PROCESS UPSET DETECTION IN WASTEWATER TREATMENT PLANTS USING CLUSTERING AND DEEP LEARNING

Abstract

The main objective of this study is to classify Wastewater Treatment Plants (WWTP) operational state in-order to predict process upsets at various stages of the treatment process based on influent data. Prediction of WWTP process upset detection is important for WWTP operators to predict and troubleshoot potential process upsets that may lead to effluent off-specification. This research will utilize Machine Lachine Learning (ML) methods including various clustering techniques created in python. The model will predict a process upsets condition using clustering by distinguishing data entries with upsets conditions from normal conditions.

Keywords: Wastewater Treatment Plant, WWTP, Clustering, Machine Learning, Model, Prediction.

Author

Mohammad Manzoor
Saudi Aramco
Dhahran, Saudi Arabia
moh.manzoor@googlemail.com

I. INTRODUCTION

Prior to being disposed of into rivers or other water bodies, water collected from homes or businesses must be cleaned. According to this perspective on environmental contamination, Wastewater Treatment Plant's (WWTPs) play an important part by removing contaminants and reusing waste water. In contrast, WWTPs are exceedingly complicated systems that must operate at a high level regardless of seasonal or human activity changes. The treatment process must be monitored in real time, which is costly and necessitates specialized equipment in order to administer a WWTP safely and efficiently. Ammonia, dissolved solids, ions, suspended particles, and organic matter are all monitored by sensors in WWTP influents. Deploying fully working sensors, having human operators oversee them, or even modifying sensor position are all nearly difficult. As a result, a major research priority is to accurately identify sensor failures. Input monitors, notably ammonia detection sensors in nitrification oxidative tanks, can have a variety of faults, although current work focuses on error detection in them. Using machine-learning techniques and algorithms to automatically analyze the data generated by WWTPs, a potential solution can be found in the automatic detection of such system flaws. Ambient expert systems can thus be linked to WWTPs to maintain high efficiency and low outputs at all times, and where faults can be dealt with quickly. Even before it was released into the environment, water had to go through a series of operations before it was safe to release. According to water source and progress, there are a variety of water systems for wastewater. Commercial wastewater is produced by businesses, such as retail establishments, markets, office eateries, buildings, hotels, and hospitals (but not by manufacturing and construction). Contaminated waste from traditional sources must be treated.

WWTP are used to remove contaminants or substrates such as nitrates and phosphates, as well as reduce indicators such as Biological Oxygen Demand (BOD) and Chemical Oxygen Demand (COD).

The most common configuration of a wastewater plant generally compromises of 4 stages, summarized below:

- **Screening & Grit removal**: for removal of debris, grit and large objects,
- **Primary Clarifiers:** removal of suspended solids including organic and inorganic matter.
- **Biological Treatment:** (e.g. Activated Sludge Process - ASP) -reduction of BOD, COD, Ammonia, Nitrates, and Phosphates,
- **Secondary Clarification:** Finer solids from the biological process are removed. This is also an integral part of the ASP biological process where activated sludge (bacteria) is recycles back to the biological process to ensure healthiness of the process.

Governments and their environmental protection agencies set strict standards for receiving watercourses that vary between localities and receiving watercourses. These standards interpreted teed into wastewater effluent specifications that wastewater process plants must adhere too. Measurements of specific nutrients and indicators provide an insight into how the wastewater process plants are operating, and to whether they will or are adhering to local governmental requirements

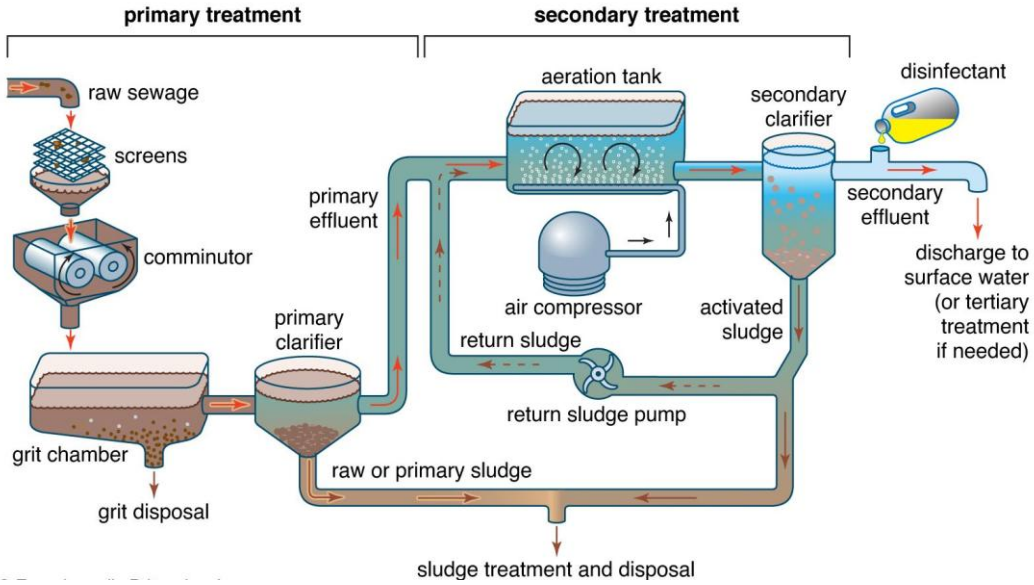


Figure 1: Typical ASP WWTP Treatment Process Configuration [15]

II. AIMS AND OBJECTIVES

The proposed research primarily focuses on achieving the following aims:

- Classify the WWTP operational state to predict process upsets through process variables of the plant at each stage of the treatment process.
- Implement K Means and Hierarchical Agglomerative clustering algorithm on the dataset
- Implement an Autoencoder model and use clustering on the bottleneck layer generated by the model.
- Do comparative studies between the Autoencoder based clustering and traditional clustering.

The main objective of the proposed research is classifying WWTP operational state to predict process upsets through process variables of the plant at each treatment stage of the WWTP process.

To achieve this research a use clustering to divide the instances of data present in the dataset into separate clusters i.e. to distinguish the entries with process upset from the one without process upsets.

1. **Scope of Study:** The scope of the study is limited to the following:

- Classify the WWTP operational state to predict process upsets.
- WWTP with Activated Sludge Processes.
- Domestic waste water treatment.

2. **Significance of the Study:** Upsets and faults can generally be divided into three types:

- Individual errors, which are instances of a single data point that are unexpected compared to other data;
- Numerous examples that are out of place in one context but are not when considered as a whole;
- Collective flaws appear as an uneven accumulation of cases concerning other data patterns [1].

Collective flaws are not necessarily uncommon in and of themselves, but a particular sequence of them is. To put it another way, it's termed a collective fault when the measured values of a succession occur in an unanticipated order or an unsatisfactory combination. Machine-learning techniques have been used in WWTP sensors to identify the first two types of failures, but the third and most difficult one—collective faults—haven't been given enough consideration.

The discussed approaches did not detect process upsets without the use of external information, and it is clear that there exists a gap when it comes to the existence of an approach which uses data-based methods to correctly investigate the process upsets in a WWTP.

III. RESOURCES

1. **The Dataset:** To adhere to effluent requirements, it is important to monitor and check wastewater parameters that can indicate whether the treatment process is operating properly. Data of key parameters at the different stages of the process can indicate the healthiness of each stage and help predict potential process upsets, hence allowing operators and engineers to fix them prior to the process going beyond its control limits. Early detection or prediction is essential as biological treatment such as an ASP has a sludge age of 8 plus days. With any large and complex biological system, early detection is important, as usually when effluent quality is bad, it's normally due to a build-up of upsets resulting in pushing the system beyond its optimum operational envelope.

   Analysis of data is important for early process upset detection and troubleshooting, and usually involved analyzing vast amounts of data and many different variables – this can be beyond the comprehension of engineers and operations that normally rely on a limited amount of data and key indicators. Machine learning provides the advantage of analyzing and identifying trends from a vast amount of data, and many variables from varied sources with varied formats.

   Thus to demonstrate the above approach, the WWTP dataset (https://archive.ics.uci.edu/ml/datasets/water+treatment+plant) has been used.

   The following dataset consists of 39 variables; first one being the date the sensor data was collected making it effectively 38 variables. This sensor data has been collected for 507 days making the dataset thorough enough to use a model for making clustering possible.

The 38 attributes can be described in three categories: Input indicators, output indicators and performance indicators. The input indicators include the following

- **Flow to the plant:** it represents flow of water to the plant; it may vary in quantity and show fluctuations.
- **Ph:** it represents the concentration of H+ ions in the water implying acidity or basicity of water supplied it affects the plant's overall treatment environment as biological process operate best within a certain pH range.
- **Zinc:** is a heavy metal that can both affect ASP reaction kinetics and receiving watercourse aquatic life.
- **BOD:** is a measure of oxygen used by microorganisms to breakdown organic matter.
- **COD:** indicates the amount of oxygen required to breakdown organic material using chemical oxidation.
- **Suspended Solids:** small suspended particles remaining in the wastewater after the filtration step
- Volatile suspended solids: These are the fraction of solid that dissipates after heating and is an indicator of biological content.
- **Sediments:** These are solid particles which come because of turbulence of moving water
- **Conductivity:** It is used to monitor the operation of water purification systems by measuring **t**he ability of water to conduct electricity. It proportional to Total Dissolved Solids.

2. **Hardware and Software Resources Required:** The below hardware and software resources are required to complete the study.

Hardware Requirements

- **CPU:**
  - Windows based: 1.6 GHz or faster, 2-core
  - Mac: Intel processor
- **Memory:** 4 GB RAM;
- **Hard disk:** 4 GB of available disk space
- **Display:** 1280 x 768 screen resolution
- **Graphics:** DirectX 9 or later, with WDDM 2.0 or higher for Windows 10
- **Operating system:** Windows 10 or higher, most recent versions of macOS

Software Requirements

- Web browsers: The current version of Microsoft Edge, Internet Explorer, Chrome, Safari or Firefox
- Latest version of a Python compiler including Anaconda
- Latest version of MS Office Suite.

IV. LITERATURE REVIEW

1. **Wastewater Treatment Plant (WWTP):** An effluent that may be safely discharged back into the environment after being treated for suspended solids is called a "wastewater treatment effluent" [2]. Wastewater Treatment Plants, Water resource Recovery facilities,

and sewage treatment plants are all common names for the facilities that perform the wastewater treatment process. People and the environment can be badly impacted by the pollutants included in wastewater. This necessitates the removal of these components during treatment. Wastewater contains a wide range of contaminants, including:

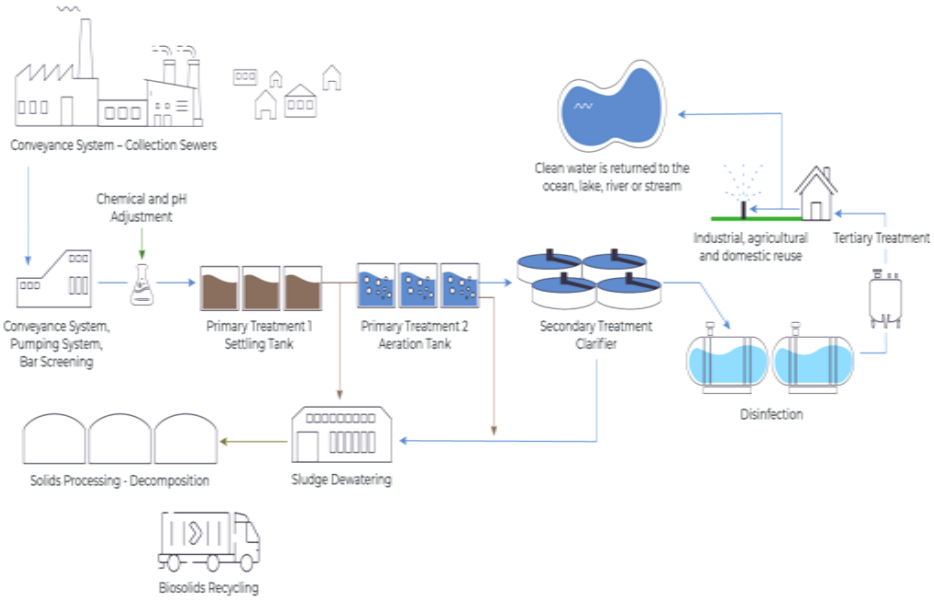
- Pathogenic microorganisms, such as viruses, bacteria, and viroids.
- Decaying vegetation and other waste.
- The term "helminth" refers to the outer layer of a creature (intestinal worms and worm-like parasites)
- It is important to be aware of the hazardous consequences of metals such as Mercury and Arsenic.
- Grease and Oils.
- Pharmaceuticals and Personal care goods.
- Toxic substances such as PAHs, PCBs, phenols, Dioxins, Furans, insecticides, and so on.
- Toxic Chlorinated compounds and Chloramines inorganic.

A multitude of activities that produce wastewater contributes to its creation. Activities such as showering, cleaning, and using the toilet in homes, restaurants, and enterprises all contribute to the generation of domestic wastewater. Mixing grit, nutrient-rich sediments, and other substances with surface rainwater generate runoff [2]. There is a lot of industrialized waste water generated because of numerous chemicals and current job outputs. The term "wastewater" refers to used wastewater that has already been subjected to commercial, industrial, or domestic processing. Domestic wastewater is difficult to clean than industrial wastewater because of its high strength [2]. This service is available for both commercial and residential customers. The clean water that is produced as a result of wastewater treatment can be utilized for a variety of industrial and agricultural purposes, including agricultural irrigation, cooling tower makeup water, and even drinking water.

Treatment processes can include filtration, which is most frequent in the tertiary stage of treatment [3]. Listed below is a brief description of each step:

- **Collection sewers:** Collecting all industrial waste from each of these processes in a single collection sewer
- **Pumping System:** Disbursement of wastewater to following treatment processes through a pumping system
- **Bar screening:** Fines, boulders, sand, and other coarse materials are screened out using a bar screen.
- **Chemical and pH adjustment:** The addition of chemicals to adjust the pH is referred to as "chemical and pH adjustment."
- **Disinfection:** UV light or chlorination can be used for disinfection.
- **Decomposition / Sludge Dewatering:** Drying the sludge following strict environmental rules for disposal or re-use as fertilizer.

An example of a conventional wastewater treatment plant is shown below, along with a description of each stage of treatment



2. Waste Water Treatment Process [Source: [3]]

Wastewater Treatment Process: The following is a typical description of the order in which wastewater treatment activities occur:

- Preliminary Treatment
- Primary Treatment
- Secondary Treatment
- Tertiary Treatment or Advanced Treatment.
-

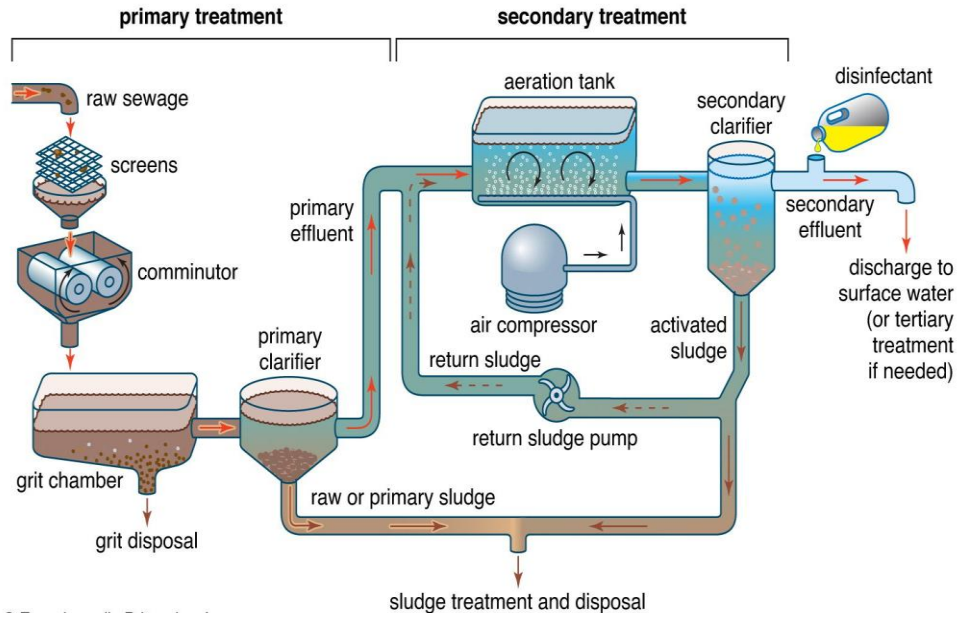


Figure 2: Different Treatment Process [Source: [4]]

- **Preliminary Treatment:** Primary treatment comes after the preliminary treatment of wastewater. Its primary goal is to keep following treatment units safe by reducing operational issues [2]. In the initial stages of wastewater treatment, the most common techniques are screening, neutralizing, equilibrating, adjusting the temperature, and removing grit.

- **Primary Treatment:** In the first step of the process, material that floats or easily settles to the bottom is removed. The processes of screenings, grit removal, combination, and sedimentation are all included in this category of processes. Metal bars spaced closely together form screens. Floating debris, such as rags, wood, and other bulky materials, can pump or clog pipes if they are allowed to float freely in the water. The screens in modern factories are cleansed mechanically, and the waste is quickly buried on the site. Discarded material can be ground and shredded using a comminutor. Flotation or sedimentation removes the shredded debris later [4].

  To remove particles like sand, coffee grinds, and eggshells from the water, long, narrow grit chambers are used. Pumps as well as other plant machinery wear out more quickly when they are in contact with grit. In communities with an integrated sewerage system, where a lot of the silt, sand, and gravel washes off roadways and land following a storm, its disposal is especially critical [4].

  Sedimentation tanks remove solids from the sewage that have passed through screenings and grit chambers. Gravity settling can take place in these tanks, which are also known as main clarifiers, for up to two hours at a time. Solids settle at the bottom of the pipes as the sewage moves slowly through them. A motorized scraper moves the solids known as raw or primary sludge up and down the tank's bottom [4]. To remove the sludge, it is placed in a hopper and then pumped out. Removal of grease as well as other floating debris is accomplished by mechanical skimming machines.

- **Secondary Treatment:** The secondary treatment eliminates the organic matter that was not removed during the main treatment process. “The amount of solids that are removed is increased as a result of this process. The organic contaminants are usually consumed by microorganisms, which turn them into water, carbon dioxide, and energy with their development and reproduction. Despite its steel and concrete construction, the sewage treatment facility provides an ideal setting for this biological activity to occur. Removed organic waste from the treated wastewater helps to maintain the dissolved oxygen balance in a receiving river, stream, or lake [4]. In terms of biological treatment, trickling filters activated sludge processes, and oxidation ponds are the three most common. Rotational biological contacted is the fourth and least prevalent way [4].

  During secondary treatment, the solubilized organic waste that evaded removal during first treatment is eliminated. As a product of this procedure, a greater volume of solids is eliminated. Microorganisms often use organic contaminants as food, changing them in to water, nitrous oxide, and energy they need to thrive. The wastewater treatment plant provides a steel-and-concrete habitat for natural biological activities. The dissolved oxygen balance of a stream, river, or lake can be protected by

removing soluble organic material from the treatment plant [4]. In biological treatment, the trickling filter, the municipal wastewater method, and indeed the oxidizing lagoon are the three most popular options available. It's also possible to use rotating biological contacts [4].

- **Trickling Filter:** A trickling filter consists of a tank packed with a thick layer of rocks. And over top of the stones, settled sewage drips down to the bottom and is gathered for the further treatment. Bacteria congregate and proliferate on the stones as wastewater flows down. To reduce the dissolved oxygen concentration (BOD), sewage must have flowed over these microorganisms constantly. As a result of air traveling upward through crevices between the stones, metabolic activities can maintain enough levels of oxygen [4].

  Secondary clarifiers, also known as settling tanks, follow the trickling filters in the filtration process. Wastewater flows through these clarifiers, which remove germs that get washed off the rocks. The sewage can be recirculated via two or more trickling filters to improve Treatment efficiencies [4].

- **Activated Sludge:** The treatment of activated sludge is finished once an aeration tank and a secondary clarifier have been utilized. The settled sewage and new sludge that comes from the secondary clarifier are both recirculated and sent to the aeration tank for processing. After that, the mixture is added with compressed air through the porous diffusers located at the bottom of the tank. As the air that has been diffused rises to the surface and bubbles, it contributes oxygen and facilitates a more rapid mixing action. There are also mechanical mixers on the surface of the tank that looks like propellers and can be used to introduce air [4].

  In such oxygen-rich settings, microorganisms can thrive, producing activated sludge as a result. Activated sludge is a healthy suspension of biological solids (mainly bacteria). The aeration tank contributes approximately six hours' worth of detention time to the overall total. This makes it possible for the bacteria to absorb the dissolved organics in the sewage, which in turn lowers the BOD. As the mixture moves from the aeration tank into the secondary clarifier, the activated sludge begins to settle towards the bottom of the device. Secondary effluent is water that has been filtered and decontaminated after being skimmed from the top of the clarifier [4]. The sludge is removed from the tank by a hopper located at the bottom of the tank. A negligible portion of the sludge is re-circulated and mixed with the primary effluent while it is being processed in the aeration tank. This recirculation is extremely important to the procedures that include activated sludge. Microorganisms that have been recycled can quickly decompose the organic material that is present in raw sewage thanks to their new environment in the sewer. Seventy percent of the secondary sludge needs to go through an appropriate treatment and disposal process. This is required.

  The activated sludge process can be modified in several ways, some of which include contact stabilization and high-purity aeration, for instance. When methods of contact stabilization are used, the initial settling step of the process is completely

skipped through. These systems are capable of properly treating sewage coming from smaller streams coming from motels, institutions, and other relatively remote sources. These two kinds of treatment are typically administered employing package plants, which are massive steel tanks that have been constructed ahead of time. In oxygen aeration systems, pure oxygen is mixed with activated sludge to produce oxygenated effluent. Utilizing oxygen at a higher concentration allows for a shorter aeration time, which can be cut from six hours down to just two [4].

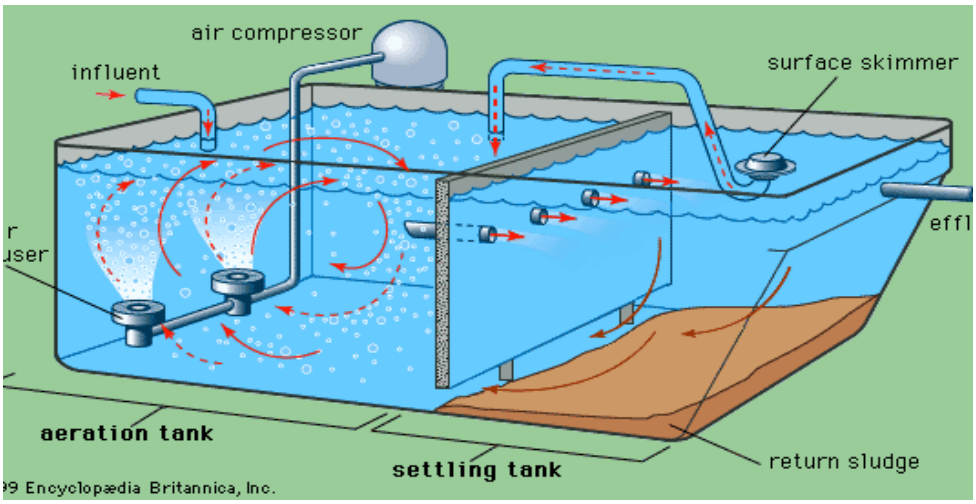


Figure 3: Package Plant [Source: [4]]

- **The Constructing of Aeration Systems:** A number of different aeration devices are currently being utilized in the process of biological treatment in wastewater treatment plants. These systems supply oxygen to the microorganisms that are found in activated sludge. Increased oxygen aeration devices are not as frequent as mechanical dispersed aeration and surface aeration (medium, coarse, and fine-bubble) [33]

- **Surface Aeration:** Surface aeration, which may be performed with Kessener brushes or rotor aerators, was a common practice right up until 2011 [34, 35, and 39]. The surface vertical and horizontal shafts aerators, brush aerators and air pump systems, are the four varieties of aerators that are utilized most frequently in modern times. The utilization of blades in these systems results in the generation of turbulence on the surface of the liquid, which is a crucial component of the process of activating sludge. Some types of biological interface oxidation systems make use of revolving blades, propellers, or filters to break down the organic matter. The creation of a turbulent flow in wastewater is beneficial to the aeration and subsequent biological processes that take place there. These techniques are currently utilized very infrequently, generally during the upgrading process of treatment plants that have shallow tanks (depths of up to 3.5 meters) or circulation ditches. The high energy demand of these processes has been demonstrated by a number of research papers [36, 37] due to the fact that the corresponding energy consumption is frequently higher than 0.7 kWh/m3 of wastewater.

- **Oxygen Transfer Efficiency :** The oxygen transfer efficiency of an aeration system is one of the most important considerations to take into account (OTE). After another method of air has diffused through a specified depth in an aeration reactor, this method estimates the oxygen content (either in percentage or weight) in the treated wastewater. This can be done in either unit of measurement. When the OTE rises, the amount of air that must be supplied to the reactors to keep up with the required volume reduces. There are a variety of factors that can influence the efficiency of oxygen transfer during fine-bubble diffusing aeration, such as [38]:

  - Substances found in Activated Sludge and Wastewater;
  - For instance: SOTE- Standard Oxygen Transfer Efficiency;
  - Dissolved Oxygen Concentration (g O2/m3);
  - Positioning depth [m] for diffusers;
  - The volume of air moving through a diffuser per unit of time [m3 /diffuser h];
  - The temperature of the wastewater;
  - The diffuser is clogged

- **Oxidation Pond:** Bacteria, Sunlight, and Algae all collaborate during the treatment of wastewater in a lagoon or stabilization pond to decompose organic debris. Algae rely on the inorganic chemicals and carbon dioxide that are created by bacteria living in water to flourish. During the process of photosynthesis, algae are responsible for producing the oxygen that Aerobic microorganisms require. The installation of mechanical air pumps, that will provide the pond with more oxygen, makes it possible to lower the size of the body of water. To eliminate the sludge accumulation in the pond, dredging will need to be done. To remove any remaining algae from the pond effluent, one of two methods filtration or a combination of settling and chemical treatment may be utilized [4].

- **Rotating Biological Contactor:** The primary wastewater is partially submerged within a collection of enormous plastic disks that have been mounted on a horizontal path. The shaft spins in such a way that it is accessible from either air or wastewater at particular intervals to make sure that a layer of bacterium grows on each disk at regular intervals [4].

- **Tertiary Treatment:** If the water that will be receiving the effluent from the secondary treatment process is highly prone to contamination, additional tertiary treatment may be necessary [4].

- **Effluent Polishing:** Polishing wastewater to remove extra suspended particles and BOD is an appropriate treatment for secondary effluent. Granular media filters, like those used to purify the drinking water, are the most common method. Backwash water is stored in tanks above the polishing filters, which are usually prefabricated units. Micro strainers, the type used to treat municipal water sources, can also be utilized for wastewater effluent polishing [4].

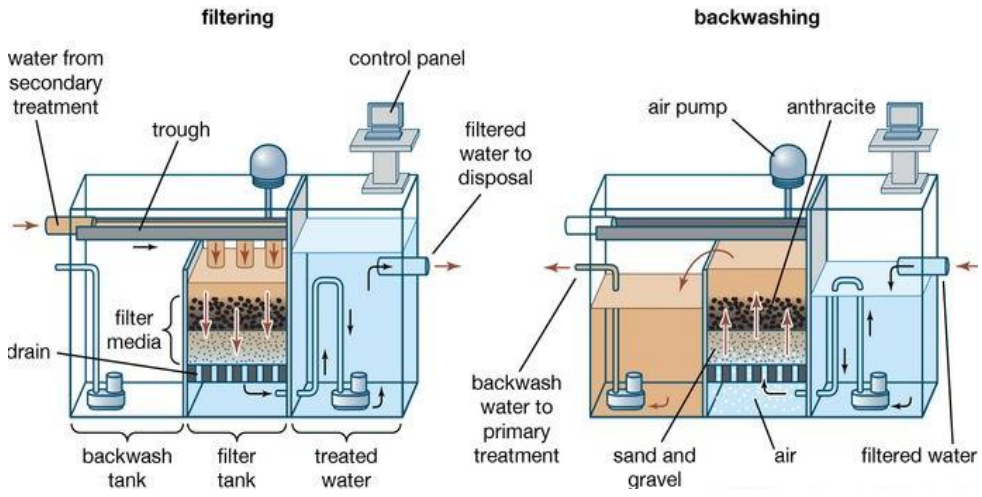


Figure 4: Tertiary treatment for wastewater plants [Source: [4]]

- **Removal of Plant Nutrients:** At this point in the treatment process, it is usual practice to remove any plant nutrients that may be present in the sewage. The removal of organic molecules and phosphates, which together account for the majority of the phosphorus found in wastewater, can be accomplished by the use of an efficient approach known as chemical precipitation. However, as a consequence of this procedure, both the volume and the weight of the sludge would increase. Two of the most important nutrients for the development of plants are found in sewage in the form of nitrogen compounds known as nitrates and ammonia. Ammonia is toxic to fish, and the process by which it is converted to nitrates reduces the amount of oxygen in the water that it enters. Nitrates, like phosphates, contribute to the eutrophication of lakes and the growth of algae in those lakes. Phosphates play a similar role. A technique that is known as nitrification-denitrification can be utilized to get rid of nitrates. Microorganisms are responsible for the first phase of the two-step biological mechanism that converts ammonia nitrogen into nitrates. In the following step, a different type of bacterium degrades the nitrates, which ultimately results in the emission of nitrogen gas into the environment. Because of this, there is a need for additional settling tanks and aeration, which drives up the overall treatment expenses [4].

  The removal of ammonia from sewage can be accomplished using a physicochemical process that is known as ammonia stripping. Chemicals are required to turn ammonia gas into a usable form. The gas is permitted to be released into the atmosphere as the sewage is allowed to fall through the tower. Nitrification-denitrification is a more cost-effective process, although it does not function as efficiently when the temperature is low.

3. **Wastewater Treatment Plants using Clustering:** In the case that it is not feasible to link individual households or components to the public sewage system, communities may prefer to go with a clustered water supply system in order to meet their requirements for wastewater treatment. These scattered treatment plants are only capable of attending to a constrained number of individual connections at a moment. Clustered Wastewater

Treatment Plant systems may employ the same innovations as traditional centralized facilities or individual on-site facilities, but this will depend on the particular applications that need to be treated or the level of treatment that is required for each application. This is a factor that is determined by the particulars of each application. It is possible to dispose of the effluent that is produced by wastewater plant systems in clusters by making use of technologies that are located either on the surface or below the surface [4].

4. **Sewage Treatment Plant (STP):** Sewage treatment plants are equipped to process both domestic and commercial sewage, in addition to the wastewater generated by enterprises [3]. As a consequence of this, sewage is composed of a significant amount of organic material, but it also has the potential to contain some inorganic garbage. The process of treatment consists of three stages: the Primary stage, the Secondary stage, and the Tertiary stage.
5. **Effluent Treatment Plant (ETP):** At the effluent treatment plant, contaminants such as oil and grease, heavy metals, phosphate, and other impurities are cleaned up (ETP). In the course of treatment, it is usual practice to make use of various biological, pharmacological, and physical therapies [3]. It is imperative that wastewater be cleaned, either for reuse or before it is released into the environment, to reduce contamination.

6. **Reverse Osmosis Water Treatment:** To drastically reduce organic pollutants, heavy metals, bacteria, dissolved solids, and other dissolved contaminants, the water is forced through semi-permeable membranes under the influence of high pressure. Safe disposal or reuse of wastewater is possible following treatment [3]. Fine filtering is used to protect RO membranes from clogging and significant damage to the surface of the membrane and layering and to limit the growth of biofilm. Usually used for industrial reuse as a final polishing phase following tertiary treatment.

7. **Faults Categorization:** Faults can generally be divided into three types:

    - Individual errors, which are instances of a single data point that are unexpected compared to other data;
    - Numerous examples that are out of place in one context but are not when considered as a whole;
    - Collective flaws appear as an uneven accumulation of cases concerning other data patterns [5].

    Collective flaws are not necessarily uncommon in and of themselves, but a particular sequence of them is. To put it another way, it's termed a collective fault when the measured values of a succession occur in an unanticipated order or an unsatisfactory combination. Machine-learning techniques have been used in WWTP sensors to identify the first two types of failures, but the third and most difficult one collective fault haven't been given enough consideration.

8. **Fault Detection Methods:** The classification of defect detection techniques can be accomplished through the use of models of time series, statistical methods, and learning models, in that order of application. When it comes to monitoring the sensor data from WWTPs, statistical methodologies have been investigated to a much greater extent than

any others. The Mann Kendall test is an uncomplicated tool for analysing data trends, whereas statistical quality control methods make use of statistical control charts to monitor the progression of the model's processes variables over time. Principle Component Analysis (PCA) and the Kernel Principal Component Analysis (Kernel PCA) charts can be utilized in either univariate or multivariate modelling. Shewhart and cumulative sum charts, as well as exponentially weighted moving averages, are all examples of univariate models [6]. When looking at the second category of learning models, a classification performance problem is taken into consideration. Support Vector Machines [7]; Random Forests [8]; Fuzzy Classification [9]; Neural Networks [9]; Random Forests [8]; and Fuzzy Classification [10], [11]. There have been several studies that have compared various statistical and learning methods making use of information from wastewater sensor devices; [12]. Neural networks, such as radial basis functions, self-organizing maps, multi-layer perception, and functional link neural networks, are the most successful methods for learning to detect errors in WWTP data. Other effective methods include self-organizing maps [13]. The categories that have been discussed previously are capable of successfully capturing not only the individual defects but also the contextual irregularities. They are not capable of accurately detecting the intricate temporal patterns that are present in the collective faults. This necessitated the introduction of time series modeling approaches such as Autoregressive Integrated Moving Average (ARIMA) and the Time Delay Neural Networks (TDNN), which were able to capture sequential trends in WWTP datasets. It is a linear algorithm that calculates the next data value based on the data sequence that came before. After that, a control chart is employed to demonstrate the complexity of the model and figure out whether or not the data is typical. It makes predictions about the nonlinear temporal dependence of signals by employing a TDNN, which is a multivariate neural network with a temporary memory structure. The model is trained on data from time-segmented windows of data [14]. The work by [15] uses eight created datasets to evaluate and contrast the linear ARIMA model with the TDNN model. The results show that TDNN has a significant advantage over ARIMA. To properly segment the data, TDNN uses a ratio that is proportionate to the window's size. By increasing the size of the windows, it is possible to attain increased network dimensions and characteristics. In the alternative, having a window size that is too small can cause critical system dynamics information to be missed.

9. **The Challenge of Fault Detection in the Nitrification Oxidation Tank:** The ammonia-to-nitrate nitrification oxidation chamber is used as a component of the process for the degradation of macro-pollutants. In this tank, nitrogen and carbon are broken down into their parts. The method is made more reliable by the introduction of air into the container. One of the most important requirements for effective purifier management and outstanding purification efficiency at a cost of energy that is affordable is accurate blower control [16]. The major method for controlling the nitrification and oxidation processes is to maintain constant oxygen set temperature and to adjust the amount of airflow necessary to keep it at that temperature. Because of this, the minimum airflow delivered by blowers in this system is greater than that which is required to maintain oxygen set point. As a consequence of this, the purifier suffers from energy losses and dissolved oxygen even when it is operating under a low load. In these tanks, a control process is used to calculate interactively the oxygen performance level that must be maintained in the tank. When the ammonia concentration drops below a predefined value (based on the

oxidation tank's ammonia nitrogen concentration), the set point is reset to zero. Even though measurements of ammonia have been used to regulate the purification process for a considerable amount of time, an inaccurate reading may fail to comply with legal discharge quality requirements or in an excessive amount of energy consumption that is not justifiable. As a result of this, the objective of the proposed initiative is to identify these errors in ammonia readings as promptly and accurately as is practical [16].

10. **Socioeconomic relevance of Detecting and Mapping Wastewater Treatment Plants:** Even though the United Nations has acknowledged that everyone has the inherent right to safe drinking water and sanitation [17], there are still hundreds of millions of people who do not have access to these basic amenities. The identification and mapping of WWTPs can be of tremendous benefit to the expansion and investment choices made by water utility companies, regardless of whether those businesses are public or private. It is impossible to deliver essential services like clean water unless those services are accessible to all people and readily available. By making an effort to incorporate such insights into the process of developing funding programs, governments and other public authorities can both profit from the information provided here. Because of this, the deep learning method based on experimental data from Earth is significantly more useful in areas with limited amounts of charted data. It has important socioeconomic ramifications to use Geo AI to detect and map WWTP since it detects undersupplied areas that can be classed as such based on population data. As a result of this, the needs to build a new JDL method for WWTP mapping that is both effective and automated, and it is this method to be examined.

11. **Geospatial Object Detection and Mapping:** Images (e.g., automobiles or buildings) can now be spotted with exceptional accuracy and speed because of advances in current imaging techniques. Many real-world applications, are including precision farming [18], animal protection [19], and even caring mapping, rely on geospatial feature detection and mapping [20]. Using a deep learning algorithm with VHR satellite data (with a resolution of sub-meters), [21] were able to detect approximately 1.8 billion native trees within West African Deserts, Sahel, and the sub humid regions. Deep learning-based geographic object recognition technologies, on the other hand, offer a new ability to monitor & map target items worldwide via a completely autonomous approach. As a result, the advancement of geographic object detection has been hindered by a lack of size of the training data [22]. As it turns out, there has been a lot of effort put into generating benchmark functions for multi-class geographic object recognition, such as the DOTA [23],and the FAIR1M [24],NWPU VHR-10 [25], the DIOR [20]. As a result, it remains an open question exactly multimodal RS data might be used to improve current object detection algorithms, especially when the single RS data fails to produce adequate performance. RS data with complementary views (such as spatial or spectral), such as multimodal RS data, have the potential here. Concerning labor & time costs, the current method utilized in benchmark creation is still too expensive for interpreting such sparse structures as WWTP.

12. **Multimodal Remote sensing Data Fusion:** The intrinsic heterogeneity of Multimodal RS data collected by many satellites, which is now becoming more widely available, creates a pressing need for improved multi-modal RS data fusion [26]. A wide range of applications, including picture pan sharpening (or resolution augmentation), multisensory

data transformation, and multimodal/cross model feature learning, have all shown a strong interest in this type of data fusion. For example, [27] present an MSDCNN for MSI picture pan sharpening (multiscale and multi-depth convolutional neural network). The next step was to create a new CNN framework for image pan sharpening based on detail injection, where MSI details were deliberately stated from beginning to end. When it comes to data fusion, Rasti and colleagues (2017) used a unique sparse and low-rank technique to recover spectroscopic and elevation information for accurate picture categorization of LiDAR data and HSI. Similar to this, [27] construct three-stream Convolutional neural networks for the integration of spectral, spatial, and elevation characteristics from LiDAR data and HSI to improve classification accuracy. The recent study also suggests the semi supervised cross modality learning approach of learnable manifold alignment (Lema) for reliable LULC classification by merging MSI data and HSI [28]. Multimodal/cross model feature learning takes its cues immediately from feature-level data fusion models, as opposed to the usual image-level data fusion methods. Prior studies [28] look into both manifold learning and deep learning methodologies for a robust and effective purpose of learning features. As a result of these and other community contributions, multimodal RS data fusion development has been promoted by offering open-access benchmarks and hosting regular data fusion competitions. Multimodal RS data has primarily been used for picture classification tasks; hence the enormous possibilities of multimodal RS data fusion for more complex tasks, such as geographic object recognition are largely untapped and have to be studied further [29].

13. **Deep Learning from VGI:** Because of the introduction of big data & crowd sourced technologies, VGI, a sort of user-generated content, has been possible to continue capturing massive geographic data from users [30]. When it comes to OSM's semantic properties (e.g. OSM tag and value), it has recently been investigated to recover customized geography objects and produce geo-referenced samples to develop successful feature extraction for geographic information [31]. There were numerous, free labels and extensive high-level semantics from which to draw for aerial photo analysis on VGI systems like OpenStreetMap [31]. Thus, deep learning using VGI is being researched by both the VGI and the RS community members. A deep learning model was utilized to extract vector data from OpenStreetMap (OSM) in order to perform supervised street recognition, and several loss functions were used to mitigate the impact of missing data and registration mistakes on OSM labels. As part of a humanitarian modelling project [31], Map Swipe volunteer input were used to train the model using an on-going learning approach called DeepVGI. As a result of this research, a new process was developed to better incorporate crowdsourcing contributions through the allocation of deep learning task. The proposed machine-assisted procedure has been shown to reduce volunteer efforts by at least 80% in Guatemala, Laos, and Malawi. [27] Used geotagged tweets & deep learning construction identification models to identify 13 OSM missing built-up zones. There have been few experiments on mapping more complicated entities in OSM using its semantic knowledge using VGI-based deep learning.

14. **Types of Wastewater Treatment Plants:** For the sake of our environment and our future, it must take Wastewater Treatment seriously. Wastewater treatment can help reduce the spread of a wide range of waterborne infections [32]. As a result, wastewater treatment plants serve a critical role in both maintaining a healthy environment and

saving a large number of lives every year. Wastewater treatment plants can be divided into three categories:

- Effluent treatment plants
- Sewage treatment plants,
- Combined Effluent treatment plants

  - **Effluent Treatment Plants:** Chemical, pharmaceutical giants and leather depend on these filters to rid water of impurities and contaminants such as pollutants, polymers, toxins, and other contaminants. Concentration, purification, incinerator for chemical processing, and effluent wastewater treatment are among the common methods employed by Effluent Treatment Plants [32].
  - **Sewage Treatment Plants:** Pollutant removal from industrial and domestic sewage systems Water and solid waste can be reused once physical, chemical, and biological techniques are used to eliminate physical, chemical, and biological impurities [32].
  - **Combined Effluent Treatment Plants:** Typically built in areas where there is a concentration of small-scale industries. A single firm's costs are reduced while pollution is prevented by such facilities [32].

A typical Wastewater treatment process flowchart is depicted in Figure 5

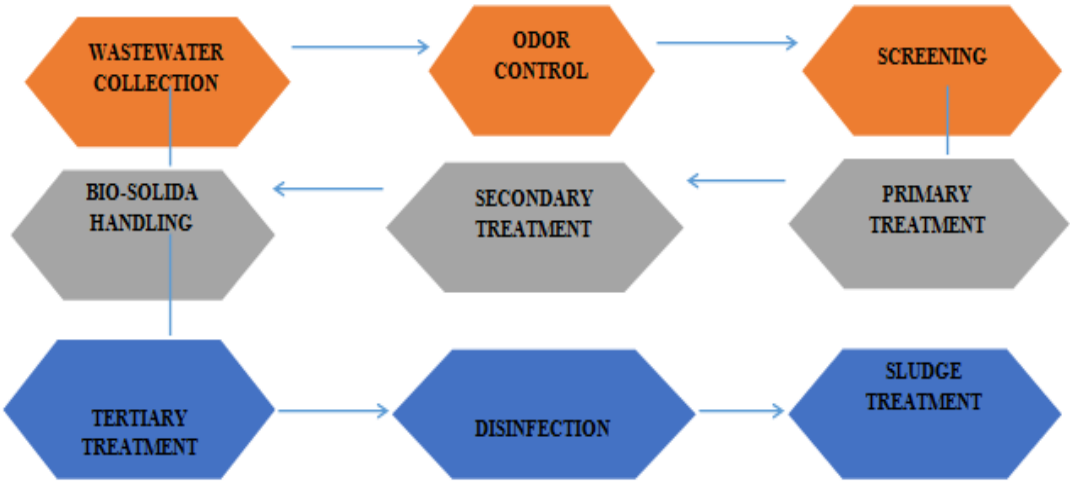


Figure 5: Flowchart for Wastewater Treatment Process [Source: Self-created]

15. **Demineralization Treatment Plants:** Demineralization, or Deionization, is a scientific process that uses ion exchange to eliminate mineral salts from water [3]. Various industries, including chemical and pharmaceutical, power plants, textiles, automobiles, and oil and gas, reuse water treated through demineralization. The anion and cation ion exchange resins are used in this process, which releases hydroxyl and hydrogen ions, respectively.

Demineralization can be accomplished in two ways:

- **Two-Bed:** A cation resin is used in one vessel, while an anion resin is used in the other. Hydrogen ions replace cations in the cation exchanger, which removes cations from the water. The anion exchange replaces anions with hydroxyl ions as the water proceeds.

- **Mixed-Bed:** The anions and cations resins are blended in a single pressure vessel in a process known as a "mixed-bed." There are times when increased levels of purification call for this method, even though it is more efficient. As a final step, the two-bed method is sometimes followed by the mixed-bed method for polishing.

16. **Concentration of Dissolved Oxygen:** Growing microorganisms depends on a high concentration of dissolved oxygen. When oxygen levels are low, biological functions can't proceed normally. Consequently, increasing the dissolved oxygen concentration above what is necessary does not enhance the effectiveness of biological processes and just increases the costs associated with aeration. Dissolved oxygen (C0) in the aerated medium exceeds the amount of oxygen in the air; therefore oxygen is transferred to water or wastewater to maintain hydrodynamic equilibrium. When no reactions require oxygen, the concentration of oxygen in an aerated medium will rise to its saturation level by a process known as diffusion (CS). The saturation concentration of oxygen in the supplied air is a function of the partial pressure of oxygen in the air and the value of which is dependent on, among other things, temperature. As the absolute pressure of the gas and the percentage of oxygen in the gas both raise, so does the rate at which oxygen is transferred between molecules. Aerating reactors with pure oxygen necessitates the use of deeper reactors than is necessary when aeration is accomplished using alternate technologies. The majority of wastewater treatment plants still use rule-based or fixed-setting proportional-integral (PI) regulators for dissolved oxygen management. They are insufficient because of the non-linearity and non-stationarity of the control problem and the diversity of the operational environments. The primary cause is the fluctuating demand for oxygen, which in turn affects the volume of air provided by the aerating system. This varies according to the desired dissolved oxygen levels, which must be established, based on the level of wastewater pollution [40].

V. METHODOLOGY & ANALYSIS

The methodology detailed in this section will be followed to achieve the aim and objectives of the study.

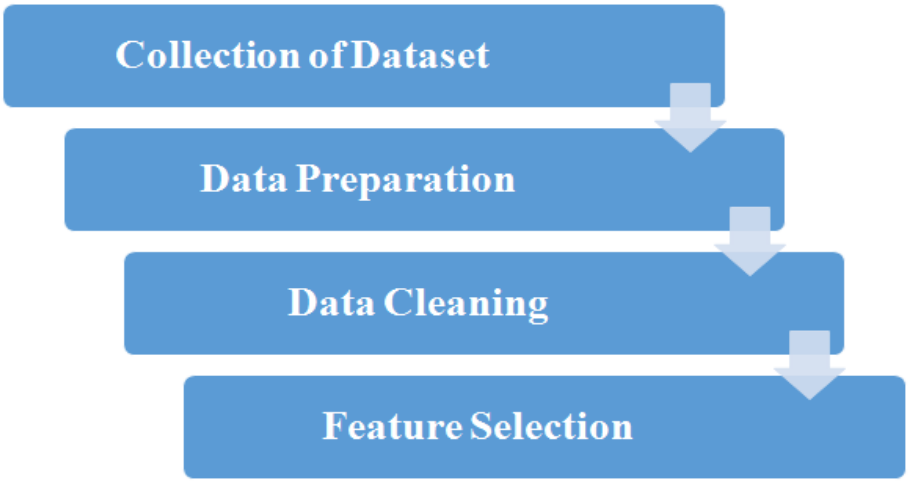


Figure 6: Methodology Block Flow Diagram

The proposed research is implemented using two approaches in the first approach two clustering algorithms K-means and Hierarchical Agglomerative are implemented The missing values in the dataset are replaced with the help KNN or k-nearest neighbour algorithm and the number of clusters for K-means are identified with the help of elbow and silhouette method.

In the latter approach an Autoencoder model is used which is an Artificial Neural network. It is used for compressing the input layer, to a bottleneck layer with reduced dimensions. This generated layer is used for applying traditional clustering algorithms such as K-means etc.

A comparison between the results obtained from both the Autoencoder based and traditional clustering algorithm is also discussed in this research.

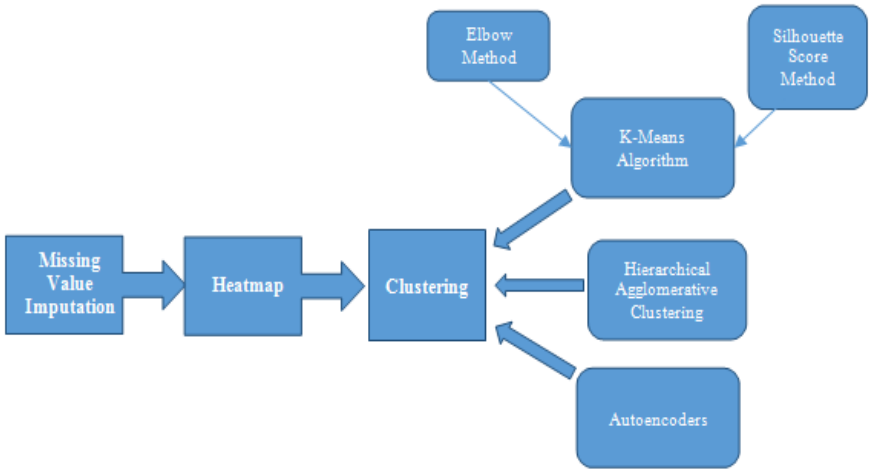


Figure 7: Data Analysis

1. **Missing Value Imputation:** The dataset used for this research contains a large number of missing values, 27 percent of observations have at least a single missing value. These missing values are also present in entries which shows 8 process upset days and if all these values are dropped from dataset it will impact the results and make them irrelevant since there are only 14 process upset days listed in the dataset and dropping 8 with leave us with about 42 percent of process upset days data.

   To deal with this issue KNN (K nearest neighbour) algorithm will be used, and then imputation of these missing values. KNN attempts to estimate the missing values in the dataset based on similarity of the statistical unit with missing entries and the closest observations. The similarity is determined using the Euclidean distance between the observations. It is an integer indicating how many nearest neighbours should be considered.

2. **Heatmap:** The dataset being used is highly multivariate and some of the features aren't necessary for further analysis, to drop these features there is a need to check the correlation between them and drop them accordingly if they have very low or no correlation. To achieve this it is required to use correlation heatmap to get a visual representation of the correlation between different pairs of features in the form of a matrix with varied levels of colors depicting the strength of relationship between the features. Cluster analysis can be applied to the following dataset to help anticipate process disturbances in WWTP once missing values are imputed and irrelevant features are removed.

3. **Clustering:** Clustering is the task of dividing a population into a number of groups so that data points in the same group are more similar to data points in other groups than data points in other groups. In a nutshell, the goal is to separate groups with similar characteristics and assign them to clusters.

   Clustering is performed based on the two approaches discussed below.

   - **K - Means Algorithm:** K-means is a popular partition clustering method, which uses unsupervised learning algorithm that divides an unlabelled dataset into different clusters. The goal of K-means is optimizing the objective function by minimizing the distance between each point from the centre of the cluster to which the point belongs. Firstly, the algorithm initializes a set of k cluster centres, followed by assigning each observation to the cluster with the closest centre, and then recalculating the centres. This process is repeated until the centres of the clusters no longer change.

     The main concern when implementing is finding k i.e. the optimum number of clusters, to compute them this research employs the use of following two methods –

     - **Elbow Method:** The elbow method is a graphical method used to determine the optimal number of clusters. The algorithm is run multiple times with increasing number of cluster options, followed by plotting a clustering score (the within-cluster SSE) as a function of cluster count. The score is calculated using mean squared distance between each instance and its closest centroid. Score will

decrease as the number of clusters increases, as the samples become closer to the centroids to which they have been assigned. The elbow method determines the value of k where the score begins to decrease most rapidly before plateauing of the curve.

Although the elbow method gives an easy visual representation of the optimum number of k, it may not give good results in certain cases where the curve decreases smoothly in such cases Silhouette Score Method is used to get the optimum number of k.

- **Silhouette Score Method:** The silhouette method computes k for K-means clustering algorithm by computing the silhouette coefficients of each point measuring how much a point is similar to its cluster compared to other clusters. Silhouette plots have an edge over elbow method clusters are evaluated on multiple criteria such as variance, skewness, and high-low differences. Hence determining the most optimal number of clusters in K-means.

4. **Hierarchical Agglomerative Clustering:** Agglomerative clustering is a type of hierarchical clustering that groups objects in clusters based on their similarity. Agglomerative clustering operates on a "bottom-up" basis i.e. at each step of the algorithm; consider each object as a singleton cluster. Two clusters the most similar are combined into a new larger cluster (nodes). This is iterated until all points belong to a single large cluster (root).

The result is a tree-based representation of the objects known as a dendrogram, which can be used to visualize the history of groupings and determine the optimal number of clusters by determining the longest vertical distance that does not intersect any of the other clusters and drawing a horizontal line at both the ends, the optimal number of clusters equals the number of vertical lines passing through that horizontal line.

In hierarchical clustering the linkage criteria determines how the distance between two clusters is calculated, there are many methods to do this i.e single, complete, average etc. The ward's method will be used.

5. **Autoencoders:** Autoencoders are a type of artificial neural network which are used for learning feature representation in unsupervised manner. It makes use of the same data for both input and output and introduces a bottleneck into the network, which forces the network to generate a compressed version of the input data, called the bottleneck layer or latent-space representation.

The code or bottleneck layer generated by this ANN can further be used for clustering as it has lower dimensions than the input layer and can be useful in cases where results are suffering because of higher dimensionality of data.

Here the example techniques under the role of the autoencoder is to try to capture the most essential features and patterns in the data and re-represent it in lower dimensions. The three components that make up an autoencoder are as follows:

- An encoder – that compresses the input and generates the code
- A code (i.e. the bottleneck layer), and
- A decoder - uses the code to reconstruct the input.

First import all the classes and functions

```
from keras.models import Model
from keras.layers import Dense, Input
from keras.preprocessing import sequence
```

Let's start by creating the validation dataset that will be used to test the efficacy of the model prior to create it.

```
## a subset from the test data
review_test = vectorizer.transform (X_test [:2000]).toarray ()
#review_test = sequence.pad_sequences (review_test, maxlen = max_len)
```

Next, we'll look at a sample model. Both the encoder and decoder will have three defined layers. And most significantly, the "bottleneck" layer, which is the compressed version of the data that will be applied later.

```
## define the encoder
inputs_dim = review_train.shape[1]
encoder = Input(shape = (inputs_dim, ))
e = Dense(1024, activation = "relu")(encoder)
e = Dense(512, activation = "relu")(e)
e = Dense(256, activation = "relu")(e)


## bottleneck layer
n_bottleneck = 10
## defining it with a name to extract it later
bottleneck_layer = "bottleneck_layer"
# can also be defined with an activation function, relu for instance
bottleneck = Dense(n_bottleneck, name = bottleneck_layer)(e)

## define the decoder (in reverse)
decoder = Dense(256, activation = "relu")(bottleneck)
decoder = Dense(512, activation = "relu")(decoder)
decoder = Dense(1024, activation = "relu")(decoder)


## output layer
output = Dense(inputs_dim)(decoder)
## model
model = Model(inputs = encoder, outputs = output)
model.summary()
```

The "bottleneck layer" is the layer of most value to us. Therefore, we will reshape our data by extracting the autoencoders trained weights from the layer it uses to encode new information.

```
## extracting the bottleneck layer we are interested in the most
## in case you haven't defined it as a layer on it own you can extract it by name
#bottleneck_encoded_layer = model.get_layer(name = bottleneck_layer).output
## the model to be used after training the autoencoder to refine the data
#encoder = Model(inputs = model.input, outputs = bottleneck_encoded_layer)
# in case you defined it as a layer as we did
encoder = Model(inputs = model.input, outputs = bottleneck)
```

6. **Collection of Dataset Information:** Continuous sensor measurements from an urban waste water treatment plant were used to compile this data collection. At each stage of the treatment process, state variables in the plant must be used to predict failures and classify the plant's operational status. This field has been referred to as an unstructured one.

7. **Attribute Information:** Everything is a number or a string of ones and zeros.

N. Attributes

1 Q - E (Input flow to Plant)

2 ZN - E (Input Zinc to Plant)

3 PH - E (Input pH to Plant)

4 DBO - E (Input Biological demand of oxygen to Plant)

5 DQO - E (Input chemical demand of oxygen to Plant)

6 SS - E (Input suspended solids to Plant)

7 SSV - E (Input volatile suspended solids to Plant)

8 SED - E (Input sediments to Plant)

9 COND - E (Input conductivity to Plant)

10 PH - P (Input pH to Primary settler)

11 DBO - P (Input Biological demand of oxygen to Primary settler)

12 SS - P (Input suspended solids to Primary settler)

13 SSV - P (Input volatile suspended solids to Primary settler)

14 SED - P (Input sediments to Primary settler)

15 COND - P (Input conductivity to Primary settler)

16 PH - D (Input pH to Secondary settler)

17 DBO - D (Input Biological demand of oxygen to Secondary settler)

18 DQO - D (Input chemical demand of oxygen to Secondary settler)

19 SS - D (Input suspended solids to Secondary settler)

20 SSV - D (Input volatile suspended solids to Secondary settler)

21 SED - D (Input sediments to Secondary settler)

22 COND - D (Input conductivity to Secondary settler)

23 PH - S (Output pH)

24 DBO - S (Output Biological demand of oxygen)

25 DQO - S (Output chemical demand of oxygen)

26 SS - S (Output suspended solids)

27 SSV - S (Output volatile suspended solids)

28 SED - S (Output sediments)

29 COND - S (Output conductivity)

30 RD-DBO - P (Performance input Biological demand of oxygen in Primary settler)

31 RD-SS - P (Performance input suspended solids to Primary settler)

32 RD-SED - P (Performance input sediments to Primary settler)

33 RD-DBO - S (Performance input Biological demand of oxygen to Secondary settler)

34 RD-DQO - S (Performance input chemical demand of oxygen to Secondary settler)

35 RD-DBO - G (Global performance input Biological demand of oxygen)

36 RD-DQO - G (Global performance input chemical demand of oxygen)

37 RD-SS - G (Global performance input suspended solids)

38 RD-SED - G (Global performance input sediments)

The proper operation of the treatment process must be monitored and checked to ensure compliance with effluent regulations. As a result of this information, operators and engineers may address problems before the process reaches its control limitations, allowing them to keep the process running at an optimal level. Because biological treatment, such as an ASP, has a Waste Activated Sludge (WAS) age of 8 days or more this indicates the ages of the bacteria that process the sewage, accurate detection or prevention is critical. Effluent quality is frequently a result of a build-up of disturbances that pushes a system beyond its optimum functioning envelope. Early diagnosis is critical in any large and complicated biological systems.

With the period the sensors data was gathered as the first variable, this dataset has 38 variables. Since this sensor data was gathered over the course of 507 days, a model for clustering can be used to analyse it. The 38 characteristics are divided into three groups:

There are three types of metrics:

- Input,
- Output, and
- Performance Indicators.

Here are the Input Indicators: -

- Flow to the Plant - This is the amount of water that is flowing to the plant, and it can fluctuate.
- pH - Acidity or basicity of water provided has an effect on the plant's overall treatment environment since biological processes operate best in a specified pH range.
- Zinc - It is important to note that zinc is a heavy element that can both impact the kinetics of ASP reactions and aquatic life in a watercourse.
- BOD - Biodegradable organic matter (BOD) is a measure of the amount of oxygen microorganisms need to break down organic matter.
- COD - Chemical oxidation of organic material (COD) is a measure of the quantity of oxygen needed to do so.
- Suspended solids - Water that has been filtered out yet has a little amount of suspended solids left in it.
- Volatile Suspended Solids - There are two types of volatile suspended solids: those that evaporate when heated, and those that don't.
- Sediments - The turbulence of flowing water produces solid particles known as sediments.
- Conductivity - Using conductivity, water purification systems can be monitored by measuring the water's ability to conduct electricity. Percentage of total dissolved solids is used.

In utility management Performance Indicators (PI) that focus on resources and other criteria are more valuable to operators. Table 1 lists WWTP operative indicators.

Table 1: Performance Indicators in WWTP

Effluent Quality PIs	Percentage removal of COD, BOD, P, TOC, N and the OCP
Effluent Quality PIs	Discharges as kg/year of BOD, COD, P, N and OCP
Sludge and Sludge Quality PIs	Sludge Production kg DS/year
Sludge and Sludge Quality PIs	Sludge Quality as mg contaminant/connected physical person/ year
Energy Performance Indicators	Purchased Energy as kWh/year
Energy Performance Indicators	Sold Energy as kWh/year
Energy Performance Indicators	Net Energy use as kWh/year
Energy Performance Indicators	Total Electrical Energy use as kWh/year

Energy Performance Indicators	Electrical Energy use for aeration as kWh/year
Energy Performance Indicators	Electrical Energy use for aeration as kWh/kg oxygen demand
Energy Performance Indicators	Other Electrical Energy use in biological step as kWh/year
Energy Performance Indicators	Biogas Production as kWh/year
Energy Performance Indicators	Biogas Production as kWh/kg COD received at WWTP Biogas Production as kWh/kg COD fed to digesters

BOD, COD, and TSS are the three most common wastewater treatment plant quality indicators. Standard procedure is to analyse nitrogen and phosphorus in degrading environments. Wastewater hazards must be analysed.

8. **Data Preparation**: Consider making very few data checks. Variations were left to avoid eradicating anything at random for this experiment. If necessary, this could be modified at a later time.

9. **Data Cleaning:** Missing data are critical in formulating a strategy. As expected, many of the features have blank values. For this plant, input flow rate is the most important criterion for success. Since this was the primary process stream, for the time being, decided to stop keeping a close eye on any deviations from the expected flow. If the current attribute is strong (red) or weak (blue) in correlation, scan the x-axis to determine what other attributes have a strong or weak association with it.
   dataset. dropna (axis=0, how='any', subset= [1], in place=True)

10. **Feature Selection:** The correlations section will no longer include ZN-E, PH, or COND because they operated as both inputs and outputs in the correlation process.

    The "performance attributes," which are nothing more than a mathematical link between the attributes at different phases of development ($A 0, $A 1, $A 2), were also left out. The equations that can be used to derive the correlation points are discussed in detail.

$$\mathrm{P}\ \% = \left(1 - \frac{A_1}{A_0}\right) \times 100\%$$

$$\text{RD-DBO-G} = \left(1 - \frac{\text{DBO-S}}{\text{DBO-E}}\right) \times 100\%$$

As an example, let's take a look at the data set's mean values:

$$\text{RD-DBO-G} = \left(1 - \frac{20}{180}\right) \times 100\% = 89\%$$

11. **Feature Scaling:** Because input flow and attribute values have a broad range, scaling the input is important to prevent algorithms from focusing on it.

    Feature scaling can reduce strained clusters and accelerate the rate in clustering algorithms. Theoretically, clustering will be more uniformly dispersed, but not ensured.

12. **Short-listing and Fine-Tuning Models:** In this study, a pair of widely-used clustering methods were employed, K-means and Hierarchical. The sample size wasn't looked at because to ensure time efficiency when fine-tuning the hyper-parameters, although doing so could be useful if there was more data. Without understanding how or why the parameters produce the observed values, fine-tuning the dataset is seen as challenging. In physical theory, one of the most fundamental rules is that parameters shouldn't be tweaked too precisely. Naturalness describes this perfectly.

13. **Data Analysis:** Data analysis is crucial for recognizing and diagnosing process problems before they can become significant problems, but operators and engineers who rely on a limited amount of data and a few important indicators may struggle. Machine learning can effectively examine enormous datasets from diverse sources and formats. Input flow to plant, Elbow Plot, Silhouette with the Dedrogram, and along Hierarchical Agglomerative Clustering have been used to evaluate the data. Scatter plots employ X, Y coordinates. Missing value imputations, Heat maps, Clusters, K-means Algorithm, and Hierarchical Agglomerative clustering improve input and output.

14. **Correlation Heatmap with Selected Features:** Because of this, it is important to examine the correlation of various features in the dataset and eliminate those with no causality. By using a correlation heat map, it can see how strong a link there is between distinct feature pairings by looking at a color-coded matrix. Null values have been interpolated and unused variables have been removed to provide a consistent color-coded matrix, the resulting dataset can be used to detect WWTP process interruptions.

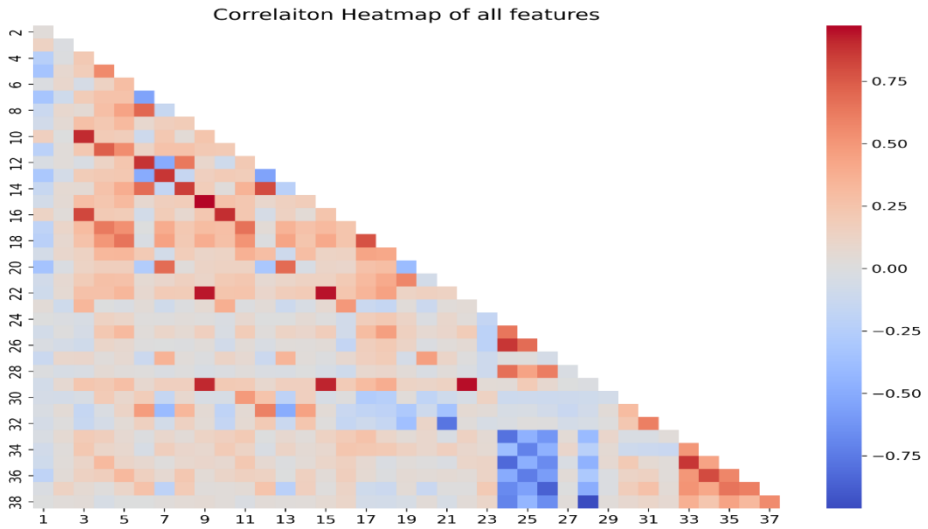


Figure 8: Correlation Heat Map for all features.

Figure 8 indicates the correlation heat map for all features in waste water treatment plan with wide range of values.

Features

- From left to right, the x-axis can be used to determine the strength or weakness of a correlation between several features.
- A decision will be made based on the results of these correlations on the removal of extraneous elements:
- (2) ZN-E likewise displays low correlations. Correlations between ZN and E are rare. Because Zn was the only input. Others to be dropped.
- There are many interconnections between the pH values (3) PH-E, (10) PH-Ps, and (23) PH-S. This is likely due to there being only a small pH gap between the influent and effluent.
- (9) COND-E and with (15) COND-P, along with (22) COND-S and the (29) COND-S, have a high correlation. This is likely due to conductivity of the plant remaining constant throughout the process.

The following number were maintained and not removed at this stage : 1, 4, 5, 6, 7, 8, 11, 12, 13, 14, 17, 18, 19, 20, 21, 24, 25, 26, 27, and 28.

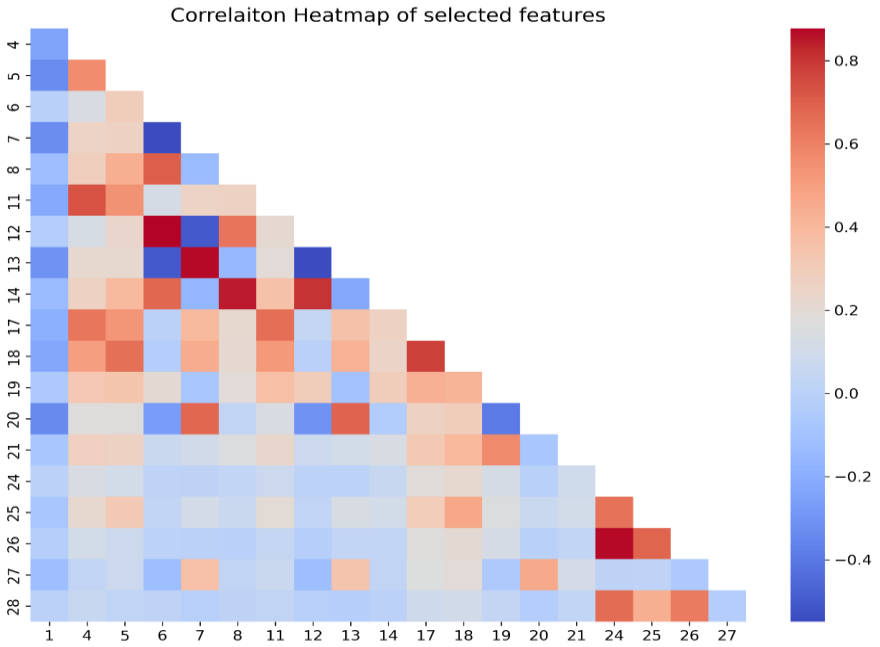


Figure 9: Correlation Heat Map of Selected Features.

The attribute numbers indicated here are the ones that need to be used in order to view the clear correlation heat map analysis.

15. **Daily Input Flow to Plant:** The flow patterns in a new plant layout are normally decided upon before the design of the overall system of facilities (equipment, materials,

designing, etc.) is initiated. This is because determining the flow patterns in a new plant layout can be time-consuming. A basic design allows for the simplest monitoring and administration of any flow pattern. Hydraulic design can have a major impact of plant performance, and is a simple indicator for prediction of plant performance. Figure 12 shows the influent flow of the plant and how it varies across the data.

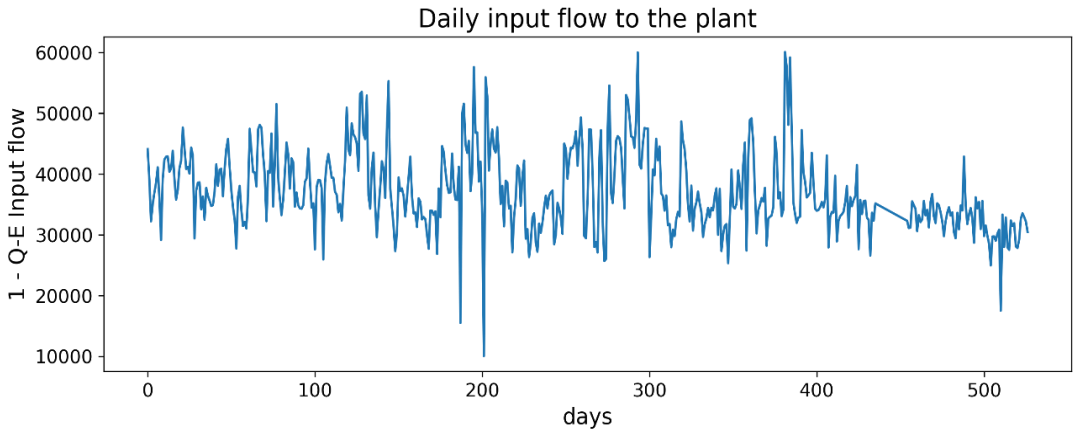


Figure 10: Influent flow to the Plant

16. **K Plot Elbow Method:** The dataset is run via k-means clustering (K plot) for a range of k values (i.e., 1-10), and an overall average for all clusters is computed for each value of k in the elbow technique. Diffusion score is calculated by default and it is made up of square distances between each point and its designated centre. Under its null hypothesis of independence, the order statistics are shown against the empirical random plots. Ranks are created from the data. An approach to unsupervised learning that is well-known for its use in the partitioning of unlabelled datasets is called partition clustering using K-means. K-means tries to minimize the distance that separates each item in a cluster from the centre of the cluster so that it can fulfil the goal function to its fullest potential. After that, the centre values are recalculated after every observation is put into the group that has the centre that is the most similar to it. This is how each and every one of the k cluster centres gets initialized. Repeating this procedure is necessary after the cores of the clusters have become stable.

Utilizing the elbow approach is a viable option for determining the optimum amount of clusters to create. After numerous iterations of the procedure, during which increasing numbers of clusters were used as alternatives, the clustering rating, also known as the within-cluster SSE, was compared to the total number of clusters. When figuring out the result, squared distances between occurrences and the centroids of their nearest neighbours are taken into account. As the number of clusters develops, the samples that are distributed to each cluster will move closer to their respective centroids. As a result, the score will decrease. The value of k can be determined by using a technique known as the "elbow approach," which involves searching for the point at which the score begins to drop significantly before levelling off.

It's likely that a straightforward visual representation of the optimal k quantity to use in order to attain the best possible k won't produce satisfactory results in other contexts when curves flatten out gradually.

17. **Silhouette vs. Elbow Method:** The weighted Silhouette values of each spot in the K-means (K plot) clustering algorithm are used to compute k, which measures how related a point is to its clusters relative to other clusters in the dataset. Silhouette method performs elbow method clusters in terms of variance, skewness, and high-to-low discrepancies, to name just a few. In K-means, this is done by determining the number of clusters that are most effective.

18. **Hierarchical Clustering Value Counts**: Using a hierarchical clustering approach known as agglomerative clustering, objects are organized into groups based on how closely they resemble one another. Agglomerative clustering is performed "bottom-up" because objects are treated as singletons at each stage of the algorithm. By joining two smaller clusters that are most similar, a new, larger one is created (nodes). It takes a number of iterations before a massive collection of points is produced (root).The best clustering procedure can be calculated by determining the longest gradient that does not overlap any other groups, and then drawing a black line at both ends, using a Dendrogram (a Tree-based Recognition of the items). There should be no more than the number of dotted bars that cross this horizontal line.

    The distance between two hierarchically clustered clusters can be calculated in several ways, including with a single method, an average method or a comprehensive method. This research will make use of Ward's method. In contrast, Ward's method can be utilized to perform cluster analysis. Instead of focusing on distance metrics or measures of association, it tends to treat cluster analysis as an analysis of variance problem. An agglomerative clustering algorithm is used in this procedure. Beginning from the leaf base, it will eventually make its way up to the tree's main stem. When it finds clumps of leaves, it follows the growth of those leaves into limbs, and those limbs continue to grow into a trunk. Starting with n clusters of size 1, Ward's approach continues until all observations are grouped together into a single cluster. For quantitative variables, this technique is perfectly suitable.

VI. RESULTS AND DISCUSSION

1. **K-means Clustering:** The ideal number of clusters is determined by a kink or elbow that begins suddenly at the beginning and gradually decreases in inertia. Figure 13 shows by applying the model with a variety of values of 'K' (number of clusters), the elbow method assists in selecting the optimal value of K. Here the optimal value of K was identified as 3, suggesting that the optimum number of clusters is 3.

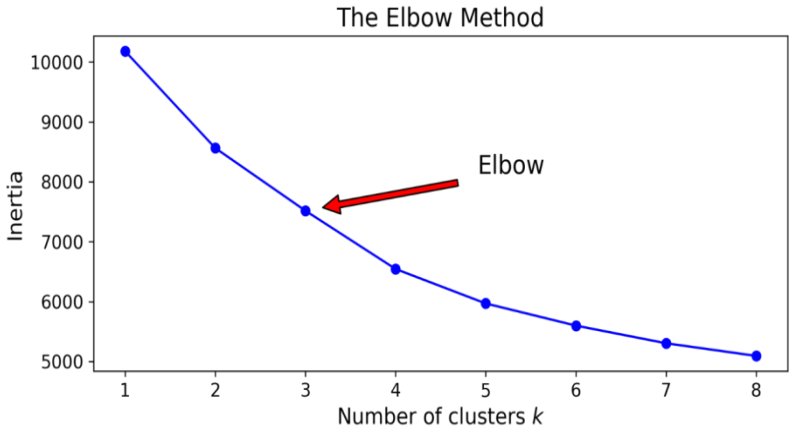


Figure 11: Elbow Method

However, the curve is fairly smooth, making this distinction difficult to decipher. Hence the alternative verification would be to plot the silhouette scores on a graph and look for the peak that would indicate the optimum number of clusters if the data were different.

The Silhouettes Score Method here suggests the peak is 4, in comparison to the Elbow method which suggests 3 clusters as optimum. Hence further analysis was done using silhouettes analysis plot of various clusters. Using this silhouettes analysis plot for various K values the optimum number of clusters is determined.

Figure12: Silhouette Score Method

The silhouettes analysis plots were used to indicates the clusters' height and width of distributions (the wider the cluster, the better). The silhouette score is indicated by the dashed vertical line. Consequently, during the analysis using the silhouettes analysis plots the aim to find the optimum number of clusters the below rules were followed:

- Make sure all clusters are extended further left than you think they are in order for it to work.
- A vertical line that extends farthest right is known as a horizontal line (score)
- For a uniform distribution, the height of the each cluster is the same.

Figure 15 shows the silhouettes analysis plots for K = 3 to 6, from these plot we can clearly see that k=3 show the highest score, which confirms that the third cluster is the best number to display the inertia's middle region.

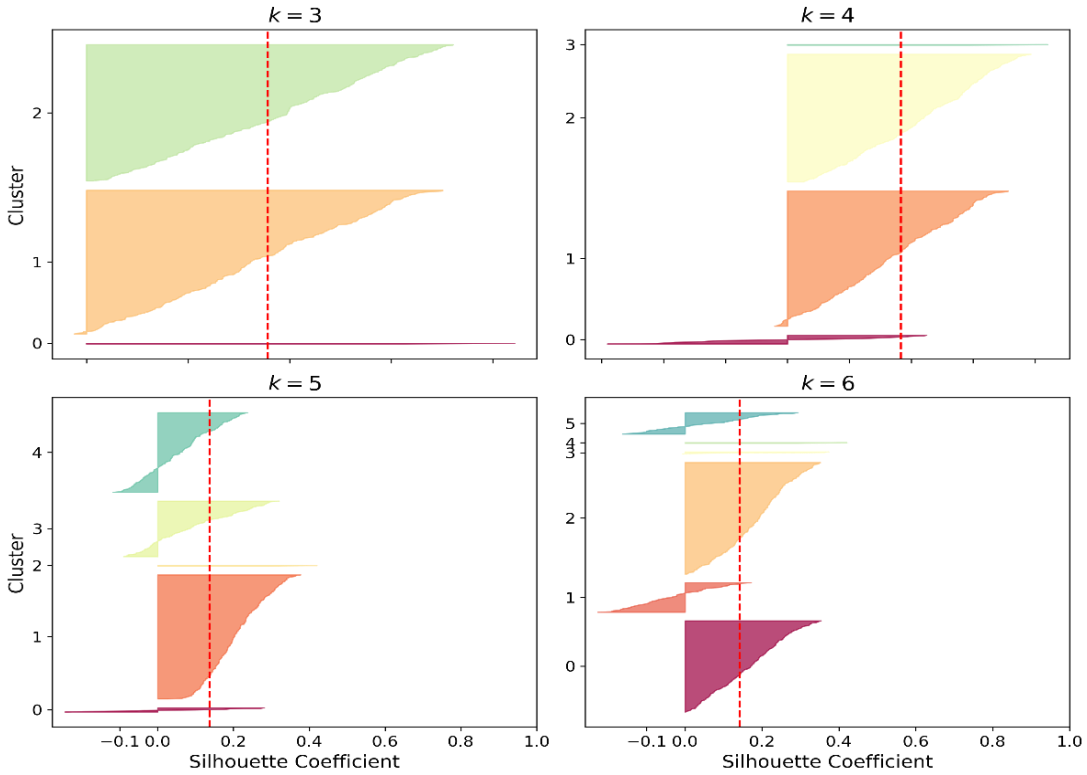


Figure 13: Silhouette Analysis Plot

2. **Hierarchical Agglomerative Clustering:** For each observation, the Dendrogram allocates it to a cluster at the bottom of the graph. Vertical lines in each pair reflect the distance between the nearest points in each cluster. A horizontal line connects these points. A single final cluster containing all of the observations is reached by going up one more level in this manner.

Locate the maximum altitude of a vertical line calculated between any Horizontal lines drawn across entirely from one side to the other in the Dendrogram to get the optimal number of clusters.

Between the 40 and 45 Euclidean distance Horizontal lines, the highest vertical height is found as depicted in Figure 16 the following figure with the Dendrogram. As a result, if a horizontal line was drawn right above 40, it would divide three (3) vertical lines, which is the ideal amount of clusters.

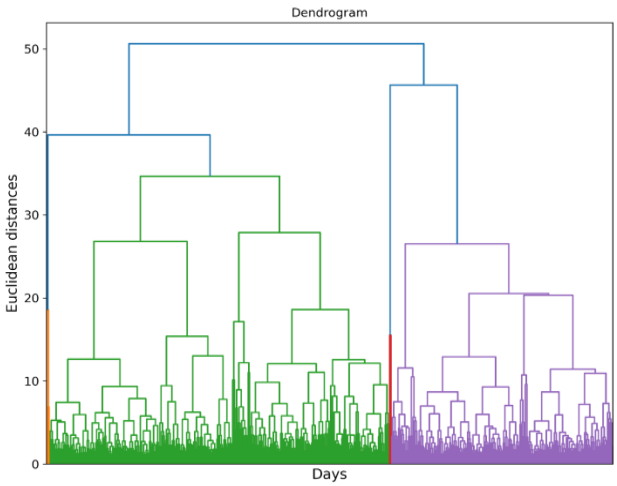


Figure 14: Dendrogram

3. **Predictions and Train the Model:** The optimum number of clustering for modelling has been verified. Clustering and K-Means can be used to train the model with the optimal number of clusters of three (3).

4. **Prediction Remapping:** The number of categories that are extracted from a dataset is what is termed to above k in the procedure, and predictive remapping analysis is dependent on this value. Using the K-means technique, the data is partitioned into k different categories. It is recommended to select representatives for each cluster at random from the dataset. For this reason, modification of the Agglomerative predictions and K-Means is completed so that indexes 1, 2, and 3 represent the clusters in increasing order of observation size, starting with index 1 for the smallest cluster (index 3).

```
# Training the K-Means model on the dataset

# Review the elbow plot, Silhouette plots, and Dendrogram and set the number of clusters.
n_clusters = 3

kmeans = KMeans(n_clusters= n_clusters, init= 'k-means++', random_state=0).fit(X)
y_kmeans = kmeans.fit_predict(X)

# Training the Hierarchical Clustering model on the dataset
from sklearn.cluster import AgglomerativeClustering
hc = AgglomerativeClustering(n_clusters = n_clusters, affinity = 'euclidean', linkage = 'ward')
y_hc = hc.fit_predict(X)
```

5. **Scatter Plot:** Due to the large number of variables in our dataset and 2D plot such as a scatter plot are unable to accurately represent them, this scatter plot is solely for demonstration purposes. Little can be deduced from the cluster formation in general. Figure 17 scatter plot shows the data points under the index 1 to 3 plots.

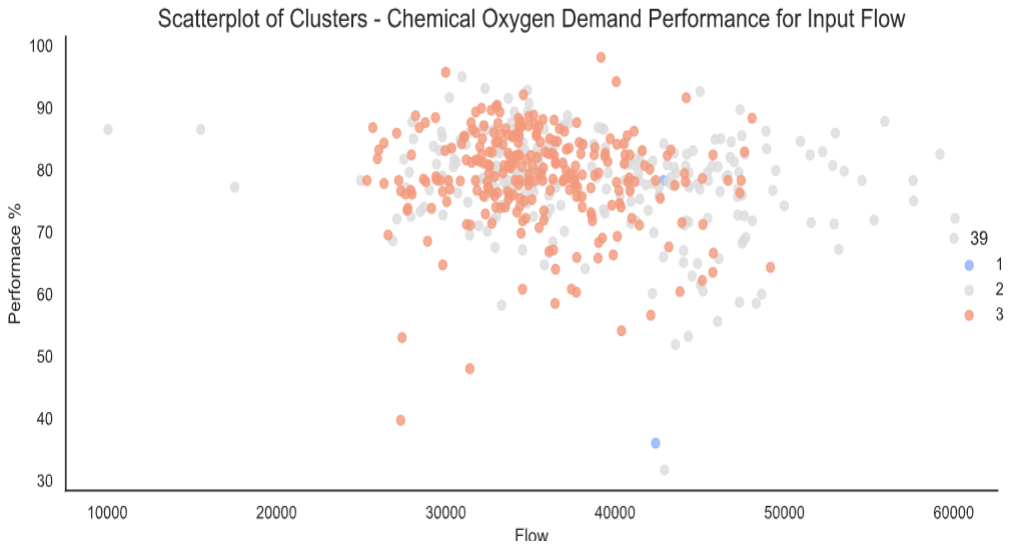


Figure 15: Scatter Plot

6. Box Plots

- **Input flow:** The Q-E, or Input flow rate, is typically around 400000 for most groups. Cluster 3 has a distribution that is more concentrated than Cluster 2, which has a less concentrated distribution. The increase in Input Flow and Zinc that results from the presence of Agglomerative Clusters is depicted in Figure 18.

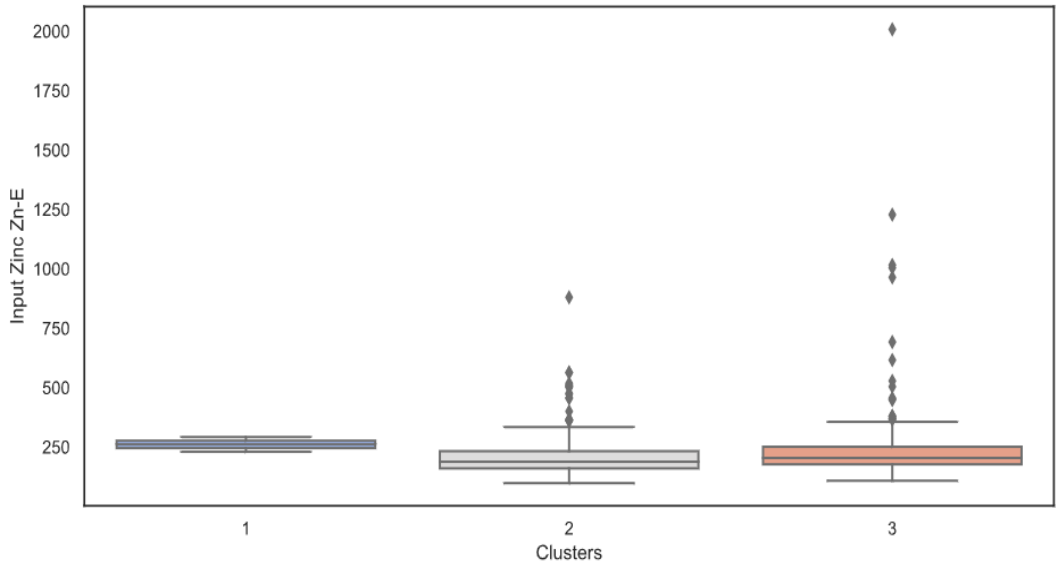


Figure 16: K- means Clusters - Input Flow and Zinc to the Plant

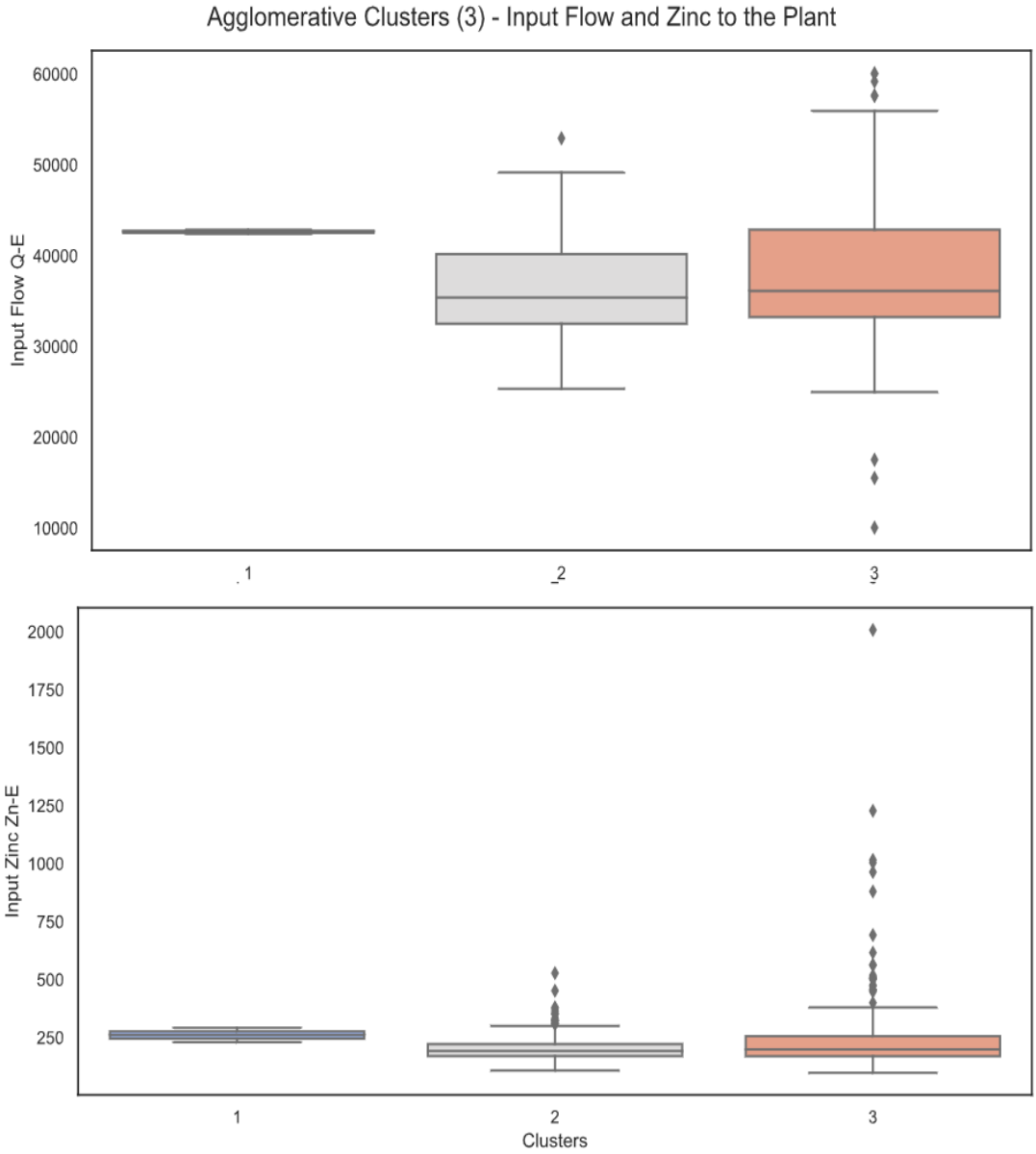


Figure 17: Agglomerative Clusters – Input Flow and Zinc to the Plant

- **Suspended Solids (SS):** The major sedimentation goal is to remove between 50 and 70 percent of the suspended particles. Effluent solids should not exceed 30 mg/L, according to the EPA 40 CFR 133.102 regulations.

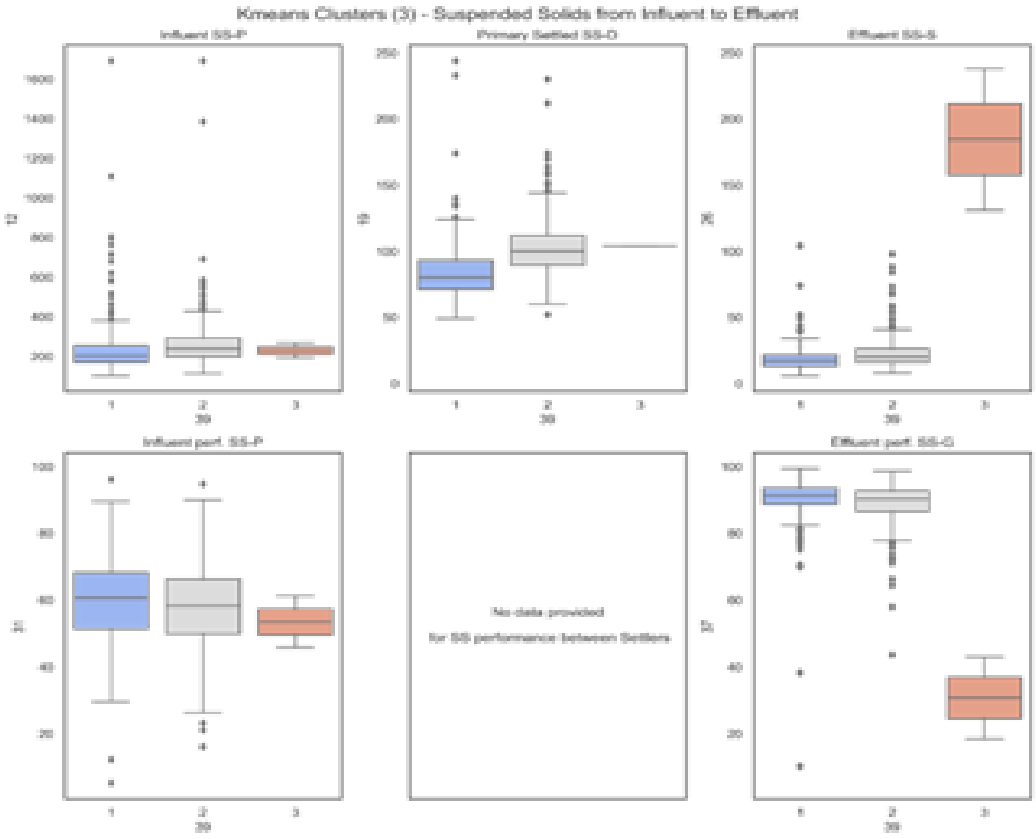


Figure 18: K- means Cluster - Suspended Solids

Accurate detection of suspended solids and water contaminants allows for the use of effective treatments like filtered water and sedimentation. It is well known that removing suspended solids from water significantly improves water quality and decreases the likelihood of water damage in industrial operating systems. The European Union's requirement for sewage discharge (SS) is 39 mg/l and this amount is often shortened to "25/39" to indicate that the effluent meets both the BOD and SS standards (BOD of 25 mg/l and SS of 39 mg/l) as shown in figure 20. Therefore, most STP packages will be built to meet a standard of '25/39' for specific demographic ranges.

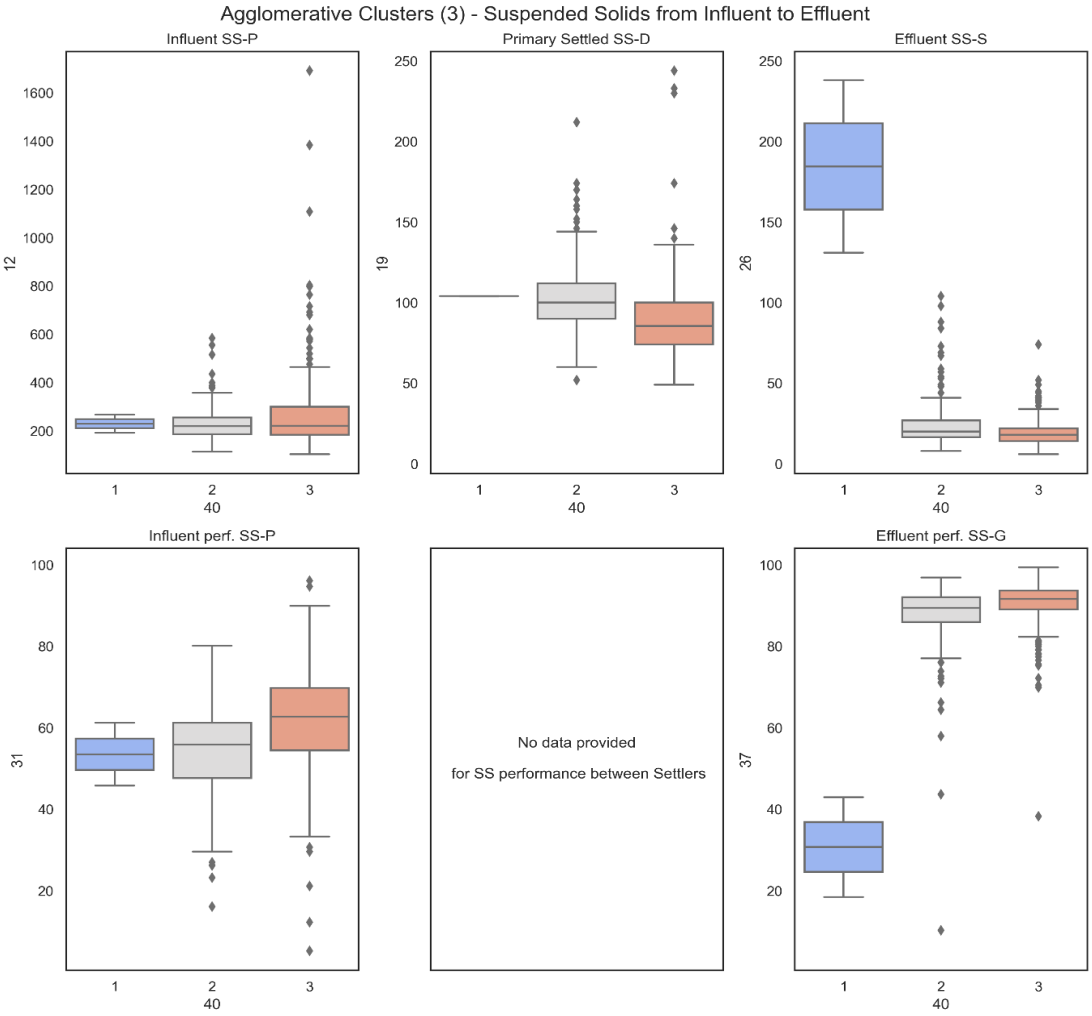


Figure 19: Agglomerative Clusters - Suspended Solids

The primary settler's SS-P input is around 200 mg/L. There are a few outliers in Cluster 3, with one exceeding 2000 mg/L.

Due to a few of big outliers at SS-P, the SS-D & SS-D y-axis has been adjusted to indicate the refined range.

The concentration of Influent SS-P in Cluster 1 SS-S Effluent remained the same at around 175 mg/L. This could be a sign of a damaged plant or flooded settlement tanks. There is a possibility that an inaccuracy in the sensor is to blame for the insufficient data that it may receive. According to Figure 21, the SS-S content in the effluent was maintained at a level that was lower than 25 mg/L in clusters 2 and 3.

- **Biological Oxygen Demand (BOD / DBO):** Primary settler has the goal of removing 25-40% of the BOD throughout treatment. This is normally associated with SS that can be settled out. A BOD concentration of less than 30 mg/L is required by EPA 40 CFR 133.102 for wastewater disposal.

The BOD (DBO-E) intake for all clusters is now around 200 mg/L. DBO-S levels in the effluent of cluster 1 remain intact, whereas Clusters 2 and 3 discharge at 25 mg/L and an efficiency or decrease of 90 percent.

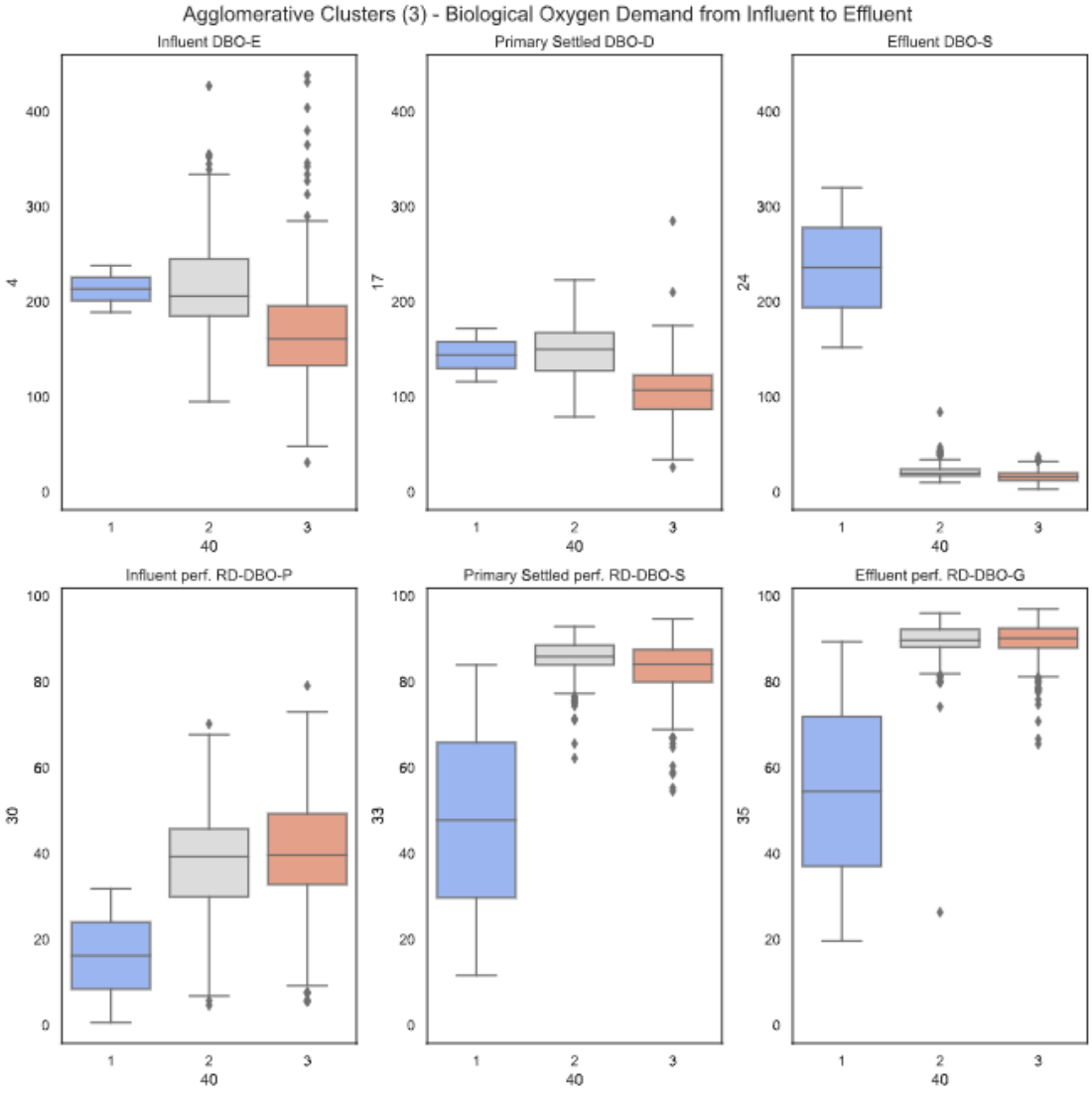


Figure 20: Agglomerative clusters - Biological Oxygen Demand

- **Chemical Oxygen Demand (COD / DQE):** The clusters have an influence COD (DQO-E) of 400 mg/L. In the absence of cluster 1, effluent DQO-S was reduced to 100 mg/L.

There were no DQO-E results in the set of data, but can determine an average of 38 percent using an equation.

$$\mathrm{RD-DQO-P\%} = \left(1 - \left|\frac{\mathrm{DQO-S}}{\mathrm{DQO-E}}\right)\right) \times 100\% = \left(1 - \left|\frac{254}{407}\right)\right) \times 100\% = 38\%$$

Clusters 1 and 2 saw a DQO-G drop of 80%.

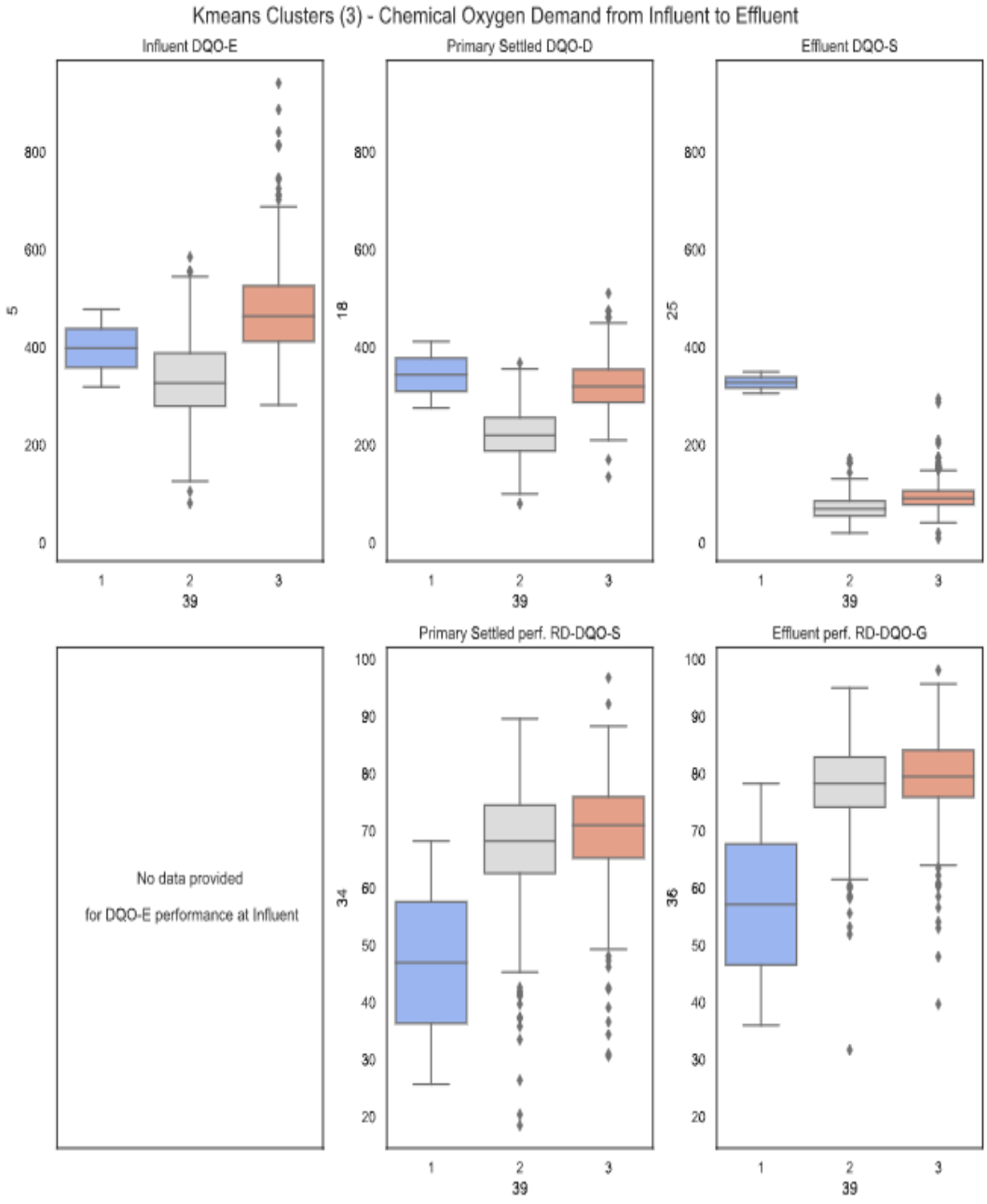


Figure 21: K-means Cluster - Chemical Oxygen Demand

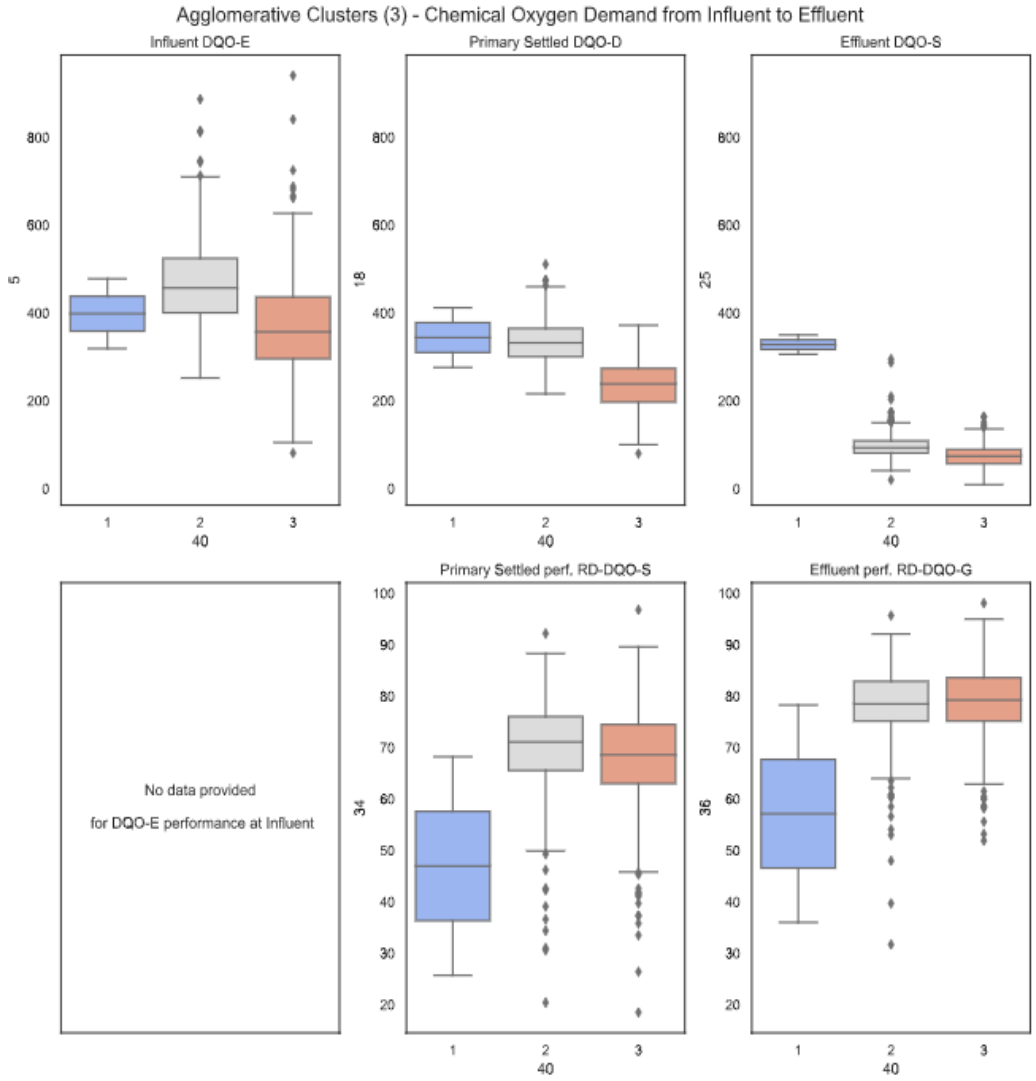


Figure 22 : Agglomerative cluster - Chemical oxygen demand

- **Sediment (SED):** As much as 90% of the sediment is removed during primary settling. Clusters 2 and 3's effluent SED was virtually eliminated by secondary settlers, who performed at 99 percent efficiency.

  Refined SED-P range is shown on the SED-D y axis because of the outliers at SED-P.

  The Average SED performance at the Secondary settler can be determined from the following equation:

$$\mathrm{RD-SED-S\%} = \left(1 - \left|\frac{\mathrm{SED-S}}{\mathrm{SED-D}}\right) \times 100\% = \left(1 - \left|\frac{0.03}{0.41}\right) \times 100\% = 93\%$$

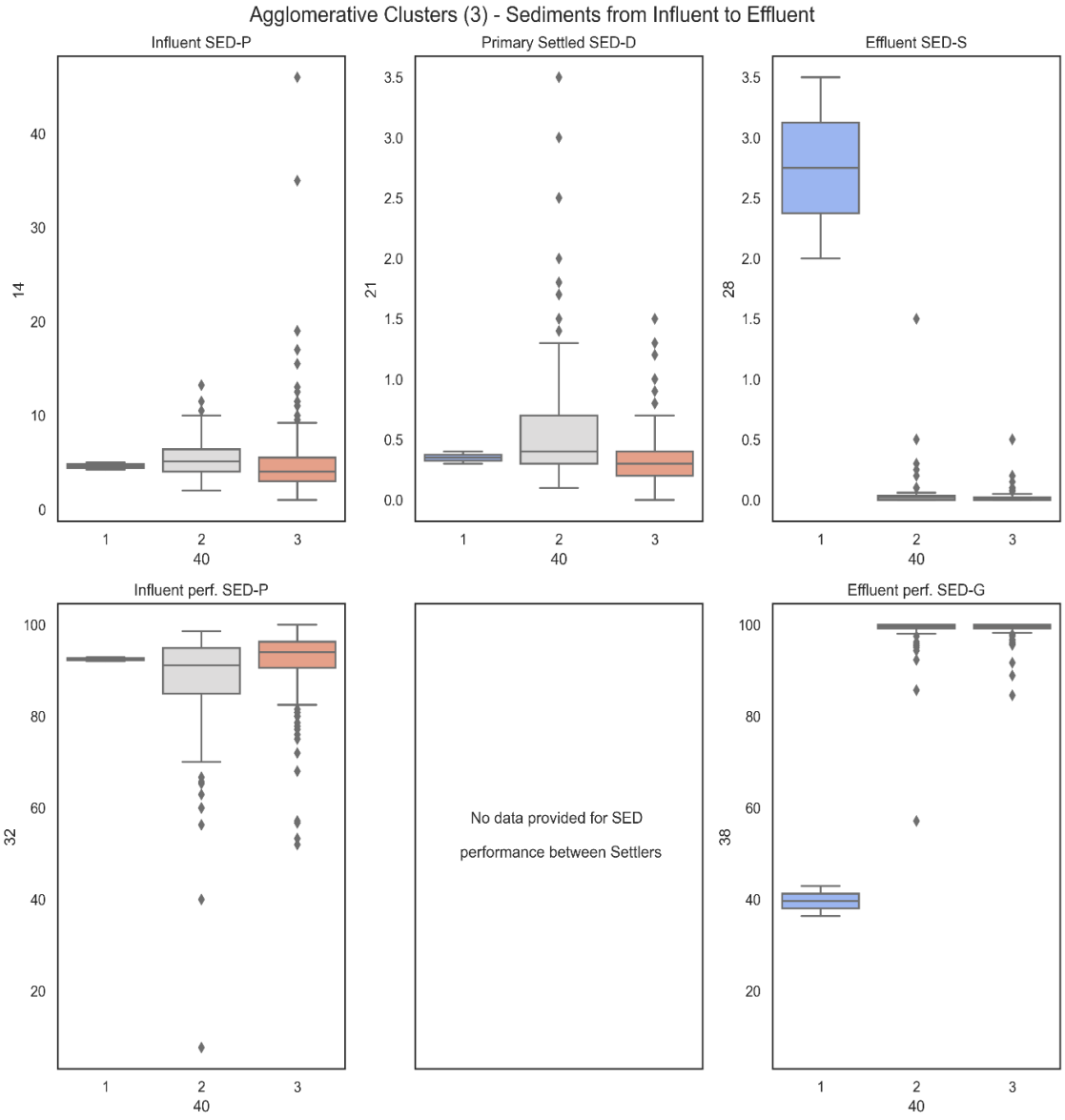


Figure 23: Agglomerative clusters – Sediments

It is possible that with more data, this plant could be expanded to include supervised learning of precise faults, which in turn could modify an overall online design that may predict possible flaws in advance under clustering methods, giving plant operators more time to prepare maintenance or repair and maintaining the plant flowing at its plated capacity.

7. **Findings and Discussion:** The wastewater from both residential and commercial establishments is sent to a wastewater treatment plant, where it is filtered, cleansed, and recycled for further use. At the sewage treatment plant, all sewage water is cleaned and treated.

The findings with respect to the aims and objectives will be discussed in this section.

- In summary these are classification of WWTP (ASP) operational state to predict process upsets using clustering, namely K Means and Hierarchical Agglomerative clustering algorithms in python. In addition implementation of an Autoencoder model and clustering on the bottleneck layers generated by this model.
- Three of the objectives have been demonstrated by the work by comparing K Means and Hierarchical Agglomerative clustering algorithms and running the Autoencoder to reduced dimensions before running on the clustering methods.

One of the objectives for this research is to evaluate K Means and Hierarchical Agglomerative clustering algorithms, including examining the strengths and drawbacks associated with each of these approaches in their own right.

In the context of this research examine whether linear or non-linear data transformations in clustering, and to answer this question within the context of the final research. It is possible that clustering data with a complicated structure isn't the best use for these methods. Hence in this research, an auto-encode ring was used prior to clustering to reduce dimensionality. The research found that using auto-encoding results in clusters that are both stable and effective. This reduces false negatives (missed defects) while increasing random errors just slightly (false positives). The results demonstrate that the suggested K-means clustering outperforms the other approaches examined in the context of Hierarchical clustering.

This is due to K-means clustering's impressive capability of modelling deep interdependencies in time series. The other methods are inadequate for modelling the interdependencies between multiple variables in time series data. This ability is especially useful in identifying cumulative flaws that display a distinctive pattern when compared to the norms of normal functioning. Furthermore, clustering is commonly observed in real-world time series data and is relatively resilient to distortion and other outliers.

There is constant pressure to enhance WWTPs' purifying capabilities while reducing their energy footprint. As a result, a greater number of measurement instruments are being installed, and the processes actually are becoming more and more automated. The detection of sensor defects is crucial to ensuring proper plant operation, and their use is on the rise not just for environmental monitoring but also as a key management tool for plants. In addition, it is challenging for a controller to detect sensor failure manually, particularly in big plants with many sensors or in small plants without any human supervision. However, there is a definite need to create procedures that can consistently detect sensor defects and allow adequate time to the plant operators, so that damage to the environment is kept to a minimum when errors do occur. The work provided in this research is a precursor to a fully automated fault detection system that can resolve issues brought on by the automatic management of WWTPs. In this research, researchers talk about how to use cluster analysis to solve research issues related to wastewater treatment facilities.

Following is an explanation of the aims that should be achieved through the utilization of K Means and Hierarchical Agglomerative clustering. It is a crucial application in cluster analysis, among a great many others in a wide variety of domains. Unsupervised machine learning algorithms, such as K-means and Hierarchical Agglomerative clustering algorithms, can be easily employed with the assistance of python support libraries to summarize and show the clusters. In this research, K-means and Agglomerative clustering algorithms were used to analyze a dataset consisting of 38 variables in order to uncover patterns contained within that dataset.

Examination of Autoencoder-based clustering and how it stacks up through more conventional methods of clustering is further discussed. Clustering is one of the simplest and most common unsupervised machine learning algorithms, and its application in this implementation of the Autoencoder model, namely on the bottleneck layer produced by the model while operating under the K-means metrics, is indicative of its growing acceptance. In this method, clustering is done using the features that each item possesses. K-means was developed to help classify data and unearth hidden patterns. In order to cluster the data, k-means searches for a predetermined number of clusters.

The elbow approach can be used to determine the number of centroids, or the number can be specified explicitly. Elbow technique operates on the premise that as 'k' increases, the cluster's size reduces, making it easier to find a subset of the cluster's elements. For the most part, a distortion decreases from the elbow out. Centre points, or centroids, are the averages of the values inside a given system.

The method decides where data points belong by computing the distance from the points to each of the centroids. Using a series of optimizations, the method identifies the optimum centre and recalculates the gap. If all the data points belong to the same group, the process ends; otherwise, it continues.

The algorithm is efficient and simple to implement. It's effective, in a sense. This method ensures data convergence at large scales. K-means works best when there is a clear separation between the data points. One of the algorithm's drawbacks is that it can only ever be as good as its beginning settings. A few extreme values can easily pull the average down.

The K-means algorithm is widely adopted for use in numerous fields, such as segmenting the market, document grouping, image segmentation, and image compression, etc. If there is significant variance in the behaviours of multiple subgroups, researchers can use a cluster-then-predict approach in which unique models are developed for each. A way to grouping based on similarities is the hierarchical cluster. Divisive hierarchical clustering and agglomerative clustering are the two main forms.

In the top-down, or divisive, clustering method, first all of the data is placed into a single cluster before splitting it into two sub clusters with the lowest similarity. Finally to do recursive operations on each cluster until there is a single cluster for each observation. Divisive algorithms are conceptually more complex than agglomerative algorithms, yet there is evidence that they yield more accurate hierarchies.

Clustering from the bottom up, or agglomerative, involves placing each observation into its own distinct cluster. The next step is to calculate the degree of similarity (e.g., distance) between each pair of clusters, and then merge them together if they are sufficiently close. A distance matrix is created by first computing the distance between each data point. Try to locate the closest pair of objects using the distance matrix. Gather them into a cluster, and then calculate the separation between that group and the rest of the data. Alter the matrix of relative positions. Move until all the data points can be placed into a single cluster.

For limited datasets, hierarchical clustering is the best alternative. The primary benefit of hierarchical clustering is that it does not necessitate any prior knowledge or specification of fields, such as determining k value, from the user. Due to computational limitations, hierarchical clustering slows down and produces unfavourable results as the dataset grows sufficiently in size.

The auto encoder is a convolutional neural network that is used in unsupervised feature learning methods. Its aim is to pick up on the input's compressed representation. Even though auto encoders are developed with supervised learning methods (self-supervised), it is considered an unsupervised learning approach because they do not depend on human supervision to improve their performance. To replicate the input, they are usually trained as part of a larger model.

Autoencoder models are generally challenging to train the data since their design is constrained to a single node in the middle of the network, where input data reconstruction occurs.

Seven distinct autoencoder varieties exist. Specifically, they are the following:

- Deep Autoencoder
- Sparse Autoencoder
- Under complete Autoencoder
- Convolutional Autoencoder
- Denoising Autoencoder
- Variation Autoencoder
- Contractive Autoencoder

Simple autoencoder is a non-recurrent neural network, feed-forward that consists of an input layer, an output layer, and optionally one or more hidden layers. In an autoencoder, the number of nodes in the output layer matches that of the input layer.

Instead of attempting to forecast the final Y value, individuals prefer to recreate their own input. In this way, autoencoder can be understood as a type of learning model. Two of the most intriguing real-world uses of autoencoders are in features extraction and dimensional reduction for data visualization.

In addition, the k-nearest neighbour, or KNN, technique is applied to the dataset in order to complete it and the elbow and silhouette method is utilized in the process of determining the appropriate number of clusters for the K-means algorithm. Hierarchical clustering is a technique that is utilized in wastewater purification plants to improve

existing clustering methods. This technique includes continually merging groups into a single entity, each of which is represented by a unique data point. As a result of this, a hierarchical structure made up of clusters is developed. Most research shows that creating a dendrogram is the most effective way to show hierarchical grouping, and therefore it is recommended. The structure of this representation is very much like a tree graph. In a dendrogram, the degree in which the different clusters can be distinguished from one another is indicated by the length of the branches.

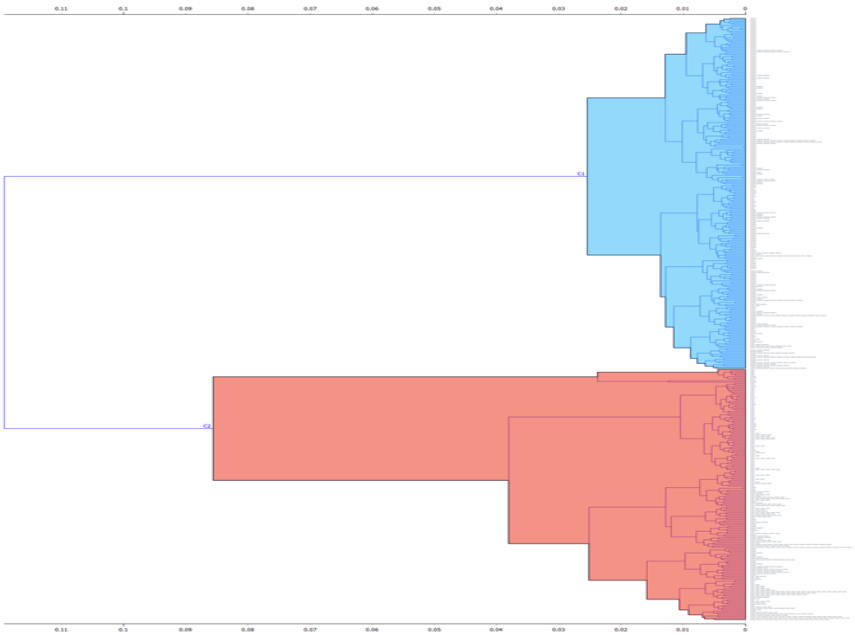


Figure 24: Dendrogram

It is useful because it gives some insight of the hierarchical clustering patterns. It is able to get it by utilizing a specialized algorithm. In addition to locating the end at the moment in time when the distances between clusters initially began to increase at an exponential rate. After that, to indicate the amount of distance that separates two sets of data instances from one another.

- Complete Linkage (Farthest pair of the points);
- Single Linkage (Closest points pair);
- Average Linkage (Average Points Distance);
- Ward linkage (a measure based on Intra-Cluster Variance).

The procedure of clustering can be approached from a number of different angles. When determining the initial cluster count, it is necessary to (k). The k-means technique group's similar data by first allocating each data point to the k-centroid that is geographically closest to it, and then randomly putting additional k-centroids.

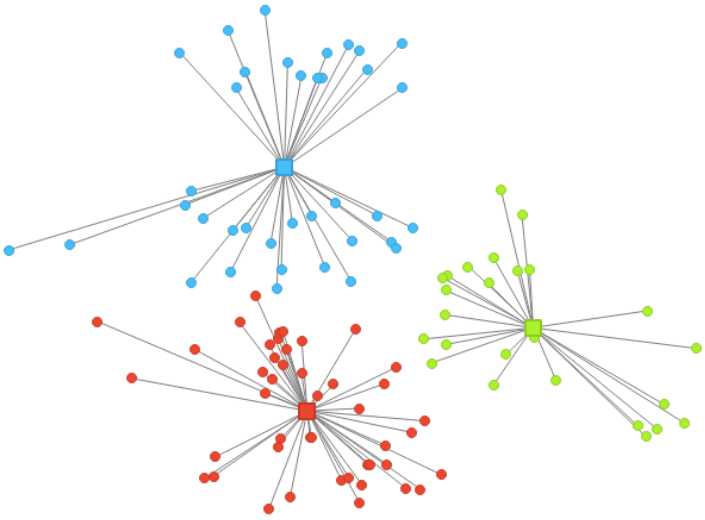


Figure 25: K-means Clustering

Each of the k-centroids is moved by the algorithm to the point of intersection of the cluster that is relatively closest to it. After then, it allocates all of the data based on where it perceives the new centers ought to be located. In addition to this, it repeats the process with the movement of the centroids.

Hierarchical clustering begins with the distances matrix, which may or may not correspond to the viewable coordinates. K-means clustering computes Euclidean distances between coordinates, whereas hierarchical clustering begins with the distances matrix. The strategy known as hierarchical clustering has the advantage of producing a more holistic view than other methods. As a result of this, it can be useful for measuring the quality of clusters and determining the optimal number of clusters to create.

The K-means method is significantly more efficient than other methods. It usually operates more quickly, with a few exceptions for areas that are less than optimal. In this particular instance, researchers have determined the process with multiple new starts results in the best overall output. It's possible that hierarchical clustering will be a slow operation because it takes a lot of memory.

VI. CONCLUSION

Wastewater management is a significant problem. There are a lot of people that don't have access to a sewer system, on the one side. However, many people who have access to a sewer system do not handle their waste properly.

Almost every municipality in the country produces wastewater, which must be treated and controlled. Here it examines the many aspects of wastewater management, including where it comes from and how to cut back on it. It discusses the good and bad points of having a wastewater treatment facility. The difficulties that can arise in a wastewater treatment facility have also been examined.

Findings from the main sections of wastewater treatment plants (WWTPs) can be expected to assist in process design and controls, reduce operational costs, improve system reliability, and encourage optimization of overall performances. Researchers propose to develop deep learning technologies as proven data-driven soft-sensors for WWTP applications to handle the dynamic, non-linear, and periodic nature of environmental data. Even though the model was developed using plant-by-plant detailed records, insufficient data still restricts the model's training. Researchers expect similar research take into account as sensor networks because of the larger dimensions and higher frequency of data.

The treatment of waste water plant is a vital component of the structure that must be present in order to maintain a clean and healthy environment. Because of the significant quantity of energy that is consumed, it is absolutely necessary that these facilities be operated in a way that maximizes the effectiveness of treatment while minimizes the amount of energy that is wasted. The findings that were enhanced under the objectives demonstrate that deep neural networks under Clustering and Hierarchical technique may be successfully utilized to solve the problem of upset conditions in wastewater treatment systems; however, this is only the beginnings of what might be a significant and productive line of research.

Classifying the units of this dataset was a challenging problem, as they represented the various days for which many variables had been recorded to detect the incidence of defects in a wastewater treatment plant. Using only clustering methods and no other data, they were unable to identify all the offending days.

Due to their extreme deviation from the standard, only a subset of them were consistently picked up by the majority of the algorithms; this indicates that on average, faults exhibit characteristics that are remarkably similar to those of operational days. The intrinsic heterogeneity of the data may have been controlled by missing data imputation, making the clustering problem more difficult to resolve. Based on our analysis of the clustering techniques' performances, it appears that the algorithms have problems accessing a stable group structure that can distinguish between the conclusions drawn on good and bad days. In general, hierarchical modal clustering performed similarly to single linkage, and its inadequate performance could be related, at least in part, to the large dimensionality of the data. Adding inter-observation dependencies did not appear to increase clustering quality in terms of cluster features or fault-detection ability. There is no reason to rule out the possibility that because not all the data in the dataset was actually utilized.

In reality, this dataset is considered an "ill-structured domain," which means that there is no clear-cut group structure and no guaranteed solution to the problem at work. Furthermore, this is typical of real-world problems, when data are contradictory and there is no unique answer present one might propose various solutions, each with their own strengths and drawbacks. Both conceptual clustering and data clustering are required in these instances. As a result, it is critical to seek out and utilize external information and expertise of context-specific knowledge. In fact, methods are extremely important, as they assist in locating the optimal solution by addressing multiple issues in a comparable environment and with similar goals.

V. FUTURE WORK

In view of the above mentioned, that have managed to identify potential avenues for additional research, all of which involve both scientific and methodological concerns. To fill in the data gaps, one needs a deeper comprehension of how these data are created. Since the missing information does not appear to have been removed at random, it is probable to require us to employ ad hoc imputation algorithms that are capable of retaining as much of the original data's variability as is practicable. External covariates, such as weather variables, whose effect on the operation of a wastewater treatment is well demonstrated in the literature, or other plant characteristics that might emerge from a more in-depth analysis of the plant's characteristics on defective days, could greatly improve under fault detection in addition to better data. From a methodological perspective, it is important to acknowledge that the potentials of Hidden Classifiers have not been fully exploited due to the limitations of the algorithm implemented in analysis, which was designed in such a way that it achieved the best possible results only in the univariate or, at most, the bivariate case.

A bivariate Deep Learning Model on the data is to be used, with the first two main components as our independent variables. This makes it easier to understand just a portion of the overall variance. Thus, in order to fit a more complex model on a bigger subset of the variables, such as the ones identified as most relevant in our analysis, it may be necessary to further refine the algorithm. This would allow for more precise consideration of the interdependence among the dataset's units, leading to improved outcomes. Alternative ML methods including Random Forest, may be used to further investigate predictions can be more accurate using the highly dimensional data. In addition to this, maybe accurate prediction of plant upset conditions is not the correct approach. Moreover, prediction of sensors is the correct approach to anticipate sensor failures that could result in missing upset conditions not being correctly detected or diagnosed.

V. LIST OF ABBREVIATIONS

1. **ANN** – Artificial Neural Network
2. **ARIMA-** Autoregressive Integrated Moving Average
3. **ASP** – Activated Sludge Process
4. **BOD** – Biological oxygen demand
5. **COD** – Chemical oxygen demand
6. **ML** – Machine Learning
7. **KNN** -K-Nearest Neighbour
8. **TDNN** - Time Delay Neural Networks
9. **WWTP** – Waste Water Treatment Plant

REFERENCE

[1] Ruiz-Shulcloper, J. and di Baja, G.S., 2013, November. Progress in pattern recognition image analysis computer vision and applications. In Proc. 18th Iberoamerican Congr.(CIARP) (pp. 1-541).

[2] Dey, A. (2022) Wastewater Treatment: Definition, Process Steps, Design Considerations, Plant Types, Whatispiping.com [online]. Available from: <https://whatispiping.com/wastewater-treatment/> [accessed 4 July 2022].

[3] POMERANZ, E. (2022) Wastewater Treatment Plants - Amiad Water Systems, Amiad Water Systems [online]. Available from: <https://amiad.com/blog/wastewater-treatment-plants/> [accessed 4 July 2022].
[4] Nathanson, J. and Ambulkar, A. (2022) wastewater treatment | Process, History, Importance, Systems, & Technologies, Encyclopedia Britannica [online]. Available from: <https://www.britannica.com/technology/wastewater-treatment> [accessed 5 July 2022].
[5] Chandola, V., Banerjee, A. and Kumar, V. (2009) Anomaly detection. ACM Computing Surveys, 41(3), pp.1-58.
[6] Zhou, B. and Gu, X. (2020) Multi-block statistics local kernel principal component analysis algorithm and its application in nonlinear process fault detection. Neurocomputing, 376, pp.222-231.
[7] Gu, Y., Zhao, W. and Wu, Z. (2011) Online adaptive least squares support vector machine and its application in utility boiler combustion optimization systems. Journal of Process Control, 21(7), pp.1040-1048.
[8] Ahmadi, F., Mehdizadeh, S. and Nourani, V. (2022) Improving the performance of random forest for estimating monthly reservoir inflow via complete ensemble empirical mode decomposition and wavelet analysis. Stochastic Environmental Research and Risk Assessment.
[9] Kweinor Tetteh, E. et al. (2020) Occurrences in Wastewater Systems: Emerging Detection and Treatment Technologies—A Review. Water, 12(10), p.2680.
[10] Du, X. et al. (2018) Parameter estimation of activated sludge process based on an improved cuckoo search algorithm. Bioresource Technology, 249, pp.447-456.
[11] Corominas, L. et al. (2018) Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. Environmental Modelling & Software, 106, pp.89-103.
[12] Avila, R. et al. (2018) Evaluating statistical model performance in water quality prediction. Journal of Environmental Management, 206, pp.910-919.
[13] Djerbouai, S. and Souag-Gamane, D. (2016) Drought Forecasting Using Neural Networks, Wavelet Neural Networks, and Stochastic Models: Case of the Algerois Basin in North Algeria. Water Resources Management, 30(7), pp.2445-2464.
[14] Wang, J. et al. (2018) Delay-dependent dynamical analysis of complex-valued memristive neural networks: Continuous-time and discrete-time cases. Neural Networks, 101, pp.33-46.
[15] Fiedler, F., Cominola, A. and Lucia, S. (2020) Economic nonlinear predictive control of water distribution networks based on surrogate modeling and automatic clustering. IFAC-PapersOnLine, 53(2), pp.16636-16643.
[16] Hajizadeh, M. and Lipsett, M. (2018) Application of interacting multiple model-based fault detection method on a hydraulic two-tank system. International Journal of Condition Monitoring, 8(2), pp.42-51.
[17] Luif, P. (2014) Der Konsens der Staaten der Europäischen Union in der Außen- und Sicherheitspolitik. Analysiert am Abstimmungsverhalten in der Generalversammlung der Vereinten Nationen. Strategie und Sicherheit, 2014(1).
[18] Sadgrove, E. et al. (2018) Real-time object detection in agricultural/remote environments using the multiple-expert colour feature extreme learning machine (MEC-ELM). Computers in Industry, 98, pp.183-191.
[19] Kellenberger, B. et al. (2019) Half a Percent of Labels is Enough: Efficient Animal Detection in UAV Imagery Using Deep CNNs and Active Learning. IEEE Transactions on Geoscience and Remote Sensing, 57(12), pp.9524-9533.
[20] Li, H. et al. (2020) Exploration of OpenStreetMap missing built-up areas using twitter hierarchical clustering and deep learning in Mozambique. ISPRS Journal of Photogrammetry and Remote Sensing, 166, pp.41-51.
[21] Brandt, M. et al. (2020) An unexpectedly large count of trees in the West African Sahara and Sahel. Nature, 587(7832), pp.78-82.
[22] Jin, R. et al. (2021) Toward Efficient Object Detection in Aerial Images Using Extreme Scale Metric Learning. IEEE Access, 9, pp.56214-56227.
[23] Tan, K., Munster, A. and Mackenzie, A. (2021) Images of the arXiv: Reconfiguring large scientific image datasets. Journal of Cultural Analytics, 6(1).
[24] Dong, Z. et al. (2022) Multi-Oriented Object Detection in High-Resolution Remote Sensing Imagery Based on Convolutional Neural Networks with Adaptive Object Orientation Features. Remote Sensing, 14(4), p.950.
[25] Yao, Q., Hu, X. and Lei, H. (2021) Multiscale Convolutional Neural Networks for Geospatial Object Detection in VHR Satellite Images. IEEE Geoscience and Remote Sensing Letters, 18(1), pp.23-27.

[26] Du, X. and Zare, A. (2020) Multiresolution Multimodal Sensor Fusion for Remote Sensing Data With Label Uncertainty. IEEE Transactions on Geoscience and Remote Sensing, 58(4), pp.2755-2769.
[27] Liu, L. et al. (2020) Shallow–Deep Convolutional Network and Spectral-Discrimination-Based Detail Injection for Multispectral Imagery Pan-Sharpening. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, pp.1772-1783.
[28] Hong, D. et al. (2019) Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification. ISPRS Journal of Photogrammetry and Remote Sensing, 147, pp.193-205.
[29] Yokoya, N. et al. (2018) Open Data for Global Multimodal Land Use Classification: Outcome of the 2017 IEEE GRSS Data Fusion Contest. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(5), pp.1363-1377.
[30] Haworth, B. (2017) Implications of Volunteered Geographic Information for Disaster Management and GIScience: A More Complex World of Volunteered Geography. Annals of the American Association of Geographers, 108(1), pp.226-240.
[31] Lewis, Q. and Park, E. (2017) Volunteered Geographic Videos in Physical Geography: Data Mining from YouTube. Annals of the American Association of Geographers, 108(1), pp.52-70.
[32] Pittmann, T. and Steinmetz, H. (2017) Polyhydroxyalkanoate Production on Waste Water Treatment Plants: Process Scheme, Operating Conditions and Potential Analysis for German and European Municipal Waste Water Treatment Plants. Bioengineering, 4(4), p.54.
[33] Wie, Y., Lee, K. and Lee, K., 2020. Physicochemical effect of the aeration rate on bloating characterizations of artificial lightweight aggregate. Construction and Building Materials, 256, p.119444.
[34] Roy, S., Tanveer, M., Gupta, D., Pareek, C. and Mal, B., 2021. Prediction of standard aeration efficiency of a propeller diffused aeration system using response surface methodology and an artificial neural network. Water Supply, 21(8), pp.4534-4547.
[35] Wei, W. and Deng, J., 2022. Free surface aeration and development dependence in chute flows. Scientific Reports, 12(1).
[36] Kent, R., 2017. Energy measurement in kWh/kg. Plastics Engineering, 73(3), pp.42-44.
[37] Mehari, A., Xu, Z. and Wang, R., 2022. Thermodynamic evaluation of three-phase absorption thermal storage in humid air with energy storage density over 600 kWh/m3. Energy Conversion and Management, 258, p.115476.
[38] Hoque, A. and Paul, A., 2022. Experimental investigation of oxygen transfer efficiency in hydraulic jumps, plunging jets, and plunging breaking waves. Water Supply, 22(4), pp.4320-4333.
[39] Serpokrylov, n., smolyanichenko, a. And lesnikov, i., 2011. Comparative analysis of aerators for wastewater purification and unified aeration criteria. Urban construction and architecture, 1(2), pp.97-100.
[40] Bang, S., Bum, B. and Kim, J., 2019. A study on the determination of a representative location for monitoring the dissolved oxygen concentration in a aeration tank of sewage treatment plant. Journal of the Korean Society of Water and Wastewater, 33(5), pp.389-394.

