Clustering Analysis of Air Pollution in South Korea

1st Ray Anthony Pranoto
Faculty of Engineering and Informatics
Universitas Multimedia Nusantara
Tangerang, Indonesia
ray.anthony@student.umn.ac.id

2nd Jovanka Suryajaya Setiawan
Faculty of Engineering and Informatics
Universitas Multimedia Nusantara
Tangerang, Indonesia
jovanka.suryajaya@student.umn.ac.id

3rd Fiena Gunawan
Faculty of Engineering and Informatics
Universitas Multimedia Nusantara
Tangerang, Indonesia
fiena.gunawan@student.umn.ac.id

4th Reva Fakhrana Athira
Faculty of Engineering and Informatics
Universitas Multimedia Nusantara
Tangerang, Indonesia
reva.fakhrana@student.umn.ac.id

5th Fayed Abdul Hakim
Faculty of Engineering and Informatics
Universitas Multimedia Nusantara
Tangerang, Indonesia
fayed.abdul@student.umn.ac.id

***Abstract*—This research discusses the issue of air pollution in South Korea, which has become an urgent public health concern. K-Means and agglomerative clustering are used to categorize air pollution levels into different categories. The analysis results show that agglomerative clustering produces better results, with clusters that are more defined, dense, and compact compared to K-Means. Model evaluation reveals that agglomerative clustering produces a higher Silhouette Score, a lower Davies-Bouldin Index, and a higher Calinski-Harabasz Index. Therefore, agglomerative clustering was identified as a more effective algorithm for this South Korean air pollution dataset, providing an important contribution to the development of more effective public health strategies in reducing the impact of air pollution on the population.**

Keywords—agglomerative clustering, air pollution, clustering, data modeling, k-means

I. Introduction

Air pollution has emerged as a pressing public health concern in South Korea, impacting the well-being of millions of residents annually. This issue is attributed to a confluence of factors, including rapid industrialization, vehicular emissions, and transboundary pollution, such as sand dust originating from China. On days with elevated pollution levels, the South Korean government issues advisories, particularly targeting vulnerable populations like the elderly and those with pre-existing medical conditions, urging them to stay indoors to mitigate health risks. The primary contributors to poor air quality in South Korea are fine particulate matter (PM2.5 and PM10), ozone (O3), nitrogen dioxide (NO2), sulfur dioxide (SO2), and carbon monoxide (CO) [1]. These pollutants have been linked to a range of adverse health outcomes, including respiratory illnesses, cardiovascular diseases, and even cancer. Furthermore, air pollution levels exhibit seasonal variations, with higher concentrations typically observed during winter and spring due to the "Yellow Dust" phenomenon, which transports dust from China and Mongolia [2]. Categorizing air pollution levels into distinct categories can enhance public awareness and facilitate the implementation of appropriate protective measures. Existing research [3] has employed the K-Means algorithm for this purpose. This study aims to conduct a comparative analysis of K-Means clustering and Agglomerative clustering to determine the most effective technique for categorizing air pollution levels in South Korea. By assessing the performance of these methods, this research seeks to contribute to the development of more effective public health strategies aimed at mitigating the impact of air pollution on the population.

II. Theoretical Basis

A. CRISP-DM

CRISP-DM, an abbreviation for Cross Industry Standard Process for Data Mining, is an independent process model not tied to any specific industry for the implementation of data mining projects [4]. Comprising six stages—business understanding, data understanding, data preparation, modeling, evaluation, and deployment—CRISP-DM is designed to be flexible, adaptable, and repeatable. Despite being a common standard in data mining, CRISP-DM faces some challenges, including the absence of a deployment phase, lack of tool support, and the need for updates [5].

B. Machine Learning

Machine learning is the ability of a system to learn from specific training data to automate the process of creating analytical models and solving related tasks [6]. It is a component of artificial intelligence that employs mathematical, statistical, and computational methods to discover patterns, trends, and relationships in data. Machine learning finds applications in various tasks such as classification, regression, prediction, recommendation, and more [7].

C. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a method to uncover insights from data, separate from formal modeling or hypothesis testing [8]. It involves summarizing the statistical characteristics of the data set, with an emphasis on measures of central tendency (mean, mode, and median), measures of dispersion (standard deviation and variance), shape of distribution, and detection of extreme values (outliers) [9].

D. K-Means

Clustering is a technique used in data analysis and machine learning to group similar objects together [10]. It is widely applied in various fields such as customer segmentation, image recognition, and bioinformatics. Clustering uses distance measures, validity indices, and algorithms like K-means and hierarchical clustering. The theory behind clustering involves vector spaces, probability, optimization, and techniques for dimensionality reduction and efficiently handling large datasets. By identifying patterns and structures in data, clustering helps in extracting meaningful insights and making informed decisions.

1) *K-Means Clustering*
K-Means clustering is rooted in partitioning methods, aiming to divide a dataset into K clusters by iteratively assigning data points to the nearest centroid and updating centroids until convergence

[11]. Its objective is to minimize the within-cluster sum of squares (WCSS) and assumes spherical clusters with uniform sizes. Despite its simplicity, K-Means is widely applied due to its efficiency and effectiveness in various domains like image segmentation and customer segmentation.

2) *Agglomerative Clustering*
Agglomerative clustering is a hierarchical method that groups data points by iteratively merging the closest pairs, starting with each point as its own cluster [12]. It uses various distance measures (like Euclidean distance) and linkage criteria (such as single, complete, and average linkage) to determine the similarity between clusters. This process constructs a dendrogram, a tree-like structure representing the nested grouping of points. The method's theoretical foundation includes concepts from vector spaces, graph theory, and matrix computations, providing a robust framework for clustering in diverse applications.

E. Metrics

Metrics are measures used to evaluate the performance or characteristics of a model or algorithm. They provide an understanding of how well a model or algorithm performs in completing a specific task or making predictions. In the context of data analytics and machine learning, metrics play a crucial role in understanding the effectiveness and suitability of a model for the data it encounters.

1) *Silhouette Score*
Silhouette is a clustering evaluation metric that measures how well an object fits into the cluster where it is placed [13]. This metric calculates the distance between the object and other objects in the same cluster (a) and with objects from the nearest different cluster (b), then computes the difference between these two distances. The silhouette value ranges from -1 to 1, where a higher value indicates that the object fits better into its cluster and is far from other clusters. The silhouette provides an overview of how well objects are placed in clusters, with positive values indicating good clustering, while negative values suggest that the object may be placed in the wrong cluster.

2) *Davies-Bouldin*
The Davies-Bouldin index is a clustering evaluation metric that measures the compactness and separation between clusters [14]. This metric considers the average distance between each pair of cluster centers and then compares it with the average intra-cluster distance within those clusters. The lower the Davies-Bouldin value, the better the separation between clusters and the higher the compactness within each cluster. This metric is useful for evaluating the clarity and quality of clustering, where a lower value indicates a better division between clusters.

3) *Calinski-Harabasz*
The Calinski-Harabasz index is a clustering evaluation metric that measures the compactness within clusters and the separation between clusters [15]. This metric is calculated by comparing the dispersion between clusters with the dispersion within clusters. Higher Calinski-Harabasz values indicate that the clusters are compact and well-separated. This metric is useful for assessing the quality of clustering, where higher values indicate better clustering.

F. Previous Research

Prior research has explored the use of clustering techniques to analyze and understand air quality patterns in various urban environments. One study focused on Makassar City, Indonesia, investigating the impact of air pollution on urban planning and policy-making [3]. This research utilized the k-Means clustering algorithm, visualized with Self-Organizing Maps (SOM) and Geographic Information System (GIS), to identify clusters of air pollution sources. By analyzing data on six air quality parameters (TSP, SO2, NO2, CO, O3, and Pb), the study found that traffic, industry, and construction contributed significantly to elevated CO levels, particularly during daytime hours.

Another research utilized cluster analysis to examine ambient air quality data from China's National Urban Air Quality Report [16]. The researchers applied clustering methods to the data, revealing distinct clusters of cities with varying primary pollutants and pollution levels. This finding highlights the potential of clustering to provide valuable insights for targeted interventions aimed at reducing the health impacts of air pollution.

A study in Malaysia investigated the clustering of PM10 time series data from air quality monitoring stations [17]. Employing k-Means, Partitioning Around Medoid, agglomerative hierarchical clustering, and Fuzzy k-Means algorithms, the research identified clusters of monitoring stations based on their PM10 time series patterns. Notably, the results indicated that the clusters were primarily influenced by regional and geographical factors rather than the specific characteristics or activities of the locations.

Research has extended the application of clustering to analyze the prevalence of stunting, a nutritional deficiency, in Indonesia [18]. Using the Agglomerative Hierarchical Clustering Average Linkage method, the study categorized villages into three levels of stunting prevalence based on data from toddlers. This analysis identified villages with low, moderate, and high prevalence of stunting, providing valuable information for targeted interventions and policy development aimed at addressing this health issue.

III. Research Methodology

A. Object of Research

This research focuses on air pollution in South Korea, using a dataset that contains 34,530 rows and 13 columns. The columns used for clustering include PM25, PM10, O3, NO2, SO2, and CO.

B. Methods of Collecting Data

The data collection was conducted through the Kaggle platform using the dataset named 'South Korean Pollution' [19].

C. Methods of Research

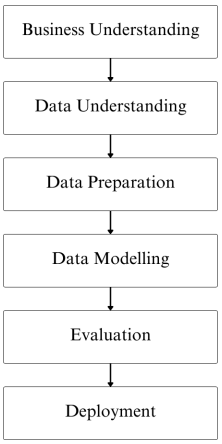

Fig. 1. Research Methodology Stages Flowchart

This research was conducted through several main phases of CRISP-DM as shown in Figure 1, which include Business Understanding, Data Understanding, Data Preparation, Data Modeling, and Evaluation. In the Business Understanding phase, the researchers identified the main problem to be solved, which is clustering air pollution levels in South Korea into low, medium, and high categories based on key pollutant variables. The Data Understanding phase involved exploring the data collected from the Kaggle platform to understand its structure, distribution, and characteristics. During Data Preparation, the researchers selected relevant variables for analysis, namely PM25, PM10, O3, NO2, SO2, and CO. Other steps included examining and handling missing values, as well as standardizing the variables to ensure that each variable is on the same scale, thus improving the accuracy of the clustering results.

In the Data Modeling phase, the researchers determined the optimal number of clusters and created models using the k-means algorithm. After the model was developed, they analyzed the characteristics of each resulting cluster to understand the differences between the groups. Subsequently, the cluster labels were assigned to the initial data to identify the pollution level category of each data point. The next phase is Evaluation, where the built model is assessed to ensure its accuracy and effectiveness in clustering air pollution data. This evaluation is crucial to determine if the model can be relied upon and used for further analysis purposes. The results of the evaluated model will be used for recommendations in the Deployment phase, to aid decision-making regarding air pollution management in South Korea.

IV. Result and Analysis

A. Business Understanding

Air pollution is a serious issue that affects air quality and public health, especially during periods of high pollution levels. According to the OECD Health Report 2023, South Korea experienced 42.7 premature deaths per 100,000 people due to air pollution in 2019. The OECD average was 28.9 deaths per 100,000 people in 2019 [20]. A lack of understanding about the distribution, trends, and patterns of air pollution can indeed hinder effective mitigation efforts and public health protection. Therefore, analyzing air pollution data is crucial for monitoring pollution levels and categorizing air pollution data into high, medium, or low population-level categories using clustering techniques. By gaining a deeper understanding of the air pollution issue, stakeholders can take appropriate actions to protect public health and maintain sustainable environmental quality. This analysis helps in identifying areas with high pollution levels, understanding the sources of pollution, and implementing targeted measures to mitigate pollution and its adverse effects on public health and the environment.

B. Data Understanding

The 'South Korean Pollution' dataset contains information about air pollution in South Korea from the year 2013 to 2022.

TABLE I. Column Details

Column Name	*Description*	*Data Type*
pm25	Particulate Matter (PM2.5) (µg/m3)	numeric
pm10	Particulate Matter (PM10) (µg/m3)	numeric
o3	Ozone (O3) (µg/m3)	numeric
no2	Nitrogen Dioxide (NO2) (ppm)	numeric
so2	Sulfur Dioxide (SO2) (ppm)	numeric
co	Carbon Monoxide (CO)(ppm)	numeric

Table 1 provides a detailed overview of the columns used in this study to analyze air pollution. The data consists of six key air pollutants:

1) *Particulate Matter (PM2.5):* This column represents the concentration of fine particulate matter with a diameter of 2.5 micrometers or less (µg/m3). PM2.5 is of particular concern due to its ability to penetrate deep into the lungs, posing significant health risks.
2) *Particulate Matter (PM10):* Similar to PM2.5, this column denotes the concentration of larger particulate matter with a diameter of 10 micrometers or less (µg/m3). While PM10 particles are generally less harmful than PM2.5, they can still contribute to respiratory problems.
3) *Ozone (O3):* This column indicates the concentration of ozone (µg/m3), a highly reactive gas that can cause respiratory irritation and damage to the lungs. Ozone is primarily formed by chemical reactions between nitrogen oxides and volatile organic compounds in the presence of sunlight.
4) *Nitrogen Dioxide (NO2):* This column displays the concentration of nitrogen dioxide (ppm), a gas primarily emitted from combustion sources like vehicles and power plants. NO2 contributes to smog formation and can cause respiratory problems.
5) *Sulfur Dioxide (SO2):* This column measures the concentration of sulfur dioxide (ppm), a gas primarily released from burning fossil fuels. SO2 contributes to acid rain and can cause respiratory problems.
6) *Carbon Monoxide (CO):* This column provides the concentration of carbon monoxide (ppm), a colorless, odorless, and poisonous gas produced by incomplete combustion of fuels. CO can reduce the

oxygen-carrying capacity of blood, leading to various health issues.

These six columns collectively provide a comprehensive snapshot of air quality, serving as the foundation for our clustering analysis aimed at identifying distinct patterns and characteristics of air pollution in the study area.

C. *Exploratory Data Analysis*

```
# Load Data
df = pd.read_csv('south-korean-pollution-data.csv')
df.head()
```

	Unnamed: 0	date	pm25	pm10	o3	no2	so2	co	Lat	Long	City	District	Country
0	0	2022/2/1	112	31	35	2	1	4	38.2089	127.9495	Bangsan-Myeon	Gangwon	South Korea
1	1	2022/2/2	92	21	35	2	1	0	38.2089	127.9495	Bangsan-Myeon	Gangwon	South Korea
2	2	2022/2/3	60	20	35	1	1	4	38.2089	127.9495	Bangsan-Myeon	Gangwon	South Korea
3	3	2022/2/4	51	27	33	1	1	4	38.2089	127.9495	Bangsan-Myeon	Gangwon	South Korea
4	4	2022/2/5	57	24	27	2	1	5	38.2089	127.9495	Bangsan-Myeon	Gangwon	South Korea

Fig. 2. Load and Reading Data

Figure 2 is an illustration of the output after importing the data. The first step involves importing all the required packages, and the data is imported into the Jupyter Notebook in the form of a CSV file. After importing the data, the next step is to display the data using the .head() code.

```
df.shape
```

```
(34530, 13)
```

Fig. 3. CheckDataset Size

Figure 3 shows that the DataFrame displays the number of rows and columns, in this case, there are 34,530 rows and 13 columns.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34530 entries, 0 to 34529
Data columns (total 13 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  34530 non-null  int64
 1   date        34530 non-null  object
 2   pm25        34530 non-null  int64
 3   pm10        34530 non-null  int64
 4   o3          34530 non-null  int64
 5   no2         34530 non-null  int64
 6   so2         34530 non-null  int64
 7   co          34530 non-null  int64
 8   Lat         34530 non-null  float64
 9   Long        34530 non-null  float64
 10  City        34530 non-null  object
 11  District    34530 non-null  object
 12  Country     34530 non-null  object
dtypes: float64(2), int64(7), object(4)
memory usage: 3.4+ MB
```

Fig. 4. Exploration of Variable Types

Figure 4 is the output of the data information. This DataFrame consists of 34,530 entries with 13 columns. There are seven columns that store numerical data with integer type, while there are two columns that store numerical data with float type. The other four columns contain text or string data.

```
df.describe()
```

	Unnamed: 0	pm25	pm10	o3	no2	so2	co	Lat	Long
count	34530.000000	34530.000000	34530.000000	34530.000000	34530.000000	34530.000000	34530.000000	34530.000000	34530.000000
mean	17264.500000	53.224616	34.757428	34.506371	14.802114	3.553084	4.548422	37.022716	127.247759
std	9968.096734	39.952008	21.046112	17.393329	11.479781	2.963608	2.966530	0.919155	0.613795
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	35.059900	126.138700
25%	8632.250000	20.000000	22.000000	23.000000	6.000000	2.000000	3.000000	36.410900	126.896900
50%	17264.500000	55.000000	32.000000	32.000000	12.000000	3.000000	4.000000	37.132400	127.059200
75%	25896.750000	78.000000	45.000000	43.000000	21.000000	5.000000	6.000000	37.757700	127.716100
max	34529.000000	220.000000	685.000000	152.000000	85.000000	110.000000	281.000000	38.208900	130.821700

Fig. 5. General Statistics of Numeric Variables

Figure 5 is a statistical summary of the DataFrame df, providing information about the data distribution in each numeric column. The mean, standard deviation, minimum value, quartiles (25%, 50%, 75%), and maximum value of each column can be used to evaluate the data spread. For instance, from the summary, we can observe that the PM25 column has a mean of approximately 53.22 with a standard deviation of around 39.95, indicating a considerable variation in the data. Additionally, the quartiles indicate that 50% of the data has PM25 values below 55, while the maximum value is 220, suggesting the presence of outliers in the data

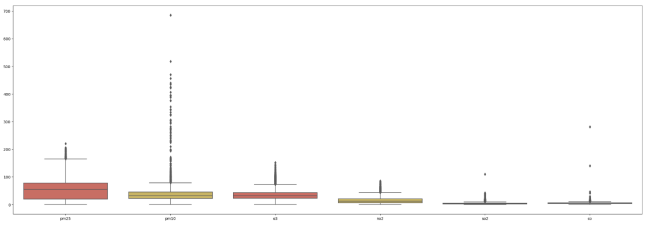

Fig. 6. Check Outliers with Boxplot

Figure 6 depicts boxplots above showing the data distribution for six air quality variables, namely, PM10, O3, NO2, SO2, and CO. Each boxplot illustrates the interquartile range (IQR) of the data, with the middle line representing the median, as well as whisker lines indicating the minimum and maximum values within specific times the IQR range from the first quartile (Q1) and third quartile (Q3). Points outside these whiskers are considered outliers. From the figure, it can be observed that the PM10 variable has many outliers, indicating significant variations in their particulate concentrations. The rest of the variables also have some outliers, but in fewer numbers compared to PM10, suggesting a more consistent data distribution.

D. *Data Preparation*

```
# Select relevant features (numeric)
features = ['pm25', 'pm10', 'o3', 'no2', 'so2', 'co']
X = df[features]
X.head()
```

	pm25	pm10	o3	no2	so2	co
0	112	31	35	2	1	4
1	92	21	35	2	1	0
2	60	20	35	1	1	4
3	51	27	33	1	1	4
4	57	24	27	2	1	5

Fig. 7. Select Variables to Use

Figure 7 shows the code used to select relevant features from a dataframe related to air quality. It only includes the numerical features that contain information of the concentrations of various pollutants. By selecting these columns from the dataframe, we can prepare the data for further analysis or for use in machine learning models.

```
# Check Missing Value
X.isnull().sum()
```

```
pm25    0
pm10    0
o3      0
no2     0
so2     0
co      0
dtype: int64
```

Fig. 8. Check for Missing Values

Figure 8 is used to check for any missing values (missing value) within the dataset. The result indicates that there are no missing values in the columns of the dataframe.

```
# Scaling
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

Fig. 9. Data Standardization

Figure 9 utilizes the StandardScaler from scikit-learn to scale the data in X, ensuring it has a mean of 0 and a standard deviation of 1, a process known as standardization, which is used to normalize the data and ensure that all features contribute equally to the clustering process. The result is stored in the variable X_scaled.

E. *Data Visualization*

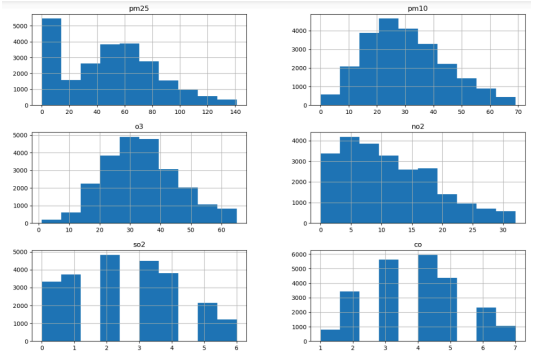

Fig. 10. Distribution Plot

Figure 10 the distributions of the air pollutants in Figure 12 exhibit varying degrees of skewness, providing valuable insights into their characteristics and potential influences.

1) *PM2.5:* The distribution of PM2.5 is skewed to the right, indicating a longer tail towards higher values. This suggests that while many data points fall within a lower range of PM2.5 concentration, a smaller portion of readings experience significantly higher levels.
2) *PM10:* Similar to PM2.5, the distribution of PM10 is also right-skewed, highlighting a similar pattern of predominantly lower values with a smaller number of higher readings. This suggests that PM10 also experiences periodic increases due to specific pollution sources or events.
3) *O3:* The distribution of O3 exhibits right skewness, with a longer tail towards higher ozone values. This suggests that while most readings fall within a lower range, there are occurrences of significantly higher ozone concentrations.
4) *NO2:* The distribution of NO2 appears close to normal, although there might be a slight right skew. This suggests that NO2 levels tend to be more evenly distributed across the range.
5) *SO2:* The distribution of SO2 is skewed to the left, with a longer tail on the left side. This indicates that most data points are concentrated towards higher values, while fewer data points have lower values. This left skew might suggest that SO2 emissions are more consistently elevated.
6) *CO:* The CO distribution also appears near the normal. This suggests that CO levels are relatively evenly distributed across the range.

Understanding these skewness patterns is crucial for selecting appropriate statistical methods for analysis, as skewed distributions can affect the validity and reliability of certain statistical tests and models.

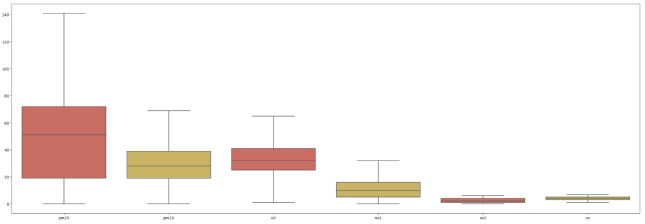

Fig. 11. Numerical Variables Boxplot

Figure 11 depicts boxplots for six air quality variables: PM25, PM10, O3, NO2, SO2, and CO, after the process of outlier removal. From the image, it can be observed that after outlier removal, the data distribution for each variable becomes more consistent. There are no points outside the whiskers of the boxplot, indicating that all values now lie within specific times the IQR range from Q1 and Q3. This suggests that the variation in particulate concentrations is now smaller compared to before.

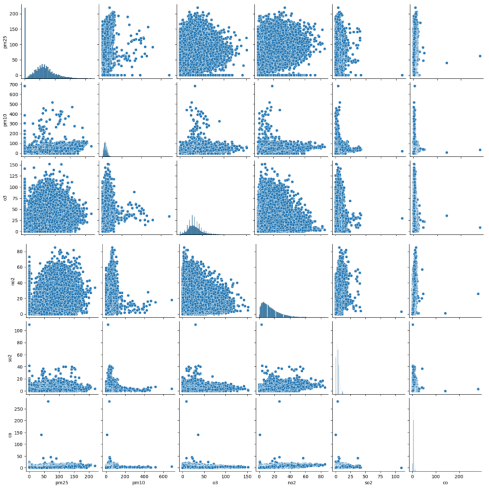

Fig. 12. Pairplot of Numeric Variables

Figure 12 displays a pairplot based on the 'Air Pollution in South Korea' dataset, depicting the relationship between each pair of variables. However, there appears to be no significant correlation among these variables. The pairplot shows the data distribution of each variable as well as scatter plots to examine the patterns of relationships between variables. Although there is variation in the distribution of each variable, no clear patterns indicating strong correlations are evident in each variable.

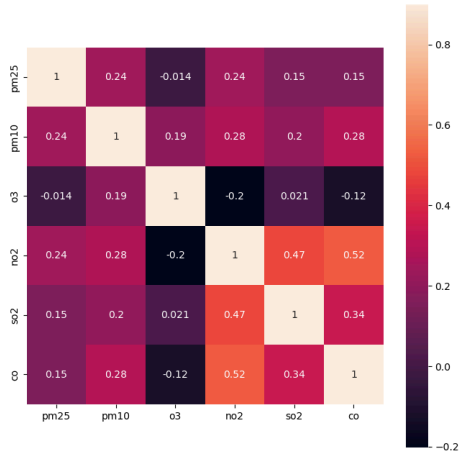

Fig. 13. Heatmap of Numeric Variables

Based on Figure 13, it can be observed that there is a strong correlation between the concentrations of carbon monoxide (CO) and particulate matter (PM2.5), approaching a value of 1. This indicates that changes in one variable are almost always followed by changes in the other variable in the same direction. This strong correlation has significant implications in air pollution analysis, aiding in the understanding of pollution sources and more effective mitigation strategies. Additionally, from the heatmap, the variables NO2 and CO exhibit the strongest positive correlation (0.52), while O3 and NO2 demonstrate the strongest negative correlation (-0.2). In conclusion, the positive correlation between NO2 and CO, along with the negative correlation between O3 and NO2, provides a more comprehensive picture of the relationship between different air pollutants.

F. Data Modelling

```
# Evaluate Silhouette Scores for Different Cluster Numbers
silhouette_scores = []
for i in range(2, 11):  # Start from 2 clusters
    kmeans = KMeans(n_clusters=i, random_state=42)
    cluster_labels = kmeans.fit_predict(X_scaled)
    silhouette_avg = silhouette_score(X_scaled, cluster_labels)
    silhouette_scores.append(silhouette_avg)

# Find the number of clusters with the best Silhouette Score
best_clusters = np.argmax(silhouette_scores) + 2  # Add 2 to account for starting at 2
print("Optimal number of clusters (Silhouette Score):", best_clusters)

plt.plot(range(2, 11), silhouette_scores)
plt.title('Silhouette Scores for Different Cluster Numbers')
plt.xlabel('Number of Clusters')
plt.ylabel('Silhouette Score')
plt.show()
```

```
Optimal number of clusters (Silhouette Score): 2
```

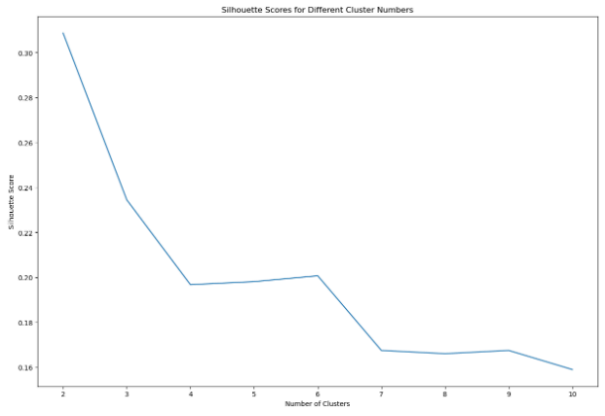

Fig. 14. Determine Cluster Numbers for K-Means Clustering

Figure 14 is the process of determining the optimal number of clusters before performing clustering in data analysis. The code utilizes the Silhouette Score to find the optimal number of clusters. K-Means clustering is performed for each number of clusters in this range, and the Silhouette Score is computed for each. The line graph shows an initial peak in the silhouette score as the number of clusters increases from 2 to 11, indicating the optimal number of clusters where the differentiation between clusters reaches its peak. After this point, the score tends to decrease, indicating that adding more clusters does not significantly contribute to better separation.

```
# K-Means Clustering
kmeans = KMeans(n_clusters=best_clusters, random_state=42)
df['cluster'] = kmeans.fit_predict(X_scaled)
```

```
# Examine the characteristics of each cluster
clusters = df.groupby('cluster')[features].mean()
clusters
```

cluster	pm25	pm10	o3	no2	so2	co
0	76.431790	47.819374	27.776595	27.610609	6.171594	6.927560
1	43.057765	29.035109	37.454625	9.190829	2.405939	3.506143

Fig. 15. Make K-Means Clustering Model

Figure 15 shows K-Means Clustering is employed to group air quality data based on air pollutant concentrations. This algorithm divides the data into groups that exhibit similar characteristics. The clustering process is used to identify hidden patterns in the data, such as similar pollutant concentration patterns across various regions. By analyzing the average pollutant concentrations for each feature within each cluster, the characteristics of each cluster can be explored more deeply. This aids in understanding the differences in air quality among regions and their potential causes.

```
# Visualize clusters using PCA for dimensionality reduction
pca = PCA(n_components=2)
principalComponents = pca.fit_transform(X_scaled)
principalDf = pd.DataFrame(data = principalComponents, columns = ['principal component 1', 'principal component 2'])
principalDf['cluster'] = df['cluster']

plt.scatter(principalDf['principal component 1'], principalDf['principal component 2'],
            c=principalDf['cluster'], cmap='viridis')
plt.title('Clusters of Air Pollution in South Korea (PCA)')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.show()
```

Fig. 16. K-Means Clustering Cluster Visualization Code

Figure 16 is utilized to visualize clusters using Principal Component Analysis (PCA) for dimensionality reduction. Firstly, PCA with two components is employed to reduce the data dimensions to two dimensions. Subsequently, a scatter plot is created using the two principal components as the x and y axes. Each point on the plot is colored based on the clusters previously determined. This plot provides a visual representation of the cluster distribution in the air pollution data in South Korea.

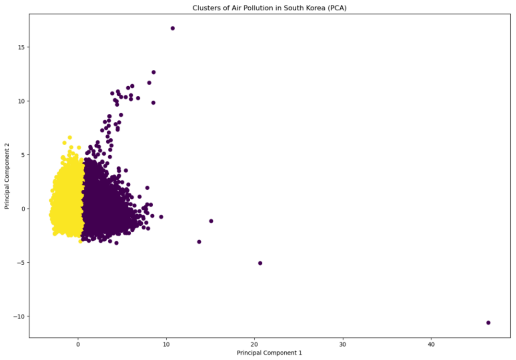

Fig. 17. K-Means Cluster Visualization

Figure 17 depicts a scatter plot that represents the results of the K-Means clustering algorithm applied to air pollution

data in South Korea. The colored points on the plot indicate different clusters identified by the K-Means algorithm. There are two clusters clearly visible in the plot, one colored purple and the other colored yellow, representing different groups of air pollution data.

```
# Create a dictionary to map cluster numbers to labels
cluster_labels = {
    0: 'High',
    1: 'Low',
}

# Add cluster labels to original dataframe
df['pollution_level'] = df['cluster'].map(cluster_labels)
df.head()
```

	Unnamed: 0	date	pm25	pm10	o3	no2	so2	co	Lat	Long	City	District	Country	cluster	pollution_level
0	0	2022/2/1	112	31	35	2	1	4	38.2089	127.9495	Bangsan-Myeon	Gangwon	South Korea	1	Low
1	1	2022/2/2	92	21	35	2	1	0	38.2089	127.9495	Bangsan-Myeon	Gangwon	South Korea	1	Low
2	2	2022/2/3	60	20	35	1	1	4	38.2089	127.9495	Bangsan-Myeon	Gangwon	South Korea	1	Low
3	3	2022/2/4	51	27	33	1	1	4	38.2089	127.9495	Bangsan-Myeon	Gangwon	South Korea	1	Low
4	4	2022/2/5	57	24	27	2	1	5	38.2089	127.9495	Bangsan-Myeon	Gangwon	South Korea	1	Low

Fig. 18. K-Means Map Cluster Label

Figure 18 shows the process of mapping cluster labels to air pollution levels and adds them to the original dataframe. Within the labels, associations are established between cluster numbers and labels ‘High’, and ‘Low’. The `.map()` method is utilized to append these labels to the ‘pollution_level’ column in the dataframe. The resulting output is a dataframe encompassing information on air quality measurement dates, geographic coordinates, as well as location details, along with the mapped pollution levels.

```
df['pollution_level'].value_counts()
```

```
Low     24011
High    10519
Name: pollution_level, dtype: int64
```

Fig. 19. Number of Values in Each Cluster in K-Means Clustering

Figure 19 is used to calculate the frequency of each value in the ‘pollution_level’ column. The results show that there are two categories of pollution levels: Low, with a frequency of 24,011; and High, with a frequency of 10,519.

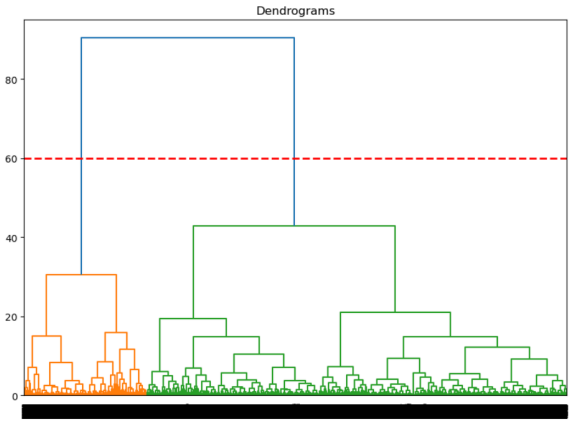

Fig. 20. Determine Cluster Numbers for Agglomerative Clustering

The dendrogram presented in Figure 20 illustrates the hierarchical clustering process applied to the air pollution data, where the vertical axis represents the distance or dissimilarity between clusters. In this dendrogram, the y-axis values indicate the linkage distance between clusters, with higher values representing greater dissimilarity. The red dashed line is drawn at a y-axis value of 60, which serves as the threshold for determining the number of clusters. By cutting the dendrogram at this level, two main clusters are identified, as represented by the number of groups below the red line. The choice of the y-axis value at 60 for the cut-off is based on the observation of significant increases in linkage distance at this point. This value effectively separates the dataset into two major clusters, capturing the primary structure of the data while maintaining a balance between detail and interpretability. If the cut-off were set lower, it would result in a larger number of smaller clusters, which may complicate the analysis without providing substantial additional insights. By forming two main clusters, this analysis allows for a clear categorization of air pollution levels.

```
# Apply Hierarchical Clustering
from sklearn.cluster import AgglomerativeClustering
cluster = AgglomerativeClustering(n_clusters=2, linkage='ward')
df1['cluster'] = cluster.fit_predict(X_scaledDendrogram)
```

```
# Examine the characteristics of each cluster
clusters = df1.groupby('cluster')[features].mean()
clusters
```

cluster	pm25	pm10	o3	no2	so2	co
0	68.557467	33.292516	33.767850	15.359277	3.590493	4.575008
1	0.619145	39.783395	37.040164	12.890543	3.424740	4.457205

Fig. 21. Make Agglomerative Clustering Model

Figure 21 shows the application of the Agglomerative Clustering model using the scikit-learn library. Two clusters (n_clusters=2) are selected using the “ward” linkage method (linkage='ward'). After the model is applied, the 'cluster' column is added to the dataframe df1, providing cluster labels for each data point. Subsequently, an evaluation of the characteristics of each cluster is conducted by calculating the average feature values using the groupby method. This results in a dataframe named clusters, where each row represents one cluster and each column represents the average feature values for that cluster. There are two clusters (cluster 0 and cluster 1), utilizing features such as O3, SO2, PM25, PM10, NO2, and CO for analysis. Each value in the clusters dataframe represents the average feature value for each cluster, providing a better understanding of the characteristics of each cluster in relation to air pollution.

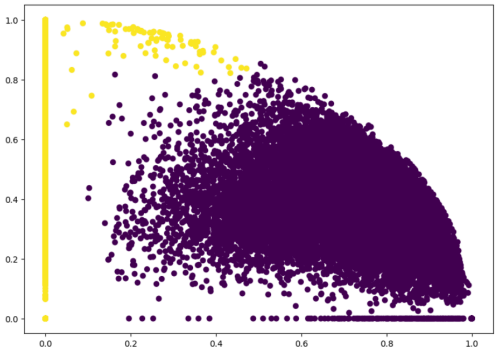

Fig. 22. Agglomerative Cluster Visualization

Figure 22 shows a scatter plot of the outcomes of the Agglomerative clustering technique applied to air quality data in South Korea. The plot displays colored points representing distinct clusters identified by the algorithm. Two clusters are prominently visible on the plot, with one group appearing purple and the other yellow, signifying different sets of air pollution data. The purple color represents a high level of pollution, and the yellow color represents a low level of pollution.

```
# Create a dictionary to map cluster numbers to labels
cluster_labels = {
    0: 'High',
    1: 'Low',
}

# Add cluster labels to original dataframe
df1['pollution_level'] = df1['cluster'].map(cluster_labels)
df1.head()
```

	Unnamed: 0	date	pm25	pm10	o3	no2	so2	co	Lat	Long	City	District	Country	cluster	pollution_level
0	0	2022/2/1	112	31	35	2	1	4	38.2089	127.9495	Bangsan-Myeon	Gangwon	South Korea	0	High
1	1	2022/2/2	92	21	35	2	1	0	38.2089	127.9495	Bangsan-Myeon	Gangwon	South Korea	0	High
2	2	2022/2/3	60	20	35	1	1	4	38.2089	127.9495	Bangsan-Myeon	Gangwon	South Korea	0	High
3	3	2022/2/4	51	27	33	1	1	4	38.2089	127.9495	Bangsan-Myeon	Gangwon	South Korea	0	High
4	4	2022/2/5	57	24	27	2	1	5	38.2089	127.9495	Bangsan-Myeon	Gangwon	South Korea	0	High

Fig. 23. Agglomerative Map Cluster Label

Figure 23 involves data analysis that includes mapping cluster numbers to labels for pollution levels. Two clusters have been defined: cluster '0' mapped as 'High' and cluster '1' as 'Low'. A column named 'pollution_level' is created in the DataFrame to map cluster numbers to corresponding labels. The table contains various air quality measurements such as pm25, pm10, o3, no2, so2, and co at specific geographic locations in South Korea. The data also includes location information such as city, district, and country.

```
df1['pollution_level'].value_counts()
```

```
High    26737
Low      7793
Name: pollution_level, dtype: int64
```

Fig. 24. Number of Values in Each Cluster in Agglomerative Clustering

Figure 24 displays the code for analyzing the 'pollution_level' column in the DataFrame using the 'value_counts()' method. Two unique values in the column, 'High' and 'Low', are shown with 'High' appearing 26,737 times and 'Low' appearing 7,793 times. This result indicates that in this dataset, pollution levels are more frequently categorized as 'High' than 'Low'.

G. Evaluation

```
# Silhouette Score
silhouette_avg = silhouette_score(X_scaled, df['pollution_level'])
print("Silhouette Score:", silhouette_avg)
```

```
Silhouette Score: 0.30871248306260723
```

```
# Davies-Bouldin Index
davies_bouldin = davies_bouldin_score(X_scaled, df['pollution_level'])
print("Davies-Bouldin Index:", davies_bouldin)
```

```
Davies-Bouldin Index: 1.4221567442473262
```

```
# Calinski-Harabasz Index
calinski_harabasz = calinski_harabasz_score(X_scaled, df['pollution_level'])
print("Calinski-Harabasz Index:", calinski_harabasz)
```

```
Calinski-Harabasz Index: 12040.730427930223
```

Fig. 25. K-Means Clustering Model Evaluation

Figure 25 provides a comprehensive evaluation of the K-Means clustering model using three commonly employed metrics: Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. The Silhouette Score, measuring the similarity of data points within a cluster compared to other clusters, achieves a value of 0.309, indicating moderately good clustering. While not exceptional, this score suggests that the clusters are reasonably well-defined, but some overlap or ambiguity in cluster assignment might exist. The Davies-Bouldin Index, evaluating cluster separation, scores 1.422, indicating moderate separation. Lower values are desirable, so this score suggests that the clusters could be better separated. Finally, the Calinski-Harabasz Index, assessing cluster separation relative to within-cluster dispersion, achieves a remarkably high score of 12040.730. This exceptionally high value indicates well-defined and compact clusters with strong separation, highlighting the model's success in grouping similar data points. Overall, the evaluation metrics suggest a reasonable clustering performance, with the Calinski-Harabasz Index indicating strong separation between clusters. However, the Silhouette Score and Davies-Bouldin Index point to some room for improvement in terms of cluster definition and separation.

```
# Silhouette Score
silhouette_avg = silhouette_score(X_scaledDendrogram, df1['pollution_level'])
print("Silhouette Score:", silhouette_avg)
```

```
Silhouette Score: 0.558493111329283
```

```
# Davies-Bouldin Index
davies_bouldin = davies_bouldin_score(X_scaledDendrogram, df1['pollution_level'])
print("Davies-Bouldin Index:", davies_bouldin)
```

```
Davies-Bouldin Index: 0.7353681566051855
```

```
# Calinski-Harabasz Index
calinski_harabasz = calinski_harabasz_score(X_scaledDendrogram, df1['pollution_level'])
print("Calinski-Harabasz Index:", calinski_harabasz)
```

```
Calinski-Harabasz Index: 40242.39735725975
```

Fig. 26. Agglomerative Clustering Model Evaluation

Figure 26 presents the evaluation of agglomerative clustering, revealing promising results. The Silhouette Score registers a strong value of 0.558, indicating a good level of cluster separation. This score suggests that the clusters are well-defined and distinct, with data points exhibiting strong similarities within their respective groups. The Davies-Bouldin Index yields a value of 0.735. This score indicates reasonably good inter-cluster division, suggesting that the clusters are adequately separated. The Calinski-Harabasz Index achieves a remarkably high value of 40242.397. This score highlights the compactness of the formed clusters, signifying that data points within each cluster are highly similar, while those in different clusters are distinctly dissimilar. The evaluation metrics for agglomerative clustering demonstrate a satisfactory performance, with strong cluster definition, adequate separation, and high compactness. These findings suggest that agglomerative clustering has effectively identified meaningful patterns within the data, yielding robust and well-structured clusters.

Comparing the performance of K-Means and agglomerative clustering on the South Korea air pollution dataset reveals distinct strengths and weaknesses. While both algorithms achieved satisfactory results, agglomerative clustering emerges as the more effective approach for this dataset. The agglomerative clustering model achieved a significantly higher Silhouette Score (0.558) compared to K-Means (0.309), indicating better-defined and more distinct clusters. This suggests that agglomerative clustering effectively grouped data points with strong similarities within each cluster. The Davies-Bouldin Index (DBI) yielded a score of 0.735 for agglomerative clustering, indicating better separation compared to the DBI of 1.422 for K-Means. Lower DBI values are preferable, signifying more distinct clusters. The Calinski-Harabasz Index showed a remarkably high score of 40242.397 for agglomerative clustering, significantly surpassing the score of 12040.730 obtained by K-Means. This further reinforces the notion that agglomerative clustering produced more compact and well-separated clusters. Based on the analysis, the evaluation metrics strongly favor agglomerative clustering as the superior algorithm for this dataset, demonstrating its ability to form more distinct, cohesive, and well-separated clusters.

H. Deployment

For the effective deployment of the research findings, it is recommended to implement a multifaceted approach addressing various aspects of air pollution management in South Korea. Firstly, the establishment of a comprehensive sensor-based air monitoring system is crucial. This system should encompass an extensive network of air quality sensors strategically placed in major urban centers and industrial zones across the country. These sensors, integrated with real-time data transmission capabilities, should feed into a centralized online platform accessible to the public. Leveraging technologies like the Internet of Things (IoT) will ensure accurate and up-to-date monitoring of air pollution levels, enabling timely interventions and informed decision-making.

The development and deployment of user-friendly air quality monitoring applications for mobile devices are essential. These applications should provide users with access to real-time data on pollutant levels in their vicinity, accompanied by alerts and recommendations for mitigating exposure to hazardous air conditions. By empowering individuals to make informed choices about their activities and movements based on air quality data, these applications can contribute significantly to public health protection efforts. Incentivizing the adoption of green practices and technologies is paramount. Implementing fiscal benefits and incentives for companies and individuals embracing sustainable practices and investing in environmentally friendly technologies will accelerate the transition towards a cleaner and more sustainable future. Moreover, supporting research and development initiatives aimed at advancing air pollution control methods and promoting renewable energy sources will drive innovation and facilitate long-term environmental sustainability.

V. Conclusion

Based on the results of the modeling analysis, it is evident that both K-Means and agglomerative clustering algorithms applied to the South Korea air pollution dataset to identify distinct patterns and characteristics within the data. For K-Means clustering, the optimal number of clusters was determined through the Silhouette Score, resulting in the identification of two clusters. The evaluation metrics revealed a Silhouette Score of 0.309, a Davies-Bouldin Index of 1.422, and a Calinski-Harabasz Index of 12040.730. These metrics indicate moderately good clustering, with reasonably well-defined clusters, but some room for improvement in terms of cluster separation.

On the other hand, agglomerative clustering yielded more promising results. This algorithm also identified two clusters, as evidenced by the dendrogram and subsequent analysis. The Silhouette Score for agglomerative clustering was notably higher at 0.558, indicating a good level of cluster separation. The Davies-Bouldin Index was lower at 0.735, suggesting better inter-cluster division compared to K-Means. Additionally, the Calinski-Harabasz Index achieved a remarkably high score of 40242.397, highlighting the compactness and distinctiveness of the formed clusters. Based on the evaluation metrics and the distinctiveness of the clusters formed, agglomerative clustering emerges as the superior algorithm for this dataset. It produced more well-defined, cohesive, and compact clusters compared to K-Means.

Acknowledgment

The authors would like to express their sincere gratitude to Mrs. Monika Evelin Johan, our esteemed instructor for Data Modeling, for her invaluable guidance and support throughout the development of this article. Her expertise and encouragement were instrumental in shaping our research and ensuring the clarity and rigor of our findings. We would also like to acknowledge the contributions of our fellow classmates, whose insightful discussions and collaborative spirit fostered a stimulating learning environment. Their support and encouragement were greatly appreciated. It is our hope that this article will contribute to a better understanding of air pollution patterns and provide valuable insights for developing effective mitigation strategies. We believe that this research will serve as a stepping stone for further investigations in this critical area, leading to a healthier and more sustainable environment for all.

References

[1] M. J. Kim, "Air Pollution, Health, and Avoidance Behavior: Evidence from South Korea," *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3361559.

[2] D. Baek, D. Altindag, and N. Mocan, "Chinese Yellow Dust and Korean Infant Health," National Bureau of Economic Research, Cambridge, MA, Oct. 2019. Accessed: May 26, 2024. [Online]. Available: http://dx.doi.org/10.3386/w21613.

[3] S. Annas, U. Uca, I. Irwan, R. H. Safei, and Z. Rais, "Using k-Means and Self Organizing Maps in Clustering Air Pollution Distribution in Makassar City, Indonesia," *Jambura Journal of Mathematics*, vol. 4, no. 1, pp. 167–176, Jan. 2022, doi: 10.34312/jjom.v4i1.11883.

[4] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[5] S. W. Chong, L. T. Jun, and Y. Chen, "A Methodological Review of Systematic Literature Reviews in Higher Education: Heterogeneity and Homogeneity," Center for Open Science, Dec. 2021. Accessed: May 26, 2024. [Online]. Available: http://dx.doi.org/10.31219/osf.io/jn84b.

[6] N. Ketkar, "Machine Learning Fundamentals," in *Deep Learning with Python*, Berkeley, CA: Apress, 2017, pp. 7–16. Accessed: May 26, 2024. [Online]. Available: http://dx.doi.org/10.1007/978-1-4842-2766-4_2

[7] B. Urooj, M. A. Shah, C. Maple, M. K. Abbasi, and S. Riasat, "Malware Detection: A Framework for Reverse Engineered Android Applications Through Machine Learning Algorithms," *IEEE Access*, vol. 10, pp. 89031–89050, 2022, doi: 10.1109/access.2022.3149053.

[8] A. Oluleye, *Exploratory Data Analysis with Python Cookbook: Over 50 recipes to analyze, visualize, and extract insights from structured and unstructured data*. Packt Publishing Ltd, 2023.

[9] "EXPLORATORY DATA ANALYSIS," *Data Science Using Python and R*, pp. 47–67, Mar. 2019, doi: 10.1002/9781119526865.ch4.

[10] "Model-Based Clustering," in *SpringerReference*, Berlin/Heidelberg: Springer-Verlag. Accessed: May 26, 2024. [Online]. Available: http://dx.doi.org/10.1007/springerreference_179266

[11] W. Huang, Y. Peng, Y. Ge, and W. Kong, "A new Kmeans clustering model and its generalization achieved by joint spectral embedding and rotation," *PeerJ Computer Science*, vol. 7, p. e450, Mar. 2021, doi: 10.7717/peerj-cs.450.

[12] J. Jumadi, Y. Yupianti, and D. Sartika, "Pengolahan Citra Digital untuk Identifikasi Objek Menggunakan Metode Hierarchical Agglomerative Clustering," *JST (Jurnal Sains dan Teknologi)*, vol. 10, no. 2, pp. 148–156, Nov. 2021, doi: 10.23887/jst-undiksha.v10i2.33636.

[13] Y. Januzaj, E. Beqiri, and A. Luma, "Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 04, pp. 174–182, Apr. 2023, doi: 10.3991/ijoe.v19i04.37059.

[14] Y. Arie Wijaya, D. Achmad Kurniady, E. Setyanto, W. Sanur Tarihoran, D. Rusmana, and R. Rahim, "Davies Bouldin Index Algorithm for Optimizing Clustering Case Studies Mapping School

Facilities," *TEM Journal*, pp. 1099–1103, Aug. 2021, doi: 10.18421/tem103-13.
[15] I. F. Ashari, E. Dwi Nugroho, R. Baraku, I. Novri Yanda, and R. Liwardana, "Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta," *Journal of Applied Informatics and Computing*, vol. 7, no. 1, pp. 89–97, Jul. 2023, doi: 10.30871/jaic.v7i1.4947.
[16] Z. Li and J. Zheng, "Research on Air Quality Analysis Based on Cluster Methods," *IOP Conference Series: Earth and Environmental Science*, vol. 898, no. 1, pp. 012024, Oct. 2021, doi: 10.1088/1755-1315/898/1/012024.
[17] F. N. A. Suris, M. A. A. Bakar, N. M. Ariff, M. S. Mohd Nadzir, and K. Ibrahim, "Malaysia PM10 Air Quality Time Series Clustering Based on Dynamic Time Warping," *Atmosphere*, vol. 13, no. 4, p. 503, Mar. 2022, doi: 10.3390/atmos13040503.
[18] M. Anggraeni, U. Yudatama, and Maimunah, "Clustering Prevalensi Stunting Balita Menggunakan Agglomerative Hierarchical Clustering," *Media Informatika Budidarma*, vol. 7, no. 1, Jan. 2023, doi: 10.30865/mib.v7i1.5501.
[19] C. Reigada, "South Korean Pollution," *Kaggle*. Accessed: May 26, 2024. [Online]. Available: https://www.kaggle.com/datasets/calebreigada/south-korean-pollution/data.
[20] "Health at a Glance 2023," *OECD iLibrary*. Accessed: May 26, 2024. [Online]. Available: https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-2023_7a7afb35-en.

Work Roles

Here are the roles of each member in the project:

1) Ray Anthony Pranoto (000000666555): Involved in building the Agglomerative Clustering, making the report (conclusion; result and analysis), and the final presentation.
2) Jovanka Suryajaya (00000069834): Involved in building the K-Means Clustering, making the report (introduction; result and analysis), and the final presentation.
3) Reva Fakhrana Athira (00000068621):Involved in creating data visualization, making the report (abstract; result and analysis), and the final presentation.
4) Fiena Gunawan (00000069579): Involved in making the report (theoretical basis; result and analysis), and the final presentation.
5) Fayed Abdul Hakim (00000068732): Involved in making the report (research methodology; result and analysis), and the final presentation.