IDENTIFYING PATTERNS IN WATER TREATMENT EFFICIENCY USING CLUSTERING TECHNIQUES

Project Report

Submitted to the Faculty of Engineering of

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY KAKINADA, KAKINADA

In partial fulfillment of the requirements for the award of the Degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

B.SRAVANTHI (21481A0543)

ABDUL KAIF (21481A0501)

B.YAGNA PRIYANKA (21481A0531)

D.SAI GANESH (21481A0561)

Under the Enviable and Esteemed Guidance of

Dr.G.SRIDEVI, M. Tech, (Ph.D)

Professor, Department of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)

SESHADR IRAO KNOWLEDGE VILLAGE

GUDLAVALLERU – 521356

ANDHRA PRADESH

2023-24

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)

SESHADRI RAO KNOWLEDGE VILLAGE, GUDLAVALLERU

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project report entitled **"IDENTIFYING PATTERNS IN WATER TREATMENT EFFICIENCY USING CLUSTERING TECHNIQUES"** is a bonafide record of work carried out by B.Sravanthi (21481A0543), Abdul kaif (21481A0501), B.Yagna Priyanka (21481A0531),D.Sai Ganesh (21481A0561), under the guidance and supervision of **Dr.G.SRIDEVI,Professor,** Computer Science and Engineering, in the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of Jawaharlal Nehru Technological University Kakinada, Kakinada during the academic year 2023-24.

Project Guide

(Dr.G.Sridevi)

Head of the Department

(Dr. M. BABU RAO)

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people who made it possible and whose constant guidance and encouragements crown all the efforts with success.

We would like to express our deep sense of gratitude and sincere thanks to **Dr.G.Sridevi , Professor** Computer Science and Engineering for her constant guidance, supervision and motivation in completing the project work.

We feel elated to express our floral gratitude and sincere thanks to **Dr. M. Babu Rao**, **Head of the Department**, Computer Science and Engineering for his encouragements all the way during analysis of the project. His annotations, insinuations and criticisms are the key behind the successful completion of the project work.

We would like to take this opportunity to thank our beloved principal **Dr. B.Karuna Kumar ,** for providing a great support for us in completing our project and giving us the opportunity for doing project.

Our Special thanks to the faculty of our department and programmers of our computer lab. Finally,we thank our family members, non-teaching staff and our friends, who had directly or indirectly helped and supported us in completing our project intime.

Team Members
B.Sravanthi(21481A0543)
Abdul kaif(21481A0501)
B.Yagna Priyanka(21481A0531)
D.Sai Ganesh (21481A0561)

INDEX

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This project aims to identify patterns in water treatment efficiency using clustering techniques on a comprehensive dataset of water quality measurements taken at various stages of the treatment process. Water treatment facilities are essential for ensuring safe and clean water supply, but the complexity of the treatment process, coupled with the variability in water quality parameters, presents significant challenges in optimizing efficiency and effectiveness.

The primary objective of this project is to apply clustering techniques to a dataset of water quality parameters measured at various stages of the treatment process to identify patterns and relationships that can enhance the efficiency and effectiveness of water treatment operations. The analysis focuses on characterizing distinct clusters of water samples that exhibit similar quality characteristics at different treatment stages, including entry, primary, secondary, and tertiary phases.

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

CLUSTERING

Cluster analysis, also known as clustering, is a statistical technique used in machine learning and data mining that involves the grouping of objects or points in such a way that objects in the same group, also known as a cluster, are more similar to each other than to those in other groups. It is a main task of exploratory data analysis and is used in various fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.There are various clustering algorithms, each with its own approach to defining clusters.

- K-Means
- DBSCAN
- Hierarchical Clustering

Data Collection:

The first step in building a clustering model is data collection. In this step, the data relevant to the problem at hand is collected. The data should be representative of the problem and should contain all the necessary attributes and labels needed for classification. The data can be collected from various sources, such as surveys, questionnaires, websites, and databases.

Data Preprocessing:

The second step in building a clustering model is data preprocessing. The collected data needsto be preprocessed to ensure its quality. This involves handling missing values, dealing with outliers, and transforming the data into a format suitable for analysis. Data preprocessing also involves converting the data into numerical form, as most classification algorithms require numerical input.

Handling Missing Values:

Missing values in the dataset can be handled by replacing them with the mean, median, or mode of the corresponding feature or by removing the entire record.

Feature Selection:

The third step in building a clustering model is feature selection. Feature selection involves identifying the most relevant attributes in the dataset for classification. This can be done using various techniques, such as correlation analysis, information gain, and principal component analysis.

Correlation Analysis: Correlation analysis involves identifying the correlation between the features in the dataset. Features that are highly correlated with each other can be removed as they do not provide additional information for classification.

Information Gain: Information gain is a measure of the amount of information that a feature provides for classification. Features with high information gain are selected for classification.

Model Selection:

The fourth step in building a clustering model is model selection. Model selection involves selecting the appropriate clustering algorithm for the problem at hand. There are several algorithms available, such as k-Means,

K-Means: The K-Means clustering algorithm to the data and outputs a new dataset in which thecluster label is added as a meta attribute. Silhouette scores of clustering results for various k arealso shown in the widget. When using the silhouette score option, the higher the silhouette score,the better the clustering.

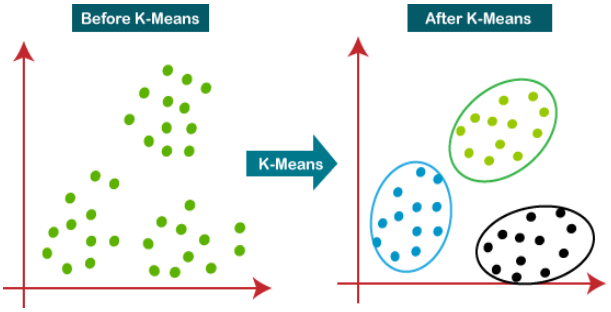


Fig 1.1.2 K-Means

DBSCAN :the DBSCAN clustering algorithm to the data and outputs a new dataset with cluster labels as a meta attribute.

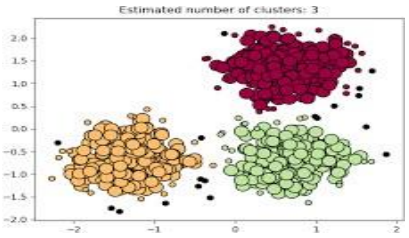


Fig 1.1.3 DBSCAN

Hierarchical clustering: Hierarchical clustering is a popular method for grouping objects. It creates groups so that objects within a group are similar toeach other and different from objects in other groups. Clusters are visually represented in a hierarchical tree called a dendrogram..

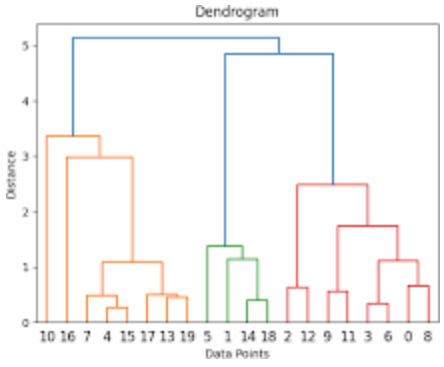


Fig 1.1.4 Hierarchical Clustering

Here's the list of Clustering Models in detail:

Model	**Description**	**Type**
K-Means	Divides data into non-overlapping subsets (clusters) such that each point belongs to only one cluster.	Centroid
Hierarchical	Constructs a tree of clusters, where each nodeis a cluster consisting of the clusters of its children nodes.	Connectivity
DBSCAN	Density-based algorithm that groups together closely packed points while marking points in low-density regions as outliers.	Density

Mean Shift	Non-parametric clustering algorithm that doesn't require prior knowledge of the number of clusters.	Density
Gaussian Mixture Models (GMM)	Assumes all data points are generated from a mixture of a finite number of Gaussian distributions.	Probability
Agglomerative	Begins with each point as a separate cluster and merges the closest pairs of clusters until only one cluster remains.	Connectivity
Spectral Clustering	Uses the eigenvalues of a similarity matrix of the data to reduce the dimensionality before clustering in fewer dimensions.	Connectivity
Affinity Propagation	Finds exemplars among data points and then identifies clusters by considering similarity between data points.	Centroid
OPTICS	Orders the database objects based on their density reachability, which reveals the density-based clustering structure.	Density
BIRCH	Hierarchical clustering method based on a tree data structure to arrange the data points into a branching structure	Connectivity

.
Table 1.1.1: Clustering Models

Model Training:

The fifth step in building a clustering model is model training. Model training involves usingthe selected clustering algorithm to learn the patterns in the data. The data is divided into a training set and a validation set. The model is trained using the training set, and its performance is evaluated on the validation set.

Model Evaluation:

The sixth step in building a clustering model is model evaluation. Model evaluation involvesassessing the performance of the trained model on a test set. This is done to ensure that the model generalizes well.

How to Choose Clustering Models?

Choosing clustering methods in data mining involves several considerations to ensure optimal model performance. First, assess the nature of your dataset, including its size, dimensionality, and the presence of noise or missing values. Next, understand the characteristics of the problem you're solving, such as the number of classes, class imbalance,and the complexity of decision boundaries. Then, evaluate the computational requirements and scalability of different algorithms based on your dataset size and available resources. Consider the interpretability of the models and whether it's essential to understand the reasoning behind predictions. Additionally, conduct experiments with multiple algorithms, using techniques like cross-validation, to compare their performance metrics such as distance, sum of squared error and Silhouette Score. Finally, consider the specific requirements and constraints of your application domain to select the most suitable clusteringmethod that balances predictive interpretability, and computational efficiency.

Challenges and Limitations of Clustering:

Imbalanced Datasets: Dealing with imbalanced datasets where one class dominates, leadingto biased models and misclassification of minority classes.

High-Dimensional Data: Handling high-dimensional data where the number of features exceeds the number of samples, causing overfitting, increased computational complexity, and interpretability issues.

Noisy or Missing Data: Difficulty in handling noisy or missing data, which can impact the model's ability to accurately learn patterns and make predictions.

Algorithm Selection: Choosing the appropriate classification algorithm is challenging and depends on factors such as dataset size, class distribution, and problem complexity.

Interpretability: Some classification models, particularly deep learning algorithms, lack interpretability, making it difficult to understand the reasoning behind their predictions.

Performance Degradation: The performance of classification models may degrade over time due to concept drift, where the statistical properties of the data change, necessitating continuous model monitoring and adaptation.

Applications of Clustering:

1. **Customer Segmentation**: Grouping customers based on similarities in demographics,behavior, or purchasing patterns for targeted marketing strategies.

2. **Image Segmentation**: Partitioning an image into distinct regions based on color, intensity, ortexture similarity for object recognition or image analysis.

3. **Anomaly Detection**: Identifying outliers or unusual patterns in data that deviate from normalbehavior, indicating potential fraud, errors, or system failures.

4. **Document Clustering**: Organizing documents into clusters based on their content similarityfor information retrieval, topic modeling, or document summarization.

5. **Recommendation Systems**: Grouping users or items with similar preferences or attributes tomake personalized recommendations for products, movies, or content.

6. **Genomic Clustering**: Classifying genes or sequences based on similarities in structure , orexpression levels for biological insights or disease diagnosis.

7. **Network Clustering**: Identifying communities or functional modules in complex networkssuch as social networks, biological networks, or transportation networks.

8. **Market Basket Analysis**: Identifying associations or patterns in customer transactions tounderstand purchasing behavior and optimize product placements or promotions.

9. **Natural Language Processing (NLP)**: Grouping text documents or words into clusters basedon semantic similarity for text categorization, summarization, or sentiment analysis.

10. **Image Compression**: Clustering similar pixel values in images to reduce redundancy andcompress image data while preserving essential features.

1.2 PROBLEM STATEMENT

By applying advanced clustering algorithms to the dataset, we seek to uncover hidden patterns and relationships among the water quality parameters. This project focuses on characterizing distinct clusters of water samples that exhibit similar quality characteristics at different treatment stages, including entry, primary, secondary, and tertiary phases. By leveraging clustering techniques, we aim to provide actionable insights that can be used to refine treatment processes, ultimately leading to improved water quality and operational challenges.

CHAPTER 2
PROPOSED METHOD

2.1 Methodology

At first the data need to be collected, then after data preprocessing is done to clean the data. Nextsome data is used for training the model, some data is used for testing the model. Finally Clustering algorithms are applied.

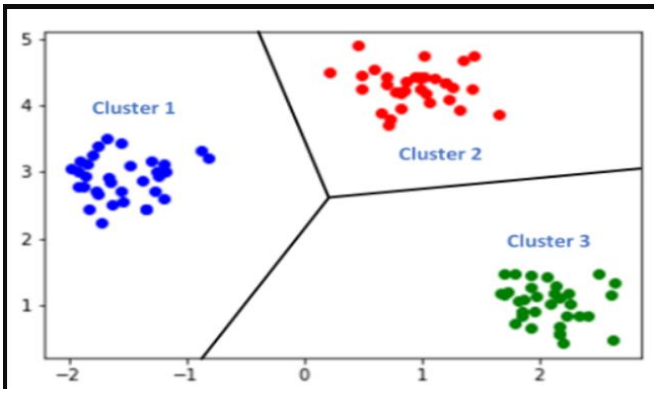


Fig 2.1.1 C l u s t e r i n g m e t h o d o l o g y

Clustering Models*:* They include K-Means, Hierarchical, DBSCAN

1.Silhouette Score: Measures how well-separated clusters are. Values range from -1 to 1, where higher values indicate better clustering.

2.Davies-Bouldin Index: Computes the average similarity between each cluster and its most similar cluster, where lower values indicate better clustering.

3. Calinski-Harabasz Index (Variance Ratio Criterion): Calculates the ratio ofbetween-cluster dispersion to within-cluster dispersion, where higher values indicate better clustering.

4.Dunn Index: Measures the ratio of the minimum inter-cluster distance to the maximum intra-cluster distance, where higher values indicate better clustering.

Block Diagram:

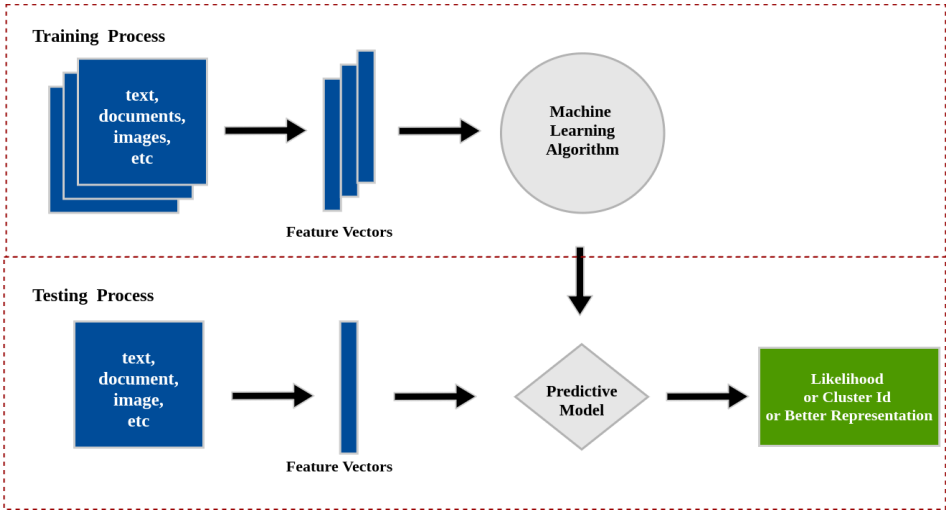


Fig : Block Diagram

Data Visualization*:* Here, In data visualization, data tables organize raw data for easy reference,scatter plots reveal relationships between variables through plotted points, and image viewers provide interactive exploration of image datasets, aiding analysis in fields like medicine and computer vision. Together, they offer powerful tools for understanding and communicating complex data.

Description of dataset:

1. Entry Point (E): Measurements taken at the initial entry of raw water into the treatment plant.

2. Primary Treatment (P): Measurements after primary treatment, which typically involves the removal of large particles and sediments.

3. Secondary Treatment (D): Measurements after secondary treatment, which often includes biological processes to remove dissolved and suspended organic matter.

4. Sludge Processing (S): Measurements related to the treatment and processing of sludge, a byproduct of water treatment.

5. General (G):Combined or overall data reductions observed across different stages of process.

Techniques and their accuracy: The evaluation results for the clustering techniques, including k-Means,DBSCAN, Hierarchical reveal distinct performance metrics. K-Means attained exceptional silhouette score and less sum of squared error. Overall, each technique showcased strengths in different aspects of clustering, emphasizing the importance of selecting the most suitable algorithm based on specific requirements and objectives. Based on the obtained silhouette score details, the K-Means appears to perform the best among the three algorithms for the given clustering task.

2.1.1 Algorithm and Explanation:

Algorithm: k-means. The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- k: the number of clusters,
- D: a data set containing n objects.

Output: A set of k clusters.

Method:

(1) arbitrarily choose k objects from D as the initial cluster centers;
(2) repeat
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) until no change;

Explanation:

K-means is a clustering algorithm used to partition a dataset into K clusters. It starts byrandomly initializing cluster centroids. Data points are then assigned to the nearest centroid, and centroids are updated based on the mean of assigned points. This process repeats until convergence. K-means converges when centroids no longer change significantly. The algorithm may converge to local optima due to random initialization.It's common to run K-means multiple times and select the best result. K-means is efficient but sensitive to initializations and assumes spherical clusters. It's widely used for its simplicity and scalability in clustering tasks**.**

Algorithm: Hierarchical Clustering

Input:

- k: the number of clusters,
- D: a data set containing n objects

Output:

- A hierarchical clustering represented as a dendrogram.

Method:

1. **Initialization:**
   - Create n clusters, each containing one of the n objects.
   - Compute the initial distance matrix D, where D[i, j] is the distance between the i-th and j-th objects.
2. **Repeat:**
   1. **Find the Closest Clusters:**
      - Identify the two clusters Ciand Cj that are closest, i.e., have the smallest distance in the distance matrix D.
   2. **Merge Clusters:**
      - Merge clusters Ci and Cj into a new cluster Cij.
   3. **Update Distance Matrix:**
      - Update the distance matrix to reflect the distance between the new cluster Cij and all other existing clusters.
      - Use a chosen linkage criterion (single, complete, average, centroid) to compute the distance between Cij and each remaining cluster.
3. **Until:**
   - All objects are merged into a single cluster or another stopping criterion is met.

Explanation:

The hierarchical clustering algorithm is a method of cluster analysis that builds a hierarchy of clusters. It starts with each data point as its own cluster and iteratively merges the closest pairs of clusters until only one cluster remains or a stopping criterion is met. At each iteration, the algorithm finds the two clusters with the smallest distance between them and merges them into a new cluster. The distance matrix is then updated to reflect the distances between this new cluster and all other existing clusters. This process continues, resulting in a tree-like structure called a dendrogram, which visually represents the nested clustering at each iteration. The choice of linkage criterion (single, complete, average, centroid) determines how the distances between clusters are computed, influencing the final clustering structure.

Data Preparation

2.1.1 Data Set Description

The "WATER_TREATMENT" dataset comprises various measurements taken at different stages of a water treatment process. Below is a detailed description of the dataset columns, including what each column represents.

1. *Q-E*: Flow rate at the entry point (initial inflow of raw water).
2. *ZN-E*: Zinc concentration at the entry point.
3. *PH-E*: pH level at the entry point.
4. *DBO-E*: Biochemical Oxygen Demand at the entry point.
5. *DQO-E*: Chemical Oxygen Demand at the entry point.
6. *SS-E*: Suspended Solids at the entry point.
7. *SSV-E*: Volatile Suspended Solids at the entry point.
8. *SED-E*: Sediments at the entry point.
9. *COND-E*: Conductivity at the entry point.
10. *PH-P*: pH level after primary treatment.
11. *DBO-P*: Biochemical Oxygen Demand after primary treatment.
12. *SS-P*: Suspended Solids after primary treatment.
13. *SSV-P*: Volatile Suspended Solids after primary treatment.
14. *SED-P*: Sediments after primary treatment.
15. *COND-P*: Conductivity after primary treatment.
16. *PH-D*: pH level after secondary treatment.
17. *DBO-D*: Biochemical Oxygen Demand after secondary treatment.
18. *DQO-D*: Chemical Oxygen Demand after secondary treatment.
19. *SS-D*: Suspended Solids after secondary treatment.
20. *SSV-D*: Volatile Suspended Solids after secondary treatment.
21. *SED-D*: Sediments after secondary treatment.
22. *COND-D*: Conductivity after secondary treatment.
23. *PH-S*: pH level related to sludge processing.
24. *DBO-S*: Biochemical Oxygen Demand related to sludge processing.
25. *DQO-S*: Chemical Oxygen Demand related to sludge processing.
26. *SS-S*: Suspended Solids related to sludge processing.
27. *SSV-S*: Volatile Suspended Solids related to sludge processing.
28. *SED-S*: Sediments related to sludge processing.
29. *COND-S*: Conductivity related to sludge processing.
30. *RD-DBO-P*: Reduction in Biochemical Oxygen Demand during the primary treatment stage.
31. *RD-SS-P*: Reduction in Suspended Solids during the primary treatment stage.

32. *RD-SED-P*: Reduction in Sediments during the primary treatment stage.
33. *RD-DBO-S*: Reduction in Biochemical Oxygen Demand during the sludge treatment stage.
34. *RD-DQO-S*: Reduction in Chemical Oxygen Demand during the sludge treatment stage.
35. *RD-DBO-G*: General reduction in Biochemical Oxygen Demand.
36. *RD-DQO-G*: General reduction in Chemical Oxygen Demand.
37. *RD-SS-G*: General reduction in Suspended Solids.
38. *RD-SSED-G*: General reduction in Suspended Solids and Sediments.

2.1.2 Data Pre-Processing

Data Validation/ Cleaning/Preparing Process:

In the data validation, cleaning, and preparation process using the Orange tool, the first step is to address missing values in the dataset. Utilizing the preprocessing module, we employ imputation techniques to handle these missing values effectively. By imputing missing values, such as those denoted by "?", with appropriate strategies like mean, median, or mode imputation, we ensure the completeness and integrity of the dataset. This step is crucial as missing data can adversely affect the performance and accuracy of downstream analysis and modeling tasks. Once missing values are imputed, the dataset undergoes further preprocessing steps, such as normalization or standardization, to ensure uniformity and comparability across features. Through these data preparation processes, we aim to create a clean and reliable dataset ready for exploratory analysis, modeling, and insights extraction, enabling effective decision-making in various domains.

CHAPTER 3
RESULTS

3.1 ORANGE tool description:

Orange is an open-source data visualization and analysis tool designed for users seeking intuitive yet powerful solutions in machine learning and data mining. Its hallmark feature is a visual programming interface, facilitating the construction of data analysis workflows through interconnected components (widgets). With this approach, users can perform various tasks seamlessly, including data preprocessing, exploratory data analysis, predictive modeling, and visualization. Orange offers an array of preprocessing techniques, allowing users to handle missing values, scale features, encode categorical variables, and select relevant features effortlessly. Moreover, its extensive collection of visualization tools enables users to explore datasets visually, uncovering relationships, distributions, and patterns. Through integration with machine learning algorithms and ensemble learning methods, Orange empowers users to train models for classification, regression, clustering, and association rule mining. Model evaluation tools further aid in assessing model performance, ensuring robust and reliable results. With its blend of usability and versatility, Orange serves as a valuable asset for data scientists, researchers, and analysts across various domains, fostering innovation and insight discovery.

3.2 Screen shots

Step 1: Download and install ORANGE

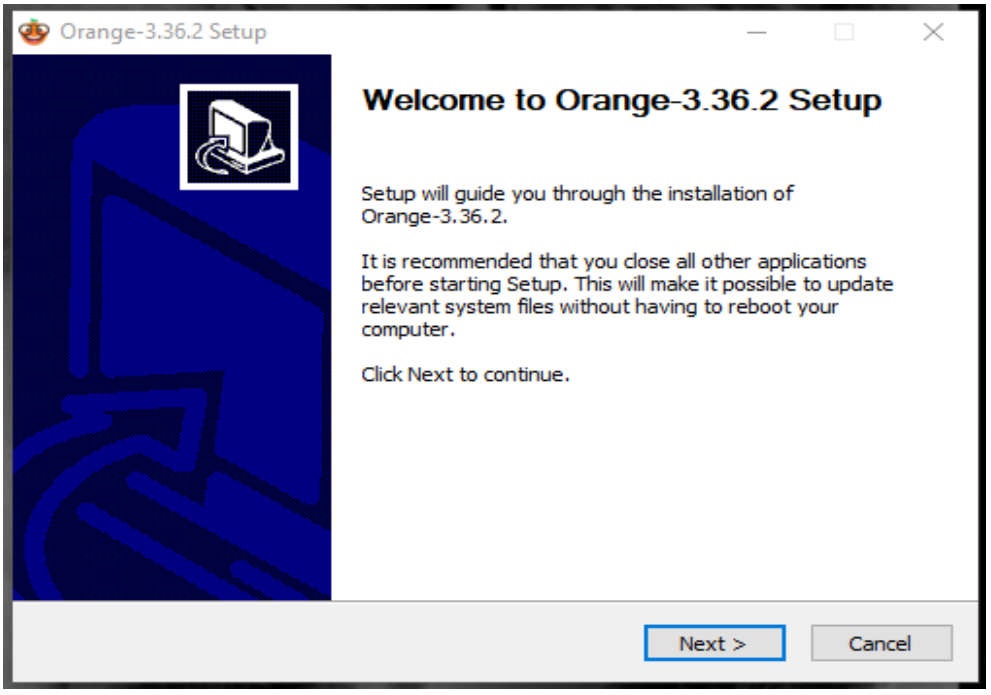


Fig 3.2.1 Download and Install Orange

Step 2: Open Orange and Select new to start a new project

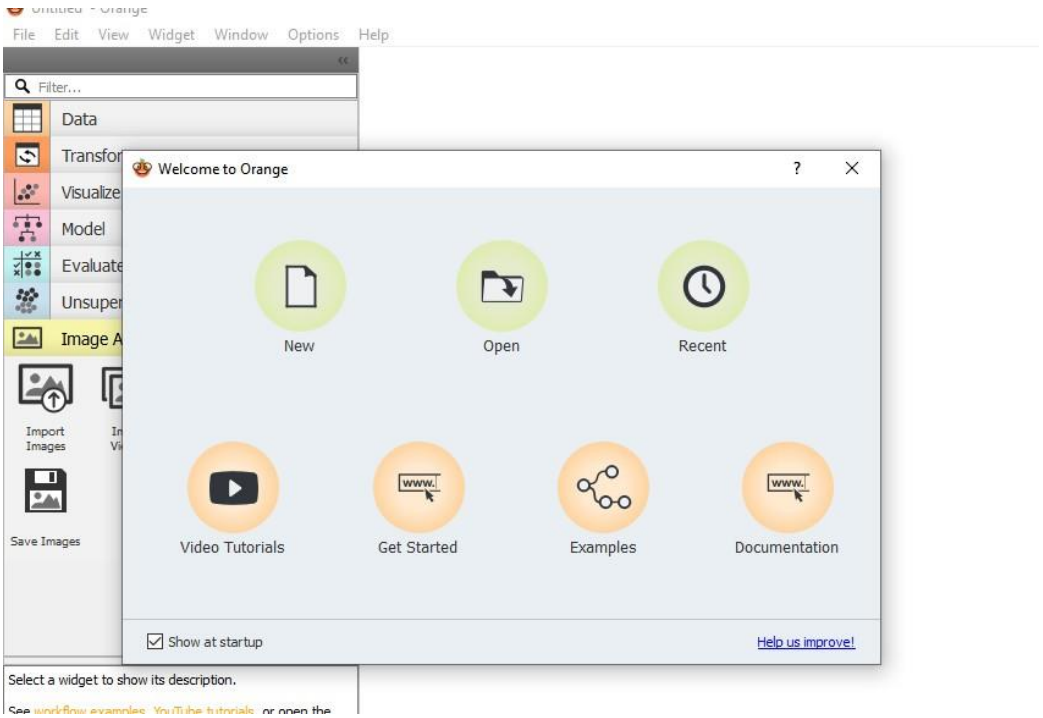


Fig 3.2..2 Open new File

Step 3: From the Data, select file. Double click on it and Load the dataset

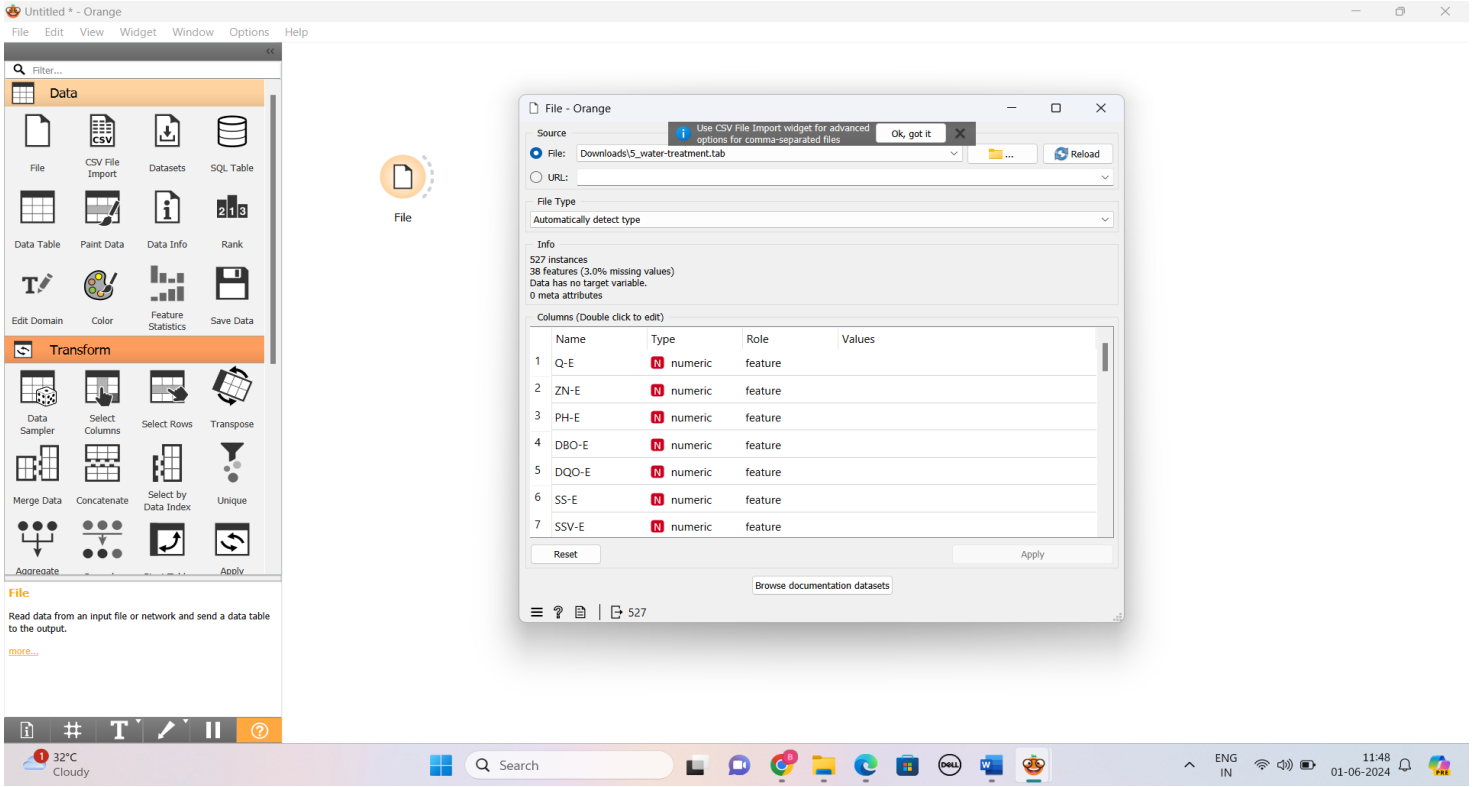


Fig 3.2.3 Load the Dataset

Step 4: The dataset can be viewed with a Data Table and its information with Data Info

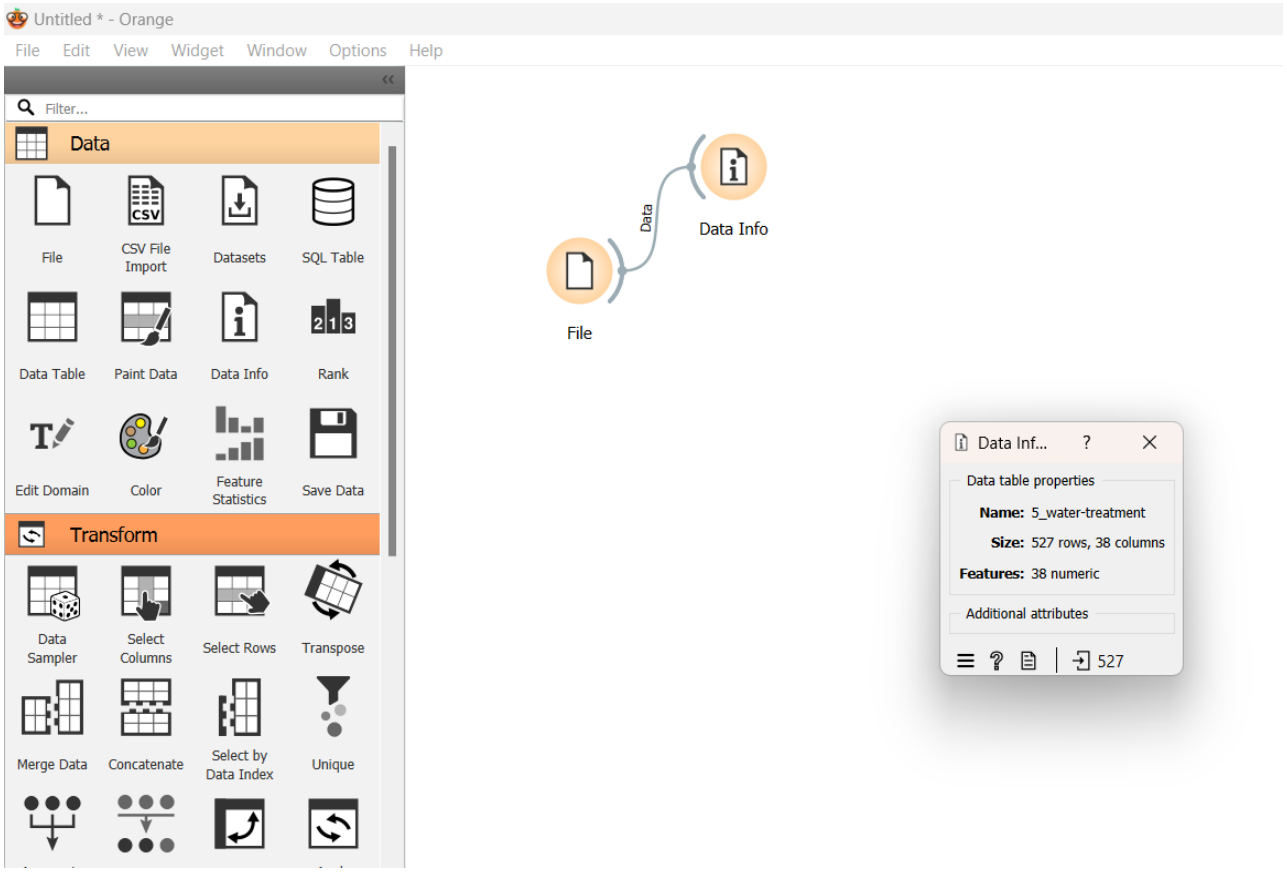


Fig 3.2.4 Data Info of datatset

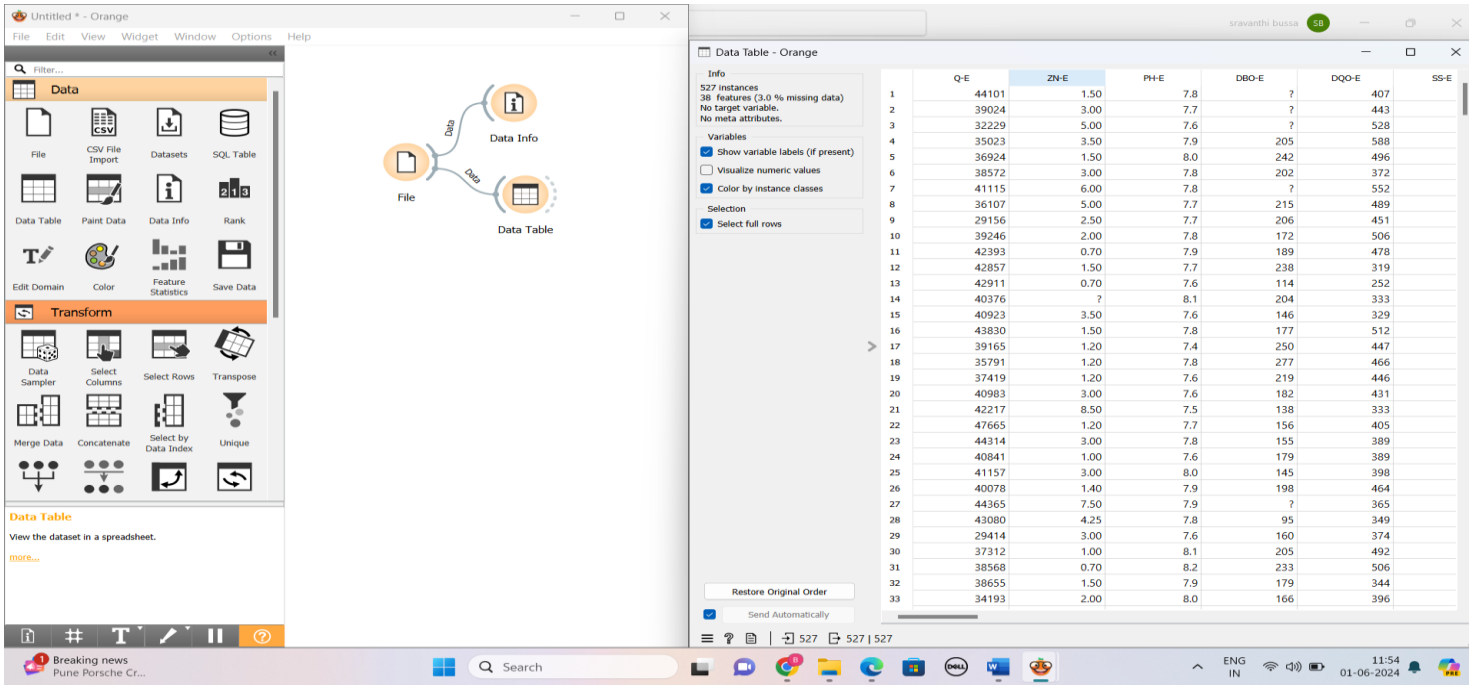


Fig 3.2.5 Data Table before Preprocessing(Normalizing)

Step 5:As we have missing values we need to preprocess the data.

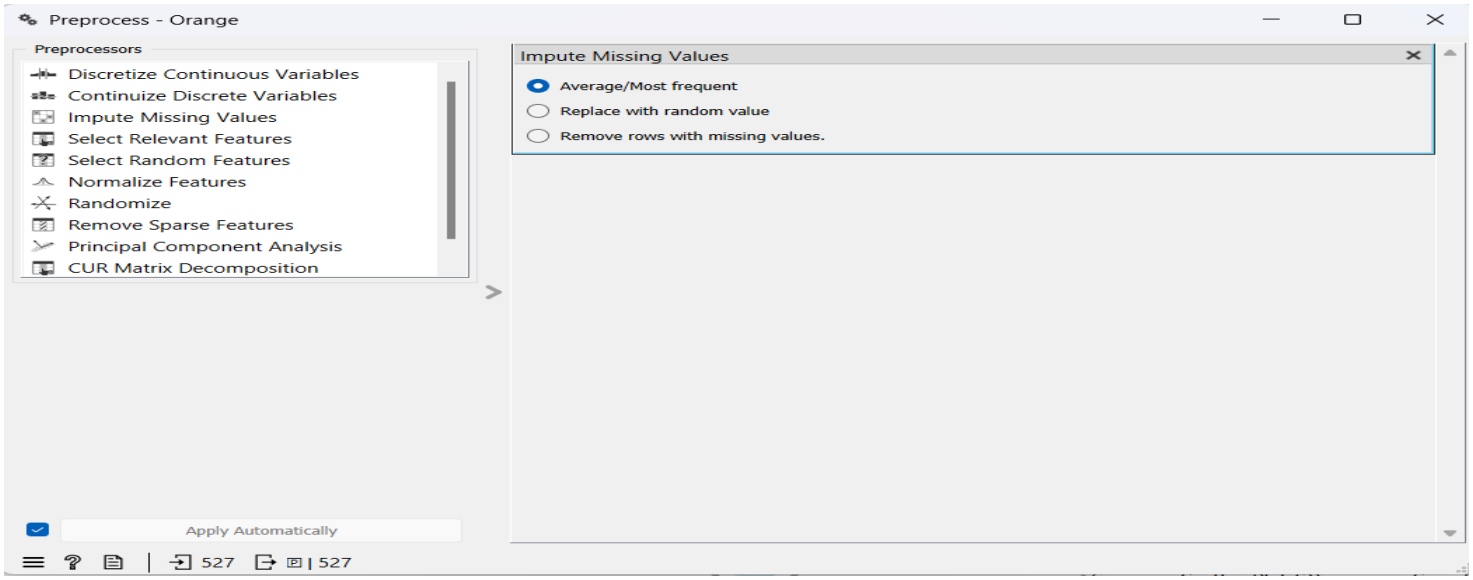
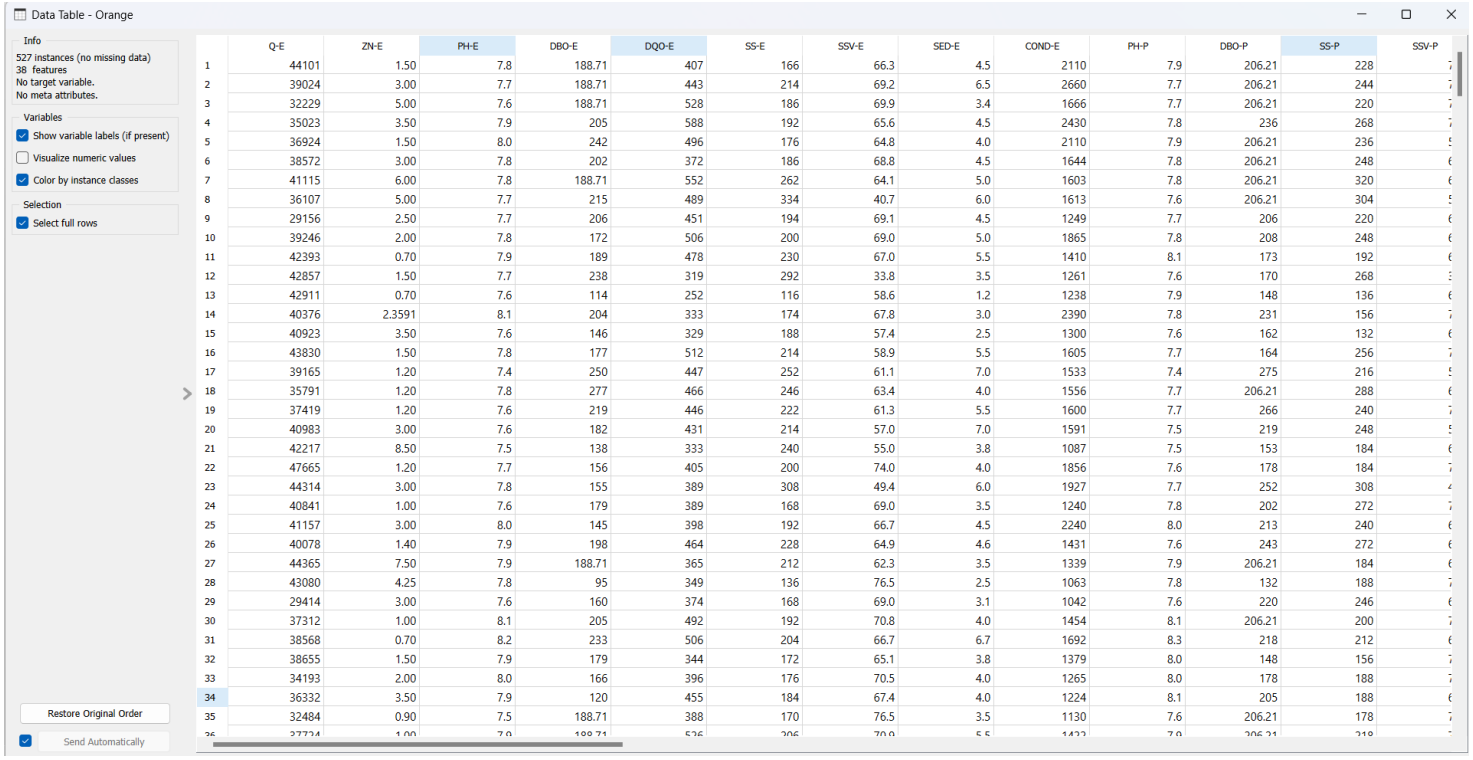


Fig 3.2.6 Preprocessing the Dataset



	Q-E	ZN-E	PH-E	DBO-E	DQO-E	SS-E	SSV-E	SED-E	COND-E	PH-P	DBO-P	SS-P	SSV-P
1	44101	1.50	7.8	188.71	407	166	66.3	4.5	2110	7.9	206.21	228	7
2	39024	3.00	7.7	188.71	443	214	69.2	6.5	2660	7.7	206.21	244	7
3	32229	5.00	7.6	188.71	528	186	69.9	3.4	1666	7.7	206.21	220	7
4	35023	3.50	7.9	205	588	192	65.6	4.5	2430	7.8	236	268	7
5	36924	1.50	8.0	242	496	176	64.8	4.0	2110	7.9	206.21	236	5
6	38572	3.00	7.8	202	372	186	68.8	4.5	1644	7.8	206.21	248	6
7	41115	6.00	7.8	188.71	552	262	64.1	5.0	1603	7.8	206.21	320	6
8	36107	5.00	7.7	215	489	334	40.7	6.0	1613	7.6	206.21	304	5
9	29156	2.50	7.7	206	451	194	69.1	4.5	1249	7.7	206	220	6
10	39246	2.00	7.8	172	506	200	69.0	5.0	1865	7.8	208	248	6
11	42393	0.70	7.9	189	478	230	67.0	5.5	1410	8.1	173	192	6
12	42857	1.50	7.7	238	319	292	33.8	3.5	1261	7.6	170	268	3
13	42911	0.70	7.6	114	252	116	58.6	1.2	1238	7.9	148	136	6
14	40376	2.3591	8.1	204	333	174	67.8	3.0	2390	7.8	231	156	7
15	40923	3.50	7.6	146	329	188	57.4	2.5	1300	7.6	162	132	6
16	43830	1.50	7.8	177	512	214	58.9	5.5	1605	7.7	164	256	7
17	39165	1.20	7.4	250	447	252	61.1	7.0	1533	7.4	275	216	5
18	35791	1.20	7.8	277	466	246	63.4	4.0	1556	7.7	206.21	288	6
19	37419	1.20	7.6	219	446	222	61.3	5.5	1600	7.7	266	240	7
20	40983	3.00	7.6	182	431	214	57.0	7.0	1591	7.5	219	248	5
21	42217	8.50	7.5	138	333	240	55.0	3.8	1087	7.5	153	184	6
22	47665	1.20	7.7	156	405	200	74.0	4.0	1856	7.6	178	184	7
23	44314	3.00	7.8	155	389	308	49.4	6.0	1927	7.7	252	308	4
24	40841	1.00	7.6	179	389	168	69.0	3.5	1240	7.8	202	272	7
25	41157	3.00	8.0	145	398	192	66.7	4.5	2240	8.0	213	240	6
26	40078	1.40	7.9	198	464	228	64.9	4.6	1431	7.6	243	272	6
27	44365	7.50	7.9	188.71	365	212	62.3	3.5	1339	7.9	206.21	184	6
28	43080	4.25	7.8	95	349	136	76.5	2.5	1063	7.8	132	188	7
29	29414	3.00	7.6	160	374	168	69.0	3.1	1042	7.6	220	246	6
30	37312	1.00	8.1	205	492	192	70.8	4.0	1454	8.1	206.21	200	7
31	38568	0.70	8.2	233	506	204	66.7	6.7	1692	8.3	218	212	6
32	38655	1.50	7.9	179	344	172	65.1	3.8	1379	8.0	148	156	7
33	34193	2.00	8.0	166	396	176	70.5	4.0	1265	8.0	178	188	7
34	36332	3.50	7.9	120	455	184	67.4	4.0	1224	8.1	205	188	6
35	32484	0.90	7.5	188.71	388	170	76.5	3.5	1130	7.6	206.21	178	7

Fig 3.2.7 Data Table after Preprocessing

Step 6:Apply Classification models on the preprocessed data.

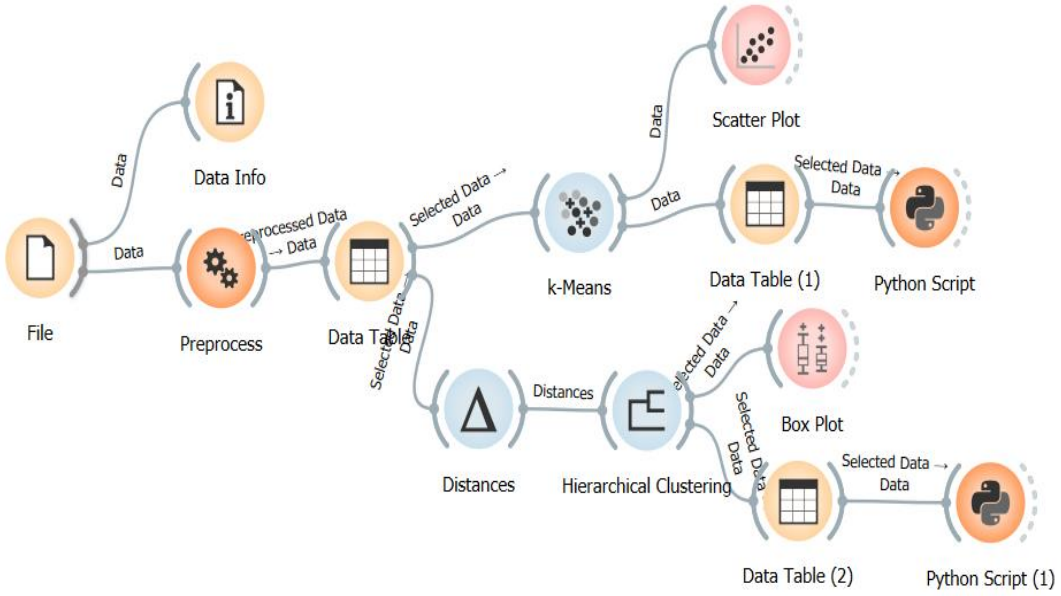


Fig 3.2.8.Applying Clustering Models before Normalization

Step 7: Evaluate the models with Silhouette Score

Data Table (1) - Orange

Info
527 instances (no missing data)
2 features
No target variable.
2 meta attributes

Variables
Show variable labels (if present)
Visualize numeric values
Color by instance classes

Selection
Select full rows

Restore Original Order
Send Automatically

	Cluster	Silhouette	ZN-E	PH-E
1	C2	0.528955	1.50	7.8
2	C2	0.549814	3.00	7.7
3	C2	0.498101	5.00	7.6
4	C2	0.566404	3.50	7.9
5	C2	0.545002	1.50	8.0
6	C2	0.537659	3.00	7.8
7	C2	0.548479	6.00	7.8
8	C2	0.501217	5.00	7.7
9	C3	0.526002	2.50	7.7
10	C2	0.549758	2.00	7.8
11	C1	0.57671	0.70	7.9
12	C1	0.553718	1.50	7.7
13	C1	0.484225	0.70	7.6
14	C2	0.521699	2.3591	8.1
15	C3	0.560393	3.50	7.6
16	C3	0.515746	1.50	7.8
17	C3	0.513985	1.20	7.4
18	C2	0.545532	1.20	7.8
19	C2	0.513523	1.20	7.6
20	C3	0.511217	3.00	7.6
21	C3	0.561415	8.50	7.5
22	C2	0.490708	1.20	7.7
23	C2	0.51314	3.00	7.8
24	C3	0.515145	1.00	7.6
25	C2	0.548925	3.00	8.0
26	C2	0.520266	1.40	7.9
27	C3	0.514558	7.50	7.9
28	C3	0.539731	4.25	7.8
29	C3	0.548623	3.00	7.6
30	C2	0.550986	1.00	8.1
31	C2	0.541468	0.70	8.2
32	C3	0.509913	1.50	7.9
33	C2	0.509092	2.00	8.0
34	C2	0.497631	3.50	7.9
35	C3	0.523514	0.90	7.5
36	C2	0.536966	1.00	7.9
37	C2	0.535283	1.00	7.7

Fig 3.2.9 Evaluating Clustering Model

Step 8: Visualize Clustering for K-Means

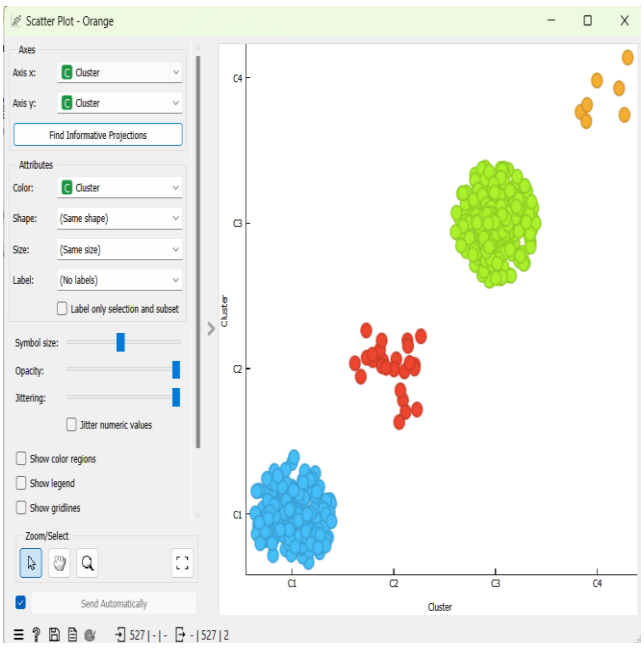


Fig:before normalization

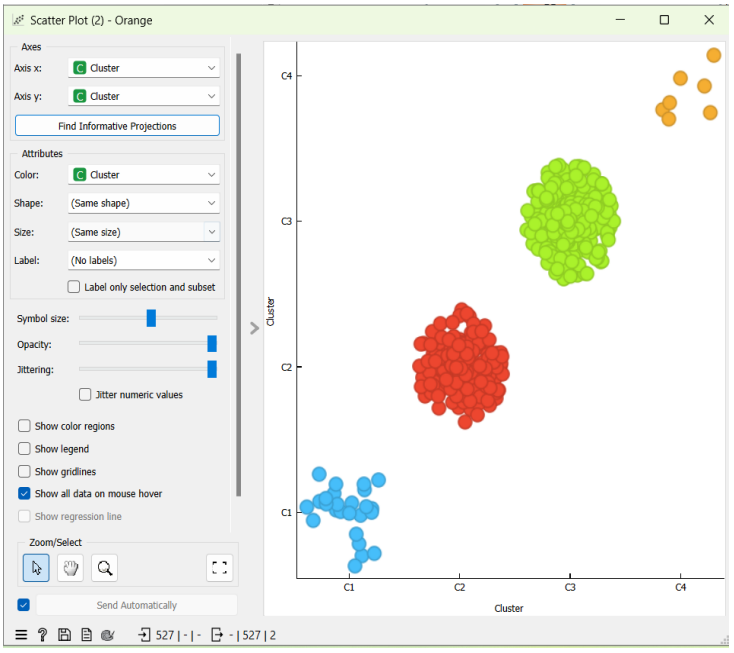


fig:After Normalization

Step 9: Visualize the output through Box Plot plot for Hierarchical clustering

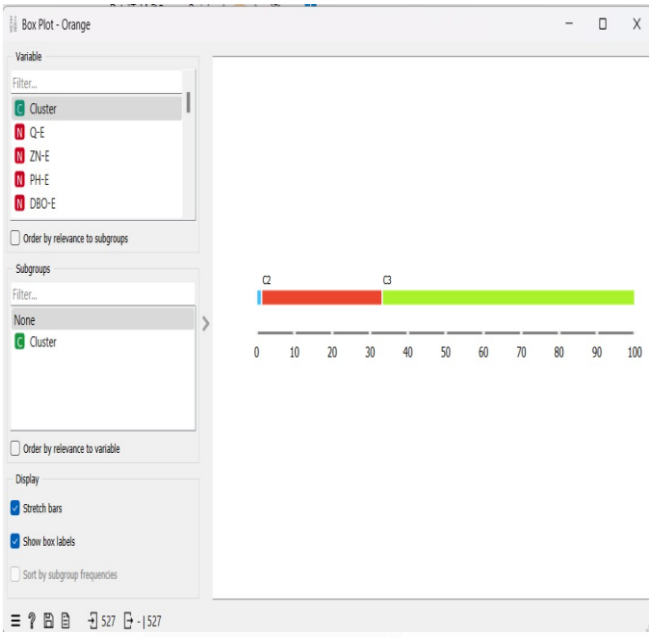


Fig: before normalization

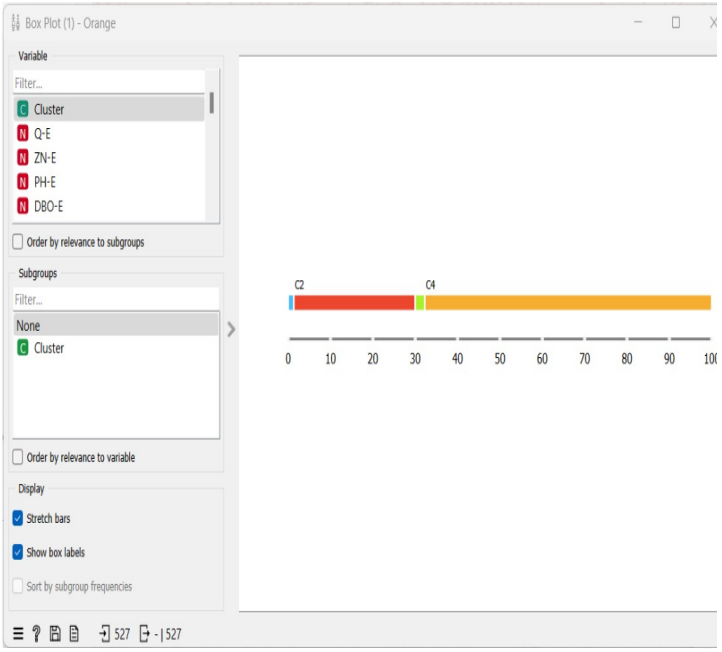


fig: After Normalization

The Overall Workflow is

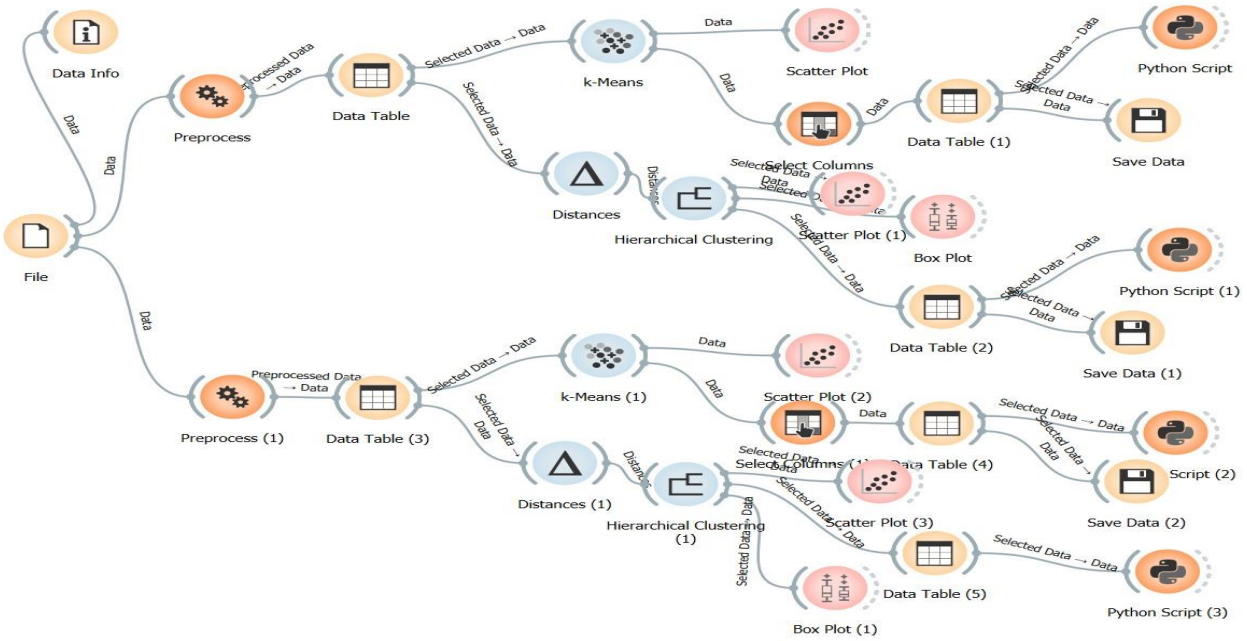


Fig 3.2.12 Overall Workflow

Python Scrips for finding Sum of Squared Error:

K-Means

```
def python_script(in_data):
```

```
from sklearn.cluster import KMeans
import pandas as pd

# Assuming X is your data
dataset =pd.read_csv(r"C:\Users\DELL\Downloads\water-1.csv")
X = dataset.iloc[:,:-1].values

# Fit K-means clustering model
kmeans = KMeans(n_clusters=6, random_state=108)
kmeans.fit(X)

# Calculate the sum of squared errors (SSE)
sse = kmeans.inertia_

print("Sum of Squared Errors (SSE):",sse)
```

Hierarchical Clustering:

```
Editor
def python_script(in_data):
```

```
import pandas as pd
from sklearn.cluster import AgglomerativeClustering
import numpy as np
def calculate_sse(X, labels, centers):
...
    sse = 0
    for i, label in enumerate(labels):
        sse += np.linalg.norm(X[i] - centers[label]) ** 2
    return sse
# Load data from CSV
data = pd.read_csv(r"C:\Users\DELL\Downloads\water-2.csv")  # Replace "your_data.csv" with your CSV file path
# Assume your data is stored in columns named 'feature1', 'feature2', etc.
X = data.values
# Perform hierarchical clustering
clustering = AgglomerativeClustering(n_clusters=6)
labels = clustering.fit_predict(X)
# Calculate cluster centers
unique_labels = np.unique(labels)
centers = np.array([X[labels == label].mean(axis=0) for label in unique_labels]
sse = calculate_sse(X, labels, centers)
print("Sum of Squared Errors (SSE):", sse)
```

```
return out_data, out_learner, out_classifier, out_object
```

OUTPUT:

Algorithm	K=2	K=3	K=4	K=5	K=6
K-Means	1814.677	890.441	580.894	309.660	198.202
Hierarchical	1899.185	934.462	644.683	373.579	236.558

Fig:Before Normalization

Algorithm	K=2	K=3	K=4	K=5	K=6
K-Means	1820.864	897.255	543.057	306.926	198.798
Hierarchical	1895.942	941.219	645.700	374.597	237.576

Fig:After Normalization

Analysis:

- By applying and comparing k-means, and hierarchical clustering to the water treatment dataset, we found after normalization at k=4 sum of squared error is reduced.
- The silhouette score value is highest at k=4 i.e. 0.126.
- After analyzing sum of squared error before and after normalization, the K-Means algorithm is best suited for the water treatment dataset rather than hierarchical clustering.
- Using K-means algorithm for clustering in the context of water treatment analysis can provide good clusters.

CHAPTER 4
CONCLUSION AND FUTURESCOPE

Conclusion

The complexity and variability of water quality parameters in water treatment processes present significant challenges for optimizing treatment efficiency and ensuring regulatory compliance. This project applied clustering techniques to a comprehensive dataset of water quality measurements taken at various stages of the treatment process, aiming to uncover hidden patterns and relationships that can inform and enhance water treatment operations.

Through the application of clustering algorithms, we successfully characterized distinct groups of water samples with similar quality characteristics. These clusters provided a nuanced understanding of water quality at different treatment stages, revealing specific profiles associated with each stage. This deeper insight into water quality dynamics is crucial for identifying areas for process improvements and tailoring treatment strategies to the unique needs of each cluster.

Additionally, the clustering approach enabled the detection of anomalies and inefficiencies within the treatment process. By recognizing patterns in the data, we were able to highlight deviations from expected performance, providing early warning signals for potential issues. This capability for early detection allows for timely interventions and corrective actions, thereby preventing minor issues from escalating into significant problems that could impact water quality and operational efficiency.

Additionally, the clustering approach enabled the detection of anomalies and inefficiencies within the treatment process. By recognizing patterns in the data, we were able to highlight deviations from expected performance, providing early warning signals for potential issues. This capability for early detection allows for timely interventions and corrective actions, thereby preventing minor issues from escalating into significant problems that could impact water quality and operational efficiency.

FutureScope

The project's future scope includes exploring advanced modeling techniques, enhancing featureengineering, integrating external data sources, prioritizing model interpretability, and optimizing scalability.

REFERENCES

1.**Burlando, B., Casagrande, L., & Galli, A. (2010). Cluster analysis for identifying wastewater treatment plant efficiency patterns.** This paper applies clustering methodologies to categorize wastewater treatment plants based on their efficiency levels, taking into account various operational and environmental factors. It provides insights into the common characteristics of highly efficient plants.

2. **Ahmad, S., & Tahar, R. M. (2014). Water quality and treatment plant performance analysis using multivariate statistical techniques: A case study.** This study uses clustering and other statistical methods to analyze water quality data and treatment plant performance, aiming to identify patterns and correlations that can help in optimizing treatment processes.

3.**Henze, M., & Comeau, Y. (2008). Wastewater characterization and treatment efficiency analysis using clustering algorithms.** The authors employ clustering techniques to analyze wastewater characteristics and treatment efficiencies across different plants. The study helps in understanding the variability and commonalities in treatment performance.

4. **Bastidas, A. A., & Kennedy, C. A. (2010). Performance benchmarking of water treatment plants: A clustering approach.** This paper benchmarks water treatment plants by clustering them based on performance indicators, identifying best practices, and common challenges among different clusters of plants.

5.**Lee, C. H., & Choi, Y. J. (2012). Identifying operational patterns of water treatment plants using clustering methods.** This research focuses on using clustering techniques to identify and analyze operational patterns in water treatment plants, aiming to improve operational efficiency and management practices.

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)
Seshadri Rao Knowledge Village, Gudlavalleru

Department of Computer Science and Engineering

Program Outcomes (POs)

Engineering Graduates will be able to:

1. **Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/development of solutions**: Design solutions for complex engineering problems anddesign system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
4. **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions., component, or software to meet the desired needs.
5. **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant tothe professional engineering practice.
7. **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need forsustainable development.
8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9. **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and effective reports and design documentation, make effective presentations, and give and receive clear instructions.
11. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
12. **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Program Specific Outcomes (PSOs)

PSO1 : Design, develop, test and maintain reliable software systems and intelligent systems.

PSO2 : Design and develop web sites, web apps and mobile apps.

PROJECT PROFORMA

Classification of Project	Application	Product	Research	Review

Note: Tick Appropriate category

Data Mining Outcomes	
Course Outcome (CO1)	Describe fundamentals, and functionalities of data mining system and data preprocessing techniques.
Course Outcome (CO2)	Illustrate the major concepts and operations of multi dimensional data models.
Course Outcome (CO3)	Analyze the performance of association rule mining algorithms for finding frequent item sets from the large databases.
Course Outcome (CO4)	Apply classification algorithms to solve classification problems.
Course Outcome (CO5)	Use clustering methods to create clusters for the given data set.

Mapping Table

CS3509 : DATA MINING															
Course Outcomes	Program Outcomes and Program Specific Outcome														
	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12		PSO 1	PSO 2
CO1	1	1										1			
CO2	1											1			
CO3	2	3	2									2		1	
CO4	2	2	3	2								2		2	
CO5	1	2	3	1								2		1	

Note: Map each Data Mining ouvtcomes with POs and PSOs with either 1 or 2 or 3 based on level of mapping as follows:

1-Slightly (Low) mapped 2-Moderately (Medium) mapped 3-Substantially (High) mapped