Anomaly Detection in Smart Water Management System

Ashok Kumar, Shreya Jingade D, Nisha Setty A R, Shreevatsa T P

BMS College of Engineering, Bengaluru, India

Corresponding author: Shreya Jingade D, Email: shreyajd.scn20@bmsce.ac.in

Freshwater available for usage is very little, so conservation and efficient usage of water plays a major role, to avoid scarcity of consumable water shortly. To help in saving water, the knowledge of where the water is getting wasted becomes important. The development of the smart city initiative taken by the government helps to preserve the natural resources and avoid unnecessary wastage. A smart water meter management system is designed and analyzed here. In this paper, the anomaly detection using a smart water meter in a house/ building is being analyzed using a few technologies like IoT, Cloud Computing, and ML Algorithms. DBSCAN, Isolation Forest, and K-Means are a few of the unsupervised clustering algorithms that are used to find outliers in the smart water consumption dataset. The goodness measure of each anomaly detection algorithm is represented by the graph showing the outliers in each of the cases.

Keywords: Smart water meter, Cloud, IoT, Anomaly, Machine Learning algorithms.

2023. In Saroj Hiranwal & Garima Mathur (eds.), *Artificial Intelligence and Communication Technologies*, 379–392. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-955020-5-9-37

1 Introduction

Almost around 71% of the earth is draped by water and it is an essential resource. In this 71%, only around 1% of the water is fresh and suitable for drinking and performing other day-to-day activities. Therefore, it becomes very important to use water in a very optimized and efficient way without much wastage. As there is a huge population and more industries in the metropolitan cities and other popular cities, the water demand always stays at the pinnacle [1]. Water is a scarce resource. As the water demand is increasing day by day, it is becoming difficult to fulfill the same[2]. In this situation, if the water goes unused due to some reasons such as lack of proper monitoring of the water consumption, leakage from the water pipe, leaving the taps unattended, causes a problematic situation. Hence, water monitoring using technology such as IoT, cloud, Machine Learning, etc., gives much focus on the conservation of water. IoT plays a major role in this, as it connects devices through the internet and sensor is part of such technology which is used in smart water meters. To conserve water in the cities, IoT smart water sensors are used. Smart cities are being developed using information from the sensor's data to build a responsible society. The model that is designed and implemented is easy and cost-efficient making it user-friendly for all the users. The water reading data is measured from the sensor. This data usually will be stored in the cloud. Publishers can publish to this topic; subscribers can receive the subscribed topic to and from the publisher through the MQTT broker. Message Queuing Telemetry Transport (MQTT) is a lightweight protocol, which transports messages between the devices which publish and subscribe to the data[3]. Most of the houses' water consumption is not measured in a real-time manner. They will have the water meter which is analog, to measure the consumption of water. Analog dials will be present in the analog water meter which will display the total water consumption reading[4]. The user has to wait till the end of the month to receive the bill. Hence by using this water meter reading which will be recorded user has to manually compare the reading and hence there is no information about the water consumption rate daily and water leakage measurement by using this analog water-meter technique. Hence this method will be difficult therefore the smart water-meter is built here, with the help of a Machine learning algorithm. This can be used to detect the anomaly in water consumption. Machine learning algorithms will not need too much intervention from people. In the supervised learning method, training data along with the output label will be present. In the unsupervised learning methods, only data without any output label will be given. As the data will not have any information about the anomalies, unsupervised learning methods can be used to detect the anomalies in the data. Unsupervised Machine learning algorithms will learn by themselves as it does not have any output label. It detects the relationship pattern in the dataset and then analyzes whether the data is anomalous or not. Cloud computing is one of the major technologies used by many people. The entire process described in the paper is done on the cloud. Cloud provides services on demand such as storage and computation[5]. To process the data, a huge amount of computational power is required and the cloud is the best option in such cases. Not only this, data and processed information can easily be retrieved.

The paper describes the analysis performed on the dataset. The water-meter reading, water consumption, DateTime parameters are considered and unsupervised learning algorithms are applied to detect whether the dataset set identified contains the data which is an anomaly. As the data does not contain any label such as anomaly and normal, an unsupervised machine learning algorithm is used. Clustering algorithms such as DBSCAN, isolation forest, k-means algorithm are used. Clustering algorithm, groups the data in the dataset as anomalous and non-anomalous by finding the pattern in the data. These algorithms' accuracy rates are also compared.

2 Review of Literature

2.1 Water consumption detection using IoT

IoT is nowadays used as the main aspect for many intelligent systems. Using IoT, objects that are connected which have sensors, actuators, processors can communicate with each other over the internet. IoT can be used as a part of water management for smart homes. A sensor can be inserted inside the water pipe, using the reading which the sensor provides further analysis can be made anomalies and water consumption patterns and better solution can be found.

Fuentes et al. [6] used smart meters, gateway, and cloud technologies for the purpose of anomaly detection in the household. Water consumption data of a house is collected through the smart water meter, and then stored in an edge gateway; also an anti-tampering system is used for the security of the device. Then this data is sent to the cloud and a leak detection algorithm is applied. Scenarios for anomalous behavior such as negative reading, a day's consumption, huge consumption of water, and the same consumption compared to other days in the house are analyzed. Many test cases detected the anomaly and obtained 100 % accuracy.

Herath [7] has developed a system that will detect the usage of water using an IOT device and detect anomalous data using machine learning techniques. DBSCAN and K-means algorithms are used for anomaly detection. The rate of flow of water in the pipe is calculated using the NodeMCU microcontroller, then the data is validated by the edge processor and sent to the server using MQTT protocol. When this system is installed in a house, based on the pattern of water used, the anomaly is detected. Sandhya et al. [8] developed a sensor for the purpose of monitoring water in the house. If the value shown by the sensor is greater than the threshold value then the alert is sent to the user from Arduino UNO using the GSM module. An accurate result is given if there is WIFI connected.

2.2 Cloud Computing

Cloud computing helps with the storage, better performance with improved speed, and also computation power for the user based on the demand. A dataset that is huge in volume takes a lot of space and requires high computation power, hence the usage of the cloud helps to solve this problem. Easily computed data can be accessed from the cloud.

For the same Rayane El Sibaiet al. [9] proposed a cloud service to offload the work to it. Using which intelligent monitoring service for the water consumption is developed.

Alshattnawi [10] has provided theoretical supplies for the execution of the IoT Smart water supply[10]. Here Internet of things and Information and Communication Technology are combined to enable efficient operation, maintenance, and management of water quality. The architecture for Smart Water Distribution System with Information and communication technology that combines Cloud and IoT technologies are explained.

The smart water meter is developed by Ray et al. [11] using IoT and cloud computing technology. For the detection of anomalies, a machine learning algorithm is used. Real-time data is collected from the NodeMCU and it will be sent to the cloud server, which makes the task of visualization and analyzing easy. The amount of water flowing from a particular pipe for a particular time through the cloud and then the water consumption is measured and differentiated between normal and anomaly behavior. For large scale deployment this method can be implemented easily.

Ali et al. [12], developed a water leakage detection model. Individual reading will be automatically taken from the water meter using raspberry Pi 4 and uploaded to the cloud using internet. Analysis of the data obtained will be made based on which the leakage alert will be sent to the user. Accurate detection and monitoring of anomalous behavior are possible by this method.

2.3 Anomaly Detection

As water is an important part of our life, wastage of water makes it a scarcer resource, hence the detection of anomalous consumption to know the wastage and use the water in an efficient way plays a major role. The reasons for the wastage of water can be due to the pipe leakage, ignorance of people by not closing the water tap after usage, or by keeping the tap on and doing multiple works, lack of knowledge about the importance of water management.

Vercruyssen et al. [13] used clustering approach for the anomaly detection. COP k-means constrained clustering algorithm is used. They followed assumptions that a data point is anomalous if it is away from the centroid, also if the centroid deviates from another one it will also be considered as an anomaly if the number of data points is less or not sufficient to form a cluster then again it is considered as anomalous data. If the data has no label, the unsupervised method is used by finding out the anomaly score on the clustering. This system is used by the Colruyt group in 20 stores to monitor the usage of water. Hence provides accurate results to the user who uses this system in the stores.

Kainzet al. [14] used the K-means algorithm for anomaly detection in water consumption. The data processing was made on water consumption prediction and non-standard situations. And the prediction was also made considering hours, days, the current month. This method uses the factor's actual and predicted value. Percentage difference between these two values is found. If the percentage difference between them is greater than the actual value then it is an anomaly. Results show the accurate prediction and fewer false positives and negatives. Vidal et al. [15], used ARIMA and HOT SAX methods for anomaly detection in the time series data. ARIMA method detects the outlier by detecting the data points which does not fit the method. HOTSAX differentiates using a heuristic. The accuracy obtained is 76%. Predescu et al. [16] used a k-means algorithm for the anomaly detection in the water consumption dataset. Data from the water distribution is collected and clustering is performed using the k-means algorithm which is an unsupervised machine learning algorithm. Conditions that are normal in water consumption for a house are obtained by this algorithm. The pattern will be calculated for different types of anomalies, which is detected by the algorithm. According to the algorithm and pattern, user will know the type of anomaly. Whenever the anomaly is detected, the alert will be generated and sent to the user.

3 Proposed Architecture of Anomaly Detection in Smart Water Management System

There are four blocks in the proposed architecture as shown in Fig. 1. First phase/block of the proposed architecture is data collection phase, this acts as publisher. Second block is MQTT Broker/server, using MQTT Protocol data is streamed to the MQTT Server. Third block is data storage which acts as subscriber, data is stored on the cloud. Fourth block and final stage is Data processing, this process is performed using few of the Machine Learning Algorithms to find anomaly in the water consumption dataset.

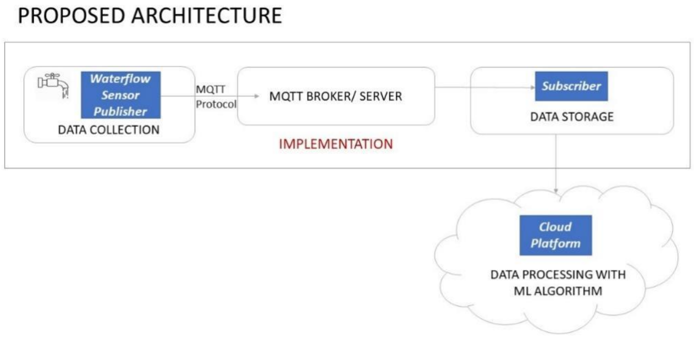

Fig. 1. Overall architecture of anomaly detection in smart water management system

3.1 Experimental Setup

(i) In Cloud Big Query, create a Table with the appropriate features.

(ii) Create a subscription and a Pub/Sub topic.

(iii) Dataset is then saved in the bucket.

(iv) With the Device ID, add the device to the Registry.

(v) Create a Virtual Machine Instance

(vi) To get the result table, write a query

3.2 Dataset

A dataset that is taken is clean data. There are 673927 data in the dataset. Dataset consists of uniqueID which is ID of the water meter, Datetime represents the date and time on which the reading is taken, meter reading indicates the current reading, water consumption indicates the difference between previous reading. Here, the date-time attribute indicates that the reading is taken from 2016 March till 2017 February which is one-year data, this data is taken on a per hour basis. In the water consumption column, 0 is the common value that indicates the consistent value in the water meter reading.

3.3 Water flow Sensor

Nowadays, the smart home concepts are being incorporated in houses, part of which is real time monitoring of water flow. This is achieved by using the water flow sensor, the sensor will be inserted in the water pipes. The sensor checks the pulse, which determines whether water which is flown through the pipe is excess or the flow is normal, which means the water usage is approximately same compared to other days.

3.4 MQTT

MQTT is a light weight protocol, it helps to transport/communicate messages between the devices.

3.5 MQTT Broker/ Server

MQTT publisher and subscribers communicate using MQTT broker/ server. Initially, connect packet will be sent to the server from MQTT client/ subscriber. Topic for the subscription will be sent to the server from the client, the reply packet is then sent back to the subscriber for the topic which was subscribed. After all the communication, disconnect packet is sent to the sever which is the indication that the packet is the last.

3.6 Data Storage

Cloud is used for the data storage purpose. Water consumption data is stored in the cloud, as cloud provides many services on demand. Data stored on the cloud can be accessed easily anytime for further processing. This data storage acts as a subscriber and it subscribes water consumption topic from the MQTT server.

3.7 Data processing

Data processing is performed on the cloud platform. Few machine learning algorithms are used to detect the anomaly in the dataset obtained.

Pre-processing: The collected dataset is cleaned data and hence there is no need for further pre-processing steps to clean the dataset. The dataset collected is very huge in volume i.e., it contains 673927 rows of data which makes the processing very slow, and performing analysis using such a dataset will be difficult and time-consuming. Hence, only unique values are considered. To make the analysis part simpler data and time are split as two different attributes.

3.8 Anomaly Detection Algorithm

DBSCAN Algorithm

As this algorithm is robust to outlier and it can identify cluster in huge datasets, DBSCAN algorithm is chosen. DBSCAN Algorithm uses the clustering technique, it considers the dense parts as groups and forms the cluster and the data point which does not belong to any cluster or if they are far from the dense region then the data point will be considered as anomaly.

Epsilon and Min point are two important parameters. Epsilon indicates the maximum distance between one data point and other which will be the neighbour data points. Distance should not be too large or too small for the clustering to be done efficiently. Min point indicates number of points required to be grouped under a single cluster or consider the data points as clusters.

Algorithm:

Step 1: To find the distance between data points

Step 2: Take a data point and check if there are min a point number of neighbours within eps range of that point

Step 3: If the data points for that point within the eps range is >= min point number of neighbours then that point is a core point

Step 4: If the data points for that point within the eps range is < min point number of neighbours then that point is a border point

Step 5: If the there are no points other than the point itself within eps range then that point is categorized as noise point or anomaly or outlier

Step 6: Repeat step 2 to 5 for all the data points present in the data set to obtain the cluster and identify anomalies

The DBSCAN Flowchart is as shown below in Fig. 2.

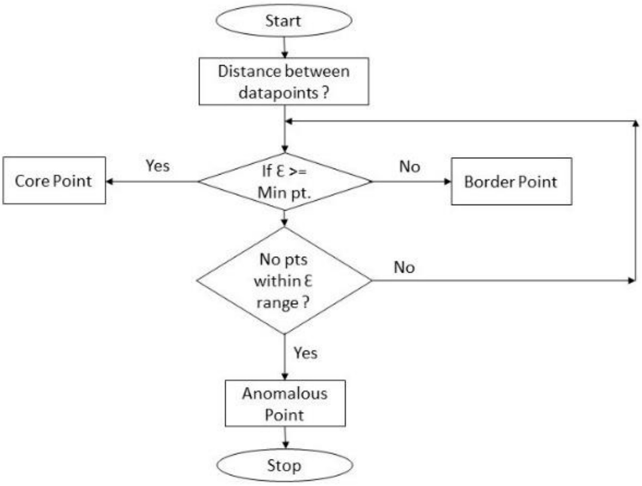

Fig. 2. DBSCAN Clustering Flowchart

K-Means Clustering

In K-means clustering the value of K (number of clusters) has to be predefined. This value of K can be found by using of the two methods mentioned below.

(i) Trial and Error method
(ii) Elbow method

Once the K value is known, the distance from K number of centroids to all the data points is measure, the data points which are nearer to each centroid is grouped under one cluster, if any data point is far from all the centroids i.e., distance is greater than boundary then such data points are taken as anomalous data points. As K-Means is used for dense datasets, this algorithm is chosen. Few threshold values can be assumed for it to be non-anomalous range of water meter readings and the data points which does not come under this range will be considered to be anomalous.

Algorithm:

Step 1: Determine K-value using Elbow method

Step 2: Assign the K number of centroids randomly

Step 3: Distance between data points and centroids are measured, according to the minimum distance, data points are assigned to the nearest centroid

Step 4: If a point is far from all the centroids, then such point is outlier point

Step 5: If the clusters are stable then, convergence

Otherwise repeat from step 3 to 4

The K-Means Clustering Flowchart is as shown below in Fig. 3.

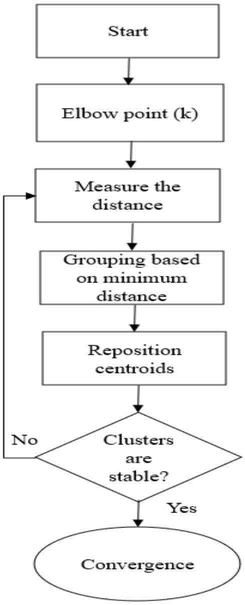

Fig. 3. K-Means Clustering Flowchart

Isolation Forest

Isolation forest is an unsupervised algorithm that is based on the decision tree. Anomalies in the dataset will be different and few. Random features are selected from the dataset and sub-sampled to form a tree. Samples are considered to be an anomaly when they have, they stop at shorter branches. It does not define any fixed normal behavior; it just takes the dataset and fix the min and max values then clustering is formed by random split between these min and max value. It does not consider any point-based calculations.

Algorithm

Step 1: Randomly sub-sample is selected from the dataset which is assigned to binary tree
Step 2: Random features are selected to create branches
Step 3: If the data point value is less when compared to the threshold value, then it goes to the left part of a branch, else right part of a branch
Step 4: Recursively repeat from step 2 till isolation of each data point is done
Step 5: To construct random binary tree Step 1 to 4 is repeated

The Isolation Forest Flowchart is as shown below in Fig. 4.

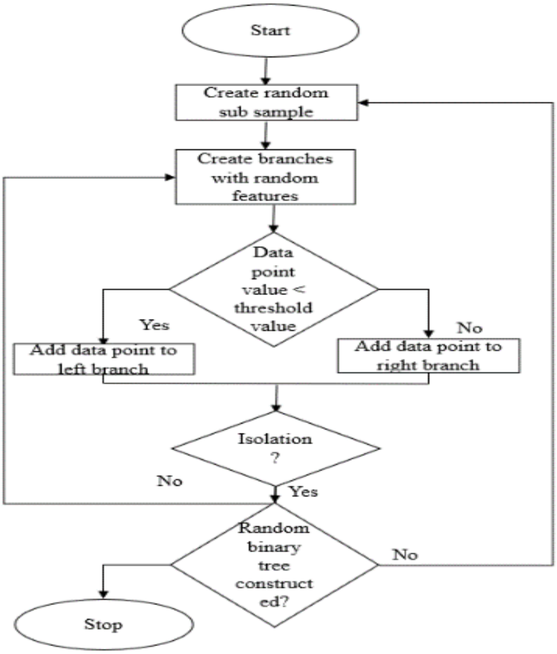
Start
Create random sub sample
Create branches with random features
Data point value < threshold value
Yes
No
Add data point to left branch
Add data point to right branch
Isolation ?
No
Yes
Random binary tree constructed?
No
Stop

Fig. 4. Isolation Forest Flowchart

4 Results and Discussion

The accuracy of the anomalies in the water consumption for a period of time is analyzed by using different anomaly detection algorithms in machine learning. The analysis is made in different phases using different machine learning algorithms and checking its anomaly.

711788 entries of data had been collected and stored in the cloud. Different ML anomaly detection algorithms were then applied to this large dataset that was initially collected, to find the outliers in the water consumption of the users of a building/ house. The anomalies may be due to several reasons, like very low/ no consumption of water as compared to the previous data, same consumption of water for two consecutive days, 24-hour water consumption, and very high consumption of water. The outliers in the water consumption dataset are then calculated and compared using different ML models.

4.1 DBSCAN

The two main parameters that are considered while applying this algorithm are Epsilon and Min. Points. Epsilon and Min. Point value is taken as 10 to detect the outliers by knowing the core point, border point, and noise point with respect to the above two parameters. It is important to keep an optimized value for these parameters to have procured an accurate clustering and find the exact outlier values.

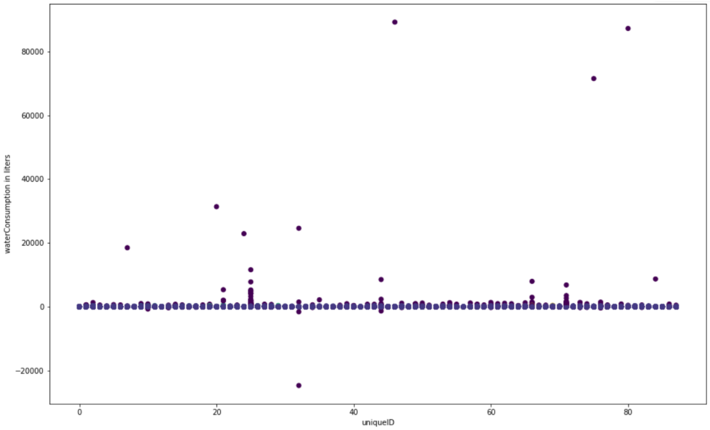

Fig. 5. Outliers found in water consumption data using DBSCAN Anomaly detection algorithm

As the dataset obtained is very large, the computation power and time consumed for processing the data becomes very high and hence the duplicate dataset which is redundant is removed. After which a redundancy cleared file is obtained on which the ML algorithm is applied. In the above Fig. 5, x-axis is the unique id for each house/device and the y-axis is the water consumption values of those devices. The water consumption nearer to zero is clustered as normal data and the very high and very low consumption of water, which is 20000 in the negative range and 20000 or above in the positive range of the dataset is considered to be anomalous, or outliers or noise points. The negative ranged water consumption data here shows that there might be a fault in the sensor device or the reading captured by the sensor is flawed. On the other hand, the positive/high ranged water consumption data here talks about the wastage or extensive use of water consumption or any pipe leakage in a house/ building. Henceforth, with the knowledge of these anomalies' measures can be taken forward in order to save and use water in a better way. The goodness of the algorithm is measured using the silhouette score or silhouette co-efficient. It ranges from -1 to +1. 0.162 is the silhouette score obtained for this algorithm.

4.2 Isolation Forest

In the below Fig. 7, the x-axis is the Unique id for each house/device and the y-axis is the water consumption values of those devices. This algorithm does not take any threshold values; the algorithm itself gives max and min values to the data by randomly selecting the feature. By these values, anomalous data and normal data can be differentiated. The model creates iTress, after which anomaly score will be assigned to the depth of the iTree. This score ranges from -1 to +1. -1 indicates anomalous data and1 indicates normal data. 77.98% accuracy is obtained by using the isolation forest model.

The below Fig. 6, shows only one negative data point that is anomalous indicating very few instances of the sensor reading being flawed. Whereas there are more anomalous data points showing high water consumption or water wastage.

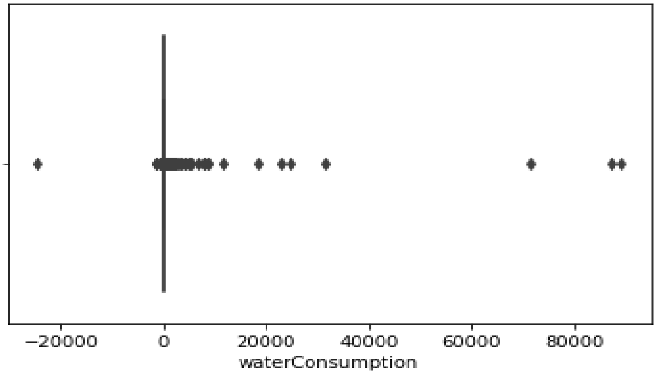

Fig. 6. Box-plot indicating the anomalous points

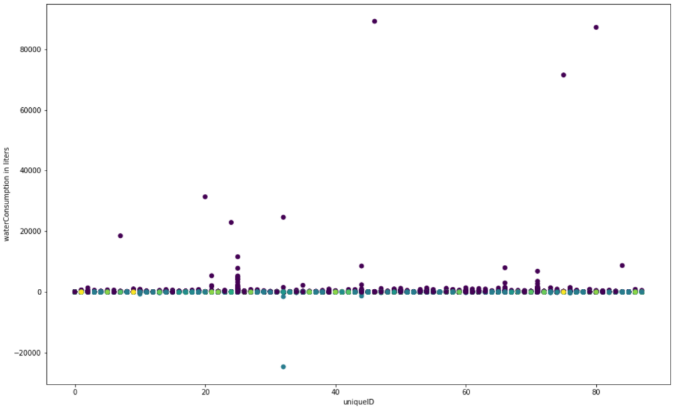

Fig. 7. Anomaly indication using Isolation Forest Algorithm

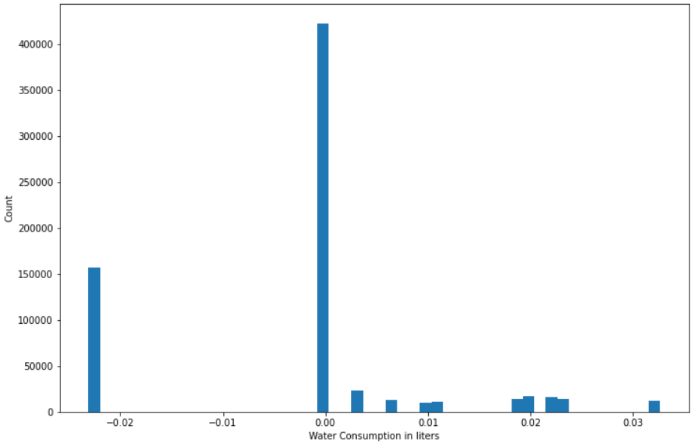

Fig. 8. Histogram Representation

In the above Fig. 8, x-axis is water consumption and y-axis are count of anomalous and non-anomalous data.

4.3 K- Means

Initially, the K (number of clusters) value is known by applying elbow method. Then the distance from the k centroids points to all the data points are measured and the data points which are nearest to each centroid is mapped. The data points which are not nearer to any of the centroids are taken as outlier points.

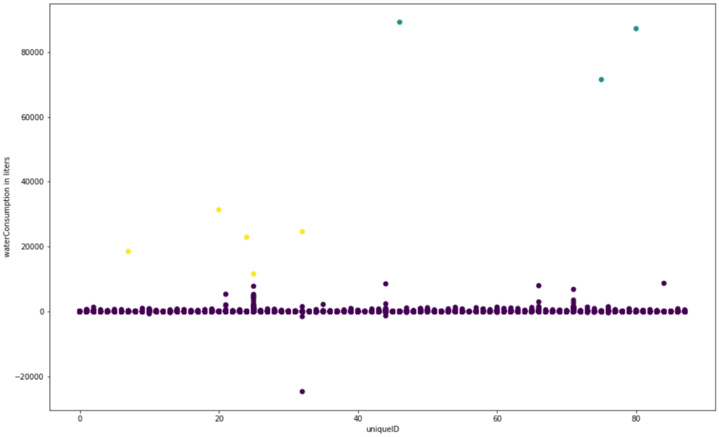

Fig. 9. Anomaly indication using K-Means Clustering Algorithm

In the above Fig. 9, the yellow-colored data points in the above graph indicates that the water consumption is very high in these datapoints compared to any other data points and does not belong to any of the clusters and hence is classified as anomalous data. The k value is taken as 3 and the goodness of the algorithm is calculated using Silhouette Coefficient which is 0.997 for K-Means algorithm.

4.4 Consolidated Table

Table 1. Consolidated result for the algorithms

Algorithm	**Result**
DBSCAN	The silhouette score obtained for this algorithm is 0.162
Isolation Forest	77.98% accuracy is obtained by using the isolation forest model.
K-Means	The k value is taken as 3 and the Silhouette Coefficient is calculated which is 0.997

The above Table 1, shows the consolidated result of DBSCAN, Isolation Forest and K-Means Algorithms.

4.5 Inference of the Paper

(i) Anomaly in the water meter dataset is detected and performance is compared using different machine learning algorithms.

(ii) Anomaly is the unusual behavior which is observed in the collected dataset, few reasons for such unusual behaviors maybe is as mentioned below:

a) 24-hour water consumption

b) Defective water meter

c) No water consumption

(iii) Three algorithms namely, DBSCAN, Isolation Forest, K-Means are used to find out such outliers.

(iv) Detection of anomaly helps in efficient use of water, as people can check the reason behind the anomalous behavior and take the decision accordingly. As usable water is a scarce resource, it is important to find a solution for such wastage of water.

5 Conclusion

As water plays major role in every one's life, it is important for careful usage of the water. Water may be wasted if the leakage occurs or because the tap is on for long time. This leakage is difficult to be calculated manually and without smart water meters. Anomaly in the water consumption is measured using the water consumption parameter which is obtained from the difference of two reading obtained from the water meter.

Anomaly reading will be obtained by using machine learning algorithms such as DBSCAN, Isolation Forest and K-Means clustering algorithms which are unsupervised learning methods. Hence by using these techniques users can obtain data stored in the cloud and can be aware of the usage of water. The goodness measure silhouette score is found for each of the anomaly detection algorithms. The silhouette coefficient for DBSCAN and K-Means are 0.162 and 0.997 respectively. The accuracy of the Isolation Forest is obtained as 77.98%. Present work concentrates more on finding the anomalous data, the future work can be continued by enhancing the present project to accommodate the feature where the cause of anomaly can be determined.

An alert system can also be added to the present project, where the system admin and the end-user can be alerted whether there is any type of anomaly in the water-meter reading. This anomaly can be caused under different circumstances like; faulty sensor reading in the meter, small leakage of water in the pipe, anomalous high consumptions, etc. This can be implemented using a user interface or any simple mobile application.

References

[1] Evans, M. (2020). Benefits of water conservation.Dotdash.https://www.thebalancesmb.com/conservation-efforts-why-should-we-save-water-3157877.

[2]Wikipedia contributors (2021).Water scarcity. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Water_scarcity&oldid=1065414927.

[3]JavaTpoint. MQTT protocol. https://www.javatpoint.com/mqtt-protocol.

[4]Verma, S. (2020). Purpose of An IoT-Based Smart Water Meter. https://www.wateronline.com/doc/purpose-of-an-iot-based-smart-water-meter-0001.

[5]Wikipedia contributors (2021). Cloud computing. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Cloud_computing.

[6]Fuentes, H. and Mauricio, D. (2020). Smart water consumption measurement system for houses using IoT and cloud computing. *Environmental Monitoring and Assessment*, 192: 602.

[7] Herath, I. S. (2019). Smart Water Buddy: IoT based Intelligent Domestic Water Management System. In *International Conference on Advancements in Computing (ICAC).*

[8] Kulkarni, S. A. et al. (2020). Intelligent Water Level Monitoring System Using IoT. In *IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC).*

[9] Sibai, R. E. et al. (2020). Cloud-based foundational infrastructure for water management ecosystem. In *5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech).*

[10]Alshattnawi, S. (2017). Smart Water Distribution Management System Architecture Based on Internet of Things and Cloud Computing. In *International Conference on New Trends in Computing Sciences (ICTCS).*

[11] Ray, A. and Goswami, S. (2020). IoT and Cloud Computing based Smart Water Metering System. In *International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC).*

[12]Ali, F. and Saidi, H. F. H. (2021). Water Leakage Detection based on Automatic Meter Reading. In *15th International Conference on Ubiquitous Information Management and Communication (IMCOM).*

[13]Vercruyssen, V. et al. (2018). Semi-Supervised Anomaly Detection with an Application to Water Analytics. In *IEEE International Conference on Data Mining (ICDM).*

[14]Kainz, O. et al. (2019). Detection of the non-standard situation in smart water metering. In *IEEE 15th International Scientific Conference on Informatics.*

[15]Gonzalez-Vidal, A. et al. (2019). IoT for Water Management: Towards Intelligent Anomaly Detection. In *IEEE 5th World Forum on Internet of Things (WF-IoT).*

[16]Predescu, A., Mocanu, M. and Lupu, C. (2018). A fault sensitivity analysis for anomaly detection in water distribution systems using Machine Learning algorithms. In *IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP).*

[17]Gchatzi (2017). swm_trialA_clean. DAIAD Github.