International Journal of Engineering & Technology, *7 (3.29) (2018) 655-659*

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Analysis of Waste Water Treatment plant to enhance the Eco-friendly Environment using Data Mining Techniques

Meghana. V[1],MamathaBai.B.G[2],Jharna Majumdar[3]

[1]MTech Student, [2]Assistant Professor, [3]Dean and HOD
[1, 2, 3] Department of M. Tech CSE, NitteMeenakshi Institute and Technology, Bengaluru, India
**Corresponding author E-mail: meghana2294@gmail.com*

Abstract

E-commerce is one of the rapidly booming sectors in India today, thanks to the rising internet user base and faster mobile penetration. The E-commerce industry is a complex ecosystem as it involves huge transaction volumes, complex procurement and logistics systems and reliance on new technologies for customer access and payment transactions. This complexity has given rise to frauds and revenue leakages which is impacting the revenue for the ecommerce companies. Hence the major concern facing the Ecommerce sector today is how to mitigate the revenue loss. Very few studies have been done in academic literature in this area hence the objective of this study is to understand the sources of revenue leakage in the ecommerce sector and propose solutions for mitigating these revenue leakages. The study focuses on 2 major areas of revenue leakage viz. Customer side, Vendor side. The proposed revenue assurance model will be helpful to Ecommerce companies for detecting the sources of revenue leakages in the abovementioned areas and plugging the same thereby reducing losses. The study can also be helpful for consulting companies who are in the business of revenue assurance and fraud management for the ecommerce companies.

Keywords: *Data Mining, Waste Water Treatment Plant, G-means, CLARA, DBSCAN*

I. Introduction

Data mining is a process that discovers patterns in massive datasets which contributes to machine learning, statistics and information systems. Data processing is finding of hid discerning information from immense data collections, as it's an intense new innovation with impressive potential to change organizations to focus on the foremost imperative information in their information distribution centers. Data processing is that the manner toward work data from alternate points of read and compression it into useful information which will be used to create financial gain, cuts costs, or both.. It permits shoppers to interrupt down data from a good vary of measurements or edges, classify it, and description the connections distinguished. In fact, data processing is that the manner toward discovering connections or examples among several fields in immense social databases. Data mining is otherwise known as data Discovery in info, alludes to finding or "mining" learning from plenty of knowledge. Data processing systems are units used to figure on large volumes of knowledge to seek out the examples [1]. Data mining are units used as a district of various fields like human services, farming, business, instruction and then on.

1.1Waste Water Treatment Plant

Water is the essential need for living and according to the statistical studies it says that nearly 3.21x10^5 million gallons of surface water and 7.7x10^4 million gallons of ground water are used by people daily. As the result of the fast growing technology water resources are polluted. Due to these conditions the quantity of the polluted water has increased. Keeping in mind the end goal to tackle this issue a WTTP's that is a Wastewater treatment plants are huge non-direct frameworks subject to expansive annoyances in influent stream rate and toxin stack, together with vulnerabilities concerning the organization of the approaching wastewater. In any case these plants must be worked constantly, meeting stricter and stricter directions.

2. Literature Survey

A data set with ‘n’ objects produces k number clusters. An objective dividing is required to form the clusters of similar and dissimilar objects as clusters. K-means clustering is introduced by predicting a whole of N data samples present in dataset which consists of K number cluster. Therefore, for step 1 we select the K data sample and then it is unevenly separated to form K classes. In upcoming step we consider the calculation of Euclidean distance for each partition of the sample. For the similarity measurement this calculated distance is used. Within every sample of data and among cluster center with minimum distance, on behalf of the class the clustering datacenters will be divided. For the divided classes, recomputation of the arithmetic mean for each data is taken under a new center. Based on to the new centers obtained, all the following steps are recalled, to re-divide the obtained clusters into further new cluster ones as shown in [2]. The main disadvantage is that it is used only when the mean is known and user is supposed to specify k value.

K-means has some disadvantages like the k value must be specified and the efficiency in case of large datasets is less to overcome this problem the improvised versions of k-means was introduced. G-means (Gaussian –means) was introduced to solve the problem of k-value specification. For first k-means algorithm is used to get the first set of cluster center and it checks is each cluster points obeys the Gaussian distribution through Anderson-Darling test. If it doesn't obey Gaussian distribution then the centers are split into new centers and k-means is applied on them. If it obeys the Gaussian distribution then obtained clusters are retained and the process stops. K-means algorithm runs multiple times to obtain optimal solution for G-means referred from [3, 4].

The objective of K-Medoid is to consider the high intracluster and low intercluster correspondence between the objects and simultaneously reducing the unlikeness of all the medoids in their respective clusters. Compared to previous method, K-Medoids chooses data points as centers of clusters this makes it more robust in handling noise and outliers as related in [5]. It is also known as PAM (Partitioning around Medoids) algorithm [6] it doesn't scale good for large datasets.

For large datasets we consider CLARA(Clustering Large Applications) algorithm where rather than finding similar objects for the complete knowledge set, it attracts an instance from the info set by applying a formula on the samples and finds the medoids for it. Supposedly if an instance is chosen in an arbitrary manner then obtained medoids of the sample would estimate the whole medoids of the dataset. For better approximations we use CLARA as it takes many samples and the best clusters are given as an output. For accuracy, quality measure is performed based on the common unlikeness of objects in the provided dataset as referred from [7, 8]. It chooses a sample of nodes at first and then neighbors are selected from the obtained sample. It restricts the search to a specific area of the original data.
In Density-Based Method, objects are clustered depending upon the distance between them. Arbitrary shaped clusters are formed in this type of methods. The basic idea here is to continue clustering process till the number of objects in the neighborhood exceeds the given threshold. This method is used to eliminate outliers [9, 10]. Clusters are formed according to a density-based connectivity approach. This method gets the regions which have high density and finds out the clusters of arbitrary shape in the given database with outliers [11].

3. Proposed Methodology

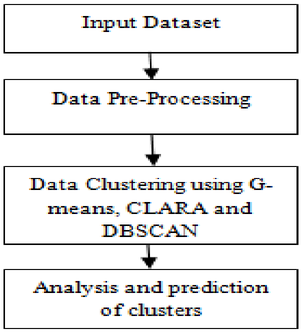

Input Dataset
Data Pre-Processing
Data Clustering using G-means, CLARA and DBSCAN
Analysis and prediction of clusters

Fig 1: Proposed flow

3.2algorithms Used

3.2.1 Gaussian- means (G-means)

Input: Waste Water Treatment Plant as input
Output: Formation of clusters
Step 1: Let C is said to be a set where, (C $\leftarrow\{\bar{x}\}$).
Step 2: By applying k-means,C$\leftarrow$kmeans(C, X).
Step 3: Data point assigned to cj is {xi|class (xi) = j}.
Step 4: Statistical test is used to find whether each {xi|class(xi) = j} use Gaussian distribution.
Step 5: If the value obtained is Gaussian then keep the value as cj, else cj must be replaced with new formed centers.
Step 6: Recall the step 2, until the number of clusters remains same.

In g-means algorithm, we initialize the centers randomly and run k-means algorithm to obtain the new centers and on the obtained centers the Gaussian distribution is applied using Anderson-darling test. If the test obtained is more than the critical value then the centers are splitted and a new center is obtained. If the test is less than the critical value then centers remains same and the clusters are obtained and the process stops.

3.2.2 Clustering Large Applications (CLARA) algorithm.

Input: Waste Water Treatment Plant as input
Output: Formation of clusters
Step 1: Loop from i = 1 to 5, repeat.
Step 2: Instances of K is retrieved randomly from whole dataset
Step 3: Then find k medoids by calling PAM Algorithm.
Step 4: k medoids similar to Oj is determined for each object in the dataset.
Step 5: The average dissimilarity is calculated in the above step.
Step 6: If obtained value in step 4 is less then set the value as current minimum.
Step 7: Then fetch k medoids found in Step 2 as the finest medoids.
Step 8: To start next level of iteration, Loop back to Step 1.
In above algorithm, instead of taking random samples directly here we first take the many small samples from the dataset and then take the random samples from that sample. But the sample is chosen which is similar to the original dataset. The next process is to calculate the Euclidian distance between the two medoids and to calculate minimal distance between them. Once the cluster is obtained compare it between the other samples is performed and with the help of the dissimilarity measures a best cluster will be returned.

3.2.3 Density Based Spatial Clustering of Application with Noise (DBSCAN)

Input: Waste Water Treatment Plant as input
Output: Formation of clusters
Step 1: Select a point as p
Step 2: Retrieve all points which are density reachable from point p with respect to ε and MinPts.
Step 3: Cluster is formed when p value is a core point.
Step 4: A condition when p is border point then traverse to the next point in the data.
Step 5: Loops till all the points are traversed.

In DBSCAN algorithm, it can find any shaped cluster. A point is selected and a condition is checked whether it is density reachable with respect to other clusters or not. If the chosen point is a core point then a cluster formation is performed else it is considered as a border point and it traverses the next point available in the database. The method continues till the points are visited.

3.3. Input Dataset

The dataset consist of Waste Water Treatment Plant records which have been considered for the analysis. It is taken from the UCI Machine learning repository [12]. Totally 38 attributes are there in the dataset where, each attributes represent the amount of components present at different levels of water purification.

Table 1: Attributes from the dataset

Sl.No	Attributes	Sl.No	Attributes
1	Input flow to the plant (Q-E)	20	Secondary settler - Volatile suspended solids (SSVD)
2	Zinc content to the plant (Zn-E)	21	Secondary settler - Sediments (SED-D)
3	pH value for the plant (Ph-E)	22	Secondary settler - Conductivity (COND-D)
4	Biological oxygen demand (DBO-E)	23	Output pH from the plant (Ph-S)
5	Chemical composition of Oxygen (DQO-E)	24	Output - Biological oxygen demand (DBO-S)
6	Suspended solids (SS-E)	25	Output - Chemical composition of oxygen (DQO-S)
7	Volatile suspended solids to the plant (SSV-E)	26	Output - Suspended solids (SS-S)
8	Sediments (SED-E)	27	Output volatile suspended of solids (SSV-S)
9	Input conductivity to the plant (COND-E)	28	Output sediments from plant (SED-S)
10	Primary settler - pH (Ph-P)	29	Output conductivity from plant (COND-S)
11	Primary settler - Biological oxygen demand (DBO-P)	30	Performance - Biological oxygen demand - Primary settler (RD-DBO)
12	Primary settler - Suspended solids (SS-P)	31	Performance - Suspended solids - Primary settler (RD-SS-P)
13	Primary settler - Volatile suspended solids (SSV-P)	32	Performance - Sediments - Primary settler (RD-SED-P)
14	Primary settler - Sediments (SED-P)	33	Performance -Biological oxygen demand - Secondary settler (RD-DBO-S)
15	Conductivity to primary settler (COND-P)	34	Performance - Chemical composition of oxygen -Secondary settler (RD-DQO-S)
16	pH to the secondary settler (Ph-D)	35	Global performance - Biological oxygen demand (G-DBO)
17	Secondary settler - Biological oxygen demand (DBO-D)	36	Global performance - Chemical composition of oxygen (G-DQO)
18	Secondary settler - Chemical composition of Oxygen (DQO-D)	37	Global performance - Suspended solid (G-SS)
19	Secondary settler - Suspended solids (SS-D)	38	Global performance - Sediments (G-SED)

3.4 Data Preprocessing

The primary step that is considered in data mining process is data cleaning and preparation. In order to solve the problem we first identify the different types of data that are missing and decide on different approaches. Here, we are using null value method to handle the missing values, which is replacing the missing with the null value. If a column exceeds null values with more than 60-70 percent then deleting that entire column would be the best option because including it would end up in inappropriate results.

4. Experimental Results and Analysis

The purity of the waste water depend on the components present in the water such as Ph value, biological and chemical demand of oxygen, suspended substances and so on. Taking the values of these components clusters are formed for primary settler and secondary settler. Formed clusters are then analyzed and separated as polluted water, unfit water and fit water which is used for drinking purpose. The clusters formed for primary settler is shown in Fig 2a and for secondary settler in Fig 2b.

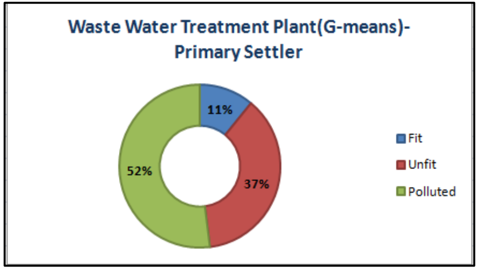


Fig 2a: Primary Settler

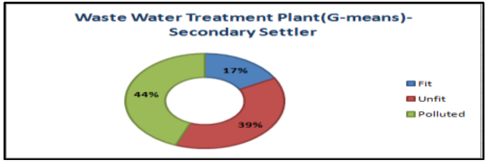


Fig 2b: Secondary Settler

Fig2: Clusters formed from the algorithm- G-means

The result obtained from the primary settler when compared to secondary settler, the fit and unfit water is more in secondary settler. As shown in the Fig 2a and b the presence of pure water is 11%, unfit is 37% and polluted is 52% whereas in secondary settler the presence of pure water is 17%, unfit is 39% and polluted is 44%.The primary settler contains more polluted water than the secondary settler. This proves that the water is more purified in secondary settler than in primary settler. As shown in Fig 3 the secondary settler contains more fit water and less polluted water when compared with the primary settler using G-means algorithm.

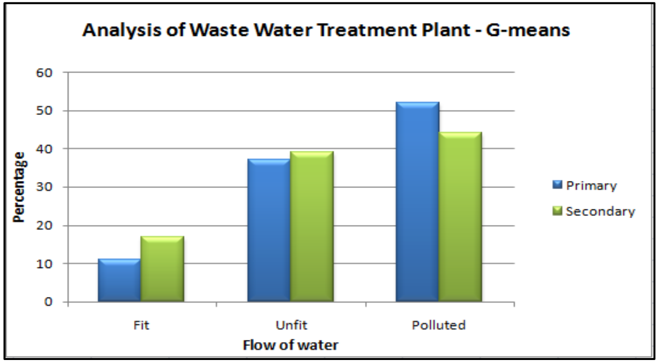


Fig 3: Analysis of Waste Water Treatment Plant – G-means

Table 2: Optimal α value calculation (for 100 instances)

Alpha value (α)	**Primary Settler**		**Secondary Settler**	
	True Positive	**True Negative**	**True Positive**	**True Negative**
0.0004	52	48	54	46
0.0003	61	39	60	40
0.0002	69	31	68	32
0.0001	72	28	71	29

The significant level for the test is chosen as α that is said to be the probability of making a Type I error and here an adjustment method is used to reduce the type I error. In order to find the final k centers we make a statistical test which means the value must not be extreme. Here we have selected 100 instances as sample and considered it further steps. We have ranged α value and have checked for each condition in order to get the optimal solution. As per our analysis and as shown in Table 2 we can say that the most optimal solution is obtained for the value α=0.0001 where the true positive value obtained is more when compared to others in both the primary and secondary settler.

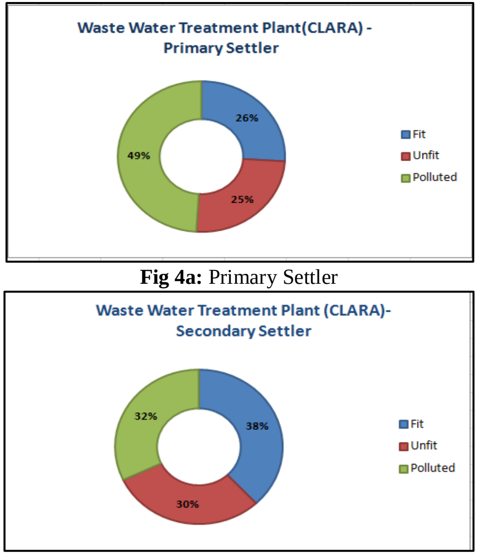


Fig 4a: Primary Settler

Fig 4b: Secondary Settler

Fig 4: Clusters formed from the algorithm- CLARA

As shown in the Fig 4a and 4b the presence of pure water is 26%, unfit is 25% and polluted is 49% whereas in secondary settler the presence of pure water is 38%, unfit is 30% and polluted is 32%.The primary settler contains more polluted water than the secondary settler. This proves that the water is more purified in secondary settler than in primary settler. As shown in Fig 5 the secondary settler contains more fit water and less polluted water when compared with the primary settler using CLARA algorithm.

Fig 4a shows the clusters formed in the primary settler and fig 4b shows the clusters formed in the secondary settler. The result obtained from the primary settler when compared to secondary settler, the fit and unfit water is more in secondary settler.

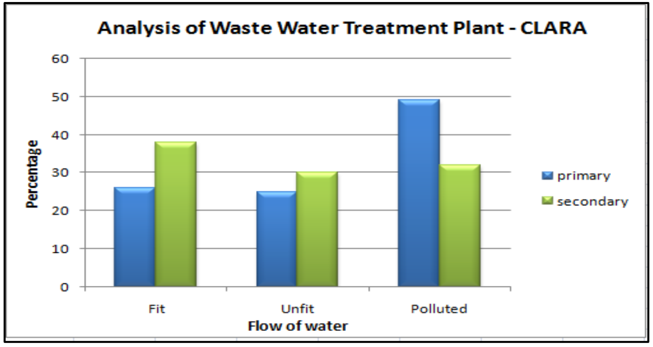


Fig 5: Analysis of Waste Water Treatment Plant using CLARA

The above analysis was for the CLARA algorithm, now we introduce DBSCAN. Here the clusters are formed based on the density of the node. If the node has high density then the more clusters are formed which is based on the noise. The clusters are formed as shown in the fig 6a and 6b

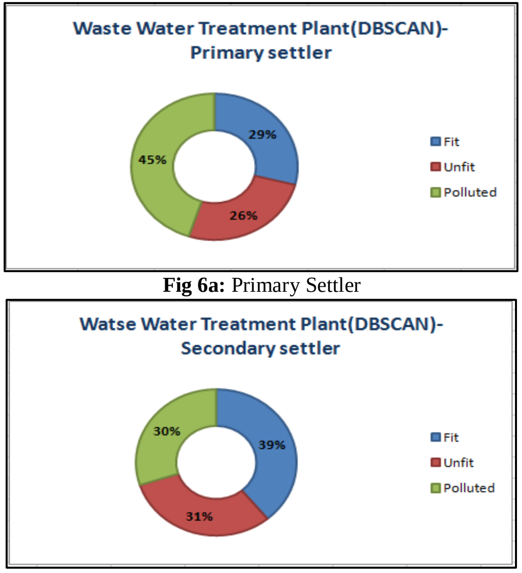


Fig 6a: Primary Settler

Fig 6b: Secondary Settler

Fig 6: Clusters formed from the algorithm- DBSCAN

As shown in the Fig 6a and 6b the presence of pure water is 29%, unfit is 26% and polluted is 45% whereas in secondary settler the presence of pure water is 39%, unfit is 31% and polluted is 30%.The primary settler contains more polluted water than the secondary settler. After comparing the clusters formed from primary and secondary settler we can say that the Waste Water Treatment Plant is working fine. As shown in Fig 7 the secondary settler contains more fit water and less polluted water when compared with the primary settler using DBSCAN algorithm.

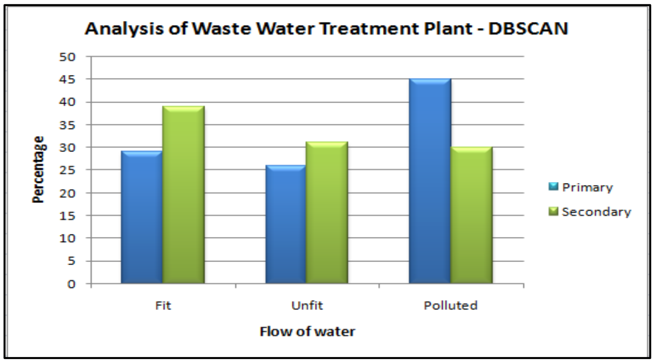
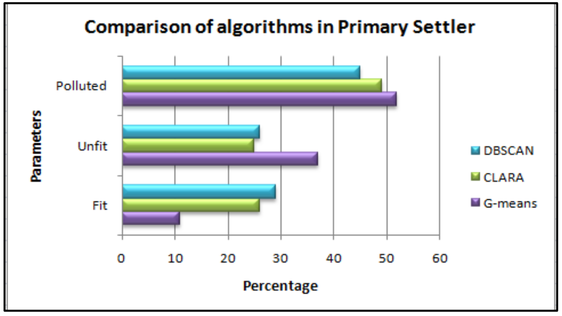
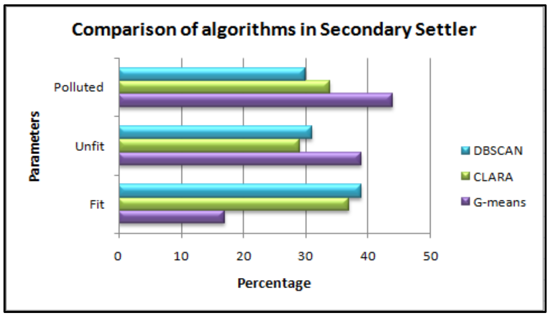


Fig 7: Analysis of Waste Water Treatment Plant using DBSCAN

By comparing the figure 8a and 8b we can say that primary settler is different from secondary settler and more fit water is present in the secondary settler. This shows that the results obtained are more approximate to the analysis performed. The figure 8a and 8b also says the performance of the particular algorithm for the given datasets which shows its efficiency. The result obtained for the below analysis for fit water in DBSCAN is 29%, CLARA is 26% and for G-means is 11%. The similar fashion is seen in secondary settler also, from this we say that DBSCAN algorithm is more efficient than CLARA and G-means for the given dataset.



8(a) Primary Settler



8(b) Secondary Settler

Fig 8: Comparison of algorithms

5. Analysis and Conclusion

The analysis says that for any conclusion the primary analysis might not be sufficient so we go for secondary analysis to obtain more efficiency. As we observe the result we say that the purification result obtained for secondary settler is more efficient than the primary settler. The major purpose is to check the working of waste water treatment plant in order to obtain the efficiency in the process of purification. G-means, CLARA and DBSCAN clustering algorithm is used to cluster the components present in water from both the primary and secondary settler from which the efficiency of the process is obtained. The experimental results say that good performance is obtained for DBSCAN than in G-means and CLARA.

Acknowledgement

The authors express their sincere gratitude to Prof. N. R Shetty, Advisor and Dr. H C Nagaraj, Principal, NitteMeenakshi Institute of Technology for giving constant encouragement and support to carry out research at NMIT.
The authors extend their thanks to Vision Group on Science and Technology (VGST), Government of Karnataka to acknowledge our research and providing financial support to setup the infrastructure required to carry out the research

7. References

[1] Deepashri.K.S, AshwiniKamath, “Survey on techniques of data mining and its applications”, International Journal of Emerging Research in Management &Technology ,February 2017,ISSN 2278-9359 ,Volume-6, Issue-2.

[2] JiaQiao, Yong Zhang, “Study on K-means Method Based on Data-Mining”, IEEE, 2015.

[3] TomislavErdelić, SilvijaVrbančić, LovroRožić, “A Model of Speed Profiles for Urban Road Networks Using G-mea-ns Clustering”, MIPRO 2015, 25-29 May 2015, Opatija, Croatia

[4] Aislan G. Foina, JuditPlanas, Rosa M. Badia, “P-Means, a Parallel Clustering Algorithm for a Heterogeneous Multi-Processor Environment”, Intelligence Research Institute (IIIA), IEEE, 2011

[5] Greg Hamerly, Charles Elkan, “Learning the k in k-means”, Department of Computer Science and Engineering, University of California

[6] Xiao Dong, Zhongnan Zhang, “Research and Implementation of PAM Algorithm with Time Constraints”, International Conference on Informative and Cybernetics for Computational Social Systems (ICCSS), 2014

[7] M. OmairShafiq, Eric Torunski, “A Parallel K-Medoids Algorithm for Clustering based on Map Reduce”, International Conference on Machine Learning and Applications, IEEE,2016

[8] Raymond T. Ng and Jiawei Han, Member, “CLARANS: A Method for Clustering Objects for Spatial Data Mining”, IEEE Transactions on knowledge and Data Engineering, September/October 2002, VOL. 14, NO. 5

[9] S.Vijayarani, S.Nithya, “An Efficient Clustering Algorithm for Outlier Detection”, International Journal of Computer Applications (0975 – 8887), October 2011, Volume 32– No.7

[10] Glory H.Shah,“An Improved DBSCAN, A Density Based Clustering Algorithm with Parameter Selection for High Dimensional Data Sets”,Nirma university international conference on engineering, 06-08december, 2012

[11] Yuchao Zhang, Hongfu Liu, Bo Deng, “Evolutionary Clustering with DBSCAN”, 2013 Ninth International Conference on Natural Computation (ICNC)

[12] https://archive.ics.uci.edu/ml/datasets/water+treatment+plant