracy changed from ( 0.51% to: 0.52%)

Thus GMB with LDA is the best.

Code of GNB with LDA:







**First applying dimensionality reduction on SVM model:**

We apply SVM on training set, and test set with LDA and PCA and find accuracies as followed:

* + With LDA(n\_components=n, random\_state=0) Accuracy changed from ( 0.12% to 0.98%)
  + With PCA(n\_components=n, random\_state=0) Accuracy changed from ( 0.12% to: 0.16%)

Thus SVM with LDA is the best.

Text

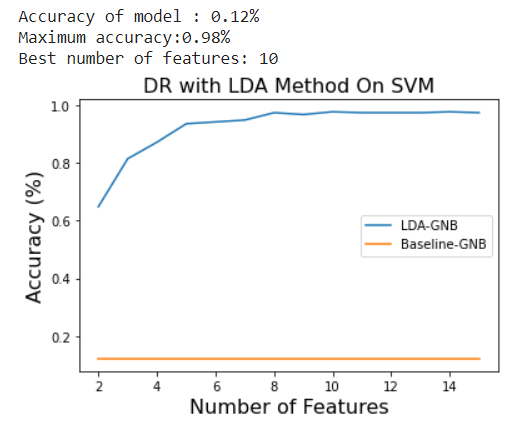
Description automatically generated

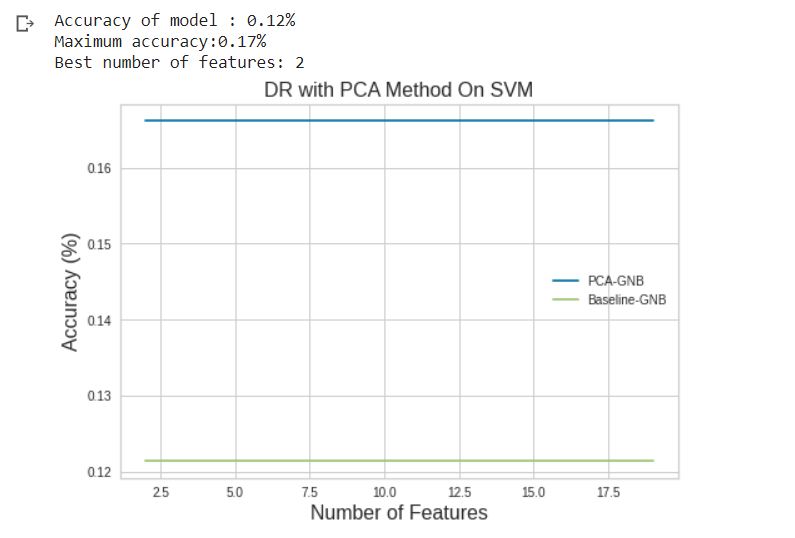
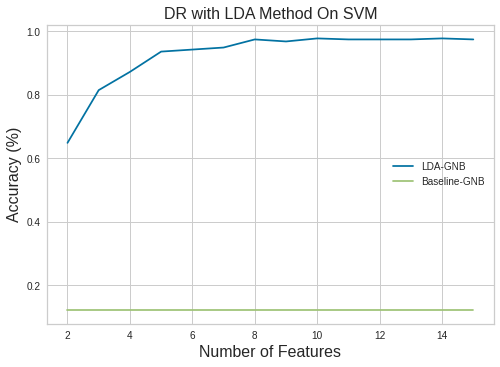
Text

Description automatically generated

Text

Description automatically generated





# **Feature Selection Methods**

We make a function to give it data and model used and feature selection method and it give us the accuracy of this model.

**And Code for selection function as follows:**

Text

Description automatically generated

## Feature selection (Filter Method) with GNB:

Apply gaussian naïve biases with LDA give accuracy 97 % and this is our baseline accuracy.

After applying filter method, we get max accuracy of 96% with number of features = 9

As shown below:

Text

Description automatically generated

Text

Description automatically generated with medium confidence

Text

Description automatically generated

Chart, line chart

Description automatically generated

## Feature selection (wrapper method) with GNB:

We apply the wrapper method on data we choose Sequential Feature Selector method and give the same accuracy with the number of feature equal 9.

Text

Description automatically generated

Text

Description automatically generated with low confidence

Text, letter

Description automatically generated

Chart, line chart

Description automatically generated

## Feature selection (Filter Method) with SVM:

Apply SVM with LDA give accuracy 98 % and this is our baseline accuracy.

After applying filter method, we get max accuracy of 98% with number of features = 10.

**Code :**

Text, letter

Description automatically generated

Text

Description automatically generated

Text, letter

Description automatically generated

Chart, line chart

Description automatically generated

## Feature selection (wrapper method) with SVM:

We apply the wrapper method on data we choose Sequential Feature Selector method and give the same accuracy of 98% with the number of feature equal 10.

**And code as followed:**

Text, letter

Description automatically generated

A picture containing application

Description automatically generated

Text

Description automatically generated

Chart

Description automatically generated with medium confidence

# **Applying elbow method to K-means**

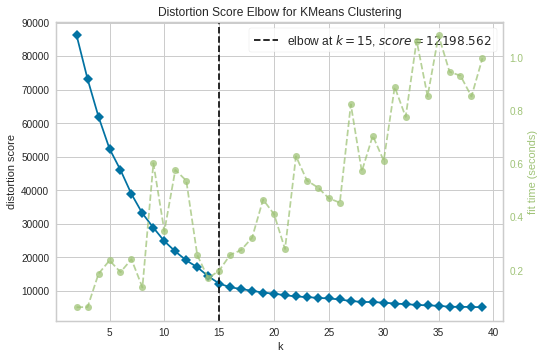
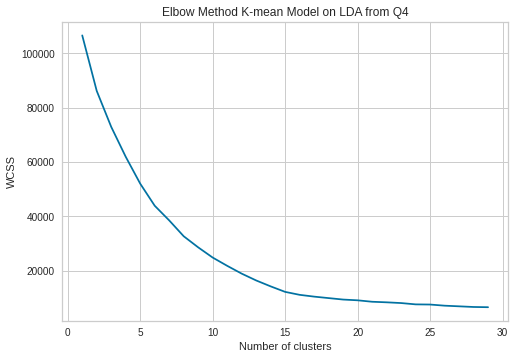
After choosing best number of clusters for k-means clustering algorithm from Q4 and Q5 we find that the best number of features equals 10 and

* we plot the elbow curve with the n clusters
* we make the number of clusters from 2 to 40 and apply the k-mean on it and

**Result as shown in the graph:** **the** **best number of clusters is 15**

Graphical user interface, text

Description automatically generated

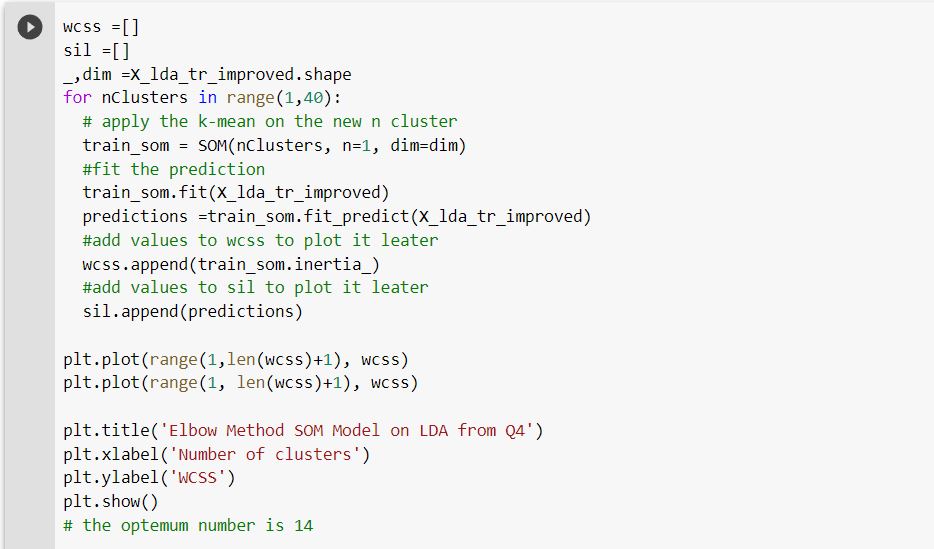


## **We determine the best number of neurons:**

we use SOM algorithm with the best features or dimensionality from Q4 and Q5, elbow rule, plot the distortion score (a.k.a inertia) vs the number of neurons to determine the optimal number of neurons for SOM we find that there are many number of optimal neurons and Plot the initial and final Neuron positions by the distance map function and color map function**. the optimum number is 14**.

* **Using the elbow rule, plot the distortion score (a.k.a inertia) vs the number of neurons (max 30**

**neurons)**

****



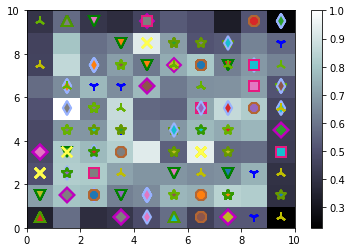
* **Determine the optimal number of neurons for SOM**

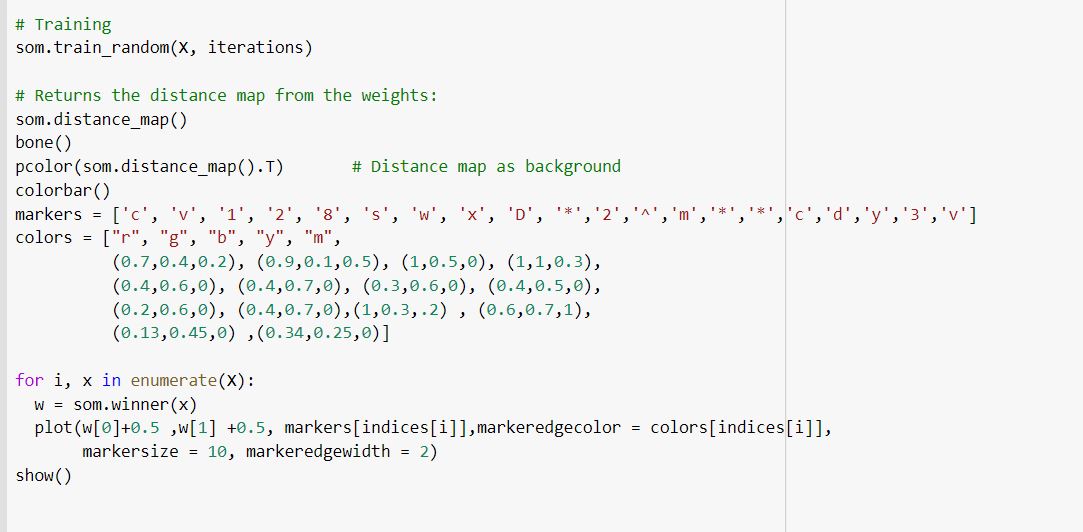
The optimal number of neurons is 14

* **Plot the initial and final Neuron positions**

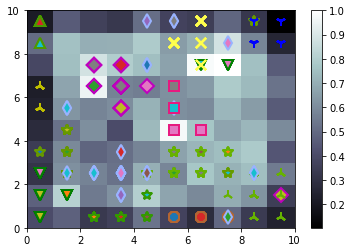


**Initial position of neurons: I plot the neuron and winner after initialization before train the model**





**Final positions of neurons after train model as we see he make cluster to data**



## **using DBSCAN and Plot the epsilon and midpoints**

As shown in the figures bellow we did the following:

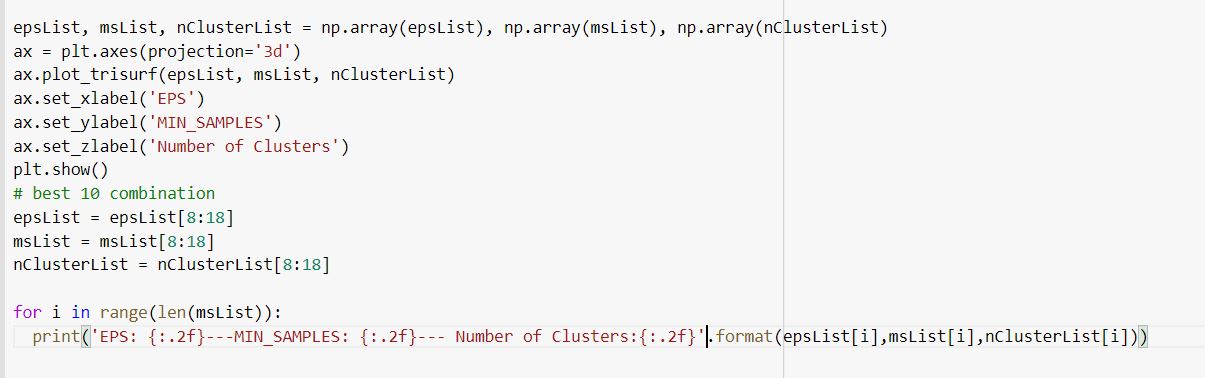
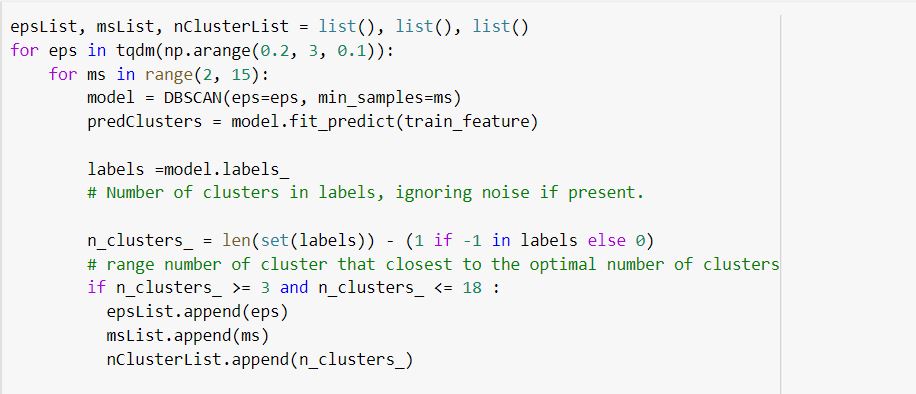
* After Tune the epsilon (0.2-3) and minpoints (2-15) values in the given intervals to obtain same number of

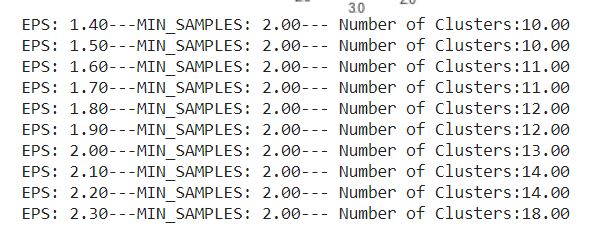
clusters in Q6 by using DBSCAN

* Plotting the epsilon and minpoints values using a 3D figure to show the best

10 combination of epsilon and minpoints that brings you closer to the desired cluster number.

**Code :**



**The output of the 10 combination of epsilon and minpoints that brings you closer to the desired cluster number**: 

* 3D figure to show the best 10 combination of epsilon and minpoints.