**Implementation:**

1. **Agglomerative clustering:** isa hierarchical clustering that is used to group objects in clusters based on their similarity. The result of the agglomerative clustering is a tree-based representation of the objects which called dendrogram.

**Dendrogram has been used to see the best number of clusters using both linkages average and ward.**

|  |  |
| --- | --- |
| Average of agglomerative clustering | Ward of agglomerative clustering |
|  |  |

To reach the best results, two experiments have been applied, one for the best silhouette score and the other for the best kappa score.

**1- The first experiment:** based on the highest silhouette score. The main idea is tunning agglomerative clustering model using ward linkage method and that gives the highest silhouette score.

The following figure indicates that, the best number of clusters = 11:

Chart, line chart

Description automatically generated

1. **The second experiment**: is based on the highest kappa score by tunning agglomerative clustering model using average linkage method.

The following graph shows that the best number of clusters is 11:

Chart, line chart

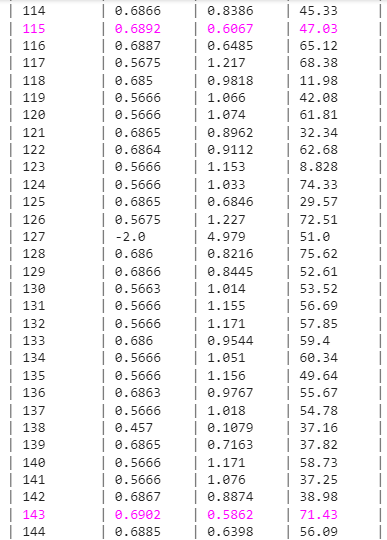
Description automatically generated

**DBSCAN clustering model:** DSCAN is one of the unsupervised clustering techniques that can divides the data points into specific groups that contain the similar points.

Two experiments have been applied by Bayesian optimization one to get the highest silhouette score and the other is to get the highest kappa score.

1. **The third experiment:** Tunning DSCAN using Bayesian optimization to get the best parameter epsilon and minimum samples based on the highest silhouette score.

Here, the Bayesian optimization is with n=250 iteration to get the best parameter of DSCAN which has Min sample = 71 and epsilon = 0.586222874 at the highest silhouette score 0.690042

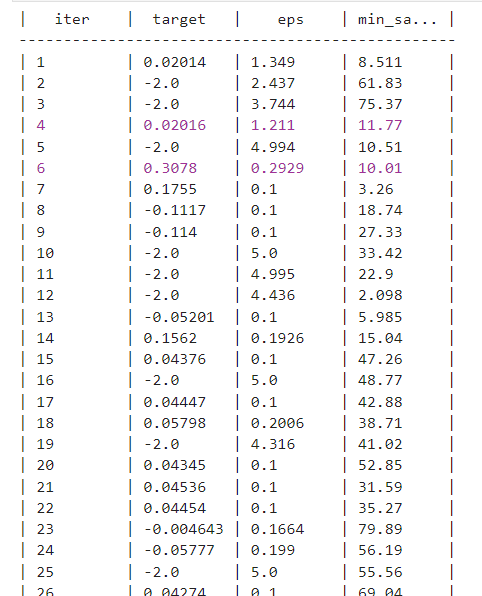
After tuning, the best number of clusters is 11, so a DBSCAN model has been built with these best hyperparameters. The following figure shows the best hyperparameters:

1. **The fourth experiment: is** tunning DBSCAN clustering model using Bayesian optimization to get the best parameter epsilon and minimum samples based on highest kappa score.

Bayesian optimization has been tried with n = 350 iterations to get the best parameter of DBSCAN which is Min sample = 10 and epsilon = 0. 292886058969 at the highest kappa score is 0. 3078, and the best number of clusters is 17

After that DBSCAN model has been built with these best hyperparameters.

**The following figure shows the best hyperparameters for DBSCAN model:**



**verification and validation:**

1. **For the first experiment:** The following table contains the evaluation metrices silhouette and kappa score for both training and testing datasets at k=11 for the ward agglomerative clustering:

|  |  |  |
| --- | --- | --- |
|  | **Evaluation metric** | **Value** |
| **Training** | Actual kappa score | 0.02303368 |
| Silhouette score | 0.7005552 |
| **Testing** | Actual kappa score | 0.00511996 |
| Silhouette score | 0.63134602 |

The following cluster distribution shows that the algorithm can isolate class five perfectly on both training and testing datasets, but the train and test distributions were not consistent.

|  |  |
| --- | --- |
| Cluster distribution for training | Cluster distribution for testing |
|  |  |

1. **For the second experiment:**

The following table contains the evaluation metrices silhouette and kappa score for both training and testing datasets at k=11 for the linkage agglomerative clustering:

|  |  |  |
| --- | --- | --- |
|  | **Evaluation metric** | **Value** |
| **Training** | Actual kappa score | 0.183672 |
| Silhouette score | 0.49738 |
| **Testing** | Actual kappa score | -0.021478 |
| Silhouette score | 0.54609 |

The following cluster distribution shows that the algorithm can isolate class5 perfectly on both training and testing datasets, also the majority of all classes are overlapped in predicted cluster 2 on the training dataset and the majority of all classes overlapped in predicted cluster 6 on the testing dataset ,so the train and test distributions were not consistent.

|  |  |
| --- | --- |
| Cluster distribution for training | Cluster distribution for testing |
|  |  |

1. **For the third experiment:** The following table contains the evaluation metrices silhouette and kappa score for both training and testing datasets, as the Bayesian has been utilized to get the highest silhouette score regardless kappa score.

|  |  |  |
| --- | --- | --- |
|  | **Evaluation metric** | **Value** |
| **Training** | Actual kappa score | 0.159561 |
| Silhouette score | **0.690042** |
| **Testing** | Actual kappa score | -0.061771 |
| Silhouette score | **0.5762** |

The following cluster distribution shows that, cluster 1 contains the majority of classes 1 and 3, and outlier cluster (-1) contains less overlapped minority of all classes on the train and the test distribution are not consistent with train because, cluster 0 in testing contains overlapped classes 1 and 3 but on the training is cluster 1.

|  |  |
| --- | --- |
| Cluster distribution for training | Cluster distribution for testing |
|  |  |

1. **For the fourth experiment:** The following table contains the evaluation metrices silhouette and kappa score for both training and testing datasets, as the Bayesian has been utilized to get the highest kappa score.

|  |  |  |
| --- | --- | --- |
|  | **Evaluation metric** | **Value** |
| **Training** | Actual kappa score | 0.3078 |
| Silhouette score | 0.638944 |
| **Testing** | Actual kappa score | -0.021478 |
| Silhouette score | 0.54609 |

The following cluster distribution shows that, cluster 1 is overlapped all points of classes 1,3 and also cluster -1 contains some points of all classes and there are no pure class and the same thing for testing.

|  |  |
| --- | --- |
| Cluster distribution for training | Cluster distribution for testing |
|  |  |

**Results and analysis:**

1. **For the first experiment:**

The following figures show a comparison between TSNE actual labels and TSNE with predicted label of agglomerative clustering model with linkage ward with k=11

|  |  |
| --- | --- |
| Actual agglomerative TSNE plot for training | Predicted agglomerative TSNE plot for training |
|  |  |
| Actual agglomerative TSNE plot for testing | Predicted agglomerative TSNE plot for testing |
|  |  |

From TSNE plot on training, cluster 8 that contains most of class1, class3 and minority of class4, so  
class5 is well separated.

From TSNE plot on testing, cluster ten that contains most of class1, class3 and minority of class4, and  
class5 is well separated.

1. For the second experiment:

The following figures show a comparison between TSNE actual labels and TSNE with predicted label of agglomerative clustering model with linkage average with k=11

|  |  |
| --- | --- |
| Actual agglomerative TSNE plot for training | Predicted agglomerative TSNE plot for training |
|  |  |
| Actual agglomerative TSNE plot for testing | Predicted agglomerative TSNE plot for testing |
|  |  |

From the TSNE plot on training, the predicted training cluster 2 contains most of the classes points except class5 which is well isolated and also well separated class5.

As we see the cluster 2 contains less overlapped points from four classes except class5, which is well isolated and well separated, train and test distribution are not consistent as cluster 6 on training contains nearly no points, while cluster 6 on the test contains most of the classes points except class5.

1. **For the third experiment:**

The following figures show a comparison between TSNE actual labels and TSNE with predicted label of DBSCAN clustering model with k=11

|  |  |
| --- | --- |
| Actual DBSCAN T-SNE plot for training | Predicted DBSCAN T-SNE plot for training |
|  |  |
| Actual DBSCAN T-SNE plot for testing | Predicted DBSCAN T-SNE plot for testing |
|  |  |

From the TSNE plot on training, the predicted training cluster 3 contains all values of classes 1, 3 and also cluster 0 contain some points of all classes as outliers.

Cluster 2 contains all points of classes 1 and 3 and cluster -1 contains some points of all clusters, so the train and test distribution are not consistent.

1. **For the fourth experiment:**

The following figures show a comparison between TSNE actual labels and TSNE with predicted label of DBSCAN clustering model with k=17

|  |  |
| --- | --- |
| Actual DBSCAN TSNE plot for training | Predicted DBSCAN T-SNE plot for training |
|  |  |
| Actual DBSCAN T-SNE plot for testing | Predicted DBSCAN T-SNE plot for testing |
| Chart, scatter chart  Description automatically generated |  |

From the TSNE plot on training and testing, cluster 3 contains all classes 1,3 there are not purely classes.