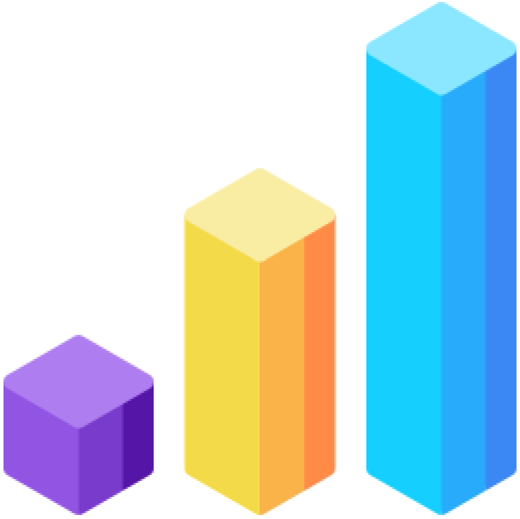


SIT720 Machine Learning

Task 6.2

Michael Rideout

Student Id: 225065259

Introduction

This report provides an analysis of clustering algorithms with the main goal of finding the optimal number of groups or clusters, related to the target variable 'SensorLocation'. The dataset utilised is the Microclimate sensors data, provided by the city of Melbourne. It is a time series dataset for a group of sensors (11 sensors in total), measuring environmental factors such as wind speed, particulate matter, nose etc.

Data Preprocessing

The following dataset preprocessing steps were performed:

- All rows for the features 'SensorLocation' and 'LatLong' were dropped
- 'LatLong' was split into two features 'Latitude' and 'Longitude' and then dropped.
- Any missing numeric features were replaced with their mean
- Any missing categorical features were

Item 1

Sub Item 1-A

To determine the optimal number of clusters I employed two approaches and then using kmeans on the resulting number from each method, the optimal number of clusters was achieved. Those methods were:

- **Ground Truth Unique Class Count** - The target feature 'SensorLocation' is a categorical feature that directly provides the ground truth for the number of distinct classes and therefore groups. This number represents the ideal number of clusters that would be discovered by any other automated means. This approach determined that there are 11 unique clusters.
- **Elbow Method -** The elbow method is an automated mechanism for determining the optimal number of clusters. It is an exploratory technique which attempts to find natural groupings based on variance. It plots the Within Cluster Sum of Squares against the cluster count and pinpoints the number of clusters whereby adding more clusters gives diminishing returns for minimising WCSS. In this context, the optimal number of clusters was 6. Figure 1 shows the

number of clusters plotted against the WCSS

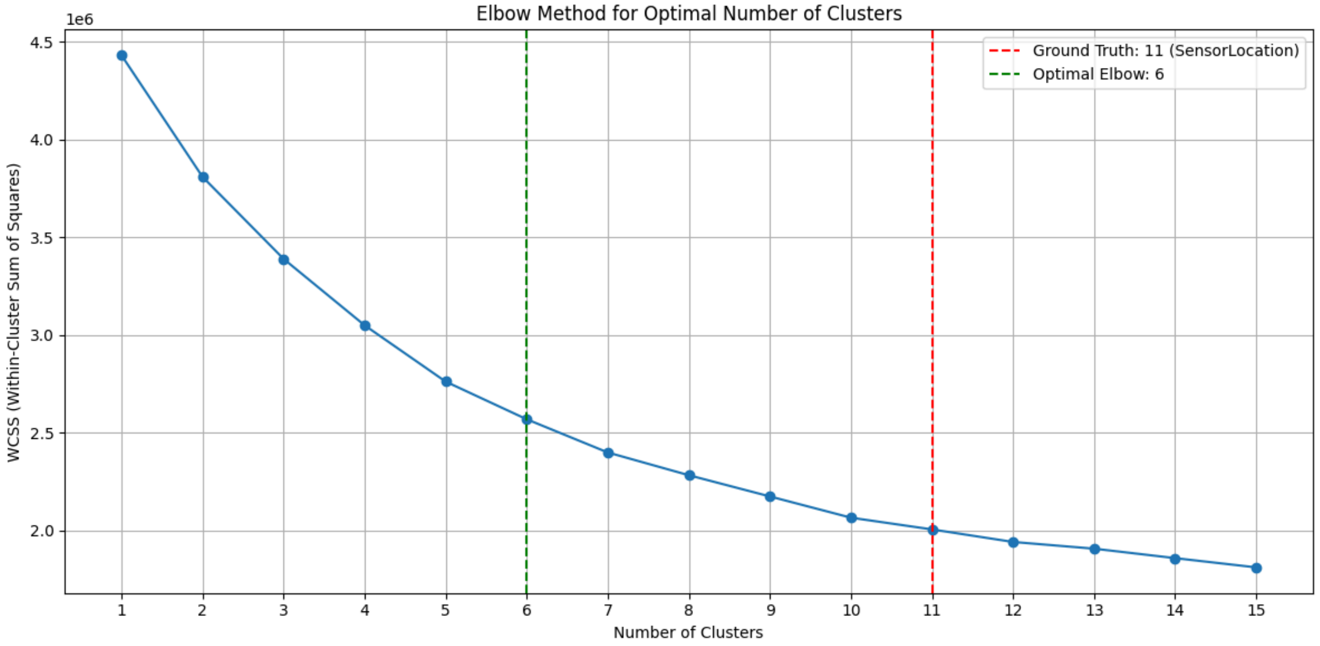


Figure 1 - Elbow Method for Optimal Number of Clusters

Kmeans clustering was performed on the dataset twice, once per optimal cluster number strategy, and are as follows:

Strategy	Number of Clusters	Adjusted Rand Index	Normalised Mutual Information	Purity Score
Elbow	6	0.0656	0.1818	0.2453
Ground Truth	11	**0.1939**	**0.3307**	**0.3597**

Table 1 - Optimal Cluster Strategy Results

As evidenced by this, the ground truth optimal cluster count produced better clustering metrics across all evaluation measures.

Sub Item 1-B

It is most definitely possible to reduce the number of features used in clustering models, and in some instances this may be desirable. Excessive dimensions in data may introduce noise, increase computational costs and may lead to less interpretable results. Two approaches were taken to investigate if dimensionality reduction would increase cluster cohesion. They were:

Automatic Feature Selection

Scikit Learn's SelectKBest strategy for feature selection bases feature importance based on univariate statistical tests. Essentially the test measures the dependency between the feature in question and the target. Feature counts from 2 to 14 were evaluated using kmeans with the ARI metric scoring each iteration. Figure 2 plots the results.

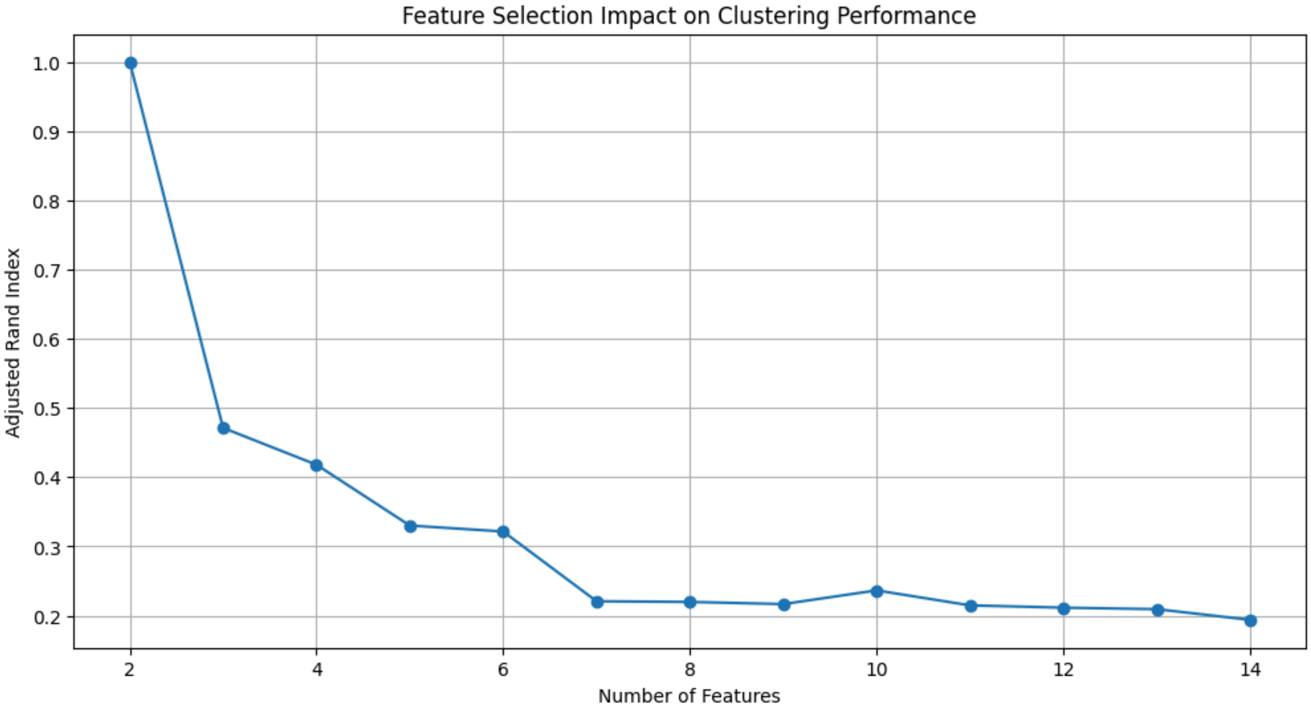


Figure 2 - Automatic Feature Selection

A feature count of 2 attributes produce an ARI score of 1, which is 100% accuracy. Therefore the optimal feature count is 2, and those two features were 'Latitude' and 'Longitude'.

Principal Component Analysis

PCA was used to investigate if dimensionality reduction through variance capture would result in comparable. Four components were used and the total cumulative variance captures was 61.5%. The results were poor in that an ARI score of only 0.07 was achieved.

Sub Item 1-C

The optimal number of clusters was found to be 11, using the strategy of counting the unique classes of the ground truth target feature. On top of this, through automatic feature reduction, the most cohesive and accurate clustering was achieved with only 2 features, those being ‘Latitude’ and ‘Longitude’. Aligning the number of clusters with the class count of the target variable is standard practice and intuitively provides the best results. As this is a time series data of environmental sensor readings, these features introduce noise into the clustering similarity measures. As the target feature is a location, it stands to reason that the two features that represent a location would provide the only data needed to enable a perfect ARI score. Figure 3 shows the cluster assignments in 2 dimensional space with the x-cordinates representing Latitude and the y-cordinates Longitude (scaled values not actual) for Kmeans 11 clusters.

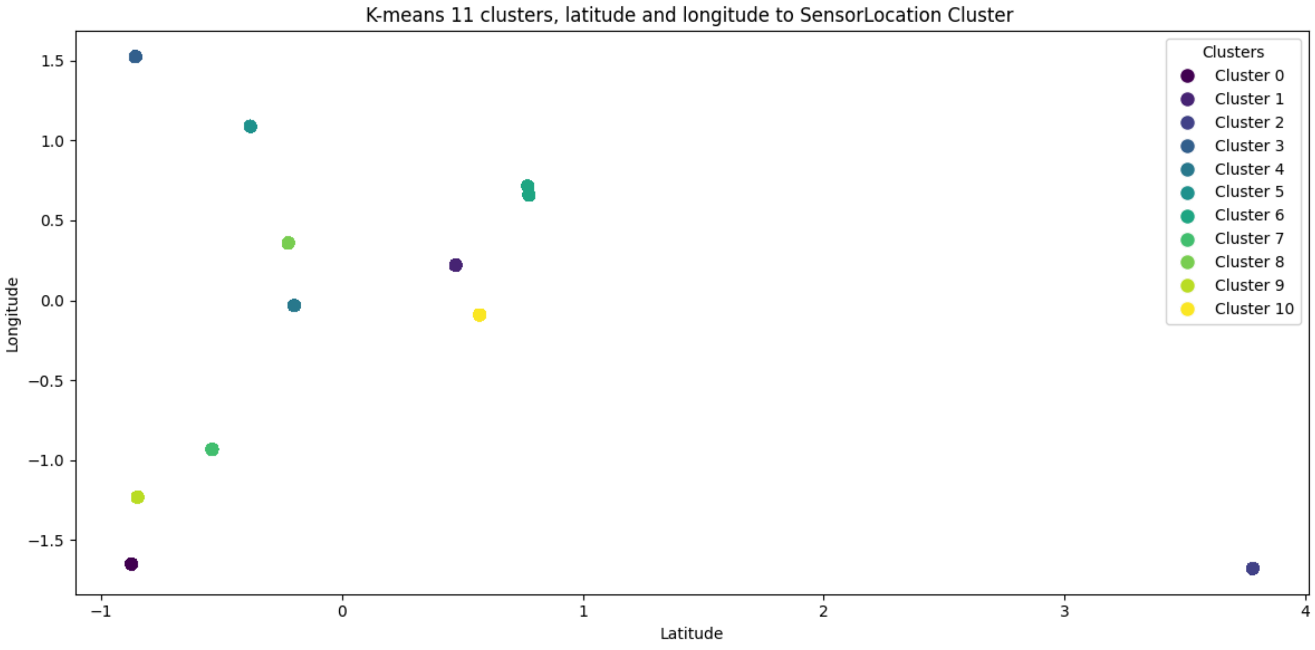


Figure 3 - Scatterplot of Kmeans 11 Clusters in Latitude by Longitude Space

All metrics for Kmeans 11 Clusters with 2 independent features:

Number of Clusters Assigned	Adjusted Rand Index	Normalised Mutual Information	Purity Score
11	1	1	1

Table 2 - Kmeans 11 Clusters 2 Features Evaluation Metrics

Item 2

The two non Kmeans, non shape based clustering algorithm alternatives used were:

Gaussian Mixture Model

A Gaussian Mixture Model is a probabilistic machine learning model in which each data point is given a probability that it is a member of each group. The model was initialised with 11 clusters and the two optimal independent features. The results were:

Number Of Clusters Assigned	Adjusted Rand Index	Normalised Mutual Information	Purity Score
11	1	1	1

Table 3 - GMM 11 Clusters 2 Features Evaluation Metrics

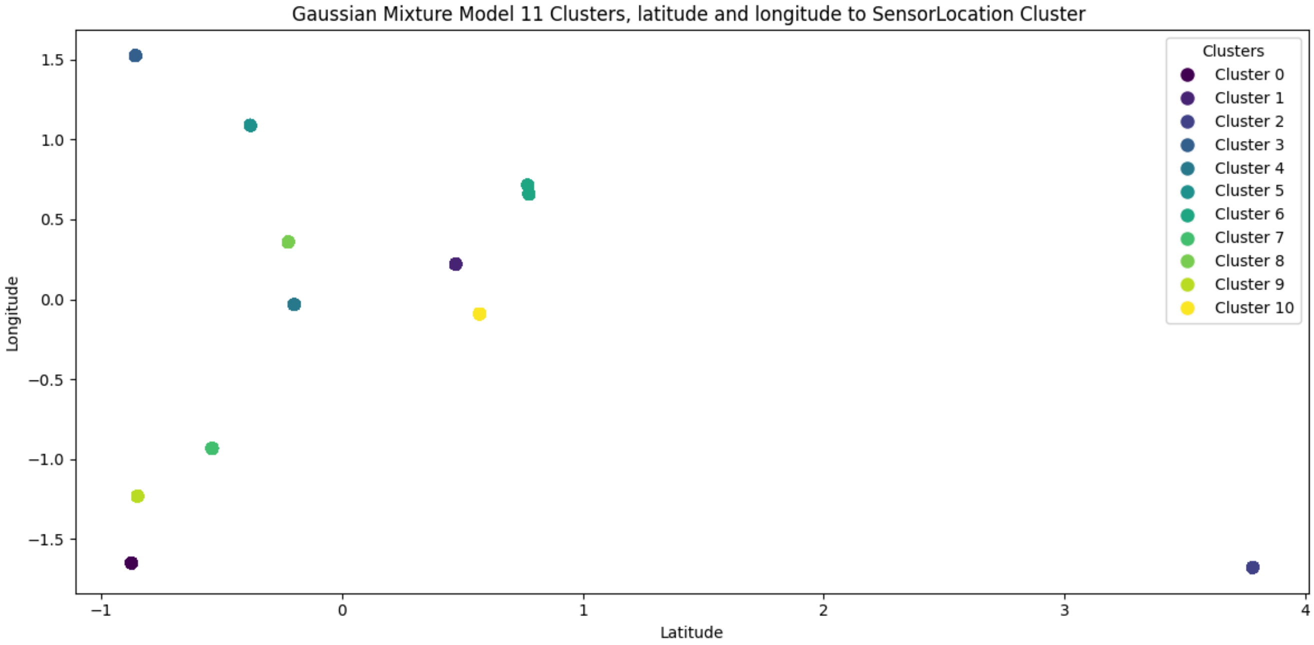


Figure 4 - Scatterplot of GMM 11 Clusters

Achieving perfect scores across the board, identically to the Kmeans 11 cluster evaluation metrics scores, GMM produced an excellent model.This shows that GMM, like Kmeans was able to use the geographical locations of the sensors to represent each geographical location.

BIRCH

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is a hierarchical clustering algorithm that builds summaries of data through the use of a tree structure. The model was initialised with 11 clusters and given the two independent features to group, however this model grouped the data into 4 clusters as seen in Figure 5. The performance of the model was quite poor with an ARI score of 0.3115.

Number Of Clusters Assigned	Adjusted Rand Index	Normalised Mutual Information	Purity Score
4	0.3115	0.6522	0.3555

Table 4 - BIRCH 4 Clusters 2 Features Evaluation Metrics

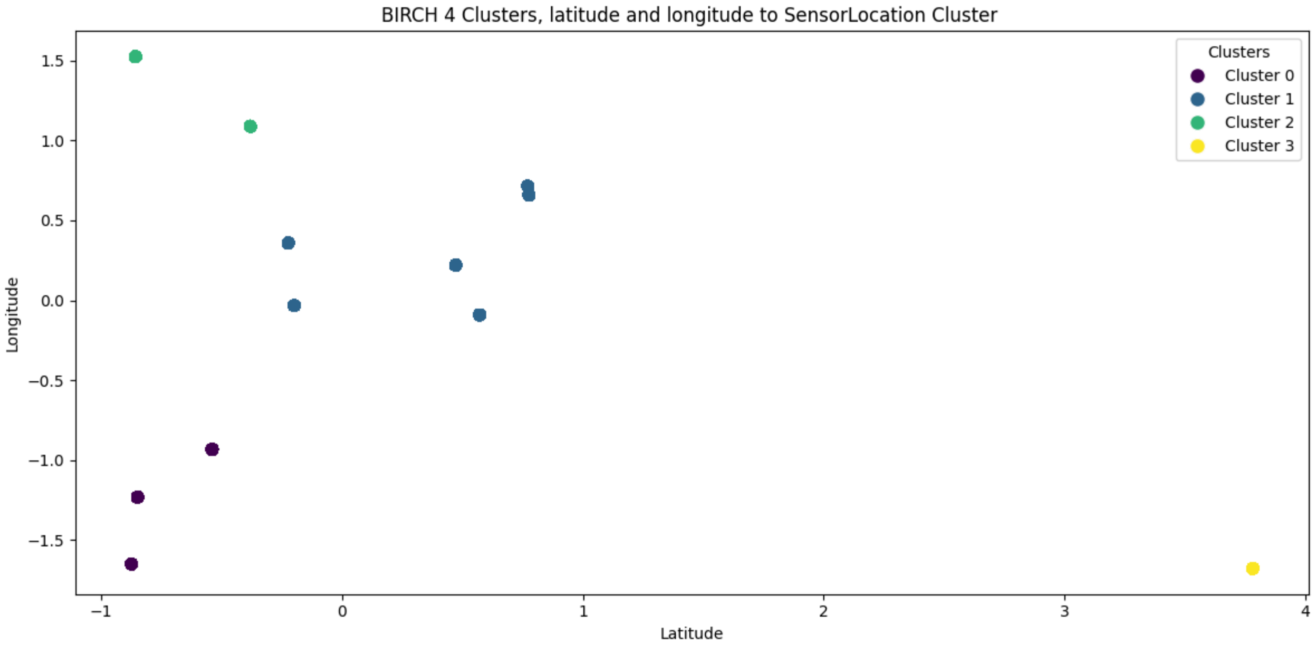


Figure 5 - Scatter Plot of Birch 4 Clusters

The GMM model performed identically to the Kmeas 11 Cluster model, having a perfect score across all measures. BIRCH on the other hand, even though it was initialised with 11 clusters, its auto cluster discovery process determined that 4 clusters to be the most optimal. One potential cause for this could be the threshold parameter limiting the size of clusters and therefore aggregate nearby clusters into one. An optimisation search of its parameters would be prudent in future investigations

Item 3

The two shape based clustering algorithms investigated were:

K-shape

K-shape is a clustering algorithm designed for time series data. It uses a shape based distance measure which is not susceptible to data scale and focuses on the shape patterns of the time series. The model was initialised with 11 clusters, however its final optimisation settled on only 8. The results were:

Number Of Clusters Assigned	Adjusted Rand Index	Normalised Mutual Information	Purity Score
8	0.7619	0.9139	0.7374

Table 5 - K-shape 8 Clusters 2 Features Evaluation Metrics

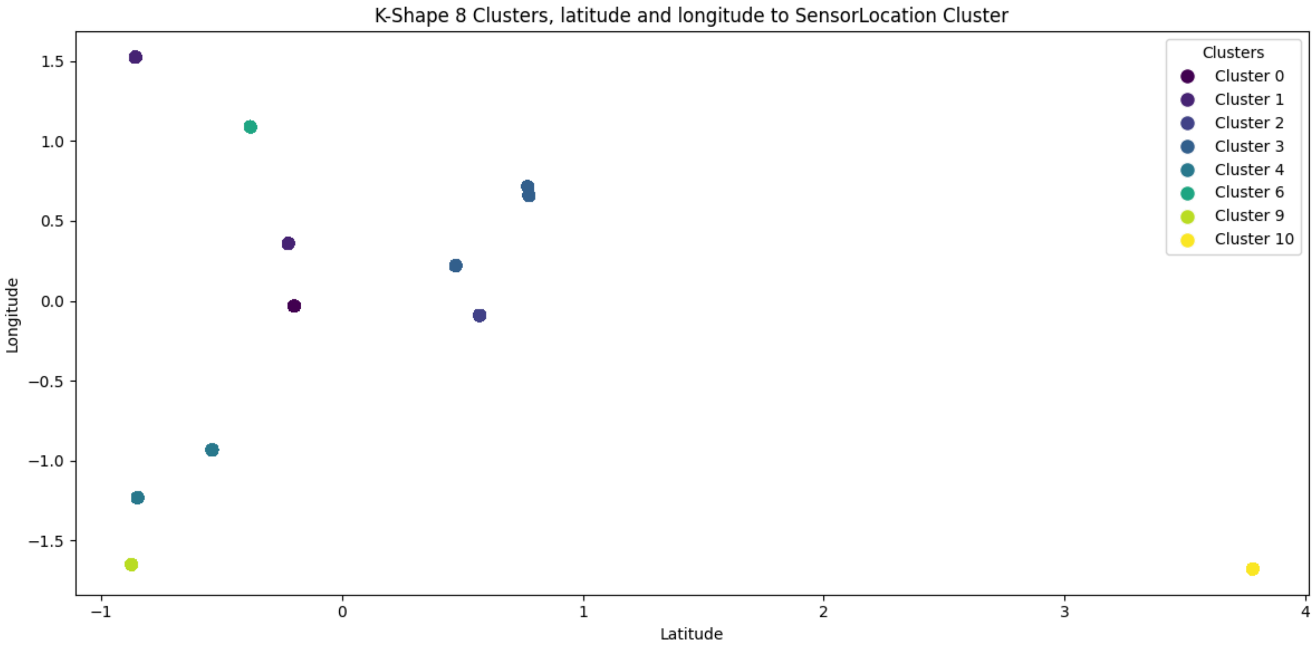


Figure 6 - Scatterplot of K-shape 8 Clusters

K-shape is an algorithm that is designed to work with temporal data. Restricting the dataset to the two features of Latitude and Longitude removed the temporal aspect of the data. Regardless of this, K-shape still managed to perform well with an ARI score of 0.7619 and a cluster count of 8.

OPTICS

Optics (Ordering Points to Identify the Clustering Structure) is a density based clustering algorithm that improves upon DBSCAN. Instead of producing clusters explicitly, it creates a reachility plot that shows how points bunch together at different density levels. The algorithm was not initialised with a cluster count so it automatically determined there to be 12 clusters according to its parameters. The evaluation metrics were:

Number Of Clusters Assigned	Adjusted Rand Index	Normalised Mutual Information	Purity Score
12	0.9717	0.9861	1

Table 6 - Optic 12 Clusters 2 Features Evaluation Metrics

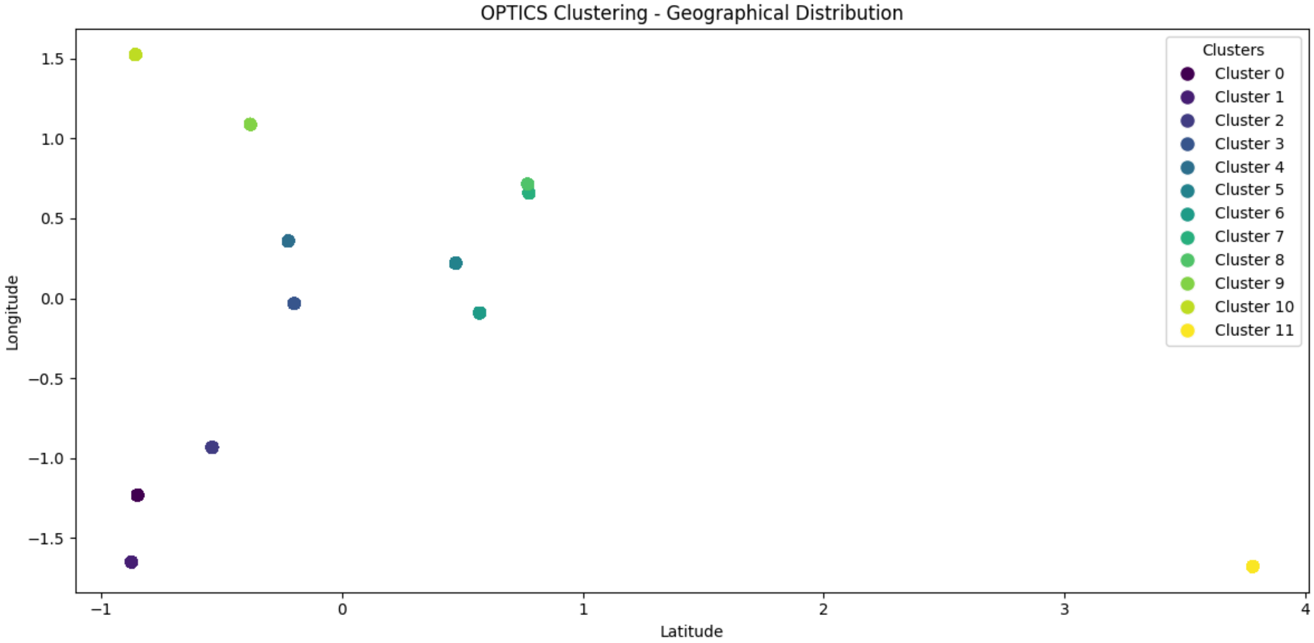


Figure 7 - Scatterplot of OPTICS 12 Clusters

Considering that the algorithm wasn't given any guidance on the desired cluster count, OPTICS performed very well, determining that there are 12 clusters (close to the ground truth class count of 11) and an ARI score of 0.9717.

Item 4

The quality of clusters determined in Q1c, Q2 and Q3 are summarised in table 7. Both K-means and GMM achieved perfect representations of the potential cluster groups as shown by all evaluation metrics (ARI, NMI and Purity) achieving a score of 1, the highest possible score.

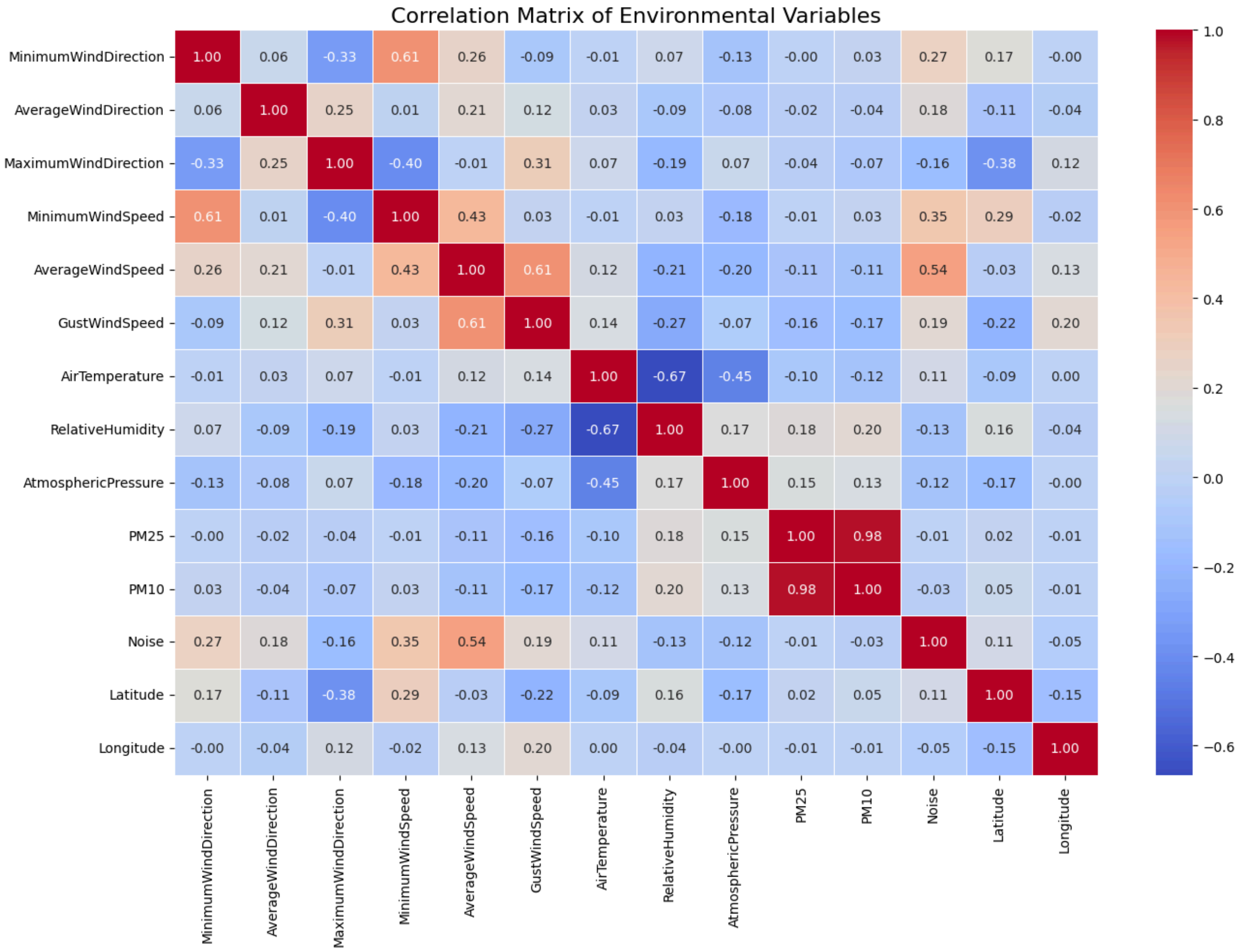
Algorithm	ARI	NMI	Purity	N Clusters Found
K-means (11 clusters)	1	1	1	11
Gaussian Mixture Model	1	1	1	11
OPTICS	0.9717	0.9861	1	12
K-Shape	0.7619	0.9138	0.7374	8
Birch	0.3115	0.6522	0.3555	4
K-means (all features, 11 clusters)	0.1939	0.3307	0.3597	11
K-means (PCA, 11 clusters)	0.076	0.1778	0.2642	11

Tabel 7 - Comparative Evaluation Metrics for All Clustering Algorithms

OPTICS has to be given a special mention as it achieved an almost perfect score across all metrics. It also managed to infer a cluster count of 12, close to the ground truth class count of 11, without any cluster count input or hints. K-shape achieved a moderately reasonable result, especially given its focus on temporal data. BIRCH didnt perform well at all. Its cluster count determination was low and its ARI score quite low. Having said that, all algorithms performed better than a K-means clustering when using the whole feature set (ARI of 0.1939) This highlights the fact that feature selection is a very important step when trying to find coherent and well formed clusters in data. Optimal cluster count, feature selection and data preprocessing are crucial steps in cluster analysis than require much effort and attention

Item 5

The relationship between independent numerical features was achieved using a correlation matrix, and visualised as a heatmap in Figure 8. The top 10 correlations between independent variables are then listed in Table 8.



	MinimumWindDirection	AverageWindDirection	MaximumWindDirection	MinimumWindSpeed	AverageWindSpeed	GustWindSpeed	AirTemperature	RelativeHumidity	AtmosphericPressure	PM25	PM10	Noise	Latitude	Longitude
MinimumWindDirection	1.00	0.06	-0.33	0.61	0.26	-0.09	-0.01	0.07	-0.13	-0.00	0.03	0.27	0.17	-0.00
AverageWindDirection	0.06	1.00	0.25	0.01	0.21	0.12	0.03	-0.09	-0.08	-0.02	-0.04	0.18	-0.11	-0.04
MaximumWindDirection	-0.33	0.25	1.00	-0.40	-0.01	0.31	0.07	-0.19	0.07	-0.04	-0.07	-0.16	-0.38	0.12
MinimumWindSpeed	0.61	0.01	-0.40	1.00	0.43	0.03	-0.01	0.03	-0.18	-0.01	0.03	0.35	0.29	-0.02
AverageWindSpeed	0.26	0.21	-0.01	0.43	1.00	0.61	0.12	-0.21	-0.20	-0.11	-0.11	0.54	-0.03	0.13
GustWindSpeed	-0.09	0.12	0.31	0.03	0.61	1.00	0.14	-0.27	-0.07	-0.16	-0.17	0.19	-0.22	0.20
AirTemperature	-0.01	0.03	0.07	-0.01	0.12	0.14	1.00	-0.67	-0.45	-0.10	-0.12	0.11	-0.09	0.00
RelativeHumidity	0.07	-0.09	-0.19	0.03	-0.21	-0.27	-0.67	1.00	0.17	0.18	0.20	-0.13	0.16	-0.04
AtmosphericPressure	-0.13	-0.08	0.07	-0.18	-0.20	-0.07	-0.45	0.17	1.00	0.15	0.13	-0.12	-0.17	-0.00
PM25	-0.00	-0.02	-0.04	-0.01	-0.11	-0.16	-0.10	0.18	0.15	1.00	0.98	-0.01	0.02	-0.01
PM10	0.03	-0.04	-0.07	0.03	-0.11	-0.17	-0.12	0.20	0.13	0.98	1.00	-0.03	0.05	-0.01
Noise	0.27	0.18	-0.16	0.35	0.54	0.19	0.11	-0.13	-0.12	-0.01	-0.03	1.00	0.11	-0.05
Latitude	0.17	-0.11	-0.38	0.29	-0.03	-0.22	-0.09	0.16	-0.17	0.02	0.05	0.11	1.00	-0.15
Longitude	-0.00	-0.04	0.12	-0.02	0.13	0.20	0.00	-0.04	-0.00	-0.01	-0.01	-0.05	-0.15	1.00

Figure 8 - Independent Features Correlation Heat Map

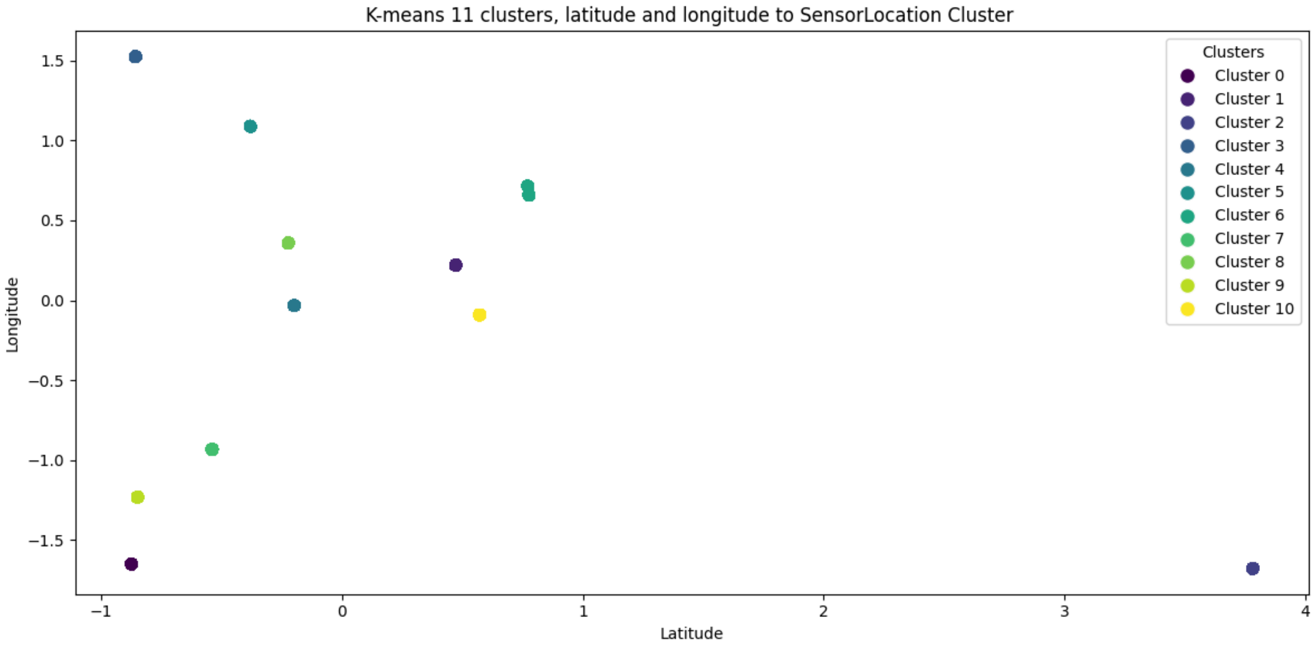
Rank	Variables	Correlation
1	PM2.5 and PM10	0.978
2	Air Temperature and Relative Humidity	-0.666
3	Average Wind Speed and Gust Wind Speed	0.614

4	Minimum Wind Direction and Minimum Wind Speed	0.611
5	Average Wind Speed and Noise	0.541
6	Air Temperature and Atmospheric Pressure	-0.452
7	Minimum Wind Speed and Average Wind Speed	0.431
8	Maximum Wind Direction and Minimum Wind Speed	-0.405
9	Maximum Wind Direction and Latitude	-0.377
10	Minimum Wind Speed and Noise	0.351

Table 8 - Top 10 Independent Variable Correlations

From the table of top correlations some interesting relationships are observed. PM2.5 and PM10 are highly correlated as expected as PM2.5 is a subset of PM10. AirTemperature and RelativeHumidity observe a well known negative correlation, in that the hotter it is relative humidity decreases. An interesting correlation is that wind speed and noise have a positive correlation. The sound of the wind may increase the noise detection.

The best visual representation of a cluster in this report is Figure 3, the scatter plot for the Kmean 11 cluster with 2 features. A copy is presented below:



The model not only achieved full accuracy, the visualisation is intuitive as the scatter plot only has two dimensions, that being latitude and longitude. All row in the dataset aligned to the 11 unique sensor locations, and this clustering was detected by kmeans perfectly.

Item 6

There are most definitely differences in the quality in the clustering solutions. There were two clustering solutions that achieved perfect scores by correctly identifying all 11 sensor locations. The visualisation in Item 5 confirms this. Shape based clustering algorithms (OPTICS and K-shape), although not perfect, their performance was at a level that would be considered good for machine learning models. OPTICS especially performed well across all metrics, even when it had to discover the optimal cluster count itself. BIRCH was the least performant clustering algorithm. Even though Kmeans (a centroid based algorithm), GMM (a distribution based algorithm) and OPTICS (a density based algorithm) all had quite differing approaches to creating clusters, they all performed well at the task at hand.