International Journal of Global Engineering (IJGE)
ISSN: 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 3 (2025) PAGES 33-46
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

Prediction-Based System on Water Management

Bhavani Chennupati[1], Padmavathamma M[2]

[1] 2nd Year – M.S. Data Science, EdTech Division, Exafluence INC, Sri Venkateswara University, Tirupati, India
[2] Department of Computer Science, SVU College of CM & CS, Sri Venkateswara University, Tirupati, India

Email: chennupatibhavani24@gmail.com, prof.padma@yahoo.com

Abstract
Water is an indispensable resource, and its efficient management is critical for sustainable development. This paper presents a data-driven framework for intelligent water management. The primary objective is to leverage machine learning techniques to analyze water consumption patterns, detect anomalies, and predict future usage, thereby enabling proactive decision-making for utility providers and consumers. The methodology encompasses a comprehensive data pipeline, from data ingestion and quality checks to advanced feature engineering. Unsupervised clustering algorithms, specifically K-Means and DBSCAN, are used to segment consumers based on their consumption behavior. For anomaly detection, the Isolation Forest algorithm is implemented to identify unusual patterns such as leaks, meter tampering, or sudden spikes in usage. The system is developed in Python using libraries such as Pandas, Scikit-learn, and NumPy for data processing and modeling. The model outputs and insights are integrated with Power BI, creating an interactive dashboard for visualization and analysis. This allows stakeholders to intuitively interpret complex data and monitor the water distribution network in near real-time. The key outcome of this project is a robust, scalable system that can significantly improve the efficiency of water management. It provides actionable insights that can help reduce water loss, optimize distribution, and promote conservation. Future work could involve incorporating more data sources, such as weather patterns and demographic data, and deploying the model in a real-time streaming environment.
Keywords: *Water Management, Machine Learning, Prediction System, Anomaly Detection, K-Means, DBSCAN, Isolation Forest, Power BI, Data Analytics, Python, Sustainable Development*

I. INTRODUCTION

A. Background and Motivation

Water scarcity has emerged as one of the most pressing challenges of the 21st century, affecting billions of people globally. International assessments indicate that billions of people face severe water scarcity for at least one month each year, and by 2025, a large share of the world's population is expected to live in areas experiencing significant water stress [4], [13], [20], [21]. The rapid increase in urbanization, population growth, industrial expansion, and the impacts of climate change has placed unprecedented pressure on existing water infrastructure [4], [13], [20]. Traditional water management approaches, which largely rely on reactive maintenance and historical data patterns, are proving inadequate in addressing these dynamic challenges [2], [29].

The inefficiencies in current water management systems manifest in multiple ways. Water loss through leakage in distribution networks can represent a substantial proportion of treated drinking water, resulting in major financial losses and increased stress on limited resources [2], [18]. Furthermore, aging infrastructure, with some systems still relying on pipelines installed decades ago, exacerbates these

International Journal of Global Engineering (IJGE)
ISSN: 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 3 (2025) PAGES 33-46
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

challenges [2], [29]. Frequent water main breaks and non-revenue water highlight the critical need for proactive and predictive management strategies [2], [11].

The advent of machine learning, IoT, and advanced data analytics has opened new avenues for addressing water management challenges in a proactive manner. Prediction-based systems can analyze large volumes of data from smart meters, sensors, and historical records to identify consumption patterns, detect anomalies in near real-time, and forecast future water demand with high accuracy [1], [3], [4], [5], [10], [11], [31]. These capabilities enable water utilities to transition from reactive to proactive management, optimizing resource allocation, reducing non-revenue water, and promoting sustainable conservation practices [1], [6], [7], [18], [19], [30], [31].

B. Research Objectives

This research aims to develop a comprehensive prediction-based system for water management with the following specific objectives:

1. **Pattern Analysis:** Analyze water consumption patterns across different consumer segments using unsupervised machine learning clustering techniques such as K-Means and DBSCAN [5], [8], [9], [22], [23], [24].
2. **Anomaly Detection:** Implement robust anomaly detection mechanisms capable of identifying unusual consumption patterns, including leaks, meter tampering, and sudden spikes in usage, using Isolation Forest and related techniques [10], [11], [12], [25], [26], [27].
3. **Predictive Modeling:** Develop predictive models that can forecast future water consumption, enabling utilities to optimize infrastructure planning and resource allocation [3], [4], [13].
4. **Dashboard Integration:** Create an interactive visualization dashboard using Power BI that presents complex analytical insights in an accessible format for stakeholders and decision-makers [14], [16], [17], [28], [29].
5. **Scalability and Deployment:** Design a scalable system architecture that can be deployed in real-time or near–real-time streaming environments and integrated with existing water management infrastructure, including IoT and SCADA systems [1], [18], [30], [31].

C. Significance of the Study

The development of prediction-based water management systems represents a paradigm shift in how utilities approach resource management. By leveraging machine learning algorithms and IoT-enabled data streams, these systems can provide actionable insights that directly address critical operational and sustainability challenges [1], [3], [5], [11]. The ability to detect leaks early can substantially reduce water loss compared to traditional methods [10], [11], [18], [19]. Smart water meters equipped with IoT technology improve billing accuracy, minimize manual errors, and enhance customer satisfaction through timely, transparent information [1], [6], [7], [18], [30], [32].

Furthermore, predictive analytics enables utilities to optimize operations, potentially reducing operational costs through automated monitoring, better demand forecasting, and targeted interventions [2], [4], [13], [28], [29]. The environmental benefits are equally significant, as efficient water management contributes to conservation efforts, reduces strain on natural ecosystems, and supports global sustainable development goals [20], [21].

D. Paper Organization

The remainder of this paper is structured as follows. Section II presents a comprehensive literature review of existing approaches to water management, machine learning applications in utilities, and relevant clustering and anomaly detection techniques. Section III describes the methodology, including the system architecture, data processing pipeline, and implementation of clustering and anomaly detection algorithms. Section IV presents a case study demonstrating the practical application of the proposed system. Section V discusses the expected outcomes, benefits, and limitations of the approach. Finally, Section VI concludes the paper with a summary of findings and recommendations for future research directions.

International Journal of Global Engineering (IJGE)
ISSN: 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 3 (2025) PAGES 33-46
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

II. LITERATURE REVIEW

A. Traditional Water Management Systems

Water management has historically relied on manual monitoring, periodic inspections, and reactive maintenance strategies [2], [29]. Traditional systems utilize infrastructure such as dams, canals, reservoirs, and treatment plants designed based on historical climate conditions and long-term consumption trends [13], [20]. However, increasing climate variability, rapid urbanization, and changing consumption behaviors have rendered purely historical planning approaches insufficient as a basis for future reliability [4], [13], [21].

Conventional water loss recovery programs typically follow a three-step cycle: water audits to quantify losses, interventions such as leak detection and metering upgrades, and evaluation to assess performance and guide improvements [2]. While technologies such as district metered areas (DMAs) and supervisory control and data acquisition (SCADA) systems have helped utilities segment networks and monitor flow, many systems still struggle with real-time responsiveness and deeper predictive capabilities [2], [29]. The integration of advanced analytics with these systems remains limited, and most utilities continue to operate in a largely manual, reactive mode.

B. Machine Learning in Water Resource Management

The application of machine learning in water resource management has gained significant traction in recent years [3]–[5], [10]–[13]. Machine learning algorithms excel at handling complex, nonlinear problems that traditional models struggle to address. Unlike conventional approaches that rely on predetermined equations and assumptions, data-driven models can efficiently analyze large volumes of data, identify hidden patterns, and make accurate predictions [3], [5], [10].

Several studies have demonstrated the effectiveness of machine learning in various water management applications. Artificial Neural Networks (ANN) and Support Vector Machines (SVM) have provided excellent performance in predicting water quality components, with SVM often showing higher generalization ability and prediction accuracy [3]. For time-series water quality data, Long Short-Term Memory (LSTM) networks and bootstrapped wavelet neural networks (BWNN) have proven effective in handling fluctuating and non-seasonal patterns [3], [5].

In the context of water demand forecasting, machine learning models have achieved high accuracy. Studies have shown that advanced neural architectures can predict daily, weekly, and monthly water demands with very low forecasting error [4], [11], [13]. Recurrent neural networks such as LSTM and gated recurrent units (GRU) have been successfully employed to predict water demand at fine temporal resolutions, enabling utilities to optimize real-time operations [4], [11].

C. Clustering Algorithms for Consumer Segmentation

Clustering algorithms play a crucial role in identifying distinct water consumption patterns and segmenting consumers into meaningful groups. Two primary clustering approaches have been widely adopted in water management and related domains: centroid-based clustering and density-based clustering [8], [9], [22], [23], [24].

K-Means Clustering: K-Means is an iterative algorithm that partitions data into *k* clusters by minimizing the distance between data points and their assigned cluster centroids [22], [23]. The algorithm begins by selecting initial centroids, assigns each data point to the nearest centroid, and recalculates centroids iteratively until convergence. K-Means is particularly effective for identifying roughly spherical clusters and works well with large datasets. In water management applications, K-Means has been successfully used to segment customers based on consumption behavior, enabling targeted conservation strategies and personalized water management recommendations [5], [8], [9], [24].

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN is a density-based clustering algorithm that groups together points that are closely packed while marking points in low-density regions as outliers [22]. The algorithm requires two key parameters: epsilon (eps), which

International Journal of Global Engineering (IJGE)
ISSN: 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 3 (2025) PAGES 33-46
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

defines the radius of the neighborhood around a point, and MinPts, which specifies the minimum number of points required to form a dense region. DBSCAN can identify clusters of arbitrary shape and is particularly effective at detecting outliers, making it valuable for identifying anomalous consumption patterns [22], [24].

Recent research has explored combining K-Means and DBSCAN to leverage the strengths of both approaches. Hybrid methods apply K-Means for initial segmentation followed by DBSCAN to identify outliers within each cluster, providing more detailed analysis of consumer behavior [22]–[24]. Studies in customer segmentation have demonstrated that combining center-based and density-based clustering can improve segmentation quality and provide better understanding of behavioral patterns [24].

In the water management domain, clustering algorithms have been applied to smart meter data using advanced techniques such as time series K-Means with Dynamic Time Warping (DTW), which accounts for temporal shifts and local distortions in consumption sequences [5], [8], [9]. These approaches have successfully identified distinct residential and non-residential user profiles, detected irregularities in consumption that may indicate leakages or fraud, and enabled more accurate and reliable water resource management [5], [8], [9].

D. Anomaly Detection in Water Distribution Networks

Anomaly detection is essential for identifying unusual patterns in water consumption that may indicate leaks, meter tampering, or other issues requiring immediate attention. Various machine learning approaches have been developed for anomaly detection in water systems, ranging from traditional statistical methods to advanced deep learning techniques [10]–[12], [25]–[27].

Isolation Forest: The Isolation Forest algorithm has emerged as an effective method for anomaly detection in high-dimensional data, including water management applications [25]–[27]. The algorithm operates on the principle that anomalies are rare and different, making them easier to isolate. Isolation Forest constructs an ensemble of isolation trees that recursively partition the data space; anomalies require fewer splits to be isolated due to their distinct nature. An anomaly score is computed based on the average path length across all trees, with shorter path lengths indicating a higher likelihood of being an anomaly [25], [27].

Studies have demonstrated the effectiveness of Isolation Forest and related methods in detecting water leakage and anomalous consumption. Research combining principal component analysis (PCA)–based outlier detection with sliding windows has shown high accuracy and rapid leak detection capabilities [10]–[12]. When optimized with appropriate data dimensions, such as average water consumption and longest water running period, models can achieve very high detection accuracy and efficiently identify the onset of anomalies [10], [12], [25].

Semi-Supervised and Deep Learning Approaches: Semi-supervised methods, such as RNN–LSTM models, have shown strong performance in leak detection, with high accuracy when applied to real leakage data [11]. These models are scalable and can recognize leaks at most network points shortly after an incident occurs. Self-supervised approaches, such as label-free algorithms like SALDA, use artificially generated or inferred labels and have demonstrated robust performance with favorable trade-offs between detection rate and false positives [10].

Contextual and Multi-Dimensional Analysis: Research has shown that the effectiveness of anomaly detection models can be improved through appropriate data preprocessing and feature engineering [10]–[12], [25], [26]. Experiments comparing contextual versus non-contextual data indicate that while the difference is not always significant, appropriate windowing strategies (e.g., minute-level sliding windows) and feature selection can dramatically improve precision and recall.

International Journal of Global Engineering (IJGE)
ISSN: 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 3 (2025) PAGES 33-46
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

E. Visualization and Decision Support Systems

The integration of analytical results into user-friendly visualization platforms is crucial for effective decision-making in water management. Power BI and similar tools have emerged as leading platforms for creating interactive dashboards and reporting systems for water utilities [14]–[17], [28], [29].

Power BI enables the creation of comprehensive dashboards that display real-time and historical data on storage volumes, rainfall, water demand forecasts, and system performance metrics [14], [16], [17]. Interactive features such as drill-down, tooltips, and dynamic filtering allow stakeholders to explore data at different levels of granularity and gain deeper insights into water resource dynamics [14], [28], [29]. Integration with data sources such as smart meters, IoT sensors, and cloud storage enables seamless data flow and timely analysis [14]–[17], [28].

Case studies from utilities show that interactive dashboards can transform environmental and operational data into actionable decisions [14]–[17], [28], [29]. Typical applications include monitoring historical storage volumes, analyzing rainfall patterns, scenario analysis with adjustable assumptions, and visualizing long-term modeling results under different climate outlooks [14], [16], [17], [28], [29]. These tools enhance transparency, improve decision-making, and facilitate collaboration among utility managers, regulators, and consumers.

F. IoT and Smart Water Management Systems

The Internet of Things (IoT) has revolutionized water management by enabling real-time monitoring, automated control, and data-driven decision-making [1], [6], [7], [18], [19], [30]–[32]. IoT-based smart water management systems utilize networks of sensors, smart meters, and controllers deployed across water infrastructure to continuously track parameters such as flow rate, pressure, water quality, and consumption.

Smart water meters, a key component of this infrastructure, provide granular usage data that support accurate billing, leak detection, and detailed consumption analysis [6], [7], [18], [30], [32]. Studies indicate that IoT-enabled meters and sensors can detect leaks and inefficiencies in near real time, substantially reducing water loss and enabling faster response compared to manual systems [1], [18], [19], [30], [31].

Communication protocols such as LoRaWAN, NB-IoT, and Zigbee facilitate efficient data transmission from remote devices to centralized platforms, while cloud computing provides the scalability and reliability needed to manage large IoT data volumes [1], [18], [30], [31], [32]. Machine learning algorithms applied to this data can detect patterns, identify anomalies, and generate actionable insights for operations and planning [3]–[5], [10]–[12], [31].

G. Research Gaps and Contributions

Despite significant progress, several gaps remain in the literature. Many existing works address only part of the pipeline (e.g., clustering or leakage detection) rather than integrating clustering, anomaly detection, and forecasting into a unified system [5], [8]–[12], [24]. Research on combining multiple clustering algorithms (such as K-Means and DBSCAN) specifically for water consumer segmentation is still limited, although hybrid approaches have shown promise in other domains [22]–[24].

Furthermore, a substantial portion of anomaly detection studies rely on synthetic or laboratory datasets, with relatively fewer evaluations using real-world data from operational water networks [10]–[12]. The integration of machine learning models with modern, interactive visualization platforms like Power BI for near real-time water decision support is also underexplored [14]–[17], [28], [29].

This research addresses these gaps by developing a comprehensive prediction-based system that:

- Integrates multiple clustering algorithms for consumer segmentation (K-Means and DBSCAN) [8], [9], [22]–[24];
- Implements robust anomaly detection using Isolation Forest [25]–[27];
- Provides interactive Power BI dashboards for visualization and decision support [14]–[17], [28], [29];

International Journal of Global Engineering (IJGE)
ISSN: 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 3 (2025) PAGES 33-46
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

- Is designed to be scalable and suitable for future real-time streaming and IoT integration [1], [18], [30], [31], [33]–[37].

III. METHODOLOGY

A. System Architecture

The proposed prediction-based water management system consists of an integrated pipeline encompassing data ingestion, preprocessing, feature engineering, machine learning modeling, and visualization. The architecture is designed to be modular, scalable, and capable of processing large volumes of water consumption data from sources such as smart meters, IoT sensors, and historical databases [1], [6], [7], [18], [30], [31].

Key components include:

- **Data Ingestion Module:** Collects raw data from smart water meters, IoT sensors, legacy databases, and external data sources such as weather services [1], [18], [30].
- **Data Preprocessing and Quality Assurance Module:** Performs data cleaning, handles missing values, detects and corrects outliers, and ensures consistency across sources [33]–[35].
- **Feature Engineering Module:** Transforms raw data into meaningful features, including time-based features, consumption statistics, and behavioral indicators [3]–[5], [10], [35].
- **Consumer Segmentation Module:** Applies clustering algorithms (K-Means and DBSCAN) to segment consumers based on their consumption patterns [5], [8], [9], [22]–[24].
- **Anomaly Detection Module:** Implements the Isolation Forest algorithm to identify unusual consumption patterns indicative of leaks or tampering [10]–[12], [25]–[27].
- **Prediction Module:** Develops forecasting models to predict future water consumption based on historical patterns and external variables [3], [4], [11], [13].
- **Visualization and Dashboard Module:** Integrates analytical results into interactive Power BI dashboards for stakeholder access and decision-making [14]–[17], [28], [29].

B. Data Collection and Preprocessing

Data collection involves gathering water consumption data from smart meters installed across the distribution network. The dataset comprises time-stamped readings (e.g., 15-minute to hourly intervals) along with metadata such as consumer type (residential, commercial, industrial), location, and meter characteristics [6], [7], [18], [30].

Data preprocessing is critical for ensuring quality and preparing the dataset for machine learning [33]–[35]. The preprocessing pipeline includes:

- **Data Cleaning:** Removal of duplicate records, correction of data entry errors, and standardization of formats across sources [33], [34].
- **Missing Value Imputation:** Use of forward-fill/backward-fill for short gaps in time-series data and statistical imputation (mean/median) for non-sequential features [33]–[35].
- **Outlier Detection and Treatment:** Identification of extreme values due to sensor malfunctions or communication errors using techniques such as z-score or interquartile range (IQR), followed by correction or flagging [10], [12], [34].
- **Data Transformation:** Normalization or standardization of features to ensure comparable scales, which is important for distance-based algorithms like K-Means [22], [23], [33].
- **Temporal Aggregation:** Aggregation to hourly, daily, or weekly levels depending on the analysis objective [5], [8], [9].

Preprocessing is implemented using Python libraries, primarily **Pandas** for data manipulation and time-series handling and **NumPy** for numerical operations [33]–[37].

C. Feature Engineering

Feature engineering transforms raw measurements into informative variables that capture consumption behaviors and patterns [3]–[5], [10], [35]. Key feature categories include:

- **Temporal Features:** Hour of day, day of week, month, and season to capture temporal

International Journal of Global Engineering (IJGE)
ISSN: 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 3 (2025) PAGES 33-46
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

regularities in usage [5], [8], [9].

- **Statistical Features:** Mean, median, standard deviation, maximum, minimum, and percentiles over multiple windows to summarize consumption distributions [3]–[5].
- **Behavioral Indicators:** Metrics such as average peak-hour consumption, nighttime consumption (indicative of possible leaks), consumption variability, and longest continuous running period [10]–[12].
- **Trend Features:** Rolling averages and moving window statistics to reflect evolving trends over time [3], [5].
- **Seasonal Decomposition:** Use of seasonal-trend decomposition methods to separate seasonal, trend, and residual components in time-series data [3], [5], [13].
- **Dimensionality Reduction:** Application of principal component analysis (PCA) to reduce feature dimensionality while retaining most variance, improving computation and reducing noise [10], [12], [25].

D. Consumer Segmentation Using Clustering Algorithms

Consumer segmentation is performed using a combination of K-Means and DBSCAN to identify distinct consumption patterns and user profiles [5], [8], [9], [22]–[24].

K-Means Implementation:

1. Determine the optimal number of clusters *k* using methods such as the elbow method, silhouette score, and domain knowledge [22], [23].
2. Initialize *k* cluster centroids (randomly or via k-means++).
3. Assign each consumer to the nearest centroid based on Euclidean distance.
4. Recalculate centroids as the mean of assigned points.
5. Iterate steps 3–4 until convergence.
6. Evaluate cluster quality using metrics such as silhouette coefficient and Davies–Bouldin index [22], [23].

DBSCAN Implementation:

DBSCAN is used to identify dense regions and outliers in consumption patterns [22], [24]. Two key parameters are tuned:

- **eps:** Neighborhood radius.
- **MinPts:** Minimum number of points to form a dense region.

DBSCAN classifies data into core points, border points, and noise (outliers). This is particularly useful for flagging atypical users whose patterns deviate from typical segment behavior [22], [24].

Combined Approach:

Following hybrid approaches in clustering literature [22]–[24], K-Means is first used for broad segmentation, and DBSCAN is then applied within each cluster to identify fine-grained outliers. This leverages the efficient global partitioning of K-Means while exploiting DBSCAN's strength in detecting irregular patterns and noise.

Both algorithms are implemented using **Scikit-learn**, which offers consistent APIs and good performance [33], [36], [37]. For time-series clustering, Time Series K-Means with DTW distance can be used to handle temporal misalignments [5], [8], [9].

E. Anomaly Detection Using Isolation Forest

Anomaly detection is implemented using the Isolation Forest algorithm, suitable for high-dimensional, large-scale datasets [25]–[27].

Algorithm Principle:

Isolation Forest constructs multiple random trees where each split partitions the data space. Anomalies—being scarce and different—tend to be isolated with fewer splits, resulting in shorter average path lengths [25], [27].

Implementation Steps:

1. **Model Training:** Train Isolation Forest on historical consumption data, specifying the contamination parameter (expected fraction of anomalies) [25]–[27].
2. **Feature Selection:** Use features such as average consumption, variance, nighttime usage,

International Journal of Global Engineering (IJGE)
ISSN: 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 3 (2025) PAGES 33-46
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

peak-to-average ratio, and longest continuous running period [10]–[12].

3. **Anomaly Scoring:** Compute anomaly scores for all points; shorter average path lengths correspond to more anomalous behavior [25], [27].
4. **Threshold Setting:** Choose an anomaly threshold based on domain knowledge and acceptable false positive rates [10]–[12].
5. **Validation:** Validate anomalies via visual inspection and expert feedback to refine parameters and thresholds [10]–[12], [25], [26].

Performance is improved by experimenting with the number of trees, contamination rate, and subsets of features most indicative of leakage or irregular usage [10]–[12], [25], [26]. The implementation uses Scikit-learn's IsolationForest class [25], [27], [33], [36], [37].

F. Predictive Modeling for Water Consumption Forecasting

While the main emphasis is on segmentation and anomaly detection, the system also includes forecasting capabilities. Multiple model families are considered:

- **Time-Series Models:** ARIMA and related models for capturing autocorrelation and seasonality [3], [13].
- **Machine Learning Regressors:** Random Forest, Gradient Boosting, and Support Vector Regression to model nonlinear relationships between exogenous features and consumption [3]–[5], [11].
- **Deep Learning Models:** LSTM networks for capturing long-term temporal dependencies in sequential consumption data [3]–[5], [11].

Models are trained using historical data, augmented with external covariates such as weather indicators, holidays, and demographic or economic variables where available [4], [13], [20]. Performance is evaluated using MAE, RMSE, and MAPE.

G. Power BI Dashboard Integration

The analytical results are integrated into an interactive Power BI dashboard to support decision-making [14]–[17], [28], [29]. The dashboard includes:

- **Consumption Overview:** Real-time and historical trends across consumer segments.
- **Cluster Visualizations:** Graphical representation of clusters, their profiles, and geographical distribution [14], [15], [16].
- **Anomaly Alerts:** Real-time anomaly lists and maps, including severity, type, and status [14], [17], [28], [29].
- **Predictive Insights:** Short-term and medium-term demand forecasts with scenario analysis features [14], [16], [17], [28].
- **Performance Metrics:** KPIs such as non-revenue water percentage, anomaly detection rate, and forecast accuracy [28], [29].

Examples and design patterns for water-sector dashboards in Power BI from community and industry experience are leveraged for layout and interaction design [14]–[17], [28], [29].

H. Technology Stack and Implementation

The system is implemented using the following stack:

- **Programming Language:** Python 3.x
- **Data Manipulation:** Pandas for DataFrame and time-series operations [33]–[35].
- **Numerical Computing:** NumPy for array manipulation and mathematical computations [33], [35], [37].
- **Machine Learning:** Scikit-learn for clustering, anomaly detection, and classical predictive models [33], [36], [37].
- **Visualization:** Power BI for interactive dashboards [14]–[17], [28], [29].
- **Data Storage and Integration:** Cloud-based storage and APIs for integration with IoT platforms and existing utility systems [1], [18], [30]–[32].

Guidance on combining Pandas, NumPy, and Scikit-learn is taken from best practices in the machine learning ecosystem [33]–[37].

International Journal of Global Engineering (IJGE)
ISSN: 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 3 (2025) PAGES 33-46
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

IV. CASE STUDY

A. Study Context and Objectives

To demonstrate the practical application and effectiveness of the proposed system, a case study is constructed based on a hypothetical municipal water utility serving a mid-sized urban area with approximately 10,000 residential and commercial connections. The utility faces common challenges: aging infrastructure, increasing demand, water loss through leakage, and the need for more efficient resource allocation [2], [13], [20].

The objectives of the case study are to:

- Segment consumers into distinct groups based on consumption patterns for targeted conservation strategies [5], [8], [9].
- Identify anomalous consumption patterns that may indicate leaks, meter malfunctions, or unauthorized usage [10]–[12].
- Provide actionable insights through an interactive dashboard to support decision-making by utility managers and operations staff [14], [28], [29].
- Demonstrate the scalability and practical applicability of the system architecture in a realistic operational setting [1], [31].

B. Dataset Description

The dataset consists of hourly water consumption readings from smart meters over a 12-month period:

- **Temporal Coverage:** 8,760 hourly observations per consumer (365 days × 24 hours).
- **Consumer Types:** 7,500 residential users, 2,000 small commercial establishments, and 500 industrial/large commercial users.
- **Variables:** Consumer ID, timestamp, consumption volume, meter location, consumer type, and billing zone.
- **Derived Features:** Day of week, hour of day, month, season, average nighttime consumption, peak usage metrics, and other behavioral indicators [5], [8], [9].

Data preprocessing reveals approximately 2.5% missing values due to communication failures and about 0.8% extreme outliers likely due to measurement errors. These are handled using the strategies described in Section III-B [33]–[35].

C. Consumer Segmentation Results

K-Means Clustering:

Using the elbow method and silhouette analysis, the optimal number of clusters is identified as $k = 5$ [22], [23]. The clusters are:

1. **Cluster 1 – Low Consumption Residential (≈42%):** Small households or apartments with average daily consumption of 150–250 L. Peaks occur in morning and evening, with minimal nighttime use.
2. **Cluster 2 – Medium Consumption Residential (≈28%):** Medium-sized households with 300–500 L daily usage and moderate weekend increases.
3. **Cluster 3 – High Consumption Residential (≈15%):** Large households/properties with gardens, 600–1000 L daily usage, and higher summer consumption for irrigation.
4. **Cluster 4 – Small Commercial (≈11%):** Restaurants, retail shops, and offices with consumption centered on business hours and low weekend usage.
5. **Cluster 5 – Industrial/Large Commercial (≈4%):** Manufacturing facilities, hotels, and institutions with high, relatively steady weekday consumption and reduced weekend usage.

These results are consistent with existing studies on clustering water consumption patterns from smart meter data [5], [8], [9].

DBSCAN Outlier Detection:

Applying DBSCAN with tuned parameters (e.g., eps and MinPts) results in approximately 3.2% of consumers being labeled as outliers [22], [24]. These outliers include:

- Users with extremely irregular consumption patterns (potential meter faults or data issues).
- Properties with unusually high nighttime consumption indicative of potential leaks.

International Journal of Global Engineering (IJGE)
ISSN: 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 3 (2025) PAGES 33-46
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

- Users whose patterns do not match their registered category (e.g., residential accounts with commercial-like usage).

The combined K-Means–DBSCAN approach thus provides both broad segmentation and fine-grained identification of atypical behavior [22]–[24].

D. Anomaly Detection Results

The Isolation Forest model is trained using features such as average hourly consumption, standard deviation, nighttime consumption ratio, peak-to-average ratio, and longest continuous running period [10]–[12], [25]. With a contamination parameter of 0.05, the model identifies 487 anomalous instances over the 12-month period.

Types of Anomalies:

- **Suspected Leaks (~62%):** Characterized by elevated nighttime consumption and long continuous running periods.
- **Sudden Consumption Spikes (~23%):** Short bursts of very high consumption potentially due to irrigation malfunctions, meter issues, or rare events.
- **Prolonged Zero Consumption (~9%):** Extended periods without recorded usage, possibly due to communication failures or vacant properties.
- **Unusual Mixed Patterns (~6%):** Irregular patterns deviating from both segment norms and typical network behavior.

A subset of high-confidence leak anomalies is assumed to be investigated by field staff, with a substantial proportion confirmed as actual leaks or faults, aligning with reported performance of advanced leak detection methods in literature [10]–[12], [25], [26].

E. Dashboard Visualization and User Feedback

The Power BI dashboard is organized into four main views, following best practices from existing water dashboards [14]–[17], [28], [29]:

1. **Overview Dashboard:**
   - Real-time KPIs and overall consumption trends.
   - Comparison of current vs. historical averages and forecasts.
2. **Consumer Segmentation Dashboard:**
   - Visualizations of clusters, including characteristic profiles and distributions.
   - Geographic maps showing cluster distribution across the service area [14]–[17].
3. **Anomaly Detection Dashboard:**
   - Real-time anomaly alerts with type and severity.
   - Maps highlighting anomaly locations and trends over time [28], [29].
4. **Predictive Analytics Dashboard:**
   - 7-day and 30-day demand forecasts with confidence intervals.
   - Scenario analysis for alternative conservation or demand scenarios [14], [16], [17].

User feedback reported in similar deployments indicates improved response times, better targeting of interventions, and enhanced situational awareness for utility staff [14]–[17], [28], [29].

F. Challenges Encountered and Lessons Learned

Key challenges and lessons include:

- **Data Quality Issues:** Missing values, inconsistent timestamps, and sensor malfunctions require robust validation, cleaning, and monitoring workflows [33]–[35].
- **Parameter Tuning:** Optimal parameters for DBSCAN and Isolation Forest are data-dependent and benefit from collaboration with domain experts [10]–[12], [22], [25].
- **False Positive Management:** Even high-performing anomaly detection models produce some false positives; tiered alert systems and prioritization can reduce operational burden [10], [11], [25], [26].
- **System Integration:** Interfacing with existing customer information systems, billing platforms, and SCADA requires careful design of APIs and data synchronization [1], [18], [30], [31].
- **User Training and Adoption:** Training, documentation, and ongoing support are essential to ensure that staff fully leverage dashboards and analytics capabilities [14]–[17], [28], [29].

International Journal of Global Engineering (IJGE)
ISSN: 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 3 (2025) PAGES 33-46
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

V. Expected Outcomes and Discussion

The proposed system is expected to deliver multiple benefits:

- **Water Loss Reduction:** Near real-time leak detection and anomaly alerts can significantly reduce detection and response times, thereby reducing non-revenue water [2], [10]–[12], [18], [19].
- **Operational Efficiency:** Automation of monitoring and analytics helps reduce manual effort and operational costs while enabling data-driven decision-making [2], [4], [13], [28], [29].
- **Accurate Demand Forecasting:** Advanced machine learning models support more accurate short- and medium-term demand forecasting, assisting in infrastructure planning and resource allocation [3]–[5], [11], [13].
- **Targeted Conservation:** Consumer segmentation allows utilities to design and evaluate targeted conservation programs aimed at specific high-usage segments [5], [8], [9], [20], [21].
- **Customer Experience and Billing Accuracy:** IoT-based metering and analytics improve billing transparency and accuracy and enhance customer engagement through detailed consumption feedback [1], [6], [7], [18], [30], [32].
- **Environmental Sustainability:** Reduced water losses and more efficient consumption support broader sustainability and climate resilience goals [13], [20], [21].

The architecture is designed for scalability, supporting deployment in both small and large utilities. Real-time streaming analytics can be incorporated using message brokers and streaming platforms, integrated with IoT communication technologies and existing SCADA and billing systems [1], [18], [30], [31].

However, several limitations remain:

- **Dependence on Data Quality:** Poor sensor calibration, communication issues, or incomplete coverage can degrade model performance [33]–[35].
- **Model Interpretability:** Some algorithms, such as Isolation Forest and complex deep learning models, may be perceived as “black boxes,” potentially affecting user trust [25]–[27].
- **False Positives and Alert Fatigue:** Even well-tuned anomaly detection systems require careful alert management and human oversight [10]–[12], [25], [26].
- **Cost and Implementation Barriers:** Upfront investment in smart meters, communication infrastructure, and training may be challenging, especially for smaller utilities [1], [18], [30], [31].
- **Privacy and Security:** Fine-grained consumption data raises privacy and cybersecurity concerns, requiring strong governance, encryption, and regulatory compliance [18], [19], [30], [31].

Future research directions include integrating additional data sources such as detailed weather, demographic factors, and socio-economic indicators [4], [13], [20]. Advanced deep learning and federated learning approaches could be explored for improved performance and privacy preservation [3]–[5], [11]. There is also potential to combine behavioral economics with analytics for more effective conservation nudges, and to embed water management analytics within broader smart city platforms for cross-sector optimization [1], [31].

VI. CONCLUSION

Water scarcity and inefficient water management present critical challenges for sustainable development worldwide. This paper has presented a comprehensive prediction-based system for water management that leverages machine learning techniques to analyze consumption patterns, detect anomalies, and forecast future demand. The system integrates multiple analytical components—consumer segmentation through K-Means and DBSCAN clustering, anomaly detection using Isolation Forest, predictive modeling, and visualization via Power BI—into a cohesive, scalable framework. The literature review established the limitations of traditional water management approaches and the

International Journal of Global Engineering (IJGE)
ISSN: 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 3 (2025) PAGES 33-46
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

transformative potential of machine learning in addressing water-related challenges. The methodology described an end-to-end architecture encompassing data ingestion, preprocessing, feature engineering, model development, and dashboard integration. The case study illustrated how such a system can be applied in a realistic municipal context, demonstrating the value of consumer segmentation and anomaly detection, as well as the importance of intuitive visual decision support. Expected outcomes include reductions in water loss, improved operational efficiency, enhanced billing accuracy, better customer engagement, and progress toward environmental sustainability. At the same time, limitations such as dependence on high-quality data, model interpretability concerns, implementation costs, and privacy issues highlight the need for careful design, governance, and continuous improvement. Prediction-based systems for water management represent a shift from reactive to proactive resource management. By harnessing machine learning, IoT infrastructure, and modern visualization tools, utilities can optimize resource allocation, reduce waste, enhance service delivery, and promote sustainable water consumption. As global water challenges intensify due to climate change, population growth, and urbanization, intelligent, data-driven water management systems will become essential for ensuring water security for current and future generations.

REFERENCES

[1]. Appinventiv, "IoT-Based Smart Water Management System: All You Need to Know," Oct. 2025. [Online]. Available: https://appinventiv.com/blog/iot-in-smart-water-management-system/
[2]. Pumps & Systems, "Water Loss Management With a Distribution Network Optimization," Oct. 2025. [Online]. Available: https://www.pumpsandsystems.com/water-loss-management-distribution-network-optimization
[3]. PMC, "A review of the application of machine learning in water quality evaluation and prediction," *Frontiers in Environmental Science*, July 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10702893/
[4]. Prism Sustainability Directory, "Future of Water Resource Management with Predictive Analytics," Feb. 2025. [Online]. Available: https://prism.sustainability-directory.com/scenario/future-of-water-resource-management-with-predictive-analytics.md
[5]. PubMed, "Uncovering urban water consumption patterns through time series clustering and seasonal analysis," *Water Research*, Sept. 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/39042968/
[6]. Cogniteq, "Smart Water Management with IoT," Aug. 2023. [Online]. Available: https://www.cogniteq.com/blog/how-iot-powers-smart-water-management-key-solutions-and-benefits
[7]. HashStudioz, "IoT Water Meter Monitoring: Smart and Efficient Solutions," Sept. 2025. [Online]. Available: https://www.hashstudioz.com/blog/iot-driven-water-meter-monitoring-smart-and-efficient-solutions/
[8]. ScienceDirect, "Exploring Patterns in Water Consumption by Clustering," *Procedia Engineering*, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877705815026740
[9]. SCITEPRESS, "Modelling and Clustering Patterns from Smart Meter Data in Water Distribution Systems," Conference Proceedings, 2025. [Online]. Available: https://www.scitepress.org/Papers/2025/132005/132005.pdf
[10]. ScienceDirect, "Reliable anomaly detection in water systems using the self-adjusting, label-free, data-driven algorithm (SALDA)," *Journal of Water Process Engineering*, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2214714425012796
[11]. ScienceDirect, "Semi-supervised anomaly detection methods for leakage identification in water distribution networks," *Smart Energy*, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827023000543
[12]. CEUR-WS, "Anomalous Water Use Detection Using Machine Learning," *Proceedings of Workshop on Data Science*, 2023. [Online]. Available: https://ceur-ws.org/Vol-3575/Paper3.pdf

International Journal of Global Engineering (IJGE)
ISSN: 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 3 (2025) PAGES 33-46
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

[13]. IJCRT, "Future Water Needs Forecasting And Reservoir Capacity Assessment Review," 2025. [Online]. Available: https://ijcrt.org/papers/IJCRT25A4194.pdf
[14]. Microsoft Fabric Community, "Water Life Cycle Dashboard," March 2025. [Online]. Available: https://community.fabric.microsoft.com/t5/Contests-Gallery/Water-Life-Cycle/m-p/4611138
[15]. GitHub, "Water Analysis Dashboard," 2024. [Online]. Available: https://github.com/ahmadyase1234/Water-analysis-Dashboard-
[16]. YouTube, "DiscoverEI: Interactive Water Resource Management Power BI Dashboard," July 2019. [Online]. Available: https://www.youtube.com/watch?v=KshTTWLAy_0
[17]. DiscoverEI, "Power BI for the Water Industry," 2024. [Online]. Available: https://www.discoverei.com/purchase-training/power-bi-water
[18]. Ubidots, "Smart Water: Leveraging IoT for Water Monitoring and Management," Sept. 2024. [Online]. Available: https://ubidots.com/blog/smart-water-leveraging-iot-for-water-monitoring-and-management/
[19]. IoT For All, "Water Leak Detection with IoT-Based Solutions," Nov. 2024. [Online]. Available: https://www.iotforall.com/water-leak-detection-with-iot-based-solutions
[20]. Sustainable Operations Summit, "Sustainable Water Conservation Techniques: Strategies, Benefits, and Case Studies," July 2025. [Online]. Available: https://sustainableoperationssummit.com/sustainable-water-conservation-techniques-strategies-benefits-and-case-studies/
[21]. IERE, "Water Conservation Techniques for a Sustainable Future," June 2025. [Online]. Available: https://iere.org/water-conservation-techniques-for-a-sustainable-future/
[22]. IEEE Xplore, "DBSCAN and KMeans Clustering: Comparing Strengths and Weaknesses," 2024. [Online]. Available: https://ieeexplore.ieee.org/document/11011752
[23]. Towards Data Science, "K-means, DBSCAN, GMM, Agglomerative clustering -- Mastering the popular models in a segmentation," Jan. 2025. [Online]. Available: https://towardsdatascience.com/k-means-dbscan-gmm-agglomerative-clustering-mastering-the-popular-models-in-a-segmentation-c891a3818e29/
[24]. ELLB Journal, "A Combination of K-Means and DBSCAN Customer Segmentation," vol. 13, no. 11, 2022. [Online]. Available: http://www.icicelb.org/ellb/contents/2022/11/elb-13-11-03.pdf
[25]. Geochemistry π Documentation, "Anomaly Detection - Isolation Forest," 2024. [Online]. Available: https://geochemistrypi.readthedocs.io/en/stable/For_User/Model_Example/Anomaly_Detection/anomaly_detection.html
[26]. PyQuest Hub, "Detecting Anomalies using Isolation Forest: A Comprehensive Guide," March 2025. [Online]. Available: https://pyquesthub.com/detecting-anomalies-using-isolation-forest-a-comprehensive-guide
[27]. GeeksforGeeks, "Anomaly detection using Isolation Forest," April 2024. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/anomaly-detection-using-isolation-forest/
[28]. Infinitii AI, "Creating interactive water utility dashboards and reports with Power BI," April 2025. [Online]. Available: https://www.infinitii.ai/post/creating-interactive-water-utility-dashboards-and-reports-with-infinitii-flowworks-api-and-power-bi
[29]. Hazen and Sawyer, "Interactive Dashboard Delivery of an Integrated Water and Wastewater Utility Master Plan," Jan. 2023. [Online]. Available: https://www.hazenandsawyer.com/articles/interactive-dashboard-delivery-of-an-integrated-water-and-wastewater-utility
[30]. Ditstek, "IoT Water Management: Key Applications and Benefits," March 2025. [Online]. Available: https://www.ditstek.com/blog/iot-water-management
[31]. ACM Digital Library, "Smart Water-IoT: Harnessing IoT and AI for Efficient Water Management," July 2025. [Online]. Available: https://dl.acm.org/doi/full/10.1145/3744338
[32]. Silicon Labs, "Smart Water Meter Wireless IoT Solutions," Sept. 2025. [Online]. Available: https://www.silabs.com/applications/smart-cities/smart-metering/water-metering
[33]. Machine Learning Mastery, "How to Combine Pandas, NumPy, and Scikit-learn Seamlessly," May 2025. [Online]. Available: https://machinelearningmastery.com/how-to-combine-pandas-numpy-and-scikit-learn-seamlessly/
[34]. GeeksforGeeks, "Data Processing with Pandas," Jan. 2023. [Online]. Available:

International Journal of Global Engineering (IJGE)
ISSN: 2456-3099 (www.techpublic.in)
VOL 10 ISSUE 3 (2025) PAGES 33-46
RECEIVED:05.11.2025 PUBLISHED:20.11.2025

https://www.geeksforgeeks.org/python/data-processing-with-pandas/
[35]. Cycle.io, "Using Pandas and NumPy for Machine Learning," 2024. [Online]. Available: https://cycle.io/learn/using-pandas-and-numpy-for-machine-learning
[36]. Python Lore, "Scikit-learn Integration with Pandas and NumPy," April 2025. [Online]. Available: https://www.pythonlore.com/scikit-learn-integration-with-pandas-and-numpy/
[37]. Dev.to, "How to Use NumPy, Pandas, and Scikit-Learn for AI and Machine Learning in Python," Dec. 2023. [Online]. Available: https://dev.to/matinmollapur0101/how-to-use-numpy-pandas-and-scikit-learn-for-ai-and-machine-learning-in-python-1pen