International Journal of Applied Science; Vol. 8, No. 2; 2025
ISSN 2576-7240 E-ISSN 2576-7259
https://doi.org/10.30560/ijas.v8n2p22

An Enhanced Framework for Urban Water Consumption Analysis: Feature Clustering with Ensemble Methods

Faye F.F. Jiang[1]

[1] Scholar, Hong Kong

Correspondence: Faye F.F. Jiang, Scholar, Hong Kong. E-mail: jiangfeifengoffice@163.com

Received: March 20, 2025; Accepted: April 5, 2025; Published: April 6, 2025

Abstract

Urban water consumption analysis presents significant challenges due to the complex interplay of socioeconomic, demographic, and built environment factors. This paper introduces a novel Feature Clustering Framework of TopK and Threshold with Ensemble Method (FCTTE) specifically designed to address high-dimensional urban datasets. We evaluate this framework using a comprehensive dataset of 1,120 features across eight domains related to New York City's urban environment. Our experiments demonstrate that FCTTE significantly outperforms conventional feature selection methods, improving LightGBM classification accuracy by 4.6% compared to baseline, while traditional methods achieved only 1% improvement. The framework identified median family income, energy usage intensity, adult male population, greenhouse gas emissions, and commercial building characteristics as the most influential factors affecting water consumption. By effectively managing feature redundancy through hierarchical clustering and strategic selection, FCTTE provides urban planners with interpretable insights for water resource management while maintaining superior predictive performance. This integrated approach bridges the gap between fragmented analyses of individual urban factors and the need for holistic understanding of water consumption patterns in complex urban environments.

Keywords: urban water, feature clustering, LightGBM, machine learning, urban planning

1. Introduction

Water resource management has become increasingly critical as urbanization accelerates worldwide [1]. Urban areas, with their concentrated populations and diverse water usage patterns, present unique challenges and opportunities for water conservation [2]. Decision-makers in urban planning and environmental policy require comprehensive data-driven approaches to effectively manage these precious resources [3–5].

1.2 Literature Review

Previous research on urban water consumption has primarily relied on limited datasets with relatively few features. For instance, [6] used demographic variables alone to predict residential water consumption in Kurdistan, achieving moderate accuracy but failing to capture the complexity of urban water usage patterns. Similarly, [7] examined the relationship between economic factors and water consumption, employing linear regression models on datasets with fewer than 20 features.

These studies, while valuable, have typically employed linear modeling approaches due to their computational simplicity and the limited number of features available. [8] achieved an high R^2 using multiple linear regression with different socioeconomic variables, while [9] incorporated climate variables to achieve marginally results. The linear nature of these models inherently limits their ability to capture complex, non-linear relationships among the multitude of factors influencing urban water consumption.

Furthermore, most existing studies have focused on specific aspects of urban environments—such as residential patterns [10], economic indicators [11], or building characteristics [12]—rather than integrating these diverse factors into a unified analytical framework. This fragmented approach has resulted in models that, while providing reasonable performance metrics, fail to offer comprehensive insights that urban planners need for holistic water resource management [13–15].

1.3 Research Gap and Objectives

The limitations of existing approaches present several important research gaps:

1. Most studies utilize datasets with limited features, failing to capture the multidimensional nature of urban

water consumption
2. Linear modeling approaches predominate, potentially missing complex non-linear relationships
3. Fragmented analysis of different urban factors prevents comprehensive understanding
4. High-dimensional data processing challenges have not been adequately addressed in this domain

To address these gaps, this paper proposes an integrated approach using machine learning methods on a comprehensive, high-dimensional dataset encompassing eight major categories of urban factors: energy, population, education, social factors, economic indicators, housing characteristics, geographic information systems (GIS) data, and additional relevant parameters [16]. Our dataset thoroughly covers various factors potentially influencing water resource consumption, resulting in over one thousand features for analysis.

The primary contributions of this paper are:

1. Development of a novel feature selection framework—Feature Clustering Framework of TopK and Threshold with Ensemble Method (FCTTE)—specifically designed to handle high-dimensional urban datasets
2. Implementation and evaluation of this framework using the LightGBM algorithm, demonstrating superior performance compared to conventional feature selection methods
3. Identification of the most influential factors affecting urban water consumption through comprehensive analysis
4. Provision of data-driven insights to guide urban planning and water conservation policy

By leveraging advanced machine learning techniques on this extensive dataset, we aim to provide urban decision-makers with more comprehensive insights into water consumption patterns, facilitating better resource planning and conservation measures.

2. Methodology

2.1 Problem Statement and Overview

When analyzing urban water consumption patterns, researchers face challenges with high-dimensional datasets containing numerous potentially relevant features across diverse domains (demographic, economic, geographic, etc.). Effective feature selection becomes critical for both computational efficiency and model performance. This section introduces our novel Feature Clustering Framework of TopK and Threshold with Ensemble Method (FCTTE), which addresses these challenges through a systematic approach to feature selection and classification.

Traditional feature selection methods face limitations when applied to high-dimensional urban datasets. For instance, correlation-based methods only examine relationships between individual features and the target variable, potentially discarding features with complex interactions [17]. While Principal Component Analysis (PCA) preserves information while reducing dimensionality, it transforms features into components that lack direct interpretability—a critical consideration for urban planners and policymakers who need actionable insights.

Our FCTTE framework overcomes these limitations through hierarchical clustering of features followed by strategic selection and ensemble learning. Figure 1 presents the complete framework:

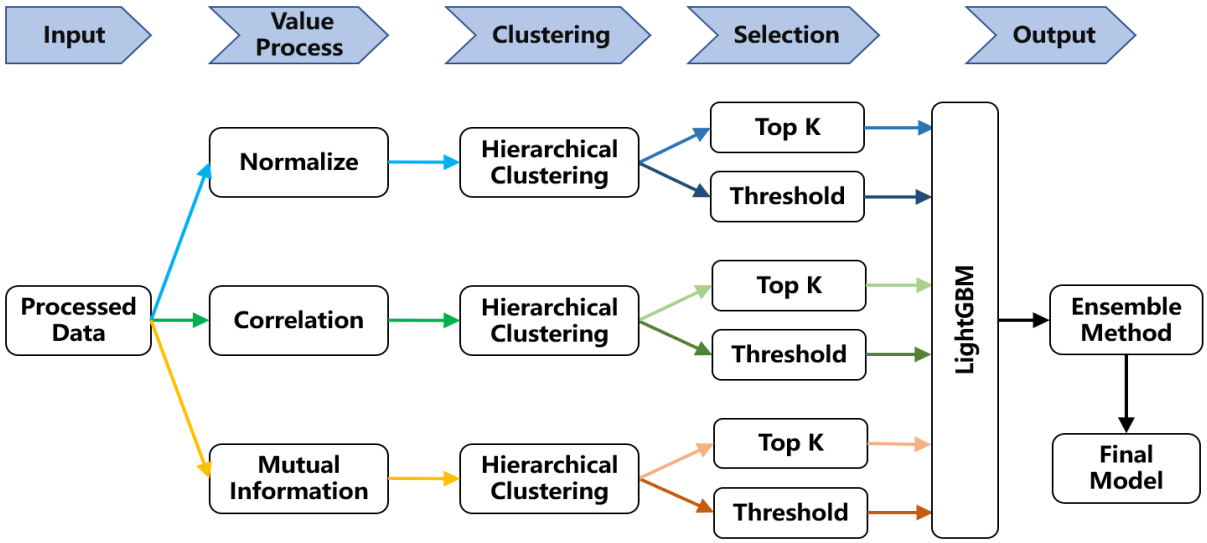

Figure 1. FCTTE Framework

2.2 FCTTE Framework Components

The FCTTE framework consists of four principal components:

2.2.1 Input Processing and Value Transformation

The framework begins with a preprocessed dataset containing all potential features [18,19]. These features undergo transformation using one of three methods:

- Normalization: Features are standardized without altering their fundamental relationships, providing a baseline approach that maintains the original dataset structure.
- Correlation Distance Matrix: Features are transformed into a distance matrix based on correlation coefficients, shifting the subsequent clustering to identify relationships based on linear feature associations [20].
- Mutual Information Distance Matrix: Features are transformed into a distance matrix based on mutual information, which captures both linear and non-linear relationships between features.

Each transformation method offers distinct advantages: Normalization preserves original relationships, Correlation focuses on linear dependencies, and Mutual Information captures more complex interactions. Their computational efficiency and representative capabilities make them ideal for our framework.

2.2.2 Feature Clustering

After transformation, we apply Hierarchical Clustering (HC) to organize features into meaningful groups. HC was selected over alternative clustering approaches for several key advantages:

- It establishes hierarchical relationships between features, facilitating the identification of feature importance structures
- It does not require a predefined number of clusters, allowing for flexible adjustment based on results
- It enables intuitive visualization of feature relationships through dendrograms

The clustering process follows these steps:

1. Initially treat each feature as an individual cluster
2. Calculate the distance/similarity between clusters using Ward's Method
3. Merge the two most similar clusters
4. Repeat until a single cluster remains

Ward's Method minimizes the increase in the sum of squared differences when merging clusters. It tends to create more balanced clusters by preferring to merge clusters with fewer samples when centroids are equidistant [21,22].

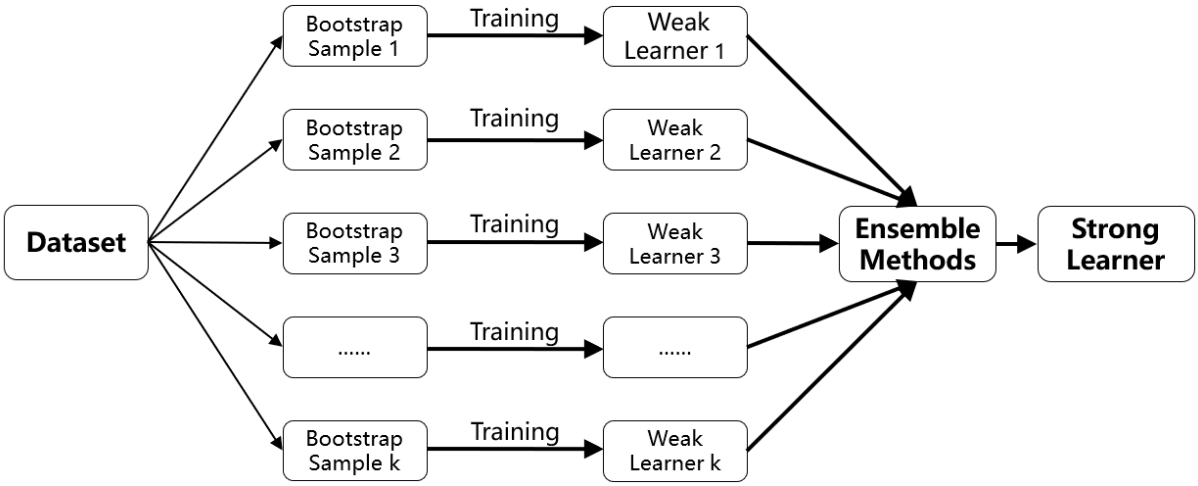

Figure 2. Ensemble Method Flow

2.2.3 Feature Selection Strategies

After clustering, we apply two complementary selection strategies to identify the most informative features [23,24]:

- Top K: Select the K most important features from each cluster
- Threshold: Select features whose importance exceeds a specified percentile threshold within each cluster

These strategies balance between cluster representation (ensuring features from each identified group are included) and feature importance (prioritizing the most predictive features). By implementing both approaches across different data transformations, we generate six distinct feature subsets.

2.2.4 Ensemble Learning

The final component of our framework employs ensemble learning to combine multiple models into a single, more robust classifier. For our urban water consumption classification problem, we use a majority voting strategy across the six models derived from our different feature selection approaches.

The ensemble approach, pioneered by [25] leverages the strengths of each individual model while mitigating their respective weaknesses. In our implementation, we generate six different LightGBM models—one for each feature subset—and combine their predictions through majority voting, as illustrated in Figure 2:

2.3 LightGBM Algorithm

For our modeling component, we selected the LightGBM algorithm for its exceptional performance with high-dimensional data. Developed by Microsoft [26], LightGBM incorporates several innovations that make it particularly well-suited for our urban water consumption analysis:

- Gradient-based One-Side Sampling (GOSS): Focuses computational resources on the most informative instances by retaining high-gradient samples and randomly sampling low-gradient ones, significantly reducing computation time without sacrificing accuracy.
- Exclusive Feature Bundling (EFB): Combines mutually exclusive features (those that rarely take non-zero values simultaneously) into single features, effectively reducing dimensionality without losing information.
- Leaf-Wise Growth Strategy: Unlike traditional level-wise tree growth that expands all nodes at the same level, LightGBM's leaf-wise approach always chooses the leaf with maximum delta loss to grow, resulting in more efficient models with the same number of splits.

Additional advantages of LightGBM include reduced memory consumption, native support for categorical features, and distributed computing capabilities, making it particularly well-suited for analyzing complex urban systems with high-dimensional data.

This comprehensive methodological framework provides a robust approach to identifying the most significant factors influencing urban water consumption patterns, while maintaining interpretability that is crucial for urban planning and policy applications.

3. Case Study

This section details our application of the FCTTE framework to analyze urban water consumption patterns in New York City. We describe the diverse data sources incorporated, data integration methodologies, and preprocessing techniques employed to prepare for modeling.

3.1 Data Sources

Our study leverages eight distinct datasets from New York City, collectively providing a comprehensive view of the urban environment. Table 1 summarizes these datasets, which together comprise 1,120 features across multiple domains.

Table 1. The collected 1120 features across multiple domains

Datasets Description					
Water Consumption		Social		POI	
Features Category	Counts	Features Category	Counts	Features Category	Counts
Area	5	Households by Type	16	Subway	3
Energy	20	Relationship	7	Theater	3
Position	2	Marital Status	12	School	3
Features with Dummy Variables	44	Fertility	7	Financial	3
Other	6	Grandparents	9	Hospital	3
Total	77	School Enrollment	6	Police	3
Housing		Education Attainment	8	HEC	3
Features Category	Counts	Veteran Status	2	Art	3
Housing Occupancy	5	Disability Status	8	News	3

Units in Structure	10	Residence One Year Ago	8	Museum	3
Year Construction Built	11	Place of Birth	7	WIFI	3
Rooms	11	Citizenship Status	3	Bus	3
Bedrooms	7	Year of Entry	7	Total	36
Housing Tenure	5	Birth of Foreign Born	7	PLUTO	
Year Householder Moved into Unit	7	Language Spoken at Home	12	Features Category	Counts
Vehicles Available	5	Ancestry	28	Area	10
House Heating Fuel	10	Computers and Internet Use	3	Location	10
Selected Characteristics	4	Total	150	Facilities	7
Occupants per Room	4	Economic		Residential Units Information	11
Owner Occupied Units	10	Features Category	Counts	Land Information	6
Mortgage Status	3	Employment Status	16	Maximum Allowable Area Ratio	4
Selected Monthly Owner Costs	33	Commuting to Work	8	Features with Dummy Variables	344
Gross Rent	18	Occupation	6	Other	10
Total	143	Industry	12	Total	402
Education		Class of Worker	5	Demographic	
Features Category	Counts	Income and Benefits	44	Features Category	Counts
Education Population by Sex & Age	75	Health Insurance Coverage	24	Sex and Age	29
Education Population by Race	72	Total	115	Citizen Voting Age Population	3
Median Earnings by Education	18			Total	32
Total	165				
ALL Features: 1120					

3.1.1 Socioeconomic and Demographic Data

Five primary datasets (Economic, Demographic, Education, Housing, and Social) were obtained from the Department of City Planning (DCP) in New York City. These datasets capture fundamental aspects of urban life that potentially influence water consumption patterns:

- Economic factors: Employment status, income levels, occupation types, industry sectors, health insurance coverage, and commuting patterns (115 features)
- Demographic characteristics: Population by age, sex, and citizenship status (32 features)
- Educational attributes: Educational attainment, school enrollment, and median earnings by education level (165 features)
- Housing parameters: Housing occupancy, unit structure, construction year, housing tenure, heating fuel types, and mortgage status (143 features)
- Social indicators: Household composition, relationship status, fertility rates, disability status, language preferences, and ancestry information (150 features)

These datasets provide essential context for understanding how socioeconomic and demographic characteristics might influence water consumption behaviors at the community level.

3.1.2 Built Environment and Geographic Information

Three additional datasets capture the physical urban landscape:

- POI (Points of Interest): Obtained from the Department of Information Technology & Telecommunications (DOITT), this dataset maps the spatial distribution of urban amenities including transportation hubs, educational facilities, commercial establishments, cultural venues, and public services (36 features)
- PLUTO (Primary Land Use Tax Lot Output): Also from DCP, this dataset provides detailed land use and geographic information at the tax lot level, recording building characteristics, land values, zoning information, and other property attributes (402 features) [27]
- Water Consumption: From the Mayor's Office of Sustainability (MOS), this dataset contains water usage metrics along with related energy consumption data including electricity usage, natural gas consumption, and carbon dioxide emissions (77 features)

The integration of these diverse datasets enables us to investigate relationships between water consumption and a wide spectrum of urban factors that might not be immediately apparent in more narrowly focused studies.

3.2 Combining Datasets

A significant methodological challenge was the integration of datasets collected at different spatial units. This section details our approach to data integration.

3.2.1 Spatial Unit Reconciliation

Two primary spatial units were present across our datasets:

- BBL (Borough, Block, and Lot): A unique identifier assigned by DCP to each property in New York City
- Census Tract: Geographic units defined by the U.S. Census Bureau for population enumeration

Our integration followed a three-step process [28], as illustrated in Figure 3:

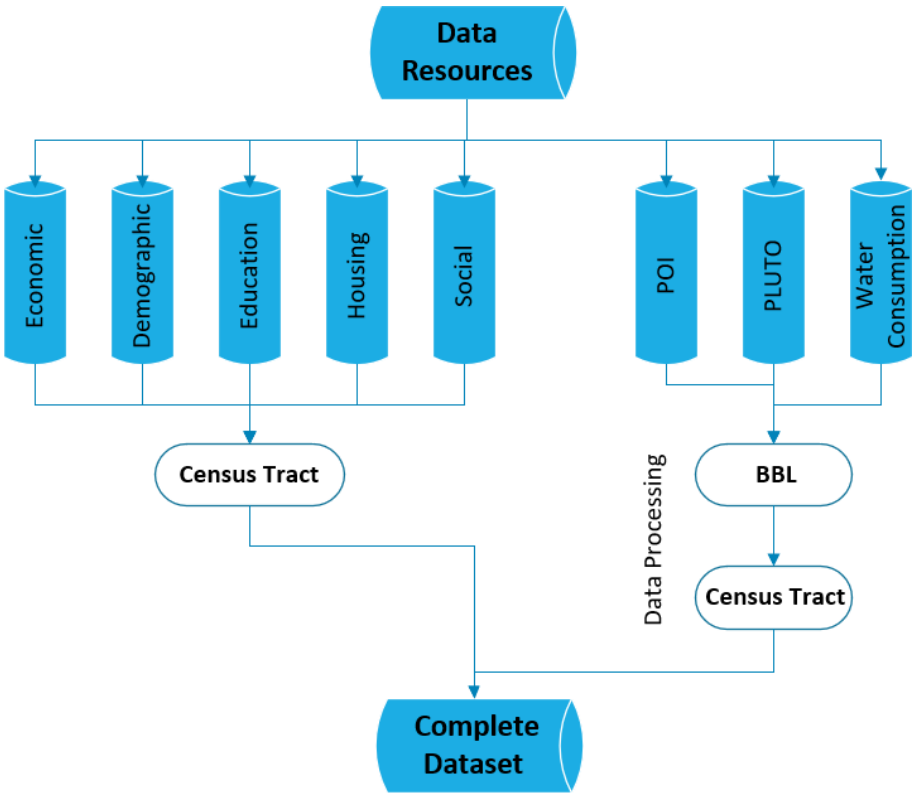

Figure 3. Data Integration Process

1. BBL-Level Integration: POI, PLUTO, and Water Consumption datasets shared BBL as their common identifier. Using Water Consumption as the base, we performed a left join with the other BBL-level datasets.
2. Census Tract Integration: Economic, Demographic, Education, Housing, and Social datasets used Census Tract as their identifier. Using Demographic as the base, we performed a left join with the other Census Tract-level datasets.
3. BBL to Census Tract Aggregation: To create a unified dataset, we converted BBL-level data to Census Tract level through aggregation:
    a. Numerical features were aggregated using median values within each Census Tract
    b. Categorical features were aggregated using mode values within each Census Tract

This aggregation process inevitably resulted in some information loss, as detailed property-level variations were consolidated into tract-level summaries. This limitation is acknowledged as a constraint of our study, necessitated by the unavailability of BBL-level socioeconomic data. However, the Census Tract represents a meaningful unit for urban analysis, balancing granularity with practical data availability constraints.

3.3 Data Preprocessing

After combining our diverse datasets, we implemented a systematic preprocessing pipeline to prepare our data for modeling [29,30]. Our target variable, Water Intensity (WI), required particular attention due to its distribution characteristics and the transformation of our analytical approach from regression to classification.

Water intensity (WI) refers to the amount of water consumed per unit area, typically measured in gallons per square foot (gal/ft²) or similar units. In the context of urban buildings and neighborhoods, water intensity provides a normalized measure of water usage that allows for meaningful comparisons across buildings or areas of different sizes.

3.3.1 Spatial Distribution Analysis

We first examined the spatial distribution of water intensity across New York City census tracts (Figure 4). The GIS map reveals distinct usage patterns, with high water consumption areas concentrated primarily in northern Manhattan and portions of the Bronx, while other high-usage zones appear scattered throughout the city. Gray areas represent locations with missing data. This spatial visualization provided initial insights into potential neighborhood-level factors influencing water consumption.

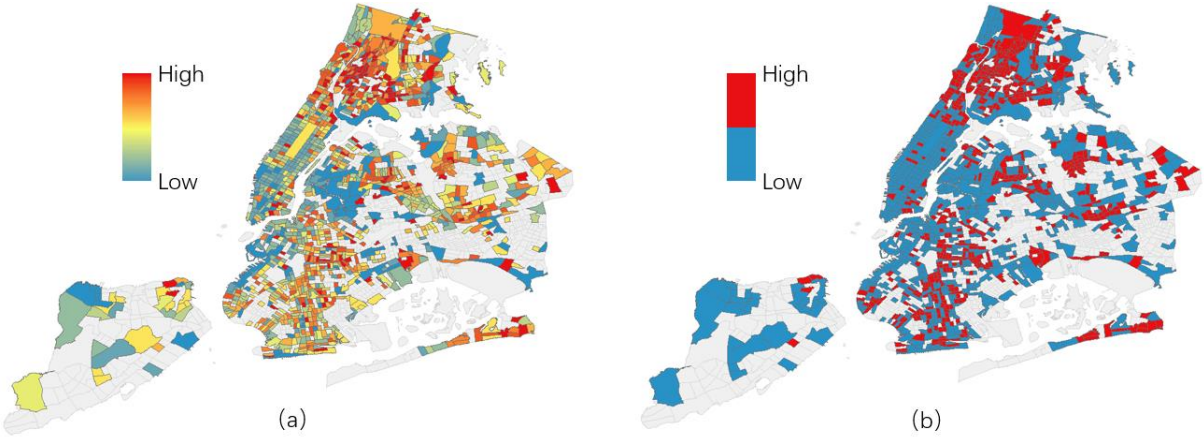

Figure 4. NYC water intensity GIS map

3.3.2 Target Variable Processing

Given our aggregation to the Census Tract level and the complexity of the factors involved, we converted the water intensity prediction from a regression to a binary classification problem. This approach allows us to identify areas with disproportionately high water usage and the features that most significantly contribute to this classification.

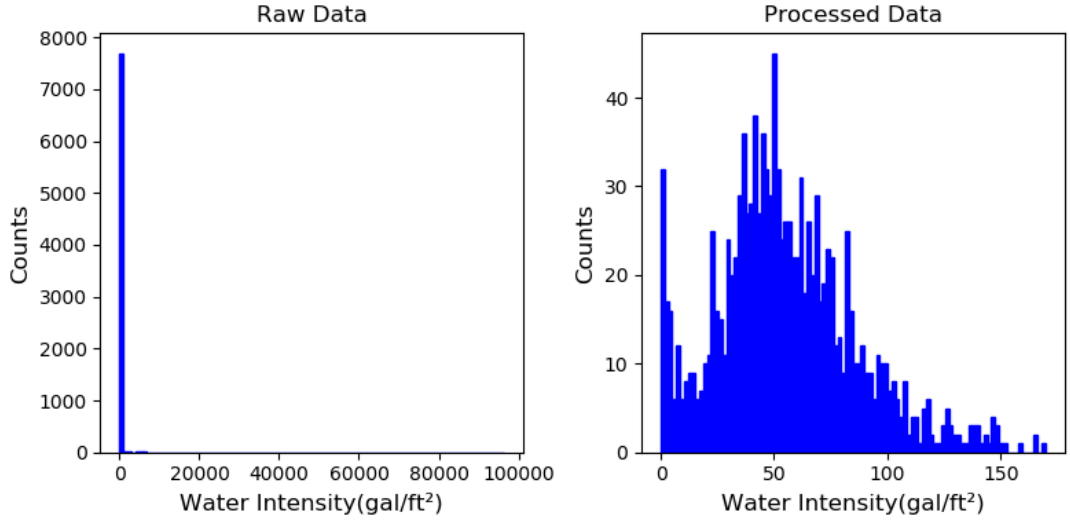

Figure 5. Water Intensity Histogram Before and After Outlier Removal

The transformation process involved several carefully sequenced steps:

1. Initial Distribution Analysis: We generated a histogram of the raw WI values using 100 bins (Figure 5-a), which revealed a highly skewed distribution with numerous outliers.
2. Outlier Management: Statistical analysis confirmed that many extreme values likely represented

measurement errors or truly anomalous consumption patterns. To ensure robust modeling, we applied the Interquartile Range (IQR) method, that is removing the outliers beyond boundaries at Q1 - 3*IQR and Q3 + 3*IQR. This approach preserved meaningful variability while eliminating potentially problematic extreme values that could distort our analysis.

3. Post-Outlier Removal Assessment: After outlier removal, we re-analyzed the distribution (Figure 5-b), confirming a more normalized pattern suitable for further analysis.
4. Binary Classification Transformation: We partitioned the continuous WI values into "Higher" and "Lower" categories using the median value as the threshold, creating a balanced binary classification problem.

3.3.3 Feature Preprocessing

We implemented comprehensive feature preprocessing to ensure data quality and model relevance:

1. Missing Value Treatment: For remaining features, we imputed missing values using the median for numerical features, providing a robust central tendency measure less affected by outliers than the mean.
2. Feature Filtering: We applied multiple filtering criteria to ensure data quality and relevance. That is removing features with excessive missing values (>50%), eliminating non-informative administrative features (addresses, building names, owner information), and excluding features with direct mathematical relationships to the target variable, such as absolute water usage metrics that could be derived from area calculations, which would create artificial predictive power
3. Categorical Feature Handling: Categorical variables were appropriately encoded to make them suitable for machine learning algorithms.

3.3.4 Final Dataset Characteristics

The final processed dataset contained 1,381 Census Tracts characterized by 1,121 features (1,120 predictors and 1 binary target variable). The considerable reduction in standard deviation (from 1,375.23 to 29.35) confirms the effectiveness of our outlier removal approach, creating a more stable foundation for subsequent modeling. The slight class imbalance (824 low vs. 557 high WI observations) was addressed through appropriate evaluation metrics and validation techniques in our modeling approach.

This comprehensive dataset served as the foundation for our application of the FCTTE framework, enabling us to identify the most significant factors influencing urban water consumption patterns across New York City.

4. Results and Discussion

This section presents the performance evaluation of our proposed FCTTE framework for urban water consumption classification. We begin by explaining the evaluation metrics used, compare various machine learning algorithms, assess existing feature selection methods, analyze our FCTTE framework's performance, and finally examine the most influential features affecting water intensity.

4.1 Evaluation Metrics

We employed two widely used metrics for evaluating classification model performance: Accuracy score and ROC-AUC score.

4.1.1 Accuracy Score

Accuracy score provides an intuitive measure of a model's predictive performance, calculated as the ratio of correct predictions to the total number of predictions:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (1)$$

This metric offers a straightforward assessment of how often the model correctly classifies samples.

4.1.2 ROC-AUC Score

The Receiver Operating Characteristic (ROC) curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The Area Under the Curve (AUC) quantifies the model's ability to discriminate between classes, with values typically ranging between 0.5 (random classification) and 1.0 (perfect classification). Higher AUC values indicate better model performance, making this metric valuable for comparing classification models.

4.2 Comparison of Machine Learning Classification Algorithms

To establish a performance baseline, we evaluated twelve common machine learning classification algorithms on our complete dataset (1,381 samples with 1,120 features). We employed K-fold cross-validation to ensure reliable and generalizable results, as illustrated in Figure 6.

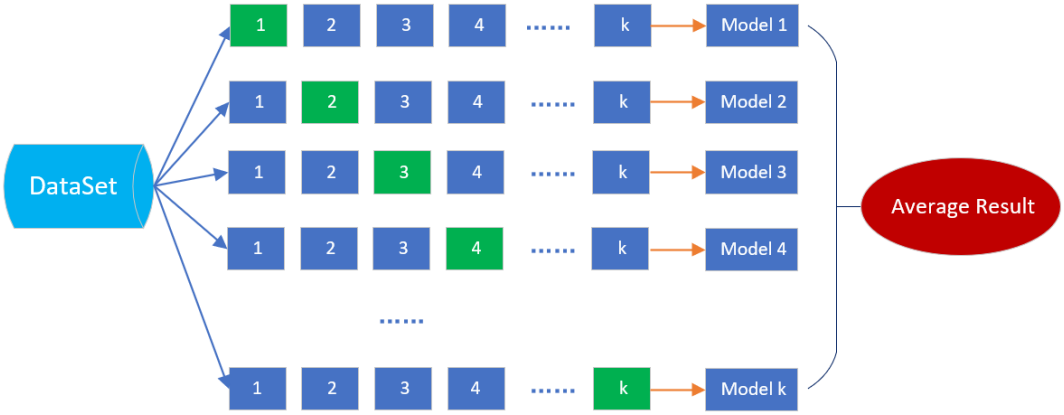

Figure 6. K-fold Cross-Validation methodology

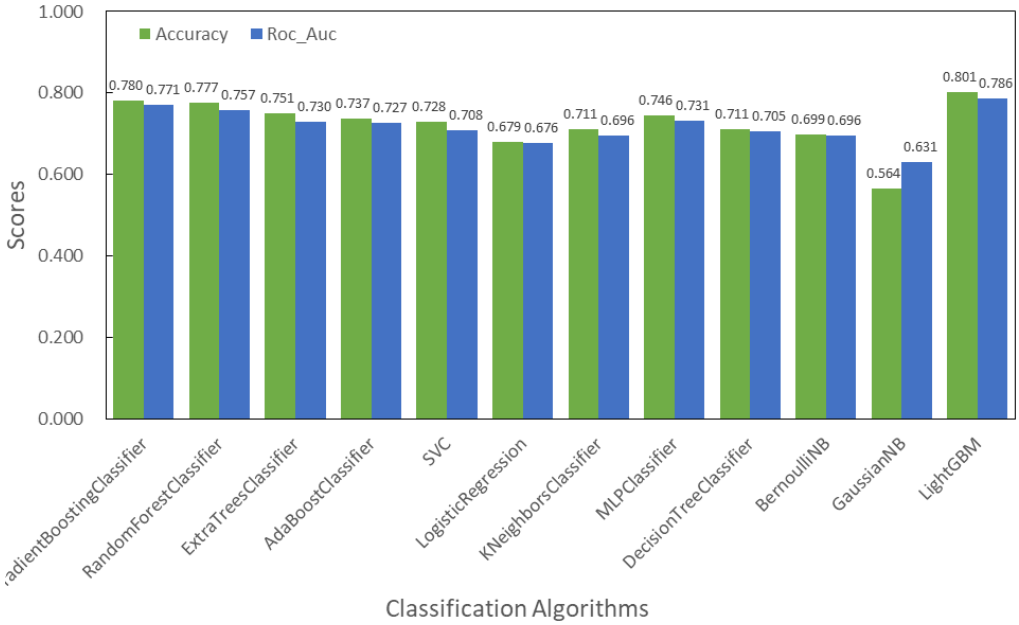

Figure 7. Performance Comparison

The performance comparison of these algorithms is presented in Figure 7. Decision tree-based ensemble methods consistently outperformed other approaches. Notably, GradientBoostingClassifier and RandomForestClassifier demonstrated superior performance compared to the basic DecisionTreeClassifier, reflecting the effectiveness of their respective optimization techniques (Gradient Boosting and Bagging).

LightGBM achieved the highest performance with an Accuracy score of 0.801 and ROC-AUC score of 0.786, outperforming the base DecisionTreeClassifier by approximately 12.6%. This superior performance can be attributed to LightGBM's advanced techniques, including Gradient-based One-Side Sampling (GOSS), Exclusive Feature Bundling (EFB), and Leaf-wise growth strategy, which efficiently handle high-dimensional data.

Support Vector Classifier (SVC) showed relatively poor performance, likely due to suboptimal kernel function selection for this high-dimensional dataset when using default parameters.

4.3 Evaluation of Conventional Feature Selection Methods

We compared five widely used feature selection approaches: Mutual Information, Correlation, Recursive Feature Elimination (RFE), Random Forest, and Extra Trees. To facilitate comparison, we used a consistent step size of 50 features and evaluated their impact on four algorithms: LightGBM, GBDT, SVC, and MLPClassifier.

Table 2 presents the optimal feature count and corresponding performance metrics for each method. Several key observations emerged:

- Varying impact across algorithms: Feature selection yielded performance improvements of approximately 1% for LightGBM, 2.2% for GBDT, 5.4% for SVC, and 4% for MLPClassifier compared to using the full feature set.
- Algorithm-specific benefits: LightGBM showed the smallest improvement from feature selection, likely because it inherently performs feature selection during training by excluding low-gain features at leaf nodes.
- Method-specific impacts: The Random Forest feature selection method provided the most substantial improvement for SVC (7.6%), while RFE yielded the greatest improvement for MLPClassifier (5.5%).
- Different convergence rates: Figure 8 illustrates how accuracy evolves with increasing feature count for each selection method. Correlation-based selection demonstrated the slowest convergence, stabilizing only after approximately 400 features, reflecting its limitation in capturing relationships beyond linear feature-target associations.
- Tree-based methods' efficiency: Extra Trees achieved optimal performance with just 151 features, followed by RFE and Random Forest. Correlation and Mutual Information required substantially more features to reach peak performance, highlighting the efficiency of tree-based methods in identifying informative feature subsets.

Based on these findings, we selected LightGBM as our base algorithm for subsequent analysis due to its superior overall performance, despite showing the least improvement from conventional feature selection.

Table 2. Performance metrics for different feature selection methods.

	Algorithms	Accuracy	Roc_Auc	Number of Features		Algorithms	Accuracy	Roc_Auc	Number of Features
LightGBM	Mutual Information	0.806	0.792	801	GBDT	Mutual Information	0.801	0.786	401
	Correlation	0.806	0.791	601		Correlation	0.786	0.774	801
	RFE	0.815	0.799	251		RFE	0.798	0.791	351
	Random Forest	0.806	0.791	251		Random Forest	0.801	0.783	51
	Extra Trees	0.815	0.803	151		Extra Trees	0.806	0.789	201
	No Feature Selection	0.801	0.786	1120		No Feature Selection	0.780	0.771	1120
SVC	Mutual Information	0.754	0.740	51	MLPClassifier	Mutual Information	0.775	0.765	451
	Correlation	0.760	0.740	401		Correlation	0.757	0.754	451
	RFE	0.772	0.764	51		RFE	0.801	0.792	201
	Random Forest	0.783	0.769	51		Random Forest	0.789	0.784	251
	Extra Trees	0.777	0.767	201		Extra Trees	0.777	0.769	201
	No Feature Selection	0.728	0.708	1120		No Feature Selection	0.746	0.731	1120

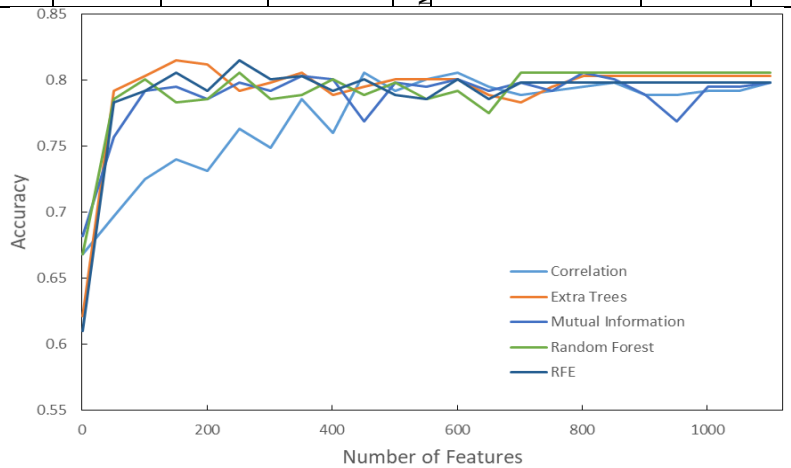

Figure 8. Performance dynamics with different number of selected features.

4.3 FCTTE Performance Evaluation

After analyzing conventional feature selection methods, we evaluated our proposed FCTTE framework using LightGBM as the base algorithm.

Figure 9 and Figure 10 illustrate how accuracy varies with cluster count for the Top K and Threshold selection strategies, respectively. The Mutual Information-based clustering consistently outperformed Normalization and Correlation approaches, with optimal performance occurring around 40 clusters. Correlation-based clustering showed the weakest performance overall.

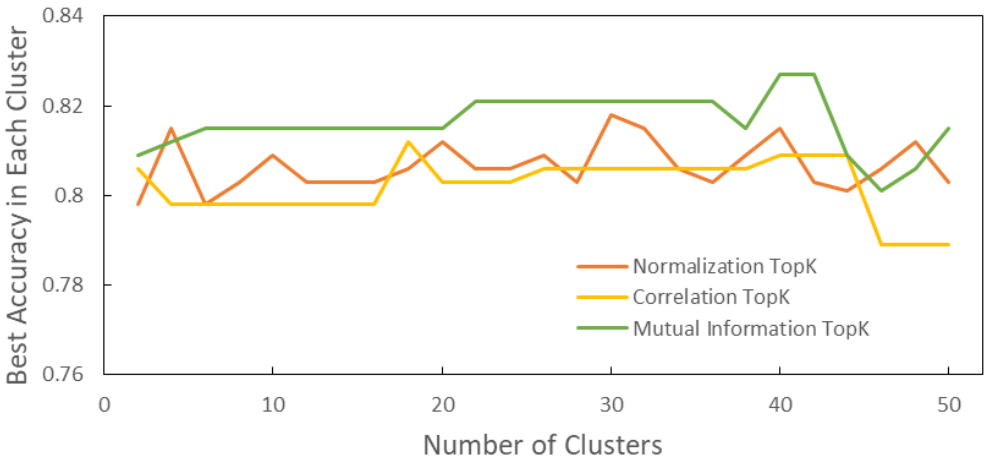

Figure 9. Accuracy vs number of clusters for TopK method

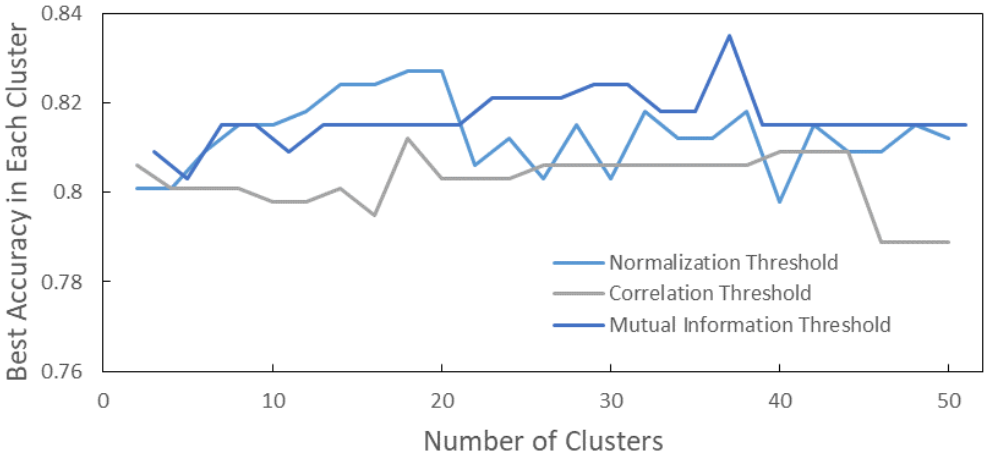

Figure 10. Accuracy vs number of clusters for Threshold method

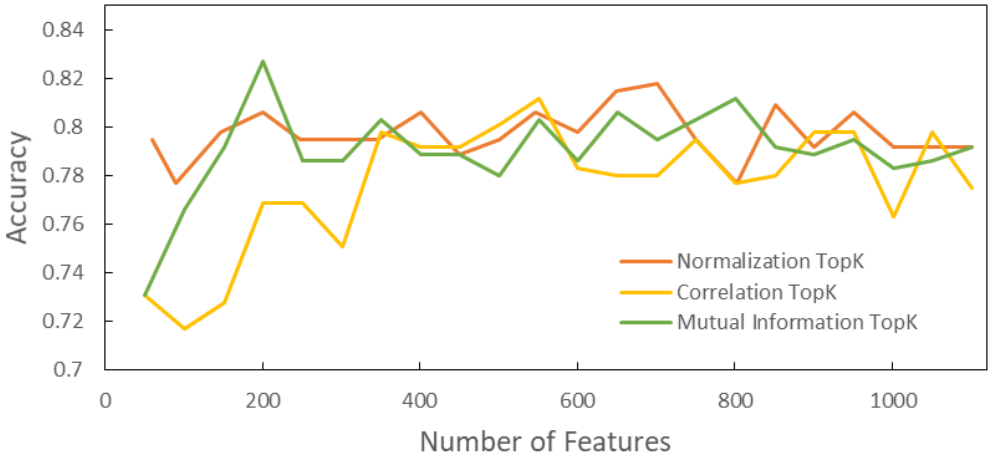

Figure 11. Feature selection process for TopK method

The feature selection progression for each method at their optimal cluster count is shown in Figure 11 and Figure 12. Notably, Correlation-based methods exhibited the slowest convergence, while the Mutual Information-based approaches achieved peak performance at approximately 200 features before stabilizing.

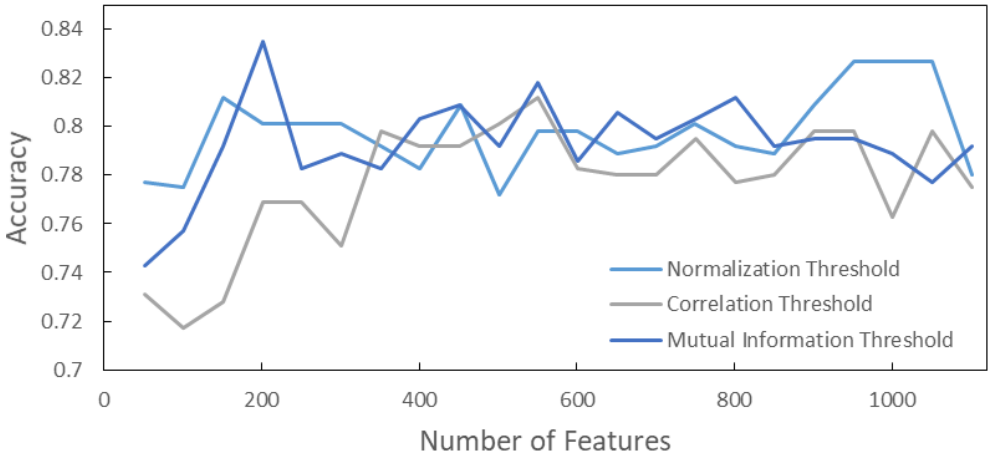

Figure 12. Feature selection process for Threshold method

Table 1 summarizes the performance of all six FCTTE variants (three transformation methods × two selection strategies). All variants demonstrated significant improvements over conventional feature selection methods. The Mutual Information transformation with Threshold selection strategy achieved the best overall performance with an Accuracy score of 0.835 and ROC-AUC score of 0.824.

Table 3. Performance summary.

Algorithms		Accuracy	Roc_Auc	Number of Features	Number of Clusters	TopK/ Threshold
Correlation	Top K	0.812	0.797	551	18	534
	Threshold	0.812	0.797	551	18	0.4835
Mutual Information	Top K	0.827	0.813	201	42	156
	Threshold	0.835	0.824	201	37	0.152
Normalization	Top K	0.818	0.804	701	30	54
	Threshold	0.827	0.807	1051	18	0.9286

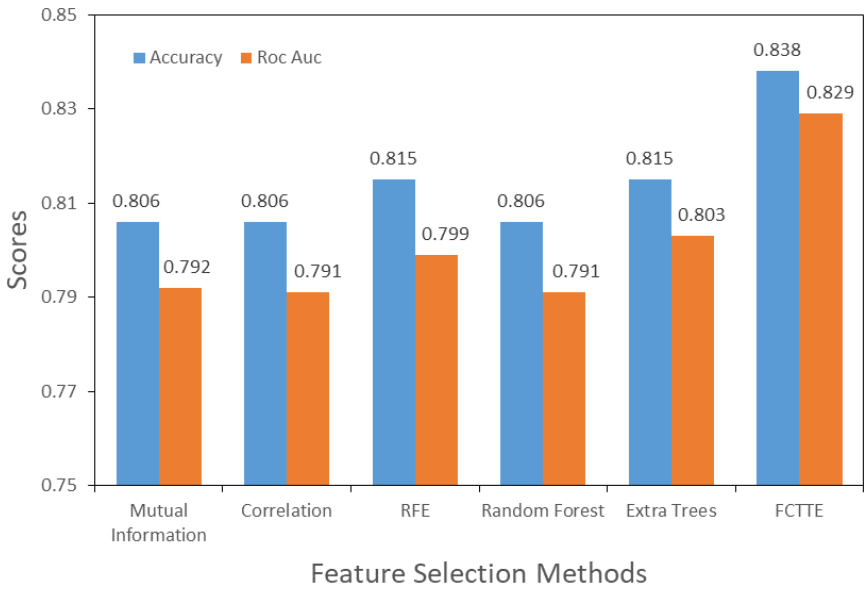

Figure 13. Performance comparison between different methods

The effectiveness of this approach stems from its ability to:

- Transform the original dataset into feature relationship measurements using Mutual Information
- Group highly redundant features through hierarchical clustering
- Select the most informative features using appropriate thresholds
- Combine these features into a comprehensive, non-redundant feature set

As shown in Figure 13, our ensemble approach combining all six FCTTE variants further improved performance, achieving approximately 3.5% higher Accuracy and 4.3% higher ROC-AUC compared to the best conventional feature selection method. This represents a 4.6% improvement over the baseline LightGBM with no feature selection—a substantial gain considering that conventional methods only improved LightGBM performance by approximately 1%.

These results convincingly demonstrate the effectiveness of our FCTTE framework for feature selection in high-dimensional urban datasets.

4.5 Feature Importance Analysis

To identify potential factors influencing water intensity (WI), we extracted the ten most important features from our FCTTE framework. As shown in Figure 14, Median Family Income (Dollars) emerged as the most influential factor, with an importance score 2.87 times higher than the second-ranked feature.

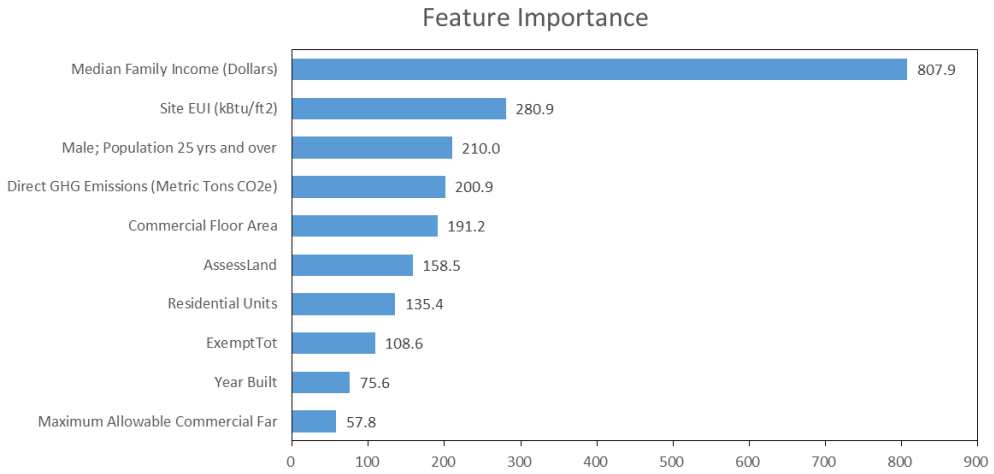

Figure 14. Feature Importance

Figure 15 illustrates the normalized values of these features across the "Lower" and "Higher" WI categories, providing insights into their relationships with water consumption patterns. Our analysis revealed several key relationships at the census tract level:

1. Median Family Income: Strong negative correlation with WI, suggesting that higher-income households typically demonstrate more water-efficient consumption patterns, possibly reflecting higher education levels and greater conservation awareness.
2. Site EUI (kBtu/ft²): Strong positive correlation with WI, confirming findings from previous research [31] that energy consumption intensity correlates with water consumption intensity.
3. Male Population 25 Years and Over: Negative correlation with WI, potentially reflecting gender-based differences in water usage behaviors. Social surveys indicate that men typically spend less time showering than women (15% of men versus 37% of women shower for over an hour), which may contribute to lower water usage in areas with higher adult male populations.
4. Direct GHG Emissions: Positive correlation with WI, likely due to the relationship between energy consumption and water usage patterns.
5. Commercial Floor Area: Negative correlation with WI, suggesting that commercial buildings typically utilize more advanced water-efficient fixtures and may house occupants with greater conservation awareness compared to residential areas.
6. AssessLand (Land Value): Negative correlation with WI, potentially indicating that higher land values correspond with commercial districts that demonstrate more efficient water usage [32].
7. Residential Units: Positive correlation with WI, supporting the observation that predominantly residential

areas tend to have higher water intensity than commercial zones.

8. ExemptTot (Tax Exemption): Negative correlation with WI, possibly reflecting the presence of educational institutions, government facilities, and community organizations that promote water conservation practices.
9. Year Built: Negative correlation with WI, indicating that newer buildings incorporate more efficient water systems and distribution designs than older structures.
10. Maximum Allowable Commercial FAR: Negative correlation with WI, consistent with the finding that commercial areas generally demonstrate lower water intensity.

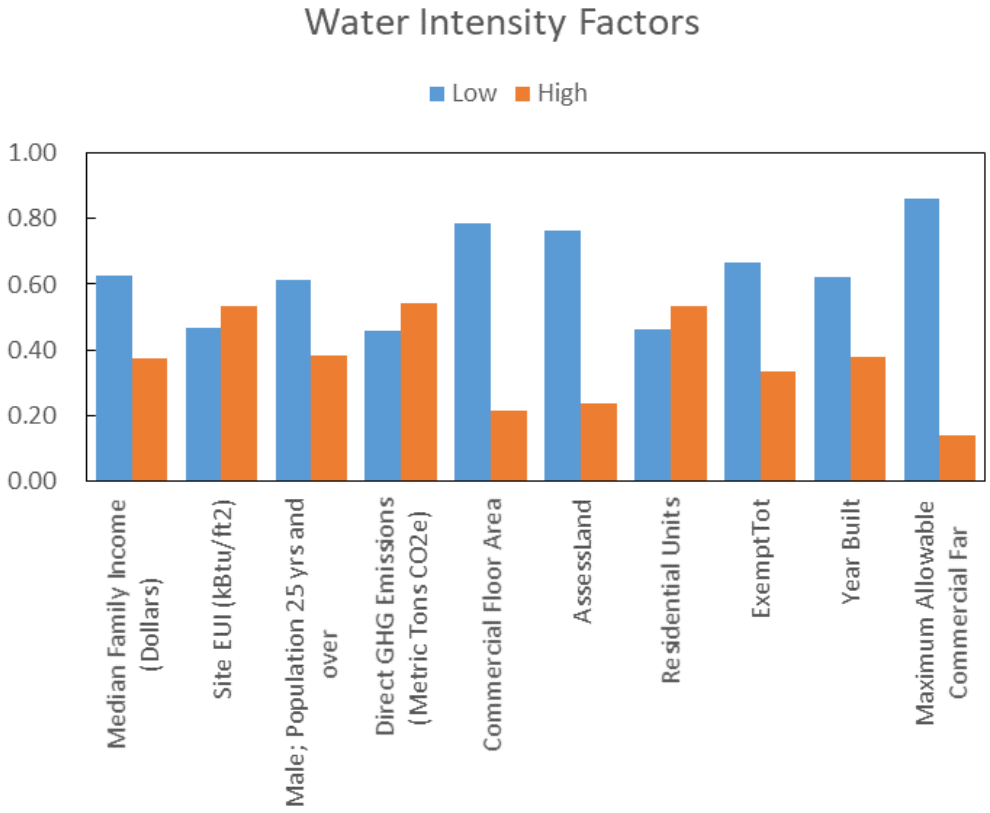

Figure 15. Feature comparisons

These insights provide valuable guidance for urban water resource management, highlighting socioeconomic, structural, and demographic factors that significantly influence consumption patterns.

5. Conclusion

Our study introduces the Feature Clustering Framework of TopK and Threshold with Ensemble Method (FCTTE), demonstrating its superior explanatory power, generalizability, and accuracy compared to conventional feature selection approaches. To validate this framework, we integrated eight diverse datasets from New York City, encompassing 1,120 features, one target variable, and 1,381 samples across multiple domains.

We evaluated various machine learning classification algorithms, with LightGBM emerging as the best performer. When applied with our FCTTE framework, LightGBM showed substantial improvements of approximately 3.5% in Accuracy score and 4.3% in ROC-AUC score compared to the best conventional feature selection methods. Most notably, FCTTE improved LightGBM performance by approximately 4.6% over baseline (no feature selection), while conventional methods only achieved approximately 1% improvement.

Our feature importance analysis identified key factors influencing urban water intensity, including household income, energy consumption, building characteristics, demographic composition, and tax policies. These insights provide valuable guidance for policymakers and future researchers working on urban water resource management.

While our study aimed to comprehensively model environmental factors affecting water intensity by integrating multiple datasets, certain limitations should be acknowledged. The use of census tract as the basic unit of analysis,

while necessary due to data availability constraints, may have introduced some distortion in the dataset. However, these limitations do not diminish the significance of the FCTTE framework's demonstrated effectiveness.

Finally, it is worth noting that maximizing predictive accuracy was not the primary objective of this research, which is why all algorithms were implemented with default parameters. The focus was instead on developing a robust feature selection framework capable of identifying the most influential factors in high-dimensional urban datasets.

References

[1] Ma, J., Ding, Y., Cheng, J. C., Jiang, F., & Xu, Z. (2020). Soft detection of 5-day BOD with sparse matrix in city harbor water using deep learning techniques. *Water Research, 170*, Article 115350. https://doi.org/10.1016/j.watres.2019.115350

[2] Saurí, D. (2013). Water conservation: Theory and evidence in urban areas of the developed world. *Annual Review of Environment and Resources, 38*, 227–248. https://doi.org/10.1146/annurev-environ-013113-142651

[3] Jiang, F., Ma, J., Li, Z., & Ding, Y. (2022). Prediction of energy use intensity of urban buildings using the semi-supervised deep learning model. *Energy, 249*, Article 123631. https://doi.org/10.1016/j.energy.2022.123631

[4] Jiang, F., & Ma, J. (2025). Predicting urban vitality at regional scales: A deep learning approach to modelling population density and pedestrian flows. *Smart Cities, 8*, 58. https://doi.org/10.3390/smartcities8020058

[5] Jiang, F., Ma, J., & Li, Z. (2022). Pedestrian volume prediction with high spatiotemporal granularity in urban areas by the enhanced learning model. *Sustainable Cities and Society, 79*, Article 103653. https://doi.org/10.1016/j.scs.2021.103653

[6] Hussien, W. A., Memon, F. A., & Savic, D. A. (2016). Assessing and modelling the influence of household characteristics on per capita water consumption. *Water Resources Management, 30*(9), 2931–2955. https://doi.org/10.1007/s11269-016-1314-x

[7] Katz, D. (2015). Water use and economic growth: Reconsidering the Environmental Kuznets Curve relationship. *Journal of Cleaner Production, 88*, 205–213. https://doi.org/10.1016/j.jclepro.2014.08.017

[8] Sant'Ana, D., & Mazzega, P. (2018). Socioeconomic analysis of domestic water end-use consumption in the Federal District, Brazil. *Sustainable Water Resources Management, 4*(4), 921–936. https://doi.org/10.1007/s40899-017-0186-4

[9] Blasco, X., Martínez, M., Herrero, J. M., Ramos, C., & Sanchis, J. (2007). Model-based predictive control of greenhouse climate for reducing energy and water consumption. *Computers and Electronics in Agriculture, 55*(1), 49–70. https://doi.org/10.1016/j.compag.2006.12.001

[10] Ma, J., & Cheng, J. C. (2016). Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests. *Applied Energy, 183*, 193–201.

[11] Ma, J., Cheng, J. C., Jiang, F., Chen, W., & Zhang, J. (2020). Analyzing driving factors of land values in urban scale based on big data and non-linear machine learning techniques. *Land Use Policy, 94*, Article 104537.

[12] Cheng, J. C., & Ma, L. J. (2015). A non-linear case-based reasoning approach for retrieval of similar cases and selection of target credits in LEED projects. *Building and Environment, 93*, 349–361.

[13] Jiang, F., Ma, J., Webster, C. J., Chiaradia, A. J. F., Zhou, Y., Zhao, Z., et al. (2024). Generative urban design: A systematic review on problem formulation, design generation, and decision-making. *Progress in Planning, 180*, Article 100795. https://doi.org/10.1016/j.progress.2023.100795

[14] Li, Z., Ma, J., & Jiang, F. (2024). Exploring the effects of 2D/3D building factors on urban energy consumption using explainable machine learning. *Journal of Building Engineering, 97*, Article 110827. https://doi.org/10.1016/j.jobe.2024.110827

[15] Zhou, J., Li, Z., Ma, J. J., & Jiang, F. (2020). Exploration of the hidden influential factors on crime activities: A big data approach. *IEEE Access, 8*, 141033–141045. https://doi.org/10.1109/ACCESS.2020.3009969

[16] Li, Z., & Ma, J. (2022). Discussing street tree planning based on pedestrian volume using machine learning and computer vision. *Building and Environment, 219*, Article 109178.

[17] Jiang, F., Ma, J., Webster, C. J., Li, X., & Gan, V. J. (2023). Building layout generation using site-embedded GAN model. *Automation in Construction, 151*, Article 104888.

[18] Jiang, F., & Ma, J. (2021). A comprehensive study of macro factors related to traffic fatality rates by XGBoost-

based model and GIS techniques. *Accident Analysis & Prevention, 163*, Article 106431. https://doi.org/10.1016/j.aap.2021.106431

[19] Jiang, F., Yuen, K. K. R., & Lee, E. W. M. (2020). A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions. *Accident Analysis & Prevention, 141*, Article 105520. https://doi.org/10.1016/j.aap.2020.105520

[20] Jiang, F., Ma, J., & Li, Z. (2022). Pedestrian volume prediction with high spatiotemporal granularity in urban areas by the enhanced learning model. *Sustainable Cities and Society, 79*, Article 103653.

[21] Jiang, F., Ma, J., Webster, C. J., Chen, W., & Wang, W. (2024). Estimating and explaining regional land value distribution using attention-enhanced deep generative models. *Computers in Industry*, 159–160, Article 104103. https://doi.org/10.1016/j.compind.2024.104103

[22] Jiang, F., & Ma, J. (2025). Environmental justice in the 15-minute city: Assessing air pollution exposure inequalities through machine learning and spatial network analysis. *Smart Cities, 8*, 53. https://doi.org/10.3390/smartcities8020053

[23] Jiang, F., Ma, J., Webster, C. J., Wang, W., & Cheng, J. C. P. (2024). Automated site planning using CAIN-GAN model. *Automation in Construction, 159*, Article 105286. https://doi.org/10.1016/j.autcon.2024.105286

[24] Jiang, F., Ma, J., Webster, C. J., Li, X., & Gan, V. J. L. (2023). Building layout generation using site-embedded GAN model. *Automation in Construction, 151*, Article 104888. https://doi.org/10.1016/j.autcon.2023.104888

[25] Dietterich, T. G. (2000). Ensemble methods in machine learning. In Multiple Classifier Systems (pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1

[26] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems: Vol. 30*. Curran Associates, Inc.

[27] Gan, V. J., Lo, I. M., Ma, J., Tse, K., Cheng, J. C., & Chan, C. (2020). Simulation optimisation towards energy efficient green buildings: Current status and future trends. *Journal of Cleaner Production, 254*, Article 120012.

[28] Jiang, F., Ma, J., Li, Z., & Ding, Y. (2022). Prediction of energy use intensity of urban buildings using the semi-supervised deep learning model. *Energy, 249*, Article 123631.

[29] Jiang, F., Yuen, K. K. R., & Lee, E. W. M. (2020). Analysis of motorcycle accidents using association rule mining-based framework with parameter optimization and GIS technology. *Journal of Safety Research, 75*, 292–309. https://doi.org/10.1016/j.jsr.2020.09.004

[30] Jiang, F., Yuen, K. K. R., Lee, E. W. M., & Ma, J. (2020). Analysis of run-off-road accidents by association rule mining and geographic information system techniques on imbalanced datasets. *Sustainability, 12*, Article 4882. https://doi.org/10.3390/su12124882

[31] Lee, M., Keller, A. A., Chiang, P.-C., Den, W., Wang, H., Hou, C.-H., et al. (2017). Water-energy nexus for urban water systems: A comparative review on energy intensity and environmental impacts in relation to global water risks. *Applied Energy, 205*, 589–601. https://doi.org/10.1016/j.apenergy.2017.08.002

[32] Jiang, F., Ma, J., Webster, C. J., Chen, W., & Wang, W. (2024). Estimating and explaining regional land value distribution using attention-enhanced deep generative models. *Computers in Industry, 159*, Article 104103.