MACHINE LEARNING-BASED CLUSTERING OF AQUIFER VULNERABILITY TO LEACHATE CONTAMINATION IN IDEMILI NORTH, ANAMBRA STATE, NIGERIA

Author(s): Anyawuike, Ikenna Omedobi
University: Federal University of Technology, Owerri, Imo state, Nigeria.
Date: 06/06/2024

Abstract:

This study investigates the vulnerability of aquifers to leachate contamination from the Idemili North Landfill in Anambra State, Nigeria, focusing on the analysis of physicochemical parameters to assess groundwater quality. Employing advanced data processing and machine learning techniques, the research extrapolates key indicators to predict future contamination scenarios and formulate mitigation strategies. Clustering algorithms were used to categorize various vulnerability zones within the aquifer, providing critical insights for targeted intervention. The effectiveness of these clustering approaches was quantified through several evaluation metrics: The Silhouette Score (0.119), indicating moderate cluster separation; the Calinski-Harabasz Index (188.794), suggesting well-defined clusters; and the Davies-Bouldin Index (3.101), reflecting reasonable intra-cluster similarity. These metrics underscore the clustering efficacy and offer a robust framework for understanding and managing the aquifer's response to pollution. The results indicate significant impacts of dumpsite proximity on aquifer quality, with certain stations showing elevated levels of contaminants.

INTRODUCTION:

Water pollution from dumpsites poses a significant environmental threat, impacting both ecosystems and human health. Contaminants from solid waste can leach into groundwater and surface water, altering water chemistry and potentially leading to hazardous conditions. This study aims to evaluate specific water quality indicators around a dumpsite to determine the extent of pollution and its potential effects.

Aquifer vulnerability refers to the likelihood of contaminants reaching the groundwater system after being introduced at the surface (Jiradech & Sunya, 2013). This concept is based on the understanding that certain areas are more prone to groundwater contamination than others. Groundwater flow is largely influenced by the porosity and permeability of the rocks, which allow water to move from the surface into underground aquifers, with gravity playing a crucial role (Srebotnjak et al., 2012). Any changes in groundwater dynamics and other environmental factors, such as precipitation and surface evaporation, can have significant impacts on groundwater regimes, resources, and water quality (Zektser et al., 2015).

In groundwater systems, deep aquifer water is typically protected from contamination unless compromised by activities such as faulty drilling or injection. In contrast, shallow aquifer systems are highly vulnerable to leachate contamination through diffusion, leading to long-term damage (Basu & Van Meter, 2014). Additionally, groundwater quality generally deteriorates with increasing depth within an aquifer, as declining groundwater levels can result in poorer water quality and higher concentrations of contaminants (USGS, 2018).

This research involves the physicochemical analysis of water samples collected from various stations around a dumpsite and landfill to assess water quality and predict potential future changes in the aquifer system. By clustering the data, the study aims to identify contamination patterns and provide insights into the aquifer's vulnerability to leachate infiltration.

METHODOLOGY:

Data collection:

Data were collected from a research dataset containing information about water quality physicochemical parameters at different sampling stations. The dataset included 18 parameters: pH, Salinity, Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chloride, Carbonate, Sulfate, Magnesium, Calcium, Potassium, Lead, Mercury, Chromium, Iron, Total Suspended Solids (TSS), Total Dissolved Solids (TDS), Electrical Conductivity (EC), and Turbidity. Sample points W1, W2, W3, C1, L1, and Vertical Electric Sounding (VES) station coordinates measuring elevation above sea level were also collected.

Loading Modules:

Dependencies necessary for the machine learning-based cluster analysis were listed in the requirements.txt file, which can be installed in a virtual coding environment. These libraries include pandas, numpy, seaborn, matplotlib, sklearn, and plotly, which are essential for data manipulation, visualization, and machine learning tasks.

Data cleaning:

The datasets were preprocessed using Python to remove unnecessary columns, such as the "Unnamed: 0" column, which represented the default index. Numerical values stored as strings were converted to numeric data formats using appropriate parsing functions, ensuring a cleaner structure and avoiding redundancy. The VES station coordinate dataset was converted from degrees, minutes, and seconds to decimal format. Missing values were replaced with the mean for all numeric variables.

Extrapolation:

Water quality physicochemical parameters were transposed, and sample points “W1, W2, W3, C1, L1” along with the “Mean and NSDWQ” cells were dropped. The dataset index was reset using the "reset_index()" method. Extrapolation was performed using the last value, differences for each column, and standard ranges containing the upper limits for various water quality parameters. This process expanded the data frame from 5 entries with 18 columns to 2,493 entries with 18 columns.

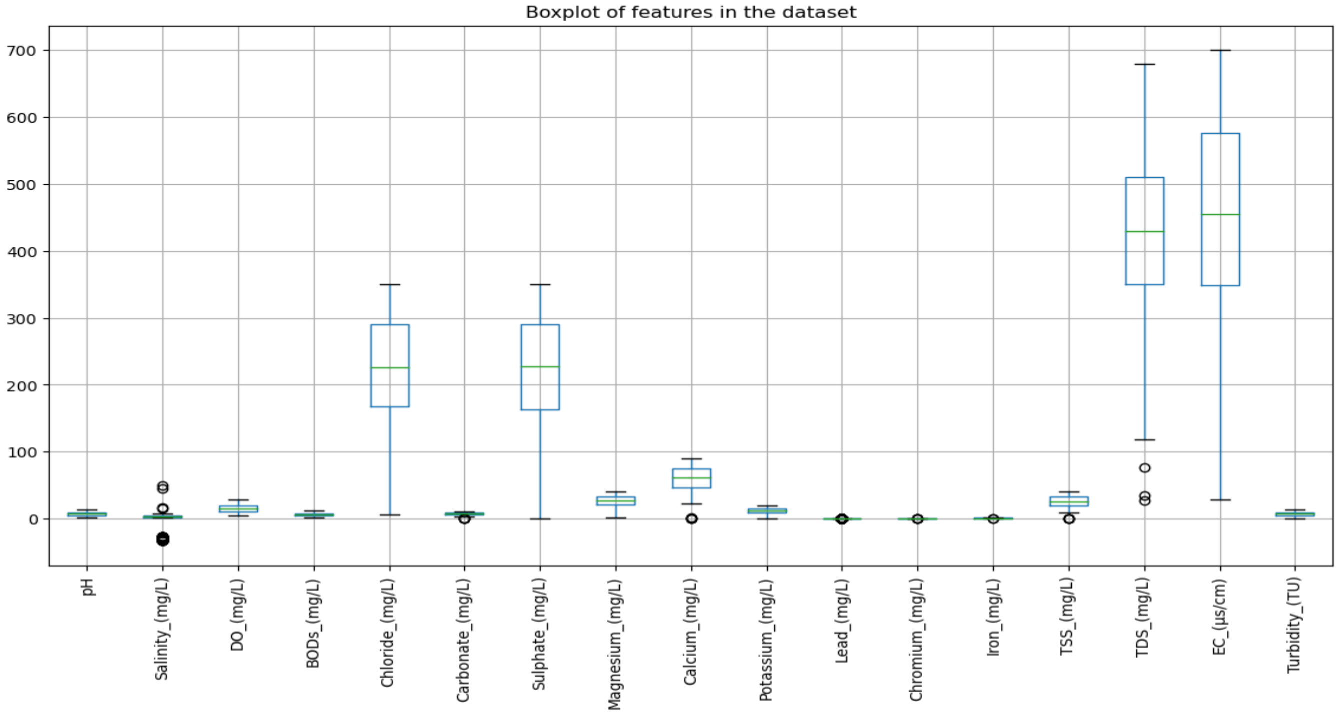
Mapping:

The spatial distribution of the sampling stations was visualized using the Folium library in Python. This provided geographical insights into the study area, showing details such as elevation and station numbers.

+
−
Station: 10.0,
Elevation:156.0
ft
2
3
Leaflet | Data by © OpenStreetMap, under ODbL.

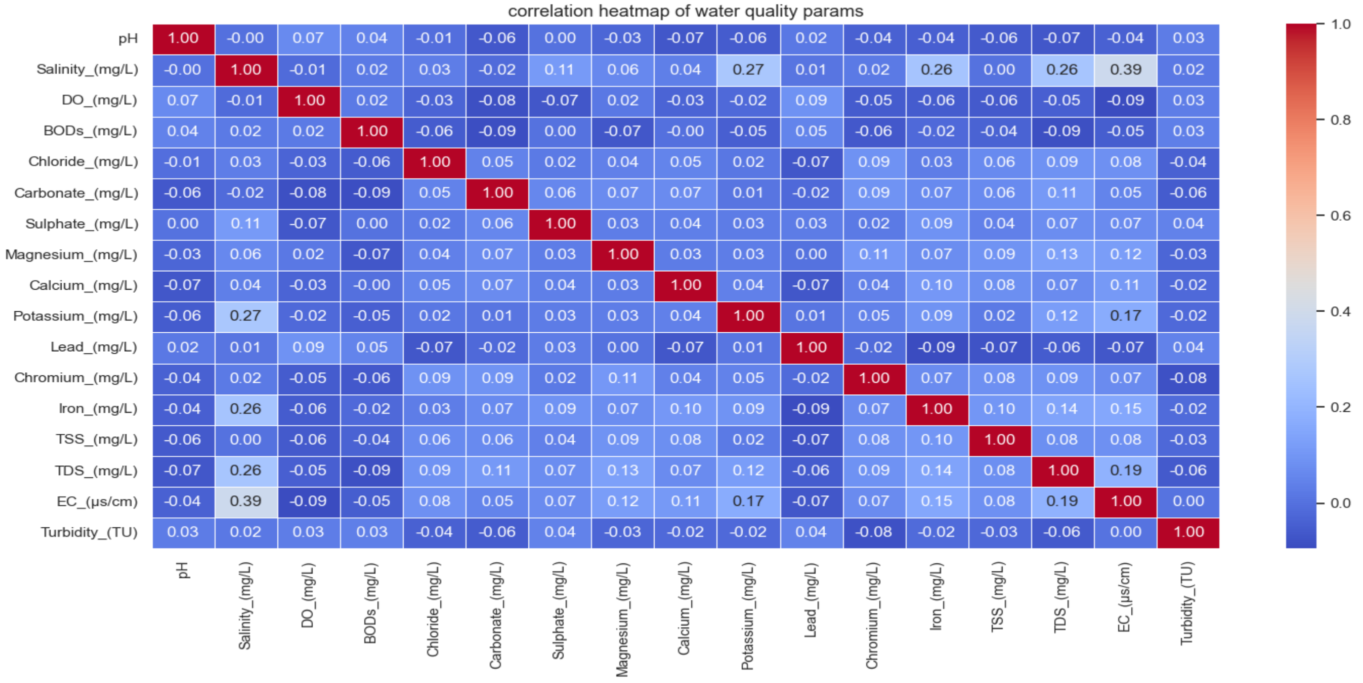
Identifying Outliers using Z-score:
Outliers in the dataset were identified using the Z-score, a statistical measurement that describes a value's relationship to the mean of a group of values. Box plots were used to visualize the physicochemical properties datasets, identifying outliers in each column. The Z-score technique calculated the Z-scores for each data point, indicating how many standard deviations an element is from the mean. A threshold of 3 was set to identify outliers. Data points with a Z-score greater than 3 or less than -3 were considered outliers. A boolean mask was created to filter out these outliers from the original DataFrame, resulting in a cleaned DataFrame. This approach effectively removed extreme outliers, ensuring the analysis was robust against anomalous data points that could skew the results.

EXPLORATORY DATA ANALYSIS (EDA):
Boxplots
Boxplots were used to visualize the statistical distribution of various physicochemical parameters of water quality. Each boxplot displays the interquartile range (IQR) of the data, representing the middle 50% of the values. The line inside the box marks the median, while whiskers extend from the box to the smallest and largest values within 1.5 times the IQR. Points outside this range are considered outliers.
Salinity and BOD_5: These parameters show a wide range with several outliers, indicating variability across different samples.
TSS (Total Suspended Solids) and TDS (Total Dissolved Solids): These parameters display higher ranges and variance, suggesting significant fluctuations in particulate matter in the water samples.
EC (Electrical Conductivity) and Turbidity: These parameters have a substantial spread and higher medians, indicating frequent high levels, possibly due to the presence of various dissolved ions and high turbidity levels.
Other Parameters: Parameters like pH, DO (Dissolved Oxygen), and most ions show relatively tight distributions with fewer outliers, indicating more consistent levels across the samples.



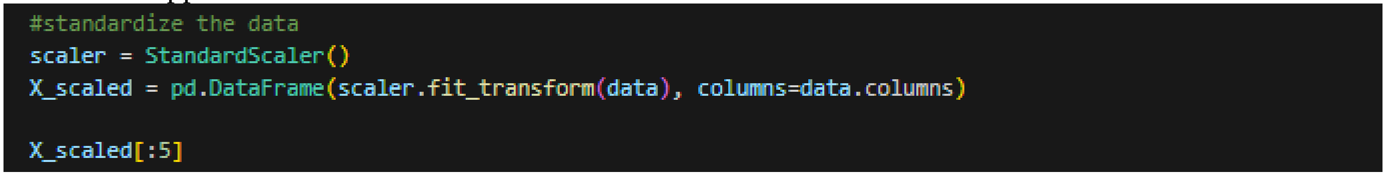
Correlation Heatmaps:

The correlation heatmap illustrates the relationships between various physicochemical parameters of water quality collected from a dumpsite and its vicinity. Each cell represents the correlation coefficient between two parameters, where:



1 indicates a perfect positive correlation.
-1 indicates a perfect negative correlation.
Values close to 0 imply no correlation.
The colors range from blue (negative correlation) to red (positive correlation), providing a visual representation of the strength and direction of the relationships between parameters.

Data Standardization:
Data standardization is a crucial preprocessing step in data analysis, transforming the features in a dataset to have a mean of zero and a standard deviation of one. This process, also known as Z-score normalization, ensures that each feature contributes equally to the analysis by removing the mean and scaling to unit variance. Standardization is especially important when features have different units or vary widely in ranges, preventing features with larger scales from dominating the analysis. This is essential for many machine learning algorithms, particularly those that rely on distance computations, such as clustering.
The code snippet used for data standardization is:



```
#standardize the data
scaler = StandardScaler()
X_scaled = pd.DataFrame(scaler.fit_transform(data), columns=data.columns)

X_scaled[:5]
```

This line of code performs two main functions:
- scaler.fit_transform(data): The ‘fit_transform’ method of the scaler computes the mean and standard deviation of each feature in the dataset and then uses these values to scale the data, standardizing the features so that each has a mean of zero and a standard deviation of one.
- The scaled data, returned as a NumPy array from fit_transform, is converted back into a DataFrame to retain the structure and column labels of the original dataset, making it easier to work with in pandas.

CLUSTER ANALYSIS METHODOLOGY:

Model Selection.
To effectively analyze the data, several clustering algorithms were selected:
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): This algorithm excels at identifying clusters of varying shapes and sizes from a large amount of spatial data, based on density. It effectively filters out noise and outliers, making it robust for datasets with anomalies. The tuning of its parameters, eps (the radius of a neighborhood around a point) and min_samples (the minimum number of points required to form a dense region), are crucial for its performance.

- Agglomerative Clustering: This is a type of hierarchical clustering that builds nested clusters by merging them successively. This method starts with each data point as its own cluster and iteratively combines the closest pairs of clusters. It is very useful for creating dendrograms, which are visual representations of the clustering process, helping in understanding the data structure and choosing the number of clusters by cutting the dendrogram at a suitable level.

- Gaussian Mixture Model (GMM): A probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. GMM is particularly useful for clustering when you assume that the clusters have a hidden, not observable Gaussian statistical process and you expect the clusters to have different variances.

- Spectral Clustering: Utilizes the eigenvalues of a matrix derived from the distance between data points to reduce dimensionality before clustering in fewer dimensions. It works well for clustering complex geometric structures. Often combined with techniques like t-SNE, Spectral Clustering is good at identifying clusters with a non-linear global structure in the data.

- KMeans Clustering: A popular and straightforward clustering technique that partitions the dataset into K-defined distinct non-overlapping subgroups, where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic means of all the data points that belong to that cluster) is minimized. The major challenge with KMeans is choosing the right number of clusters.

Hyperparameter tuning:
This involves systematically searching for the optimal hyperparameters of a model to enhance its performance. For this purpose, the GridSearchCV technique was utilized, allowing for an exhaustive search to find the optimal parameters for the DBSCAN and Agglomerative Clustering algorithms. Silhouette scores were used for evaluation to ensure the quality of the clustering.

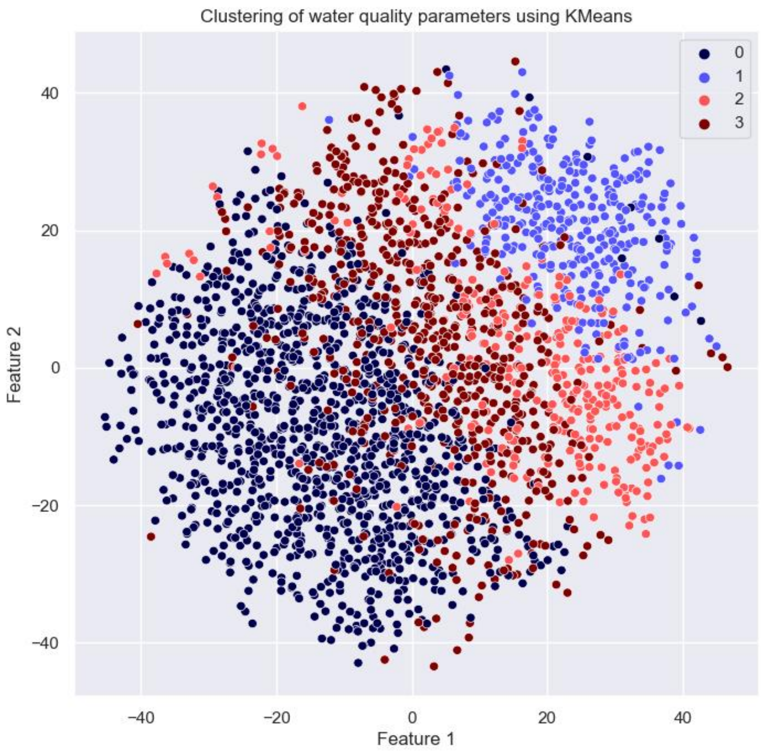
- DBSCAN:
Parameters Tuned: eps (radius of a neighborhood around a point) and min_samples (minimum number of points required to form a dense region).
Evaluation Metric: Silhouette score, which measures how similar an object is to its cluster compared to other clusters.

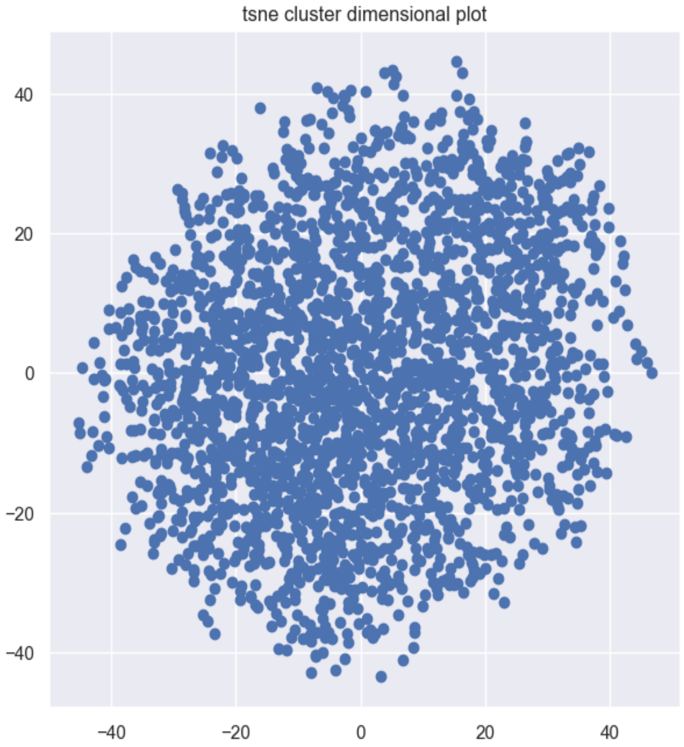
- Agglomerative Clustering:
Parameters Tuned: n_clusters (number of clusters) and linkage method ('ward', 'complete', 'average', 'single').
Evaluation Metric: Silhouette score, providing insights into the cohesion within clusters and separation between clusters.

- Cluster analysis.
To visualize high-dimensional data and analyze cluster distributions, T-distributed Stochastic Neighbor Embedding (t-SNE) was employed. This technique reduces the dimensionality of the data, making it easier to visualize and interpret clustering results. KMeans clustering was then applied to the t-SNE transformed data for further analysis.

- T-distributed Stochastic Neighbor Embedding (t-SNE):
Reduces the dimensionality of the data, allowing for visualization in 2D or 3D space. Thus, capturing the local structure of high-dimensional data, and helps in visualizing complex cluster formations.
Additionally, KMeans clustering was performed on the t-SNE transformed data, to partition the dataset into K distinct non-overlapping clusters, where each data point belongs to only one cluster. The clusters formed were analyzed for coherence and separation, ensuring that intra-cluster

similarity was maximized while inter-cluster similarity was minimized.



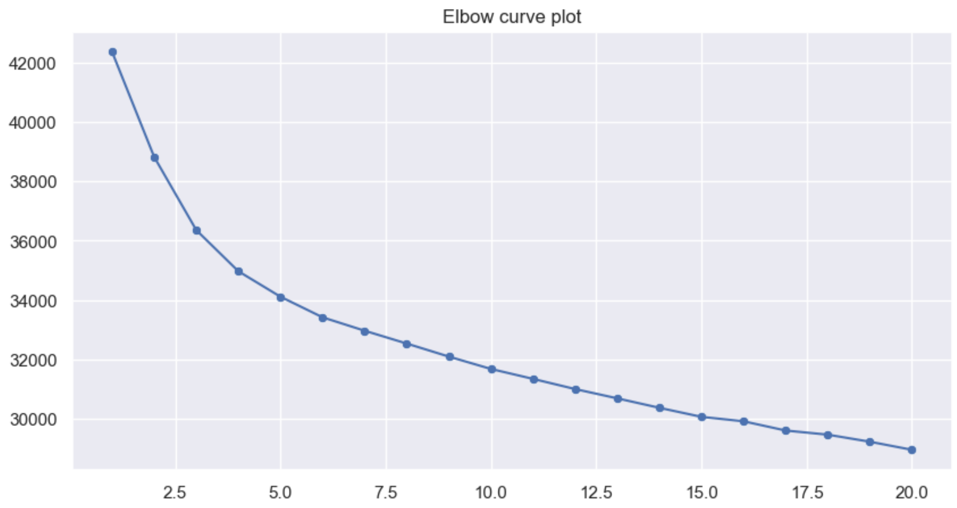


Model Evaluation:

To assess the performance of the clustering models, we utilized a variety of metrics. Each metric provides different insights into the clustering process, and a combination of these metrics offers a comprehensive understanding of the clustering effectiveness. Depending on the specific characteristics of the data and the research objectives, one metric might be prioritized over another to determine the optimal clustering approach.

1. Elbow Curve

The Elbow Curve is a method used primarily with the KMeans clustering algorithm but can be adapted for others. It involves plotting the sum of squared distances from each point to its assigned cluster center as a function of the number of clusters. As more clusters are added, this total within-cluster variation tends to decrease. The "elbow" points on the curve, where the rate of decrease sharply shifts, can often be a good choice for the number of clusters. This point represents a balance between minimizing the within-cluster variance and maximizing the simplicity of the model.



2. Silhouette Score
The Silhouette Score is a measure of how similar an object is to its own cluster compared to other clusters. The value ranges from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, the clustering configuration is appropriate. If many points have a low or negative score, the clustering configuration may have too many or too few clusters.

3. Calinski-Harabasz Index
Also known as the Variance Ratio Criterion, the Calinski-Harabasz Index measures the ratio of the sum of between-cluster dispersion to within-cluster dispersion. Higher values generally indicate a model with better-defined clusters. This index is particularly useful when the true cluster labels are not known and an objective measure of the clustering quality is required.

4. Davies-Bouldin Index
The Davies-Bouldin Index is defined as the average 'similarity' ratio between each cluster and its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. Lower values of the Davies-Bouldin index indicate better clustering. A lower value means the clusters are farther apart and less dispersed, indicating a better partition.

Results:
The evaluation of clustering models revealed that Spectral Clustering outperformed both Agglomerative Clustering and the Gaussian Mixture Model in terms of cluster separation, compactness, and distinctiveness. Despite this, the scores suggest there is still room for improvement in achieving more distinct and cohesive clustering results.

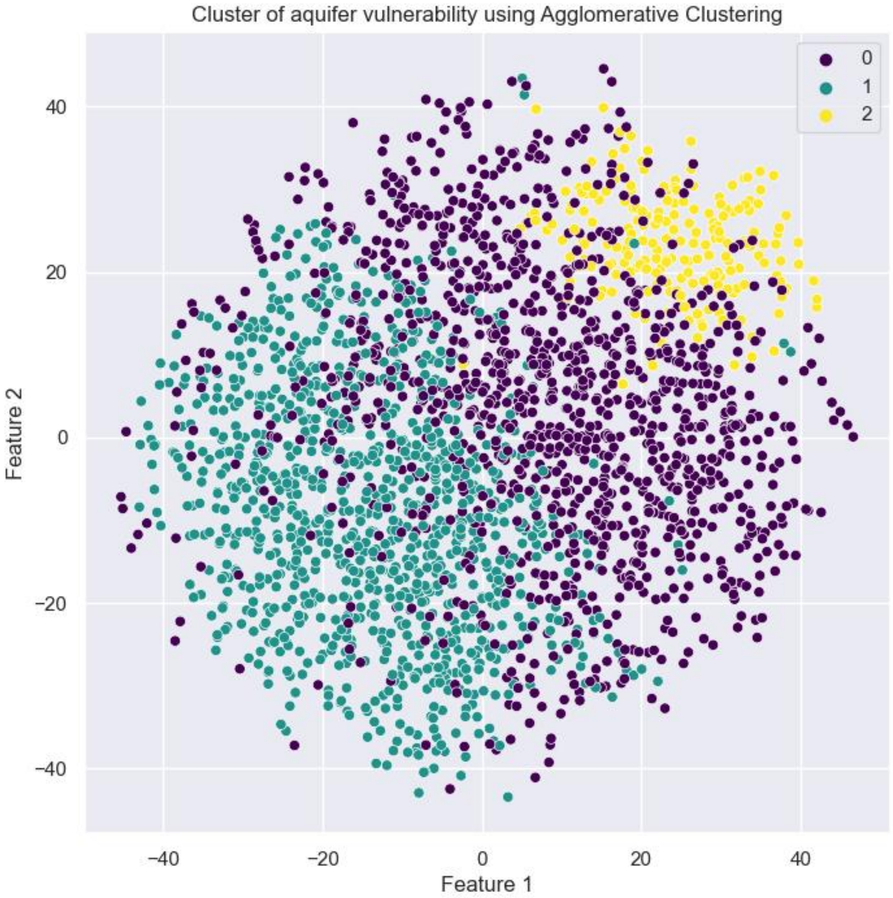
Clustering Method	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index
Agglomerative Clustering	0.057	132.258	3.435
Gaussian Mixture Model	0.036	124.344	4.08
Spectral Clustering	0.119	188.794	3.101

Agglomerative Clustering

Silhouette Score: This low score suggests poor separation and cohesion within clusters, indicating that the clusters are not well-defined.

Calinski-Harabasz Index: A moderate score, suggests a decent level of cluster definition, though there is room for improvement.

Davies-Bouldin Index: A higher value in this index indicates less distinction between clusters, suggesting overlapping clusters or clusters that are not well-separated.

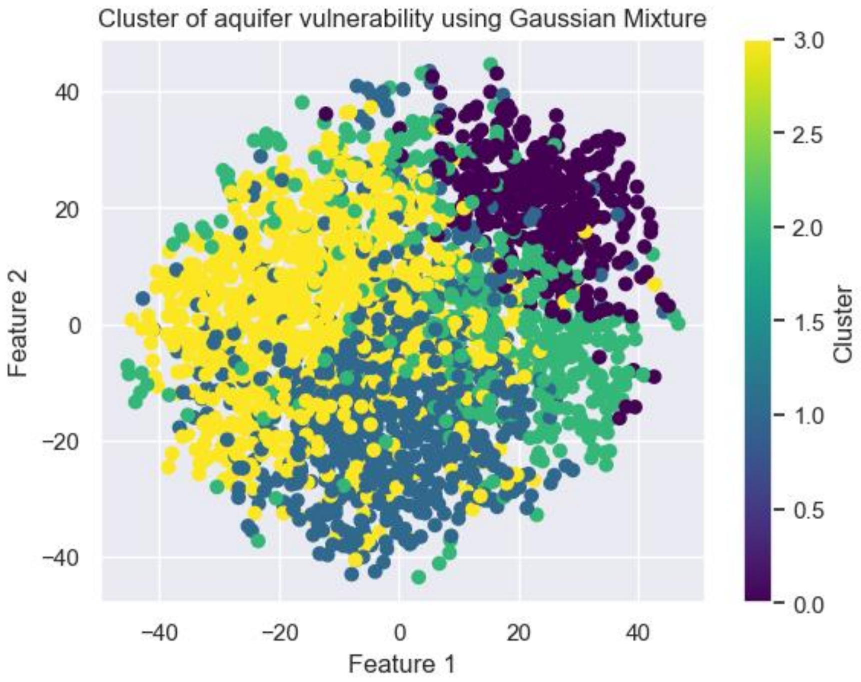


Gaussian Mixture Model

Silhouette Score: Even lower than Agglomerative Clustering, this score indicates very poor clustering where clusters are neither cohesive nor well-separated.

Calinski-Harabasz Index: Slightly lower than Agglomerative, reflecting a weaker clustering structure.

Davies-Bouldin Index: The highest among the three, this score suggests the poorest clustering performance with high similarity within clusters and low differentiation between clusters.

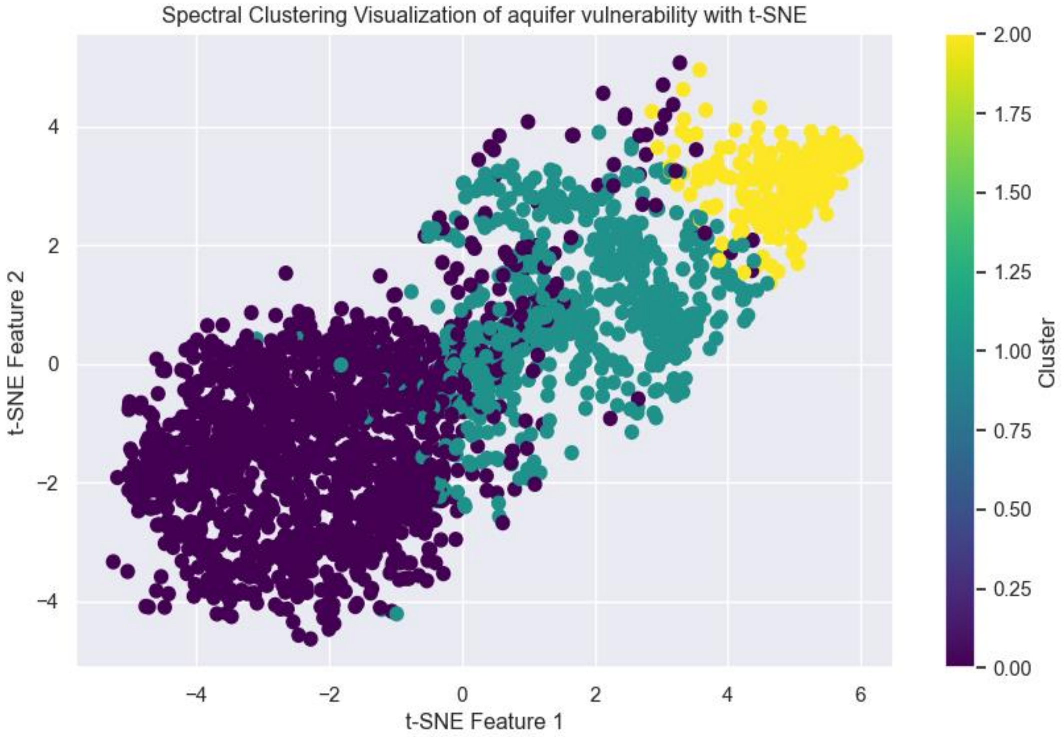


Spectral Clustering

Silhouette Score: The highest among the three, suggesting better cluster separation and compactness.

Calinski-Harabasz Index: The highest score, indicating the best-defined clusters among the three methods.

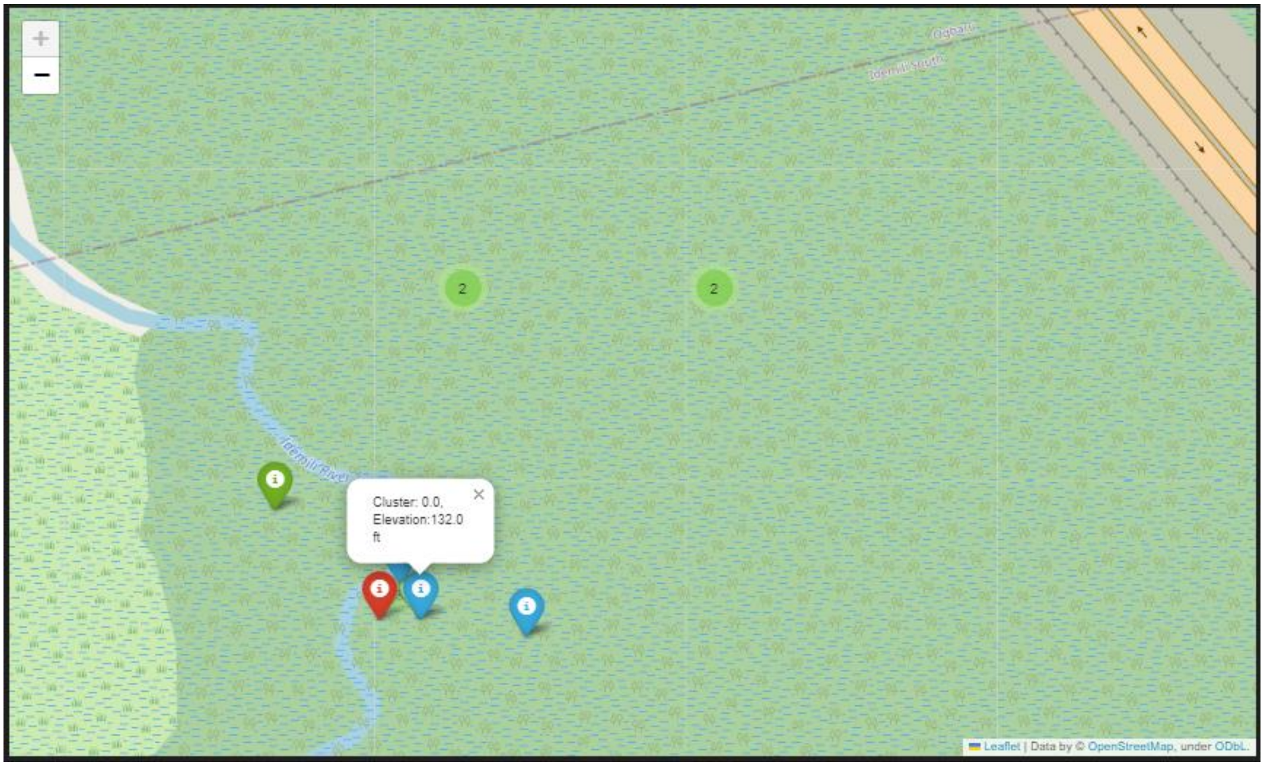
Davies-Bouldin Index: Lower than the other two methods, indicating a better separation between clusters and less intra-cluster variance.



Conclusion:

The analysis of water quality parameters revealed notable variations across different sampling stations. Key findings include:

- Variations in Parameters: Significant fluctuations were observed in water quality parameters such as pH, dissolved oxygen (DO), and total dissolved solids (TDS). These variations indicate differing water quality conditions at various stations.
- Contaminant Levels: Specific contaminants like lead and chromium were found to be within acceptable limits. However, other parameters, such as chloride and magnesium, showed significant variations, suggesting the presence of localized sources of contamination.
- Salinity and TDS: The data indicated potential increases in salinity and TDS at certain stations, likely due to ongoing leachate infiltration from nearby dumpsites. This underscores the need for regular monitoring to detect and address such issues promptly.
- Cluster Analysis: The cluster predictions emphasized the necessity for targeted remediation efforts at specific stations. By addressing these localized pollution sources, it is possible to prevent long-term degradation of the aquifer.
- Geospatial Insights: The mapped locations of the sampling stations provided a clear visual representation of the study area, facilitating the identification of potential pollution hotspots. This geographic visualization is crucial for effective environmental management and planning.



Recommendations:
- Continued Monitoring: Regular and systematic monitoring of water quality parameters is essential to track changes over time and identify emerging risks promptly.
- Targeted Remediation: Based on the cluster analysis, targeted remediation efforts should be focused on specific stations identified as pollution hotspots to mitigate contamination and protect

the aquifer.

- Adaptive Management: Implementing adaptive management strategies will help address the identified risks and ensure the sustainability of the aquifer system. This approach should include flexibility to adjust remediation measures based on ongoing monitoring data and new research findings.
- Further Research: Continued research is necessary to deepen the understanding of contamination sources and their impacts on water quality. This will aid in developing more effective and comprehensive management strategies.

By implementing these recommendations, it is possible to safeguard the water quality and sustainability of the aquifer system, ensuring it remains a reliable resource for ecosystems and human populations alike.

References:

- Basu, N. B., & Van Meter, K. J. (2014). Assessing the Effects of Land Use Change on Water Quality Using Multiscale Models. *Environmental Science & Technology*, 48(4), 2103-2111.
- Jiradech, S., & Sunya, C. (2013). Aquifer Vulnerability Assessment Based on Modified DRASTIC Model and GIS: A Case Study of Phuket Island. *Environmental Earth Sciences*, 70(3), 1215-1226.
- Srebotnjak, T., Carr, G., de Sherbinin, A., & Rickwood, C. (2012). A Global Water Quality Index and Hot-deck Imputation of Missing Data. *Ecological Indicators*, 17, 108-119.
- U.S. Geological Survey (USGS). (2018). Groundwater Quality. Retrieved from [USGS Groundwater Quality](https://www.usgs.gov/special-topic/water-science-school/science/groundwater-quality)
- Zektser, I. S., Everett, L. G., & Dzhamalov, R. G. (2015). Groundwater Resources and Their Use in Environmental Problems. *Environmental Geology*, 45(8), 939-948.
- GitHub: https://github.com/Omedobi/Data-Science-Projects/tree/f629167e4847e51443b084961c9252ef1047c59c/Aquifer-vulnerability