**Water Consumption Analysis and Prediction Using Machine Learning**

**Introduction**

This report summarizes the analysis of water consumption data, focusing on predicting current charges, identifying patterns through clustering, and conducting time series analysis. The objective is to explore the interrelationships among features while employing various statistical and machine learning techniques.

**Abstract**

**Focus:**

Integrating IoT, machine learning, and cloud infrastructure for optimized water governance.

**Goals:**

Optimize water distribution.

Reduce waste and improve accountability.

Enable real-time monitoring and predictive analytics

**Dataset Overview**

The dataset comprises **50,315** entries with **25 columns**, including features related to water consumption, charges, and account details. Key attributes include:

* **Consumption (HCF)**: The volume of water consumed.
* **Current Charges**: The billing amount associated with consumption.
* **Estimated**: Indicates whether a charge is estimated.
* **Cluster**: Derived from K-Means clustering to segment data.

**High Level Architecture**

**A diagram of data processing

AI-generated content may be incorrect.**

**Gantt Chart**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Methodology**

1. **Data Cleaning**:
   * Addressed missing values and formatted date fields.
   * Introduced new features such as Consumption\_per\_day to enhance analysis.
2. **Exploratory Data Analysis (EDA)**:
   * Visualized data distributions and relationships.
   * A screenshot of a computer screen

     AI-generated content may be incorrect.Generated a correlation matrix to understand feature interdependencies.

**Parameters/Features**

**A screenshot of a computer code

AI-generated content may be incorrect.**

**Handle Missing Values**

**A screenshot of a computer

AI-generated content may be incorrect.**

**After replacing missing and NaN values**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Visualizations**

**A graph of water consumption

AI-generated content may be incorrect.**

**Histogram of Water Consumption**

**A graph of different colored rectangular shapes

AI-generated content may be incorrect.**

**Monthly Consumption Trends**

**A screenshot of a graph

AI-generated content may be incorrect.**

**Correlation Matrix of Numerical features**

1. **Modeling Approaches**:

* Linear Regression
* Random Forest
* Gradient Boosting
* Support Vector Machine

**Linear Regression**

**A graph with blue dots and red lines

AI-generated content may be incorrect.**

Model Overview:  
Predicting Current Charges using Consumption (HCF).  
Fitted LinearRegression() model.

Evaluation Metrics:

R-squared: 0.735

Mean Squared Error (MSE): 10322.455

Mean Absolute Error (MAE): 64.767

Insights:  
Linear Regression struggles to capture data complexities, showing the lowest R² and highest errors. It may be useful as a baseline model but lacks the predictive power of other approaches.

**Random Forest**

**A graph with blue dots and red line

AI-generated content may be incorrect.**

Model Overview:  
Predicting Current Charges using Consumption (HCF).  
Fitted RandomForestRegressor() model.

Evaluation Metrics:

* R-squared: 0.982
* Mean Squared Error (MSE): 708.090
* Mean Absolute Error (MAE): 5.069

Insights:  
The Random Forest model provides strong predictive accuracy with a high R² score, indicating a good fit. Minimal error suggests robust generalization.

**Gradient Boosting**

**A graph of a graph with blue dots

AI-generated content may be incorrect.**

Model Overview:  
Predicting Current Charges using Consumption (HCF).  
Fitted GradientBoostingRegressor() model.

Evaluation Metrics:

* R-squared: 0.983
* Mean Squared Error (MSE): 653.342
* Mean Absolute Error (MAE): 6.523

Insights:  
Gradient Boosting achieves the best performance among tested models, slightly outperforming Random Forest. It captures complex patterns well but may require tuning to reduce slight MAE increase.

**Support Vector Machine**

**A blue line graph with red line

AI-generated content may be incorrect.**

Model Overview:  
Predicting Current Charges using Consumption (HCF).  
Fitted SVR() model.

Evaluation Metrics:

* R-squared: 0.905
* Mean Squared Error (MSE): 3687.852
* Mean Absolute Error (MAE): 24.480

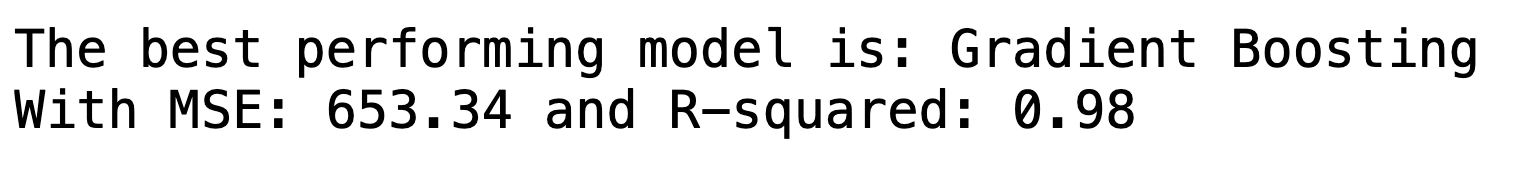
Insights:  
SVM exhibits lower predictive power compared to ensemble models, with a notable increase in error. It may benefit from kernel tuning and hyperparameter optimization.

**R-squared, MSE, and MAE  
 for Models**

**A white text with black text

AI-generated content may be incorrect.**

**Best Model**

****

**Recommendations**

1. **Model Enhancement**: Consider using more advanced regression techniques (e.g., Ridge, Lasso, and ensemble methods) to improve predictions, especially for high-charge scenarios.
2. **Feature Engineering**: Introduce additional features, such as demographic information or weather data, to enrich the dataset and improve model performance.
3. **Imbalance Handling**: Address class imbalance in the target variable through techniques like SMOTE to enhance decision tree performance.
4. **Further Analysis**: Conduct additional clustering analyses or segmentation strategies to gain insights into different consumer behaviors and preferences.

**Conclusion**

**Summary:**

* Successfully built predictive models including **Linear Regression, Random Forest, Gradient Boosting, and Support Vector Machines (SVM)**.
* Applied **ensemble learning** for improved accuracy and **evaluated model performance** using multiple metrics.

**Impact:**

* Provides insights into **consumption patterns**, **charge prediction**, and potential **cost optimizations**.
* Helps in **identifying anomalies** and **forecasting future usage trends**.
* Can support **smart city initiatives** for **efficient water resource management and governance**.