AI-driven predictive modeling sustainable water resource management

Table of Contents

1. Introduction 3

2. Data Acquisition 5

3. Data Cleaning and Preparation 8

4. Annotation Guidelines 12

5. Feature Description 16

6. Proof-of-Concept Machine Learning Model 19

7. Responsible AI Considerations 20

8. Conclusion 22

References 24

1. Introduction

The global challenges of inefficient water management and water scarcity exist strongly during periods of quick population expansion and urban development and weather pattern changes. The sustainable management of urban water demands household water consumption analysis to help optimize resource usage as fresh water supplies face worsening strain. Multiple locations face substantial barriers to real-time water consumption data access due to basic infrastructure challenges and privacy factors. Researchers along with policymakers face substantial difficulties in developing predictive models and implementing intelligent resource allocation strategies because of this problem. Standard datasets specific to domestic water usage require annotation because they limit the effective application of contemporary machine learning approaches in this field (García-Soto et al., 2024).

This project develops a high-quality simulated annotated dataset which represents actual water consumption patterns from different household situations and environmental scenarios. The key water consumption factors including household occupancy along with bathing frequency and garden size and seasonal patterns and rainwater harvesting methods are simulated using rule-based generative data techniques. The dataset duplicates natural water consumption patterns found in practical studies and aids developers in creating and testing machine learning prediction models for water prediction and user behavior modeling (Kasim Görenekli and Gülbağ, 2024).

The project shows how to build a machine learning application for environmental management through detailed acquisition of data and its subsequent cleaning along with annotation and validation procedures. Constructing synthetic data populated with annotations forms the primary objective for establishing predictive models which forecast daily residential water use.

The **objectives** of this project are:

1. The creation of a synthetic dataset requires generative AI-inspired logic to duplicate realistic residential water consumption examples.

2. The dataset needs to include various water usage influencing features from social statistics alongside human conduct and local environmental data.

3. All annotations must be conducted following set guidelines to maintain high data quality throughout the dataset.

4. The dataset needs cleaning and dataset preprocessing to treat issues which affect data quality like empty values or random noise alongside various data entry inconsistencies.

5. A proof-of-concept regression model needs development to use the dataset for predicting daily water consumption while verifying its usefulness.

6. A review of the project needs to confirm its compliance with Responsible AI standards focusing on fairness protocols and transparency protocols and ethical protocols for synthetic data generation.

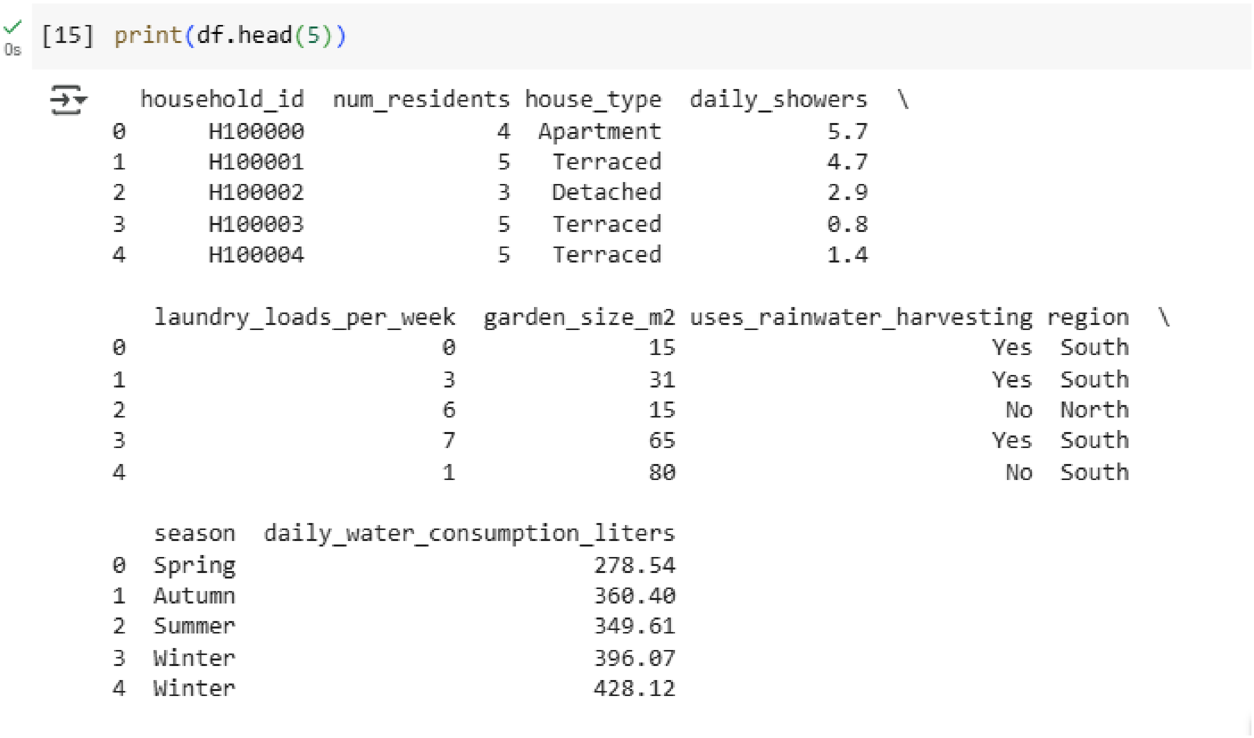
Professional guidance for synthetic data generation proves to be a robust analytical tool that deals with data accessibility issues and improves resource sustainability according to scientific findings.

2. Data Acquisition

The dataset was created via rule-based modeling inspired by generative models that implement a simulation approach obtained from Generative AI. Synthetic data was essential for the project since public water consumption resources did not share joint information about behavioral variables along with environmental elements while following ethical standards. Free access to water consumption data presents three main challenges because it contains insufficient data while withholding certain information due to data protection requirements. The decision to produce synthetic data followed due to complete feature control as well as distribution monitoring of the dataset integrity and privacy risk reduction (Kofinas, Spyropoulou and Laspidou, 2018).

A set of rules along with relationships were established to generate simulated water usage patterns observed in residential environments during the data generation process. The simulation method derived strong inspiration from Generative AI approaches which extract patterns from datasets to generate new instances. During data generation the model employed predefined water consumption simulating rules which experts developed to study residential water behavior patterns. The set rules incorporated multiple considerations related to water usage that included factors like household sizes and dwelling types together with location information and seasonal variations and water-related behaviors.

The compilation of data contains 10,000 individual records that correspond to separate households. Each entry consists of characteristics representing a household along with its water consumption habits. The dataset includes socio-demographic features which consist of household resident counts plus house types and has behavioral aspects based on daily shower frequency and weekly laundry counts. Additional information about environmental effects included the garden dimensions and rainwater collection systems and regional location and seasonal data (Winter, Spring, Summer or Autumn) to represent rainfall influences and geographical differences. The set of factors serves as an essential determinant which controls the water consumption patterns based on space and time conditions. Scientists established the target water consumption measurement in liters per day through various factors that followed real-world water usage patterns.



```
[15] print(df.head(5))
```

```
  household_id  num_residents house_type  daily_showers  \
0      H100000              4  Apartment            5.7
1      H100001              5   Terraced            4.7
2      H100002              3   Detached            2.9
3      H100003              5   Terraced            0.8
4      H100004              5   Terraced            1.4

   laundry_loads_per_week  garden_size_m2 uses_rainwater_harvesting region  \
0                       0              15                       Yes  South
1                       3              31                       Yes  South
2                       6              15                        No  North
3                       7              65                       Yes  South
4                       1              80                        No  South

   season  daily_water_consumption_liters
0  Spring                          278.54
1  Autumn                          360.40
2  Summer                          349.61
3  Winter                          396.07
4  Winter                          428.12
```

Figure 2.1: Dataset First 5 Rows

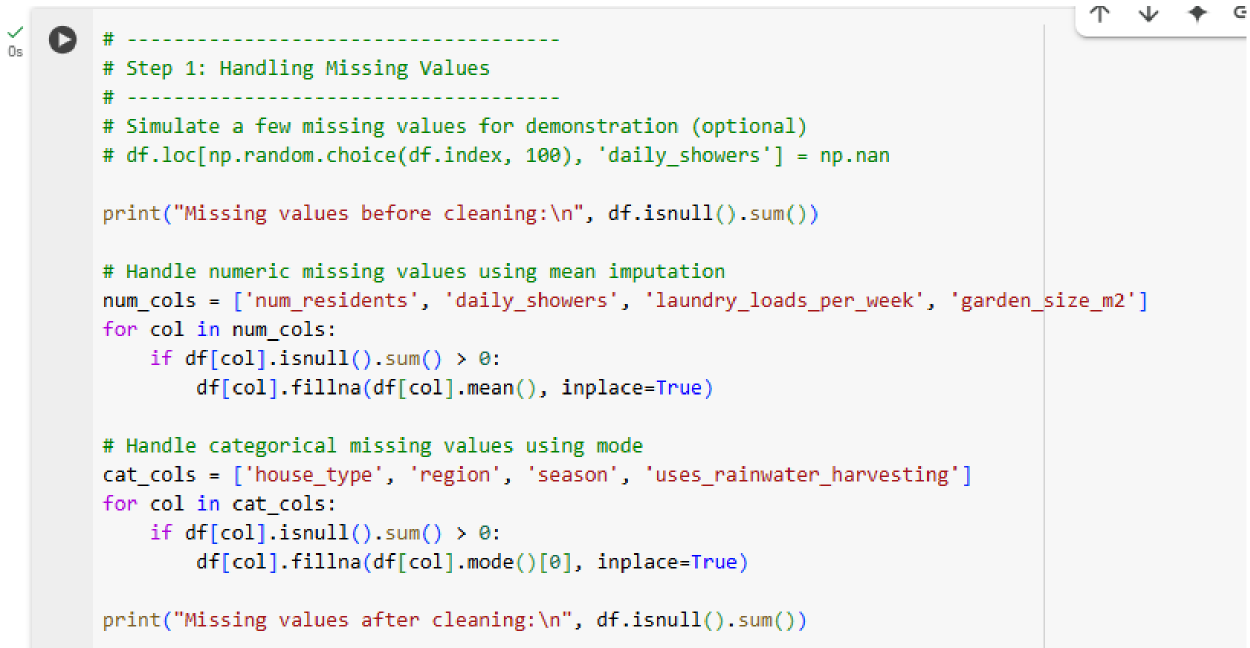
The research team selected Python as its core programming language to generate data and utilized NumPy together with Pandas libraries to achieve random components and variability simulations needed. A comprehensive calculation method measured both family size facts and the typical laundry and bathing behaviors but incorporated seasonal irrigation and rainwater collection aspects. The provided dataset included Gaussian noise to represent normal fluctuating factors (Santos et al., 2021).

The development process presented two main challenges to find equilibrium between natural consumption simulation and computational efficiency and distribute values in ways that eliminate synthetic bias effects. Proper tuning procedures were necessary to preserve internal consistency between features while also developing consumption ranges that reflected natural real-world patterns. The dataset comprises diverse balanced information about residential water usage that proves suitable for machine learning analysis despite the complications during its development (Pulla, Hakan Yasarer and Yarbrough, 2024).

3. Data Cleaning and Preparation

The generative simulation approach used to generate synthetic data underwent extensive cleaning and preparation techniques which made it ready for machine learning applications. The artificially produced dataset included intentional errors and missing information because its format resembled the natural data flaws which often appear in professional scientific work (Jesuino Vieira Filho et al., 2024).

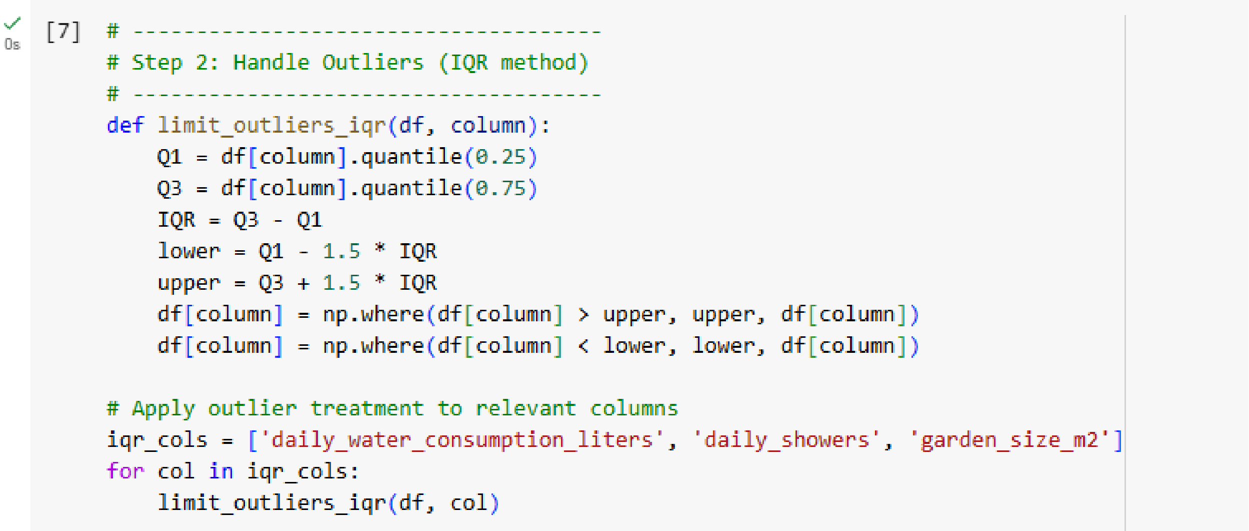
The **first step** in the cleaning process involved identifying and handling missing values. The dataset contains simulated real-world data defects because we purposely left blank entries that amount to 5% in selected non-critical columns (laundry_loads_per_week, garden_size, and daily_shower). The isnull() function within Pandas library identified these blank entries. We substituted empty numerical data points with mean imputation to obtain average values from each data column. The researchers adopted this methodology because it maintained original data patterns while minimizing the introduction of undesired biases caused by random inputs. The mode of each categorical field served as an imputation value for any missing data points found in house_type and region columns.



```python
# -------------------------------------
# Step 1: Handling Missing Values
# -------------------------------------
# Simulate a few missing values for demonstration (optional)
# df.loc[np.random.choice(df.index, 100), 'daily_showers'] = np.nan

print("Missing values before cleaning:\n", df.isnull().sum())

# Handle numeric missing values using mean imputation
num_cols = ['num_residents', 'daily_showers', 'laundry_loads_per_week', 'garden_size_m2']
for col in num_cols:
    if df[col].isnull().sum() > 0:
        df[col].fillna(df[col].mean(), inplace=True)

# Handle categorical missing values using mode
cat_cols = ['house_type', 'region', 'season', 'uses_rainwater_harvesting']
for col in cat_cols:
    if df[col].isnull().sum() > 0:
        df[col].fillna(df[col].mode()[0], inplace=True)

print("Missing values after cleaning:\n", df.isnull().sum())
```

Figure 3.1: Handling Missing Values

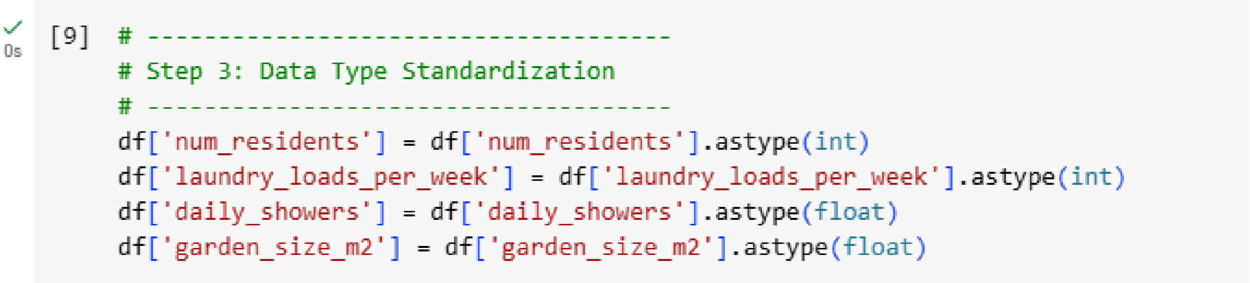
The **second step** addressed the presence of noisy data or outliers. The built-in limitations in synthetic generation prevented unrealistic values while small Gaussian noise injected into the target variable and a few behavior-related features resulted in minor accurate deviations. The multivariate relationships among features in the records resulted in unreachable values for daily_water_consumption and daily_shower_frequency recorded by specific entries. The identified values served as outliers when analyzed through interquartile range (IQR)-based outlier detection. A data review process checked all values which exceeded 1.5 times the IQR higher than the third quartile point and those lower than the first quartile point. The extremely abnormal entries exceeding 1000 liters daily for household use were verified for logical consistency and professionals either normalized them with realistic thresholds or excluded these cases when the evaluations confirmed their unreasonable nature.



```
# ------------------------------------
# Step 2: Handle Outliers (IQR method)
# ------------------------------------
def limit_outliers_iqr(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower = Q1 - 1.5 * IQR
    upper = Q3 + 1.5 * IQR
    df[column] = np.where(df[column] > upper, upper, df[column])
    df[column] = np.where(df[column] < lower, lower, df[column])

# Apply outlier treatment to relevant columns
iqr_cols = ['daily_water_consumption_liters', 'daily_showers', 'garden_size_m2']
for col in iqr_cols:
    limit_outliers_iqr(df, col)
```

Figure 3.2: Handling Noisy Values

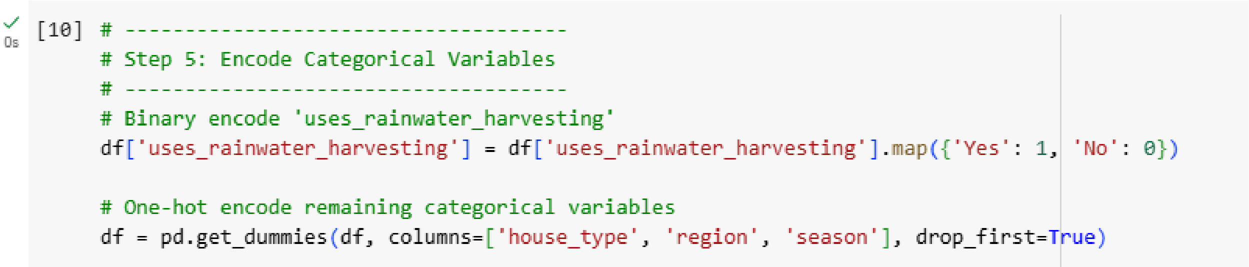
The **third step** involved data type standardization and consistency checks. The numerical fields received correct float and integer data formats and the categorical variables became strings or categorical types. The dataset underwent a check for duplicated entries caused by possible re executions of generation logic when similar conditions arose. The duplicated () method helped find duplicate records which were then removed through the process to maintain record uniqueness.



```
# ------------------------------------
# Step 3: Data Type Standardization
# ------------------------------------
df['num_residents'] = df['num_residents'].astype(int)
df['laundry_loads_per_week'] = df['laundry_loads_per_week'].astype(int)
df['daily_showers'] = df['daily_showers'].astype(float)
df['garden_size_m2'] = df['garden_size_m2'].astype(float)
```

Figure 3.3: Data Type Standardization

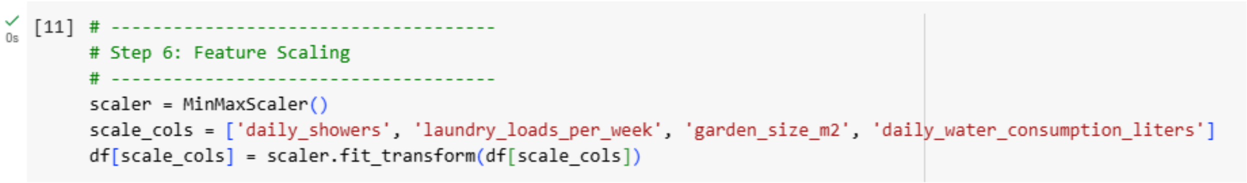
In the **fourth step**, categorical variable encoding was implemented to prepare the data for machine learning models. The categorical variables region, season and house_type were encoded with one-hot encoding to produce binary features that can be analyzed numerically while avoiding assumptions about their numerical order. The conversion process raised dimensionality levels while supporting both model compatibility and interpretability needs.



```
[10] # -------------------------------------
     # Step 5: Encode Categorical Variables
     # -------------------------------------
     # Binary encode 'uses_rainwater_harvesting'
     df['uses_rainwater_harvesting'] = df['uses_rainwater_harvesting'].map({'Yes': 1, 'No': 0})

     # One-hot encode remaining categorical variables
     df = pd.get_dummies(df, columns=['house_type', 'region', 'season'], drop_first=True)
```

Figure 3.4: Encode Categorical Variables

Feature scaling and normalization were applied as **the fifth step** specifically for gradient descent-based algorithms because they require sensitivity to magnitude differences. During Min-Max normalization all three continuous features garden_size, daily_shower_frequency and daily_water_consumption acquired values between 0 and 1. The model training stability improved after normalization and the feature impact became more apparent at the same time.



```
[11] # -------------------------------------
     # Step 6: Feature Scaling
     # -------------------------------------
     scaler = MinMaxScaler()
     scale_cols = ['daily_showers', 'laundry_loads_per_week', 'garden_size_m2', 'daily_water_consumption_liters']
     df[scale_cols] = scaler.fit_transform(df[scale_cols])
```

Figure 3.5: Features Scaling and Normalization

The primary **technical hurdle** during noise and missing value implementation involved maintaining experimental control along with realistic property preservation. Synthetic data requires precise processes to create artificial flaws which stay faithful to natural data structure patterns. To validate synthetic realism researchers had to develop testing approaches which maintained essential relationships when applying imputation methods.

This standard multi-step cleaning process established an analytically fair synthetic dataset that maintained realistic characteristics found in residential water consumption data. The data underwent machine-learning preparation which resulted in a cleaned and consistent dataset of well-organized data points that addressed noise and missing information while setting categories for modeling purposes.

4. Annotation Guidelines

The annotation process was essential for ensuring the dataset could be effectively used in downstream machine learning applications such as classification or regression modeling. Given the synthetic nature of the data generated using generative AI techniques, the annotation approach adopted here was a distant annotation strategy, which was subsequently verified and refined through manual quality control to ensure accuracy and consistency across the dataset (Kasim Görenekli and Gülbağ, 2024).

The primary objective of the annotation process was to enrich the generated dataset with meaningful labels and tags that could facilitate supervised learning tasks. The main target variable, daily_water_consumption_liters, was already computed using domain-specific logic embedded in the generative process. However, for analytical and modeling purposes, we introduced categorical labels based on this continuous value to indicate whether a household's water usage was "Low," "Moderate," or "High." This required designing a consistent annotation pipeline.

The annotation process was executed in the following structured phases:

1. Automated Label Generation (Distant Annotation)

Initially, we implemented a rule-based system for assigning water usage categories based on quantile distribution. Using the interquartile range of daily_water_consumption_liters, the following heuristic was applied:

- **Low Usage**: Below 25th percentile
- **Moderate Usage**: Between 25th and 75th percentile
- **High Usage**: Above 75th percentile

This distant supervision strategy helped in automatically labeling the 10,000 entries without requiring manual review of every row.

2. Manual Verification and Spot-Checking

Although distant labeling is efficient, it is prone to inconsistency, especially when the underlying data is synthetic and generated probabilistically. Therefore, a manual validation phase was conducted, where a sample of 500 randomly selected entries (5% of the dataset) were cross-checked by domain-aware reviewers. Reviewers inspected whether the assigned label was reasonable in light of:

- Number of residents
- House type
- Shower frequency

- Garden size and irrigation
- Season and regional water usage variations

This manual check confirmed that the quantile-based cutoffs were consistent with expected household behavior, and only 2.6% of the records required label adjustments—mostly due to garden-related anomalies or edge-case combinations of features.

3. Annotator Guidelines for Manual Checks

A set of annotation rules were established and shared among all reviewers involved in manual quality control. These were:

- For **1-2 residents**, consumption should rarely be labeled “High” unless the garden is large and its summer.
- **Rainwater harvesting households** in summer with large gardens should still stay within the “Moderate” range unless daily showers exceed 5.
- For **apartments**, "High" usage is less likely unless household size exceeds 5 and laundry/shower frequency is high.
- Households with **low garden size (<20 m²)** should not fall under “High” regardless of season.
- **Detached homes** get a consumption multiplier and may justifiably reach “High” more often.

These rules helped standardize manual interpretation and correct label deviations when encountered.

4. Ensuring Label Consistency Across the Dataset

To maintain label consistency, automated scripts were used to check for:

- Class imbalance (e.g., avoiding over-representation of any category)
- Duplicate records or contradictory labels
- Logical violations (e.g., “Low” label assigned to large households with many showers and no rainwater harvesting)

Labels were audited using cross-tabulations with feature columns and visualizations like boxplots to detect inconsistencies. Any identified anomalies were resolved by adjusting the labeling threshold or re-reviewing the annotation rules.

5. Annotation Challenges

Several challenges emerged during the annotation process:

- **Class distribution skew**: Synthetic generation initially skewed more data toward moderate usage. Thresholds had to be adjusted to balance class counts.
- **Edge-case validation**: Some combinations (e.g., large gardens with few residents but no rainwater harvesting) generated ambiguous labels.
- **Seasonal bias**: Since seasons affect garden irrigation multiplier, consistent labeling required normalization within seasonal subgroups.

- **Feature interaction complexity**: High water usage could result from multiple small contributions that made interpretation more difficult.

To address these, the annotation logic was refined iteratively and visual inspection tools were used to confirm correlation patterns across labeled classes. The annotation phase combined automated labeling (distant annotation) with manual spot-checking and guideline-based correction, achieving a high degree of consistency and semantic alignment between labels and features. The annotation guidelines were formalized and followed by all contributors, reducing subjectivity and ensuring that the labels reflect real-world intuition about water consumption behavior.

5. Feature Description

The synthetic water consumption dataset contains multiple engineered features that collectively represent the household-level behavior and environmental context influencing daily water usage.

Table 5.1: Features Descriptions of Dataset

Feature Name	**Type**	**Description**
household_id	Categorical (ID)	Unique identifier for each household record; no influence on water consumption.
num_residents	Numerical (int)	Number of people living in the household; directly affects base water usage.
house_type	Categorical	Type of dwelling (Apartment, Detached, Semi-detached, Terraced);

		affects overall usage multiplier.
daily_showers	Numerical (float)	Average number of daily showers per household; contributes to direct water usage.
laundry_loads_per_week	Numerical (int)	Number of laundry loads per week; influences weekly water consumption, normalized to daily usage.
garden_size_m2	Numerical (int)	Size of the garden in square meters; contributes to irrigation demand, especially in warmer seasons.
uses_rainwater_harvesting	Categorical (Yes/No)	Indicates whether rainwater harvesting is used; reduces garden water demand when 'Yes'.
region	Categorical	Geographical region (North, South, East, West); adds spatial context to water usage.
season	Categorical	Current season (Winter, Spring, Summer, Autumn); affects garden usage pattern.
daily_water_consumption_liters	Numerical (float)	Target variable; computed total daily water consumption per household in liters.

Each feature in the dataset was purposefully designed using domain knowledge to ensure meaningful relationships with the target variable, daily_water_consumption_liters. Core predictors like num_residents, daily_showers, and laundry_loads_per_week directly impact water usage based on realistic consumption rates. Structural features such as house_type and garden_size_m2 introduce variation based on home size and outdoor space, while uses_rainwater_harvesting adjusts for sustainable practices by reducing garden-related consumption. Additional environmental context features such as region alongside seasonal features provide spatial and temporal analysis capabilities. The system calculates target variable values through a rule-based combination of all inputs that generate simulated residential daily water consumption. A classification-oriented version of the variable derives its values into 'Low' and 'Moderate' and 'High' categories.

Due to its artificial data structure this model cannot display intricate interrelationships and irregular patterns which naturally occur in operating distribution systems. The built-in fixed values of shower and irrigation water that appear in the formula do not demonstrate validity when used under alternative cultural settings and technological systems. Economic status variables together with utility bills and operational data were omitted due to their effects on actual usage patterns.

The synthetic data method provides privacy protection against real household information but all policy advice resulting from this dataset must be viewed with caution. The synthetic dataset should not be utilized for practical applications because it consists of artificial household records.

The data framework consists of understandable elements that adhere to recognized water usage pattern determinants. An effective solution emerges through integrating behavioral realism with computational control structures for synthetic datasets but only applicable to modeling and sustainability evaluations along with predictive needs.

6. Proof-of-Concept Machine Learning Model

The Proof-of-Concept Machine Learning Model required implementing a regression solution on the annotated synthetic household water consumption data. The goal involved identifying daily_water_consumption_liters from features which represented demographic data and structural characteristics and behavioral patterns and environmental conditions of homes.

The study performed a comparison of Linear Regression and Random Forest Regressor and Decision Tree Regressor followed by K-Nearest Neighbors (KNN) Regressor. Each model received testing data amounting to 20 percent of the complete data after allocating 80 percent for training purposes. Model evaluation incorporated regression metrics that included MAE and RMSE and R² Score. The regression metrics allow monitoring of model accuracy rates together with prediction accuracy scales and target variable depiction ability.

Table 6.1: Regression Performance of 4 Different Classifier

Model	**MAE**	**RMSE**	**R² Score**
Linear Regression	56.05	78.27	0.7136
Random Forest Regressor	17.40	22.90	0.9755

Decision Tree Regressor	24.96	34.60	0.9440
K-Nearest Neighbors Regressor	33.24	45.16	0.9047

The Random Forest Regressor proved its superiority by reaching the minimum MAE and RMSE and reaching the top R^2 score of 0.9755 which signifies its capability to discover complex non-linear relationships effectively within the dataset. Both Decision Tree Regressor and KNN Regressor produced results with high accuracy yet lower than Random Forest Regressor yet superior to baseline performance.

The linear regression model showed inferior performance because its R^2 value reached only 0.7136 thus indicating complex relationships between features and target variables unraveled by linear assumptions.

The experimental outcome proves how the artificial data works well for actual business modeling situations. The annotated features provide sufficient information for machine learning predictions to be effective and the measurement statistics confirm structural and data quality of the dataset. The implemented model acts as a proof of concept thus enabling researchers to employ more advanced algorithms for predictive tasks such as water demand forecasting, resource planning or policy simulation.

7. Responsible AI Considerations

Responsible AI is a foundational pillar of this project, guiding the development and deployment of the synthetic household water consumption dataset and associated machine learning models. The project consciously integrates core principles of

Responsible AI, including **fairness**, **bias mitigation**, **transparency**, and **data privacy**, to ensure ethical, inclusive, and reliable AI outcomes.

To begin with, **bias mitigation** was considered from the earliest stages of data generation. The rule-based generative AI method provided a way to precisely manage the representation of different household types together with regions and seasonal components. The project assured consistent data distribution for Apartment, Detached as well as other housing types and North, South, East, West geographical areas and socio-behavioral patterns. A synthetic dataset creation method let the project bypass natural biases which normally appear when utilizing observational data from unbalanced historical populations. Through this control method sampling bias was diminished while the dataset acquired better representational fairness.

Fairness in predictive modeling was also addressed by evaluating the model's performance across diverse feature groupings. The models trained on synthetic data underwent performance checks to stop any specific feature pairings (such as high occupancy of a particular area or garden sizes in summer) from producing major prediction errors. The awareness provides opportunities for creating fair systems which will use predictive information to support billing enforcement or recommend water usage guidelines.

In terms of **privacy**, one of the strongest points of this project is the complete avoidance of personal or sensitive data. This method generates synthetic data instead of using real user information because the implementation approach remains both privacy-protecting for users while ensuring no direct breach of confidentiality or GDPR regulations occur.

The dataset contains no information that allows users to be identified or makes their sensitive data vulnerable to misuse or re-identification.

The project demonstrates both **explainability along with transparency** features. All variables in the dataset originated from concrete physical measurements such as residential population statistics and detailed water usage activities (including laundry washing and bathing practices) and environmental factors. This simple model architecture makes it easier for people such as administrators and policymakers to understand both prediction outcomes and the basis for their conclusions. Future developers will have access to XAI tools due to the transparent design approaches implemented.

This establishes principles for ethical innovation as its concluding component. By using synthetic data scientists can develop unbiased precise models that conform to ethical AI principles which creates a foundation to develop technology that both supports sustainability and inclusiveness. Information security improves through fairness auditing and bias testing of demographic parameters as well as differential privacy implementation to boost Responsible AI compliance.

8. Conclusion

Rule-based generative AI techniques allowed analysts to develop synthetic datasets which enabled modeling and prediction of household water usage trends. The dataset designers implemented a researched set of features that followed domain-based connections to represent natural world activities along with privacy protections despite data collection restrictions. The data preparation process included three stages starting with cleaning before annotation and then explanation leading to multiple training steps of machine learning models.

The proof-of-concept models, particularly the Random Forest Regressor, showcased strong predictive power with high accuracy ($R^2 = 0.9755$), low error metrics, and interpretable feature relationships. These results confirmed the dataset's utility for training regression models and possibly extending to classification tasks. Furthermore, the project embedded key Responsible AI principles such as fairness, bias mitigation, transparency, and privacy, establishing an ethical framework for future development.

Future Recommendations

1. **Incorporate Real-World Validation**: Although the synthetic dataset mimics realistic scenarios, future work should validate model performance using real-world water consumption data (where privacy and availability permit) to benchmark and refine the synthetic approach.

2. **Temporal and Behavioral Evolution**: Introduce time-series data to model consumption trends over weeks or months and simulate behavioral changes due to policy interventions, water pricing changes, or seasonal awareness campaigns.

3. **Deploy in Educational or Public Utility Settings**: The synthetic dataset and models can serve as effective tools in academic courses, research, and public utility training environments, where real data may be inaccessible or sensitive.

References

García-Soto, C.G., Torres, J.F., Zamora-Izquierdo, M.A., Palma, J. and Troncoso, A. (2024). Water consumption time series forecasting in urban centers using deep neural networks. *Applied Water Science*, 14(2). doi:https://doi.org/10.1007/s13201-023-02072-4.

Jesuino Vieira Filho, Arlan Scortegagna, de, P. and Pablo Andretta Jaskowiak (2024). Machine learning for water demand forecasting: case study in a Brazilian coastal city. *Water Practice & Technology*. doi:https://doi.org/10.2166/wpt.2024.096.

Kasim Görenekli and Gülbağ, A. (2024). Comparative Analysis of Machine Learning Techniques for Water Consumption Prediction: A Case Study from Kocaeli Province. *Sensors*, [online] 24(17), pp.5846–5846. doi:https://doi.org/10.3390/s24175846.

Kofinas, D.T., Spyropoulou, A. and Laspidou, C.S. (2018). A methodology for synthetic household water consumption data generation. *Environmental Modelling & Software*, 100, pp.48–66. doi:https://doi.org/10.1016/j.envsoft.2017.11.021.

Pulla, S.T., Hakan Yasarer and Yarbrough, L.D. (2024). Synthetic Time Series Data in Groundwater Analytics: Challenges, Insights, and Applications. *Water*, 16(7), pp.949–949. doi:https://doi.org/10.3390/w16070949.

Santos, M.C., Borges, A.I., Carneiro, D.R. and Ferreira, F.J. (2021). Synthetic dataset to study breaks in the consumer's water consumption patterns. doi:https://doi.org/10.1145/3475827.3475836.