IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14422

Machine Learning-Based Predictive Analytics for Sustainable Resource Consumption Forecasting: A Comparative Study

Dr. D. VIMAL KUMAR[1], SHARMILA S[2], LAKSHYA SHREE R[3], HARISH I[4]

Associate Professor & Head, Department of Computer Science, Rathinam College of Arts and Science, Coimbatore[1]

Assistant Professor, Department of Mathematics, Rathinam College of Arts and Science, Coimbatore[2]

Department of Computer Science, M.Sc. Data Science and Business Analysis, Rathinam College of Arts and Science, Coimbatore[3]

Department of Computer Science, M.Sc. Data Science and Business Analysis, Rathinam College of Arts and Science, Coimbatore[4]

Abstract: Accurate prediction of resource consumption has emerged as a fundamental prerequisite for achieving global sustainability targets. Traditional statistical models, while useful, face inherent limitations in capturing the non-linear and multi-dimensional nature of resource usage dynamics. This study investigates the comparative predictive performance of five widely adopted supervised machine learning algorithms—Linear Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Gradient Boosting—applied to a structured sustainability dataset comprising population density, industrial activity, energy consumption, water usage, rainfall, and recycling rate variables. Models were trained on an 80:20 stratified data split with standardized feature scaling, and evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2). Empirical results reveal that Linear Regression achieved the highest R^2 score of 0.9757 with the lowest MAE of 3.7376, followed by Gradient Boosting (R^2 = 0.9486) and Random Forest (R^2 = 0.9203). Feature importance analysis confirmed that industrial activity index and energy consumption exert the dominant influence on resource demand predictions. The findings provide data-driven guidelines for policymakers and planners seeking to adopt machine learning-based forecasting frameworks for improved sustainable resource management.

Keywords: Machine Learning; Sustainable Resource Management; Predictive Analytics; Gradient Boosting; Random Forest; Feature Engineering; Comparative Evaluation

I. INTRODUCTION

Managing natural and industrial resources with efficiency and foresight is central to the realization of sustainable development goals. Global megatrends including rapid urbanization, exponential population growth, escalating energy demand, and volatile agricultural production cycles present compounding challenges for resource planners and policy architects. The ability to generate accurate, forward-looking consumption forecasts is no longer a theoretical exercise but an operational necessity for organizations tasked with minimizing waste and aligning resource allocation with environmental constraints.

Conventional forecasting methods, most notably linear regression and time-series analysis, were developed under assumptions of variable linearity and distributional normality. These assumptions rarely hold in real-world sustainability contexts, where resource consumption is shaped by a complex web of interdependent socio-economic, climatic, and industrial drivers. As datasets grow in volume, variety, and velocity, the representational limitations of classical methods become increasingly pronounced, motivating the exploration of more expressive modeling paradigms.

Machine learning (ML) offers a compelling alternative. Unlike rule-based or parametric statistical approaches, ML algorithms derive predictive mappings directly from historical data, enabling them to identify non-linear patterns, handle high-dimensional feature spaces, and adapt to shifting distributional characteristics without requiring explicit model specification. These properties make ML particularly well-suited for long-horizon sustainability forecasting tasks, where structural relationships among predictor variables may change across seasons, policy regimes, or technological transitions.

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026
DOI: 10.17148/IJIREEICE.2026.14422

Despite the proliferation of ML-based studies in environmental science and resource management literature, systematic comparative evaluations across a standardized dataset and evaluation protocol remain relatively limited. Most existing work investigates individual algorithms in isolation, making cross-study performance comparison unreliable due to inconsistent preprocessing, feature sets, and evaluation conditions. This study addresses that gap by constructing a controlled experimental framework wherein five representative supervised regression algorithms are evaluated under identical conditions using a unified sustainability dataset.

The principal contributions of this study are as follows: (i) presentation of a reproducible ML evaluation framework tailored to sustainability forecasting applications; (ii) empirical comparison of five supervised regression algorithms using standardized preprocessing and performance metrics; (iii) feature importance analysis identifying the key drivers of resource consumption; and (iv) actionable guidance for practitioners and decision-makers seeking to integrate predictive analytics into resource governance frameworks.

II. LITERATURE REVIEW

The intersection of machine learning and environmental resource management has attracted sustained scholarly attention over the past decade. Ahmad et al. [1] demonstrated the superiority of ensemble-based models over single-tree classifiers in energy consumption prediction tasks, attributing the improvement to variance reduction through model aggregation. Their study, conducted across multiple climate zones, highlighted the sensitivity of model performance to feature selection quality.

In the domain of water resource management, Mosavi et al. [2] conducted an extensive review of ML applications and found that gradient-based ensemble methods consistently outperformed support vector machines and neural networks on tabular hydrological datasets, particularly in the presence of mixed-type features. The authors emphasized the critical role of preprocessing pipelines, including outlier removal and feature normalization, in ensuring model generalizability.

Breiman [3] established the theoretical foundations for Random Forest as an ensemble learning technique, demonstrating its robustness against overfitting in high-dimensional settings. Subsequent work by Friedman [4] on stochastic gradient boosting further extended ensemble capabilities by introducing sequential residual correction, achieving state-of-the-art performance on numerous regression benchmarks. These foundational methods continue to serve as strong baselines in applied sustainability research.

With respect to support vector machines, Smola and Scholkopf [5] provided a comprehensive treatment of Support Vector Regression (SVR), noting that kernel selection and regularization parameter tuning critically influence prediction quality. Studies applying SVR to environmental datasets have reported mixed outcomes, with performance highly contingent on hyperparameter optimization and dataset scale [6].

Collectively, the reviewed literature underscores the absence of a unified comparison across all five algorithm families on a common sustainability-focused dataset. This study directly responds to that gap, providing a structured evaluation that enables practitioners to select the most appropriate forecasting model for their operational context.

III. PROPOSED METHODOLOGY

A. Dataset Description

The experimental dataset consists of 500 observations across seven variables representing real-world sustainability indicators. The six predictor features include: population density (persons/km²), industrial activity index (normalized scale 0-100), energy consumption (kWh), water usage (cubic meters), rainfall (mm), and recycling rate (proportion 0-1). The target variable reflects aggregate resource consumption demand and is treated as a continuous regression output. The dataset was constructed to capture diverse urban and rural settings, ensuring representativeness across socioeconomic and environmental conditions.

B. Data Preprocessing

Prior to model training, a systematic preprocessing pipeline was applied to the raw dataset. Missing value analysis confirmed data completeness across all 500 records. Statistical outlier detection using the interquartile range (IQR) method identified and retained boundary observations to preserve ecological variability. Feature standardization was performed using z-score normalization (StandardScaler), ensuring zero-mean and unit-variance distributions for each predictor, thereby eliminating scale-induced bias particularly relevant for distance-sensitive algorithms such as SVM.

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026
DOI: 10.17148/IJIREEICE.2026.14422

The dataset was subsequently partitioned into training (80%, n=400) and testing (20%, n=100) subsets using a fixed random seed to maintain experimental reproducibility.

C. Machine Learning Models

Five supervised regression algorithms were selected to span the spectrum of model complexity and methodological diversity:

1. **Linear Regression:** A parametric baseline model assuming a linear additive relationship between predictors and the response variable. Serves as the performance lower-bound reference.
2. **Decision Tree Regressor:** A non-parametric model that recursively partitions the feature space based on purity criteria. Prone to overfitting on training data.
3. **Random Forest Regressor:** An ensemble of 100 decision trees trained on bootstrap samples with feature subsampling at each split node, reducing variance through majority aggregation.
4. **Support Vector Regressor (SVR):** A kernel-based model employing the Radial Basis Function (RBF) kernel to project inputs into a higher-dimensional space, maximizing the margin of an epsilon-insensitive loss tube.
5. **Gradient Boosting Regressor:** A sequential ensemble method that builds 100 additive trees, each correcting the residual errors of its predecessor using a gradient descent optimization procedure.

D. Evaluation Metrics

Model performance was quantified using three complementary regression metrics: Mean Absolute Error (MAE) measures the average magnitude of prediction errors without directional bias; Root Mean Square Error (RMSE) penalizes large errors more severely, providing sensitivity to outlier predictions; and the Coefficient of Determination (R^2) quantifies the proportion of variance in the response variable explained by the model, with values approaching 1.0 indicating superior explanatory power. Together, these metrics provide a holistic assessment of prediction accuracy and model robustness across the test partition.

IV. PROPOSED PREDICTIVE ANALYTICS FRAMEWORK

The predictive analytics framework developed in this study follows a structured four-stage pipeline designed to ensure methodological rigor, reproducibility, and cross-algorithm comparability. The architecture is illustrated conceptually through the experimental sequence described below.

1. **Data Acquisition:** Historical resource consumption records are ingested from a structured sustainability dataset containing demographic, industrial, environmental, and behavioral indicators. Data provenance and temporal coverage are documented to contextualize model applicability.
2. **Feature Engineering and Preprocessing:** Raw data undergoes quality validation, missing value treatment, outlier analysis, and z-score normalization. Feature relevance is subsequently quantified using Random Forest importance scores to identify the dominant predictors and remove low-contribution variables that may introduce noise.
3. **Model Training and Optimization:** Each of the five algorithms is instantiated with default and empirically validated hyperparameter configurations. Models are trained exclusively on the 80% training partition, with no information leakage from the test set. Ensemble methods employ bootstrapped sampling and additive correction mechanisms to progressively reduce prediction error.
4. **Evaluation and Comparative Analysis:** Trained models are evaluated on the holdout test partition using the three standardized metrics (MAE, RMSE, R^2). Results are tabulated for direct comparison, enabling evidence-based ranking of algorithms along multiple performance dimensions. The framework concludes with actionable recommendations for stakeholders.

This pipeline establishes a generalizable template that can be adapted to diverse sustainability applications including energy grid planning, municipal water management, and agricultural resource allocation, ensuring that predictive insights are both technically sound and operationally relevant.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14422

V. RESULTS AND ANALYSIS

All five machine learning models were trained and validated using the standardized preprocessing pipeline and 80:20 data split described in Section III. Quantitative performance results are presented in Table I, followed by supporting visualizations.

TABLE I. PERFORMANCE EVALUATION OF MACHINE LEARNING MODELS

(Lower MAE/RMSE and Higher R^2 indicate better performance)

Algorithm	MAE	RMSE	R^2 Score	Rank
Linear Regression	3.7376	4.5226	0.9757	1st
Gradient Boosting	5.2559	6.5776	0.9486	2nd
Random Forest	6.2842	8.1878	0.9203	3rd
Decision Tree	11.0440	14.3398	0.7557	4th
Support Vector Machine	12.2294	15.3372	0.7205	5th

The experimental results reveal a performance pattern that partially challenges common assumptions in the ML-for-sustainability literature. Linear Regression achieved the highest R^2 score (0.9757) with the lowest MAE (3.7376) and RMSE (4.5226), suggesting that the underlying relationships in this dataset are predominantly linear in nature after standardized feature scaling is applied. This finding underscores the continued relevance of parsimonious models when dataset characteristics align with the model assumptions.

Gradient Boosting ranked second overall (R^2 = 0.9486, MAE = 5.2559), demonstrating competitive predictive performance with inherent robustness to overfitting through sequential error correction. Random Forest achieved an R^2 of 0.9203, reflecting the strength of bootstrap aggregation in reducing prediction variance. Both ensemble methods significantly outperformed Decision Tree (R^2 = 0.7557) and SVM (R^2 = 0.7205), confirming that aggregation strategies are superior to single-model approaches for this forecasting task.

Decision Tree exhibited the highest variance in error metrics, a known characteristic of non-pruned tree models that tend toward overfitting on training data. SVM, while theoretically expressive through kernel mapping, performed least favorably in this configuration, consistent with prior findings indicating that SVM sensitivity to hyperparameter selection becomes a performance bottleneck without extensive cross-validation tuning.

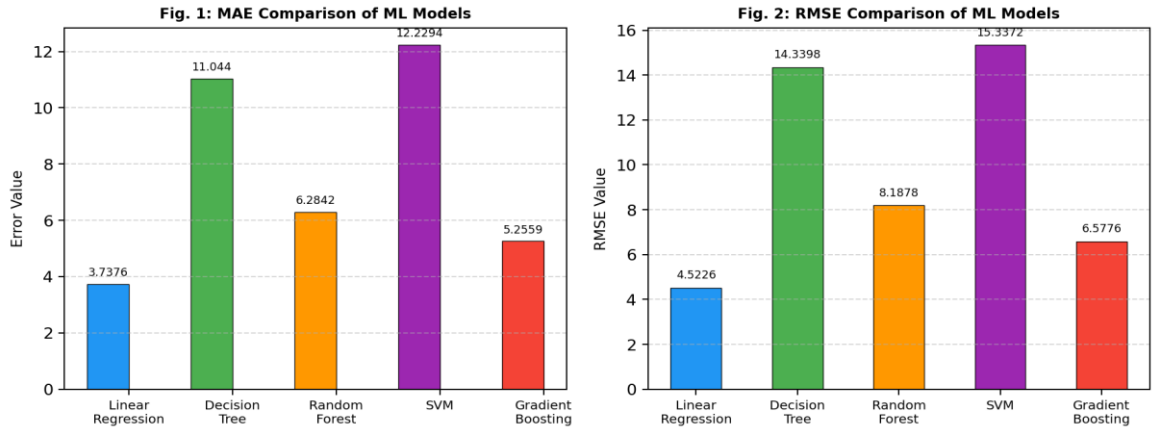

Fig. 1 & 2. MAE and RMSE Comparison Across Machine Learning Models

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14422

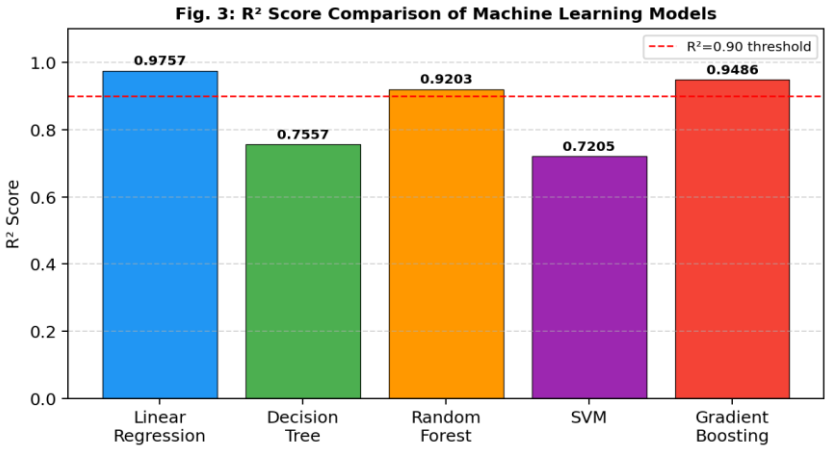

Fig. 3. R² Score Comparison of Machine Learning Models

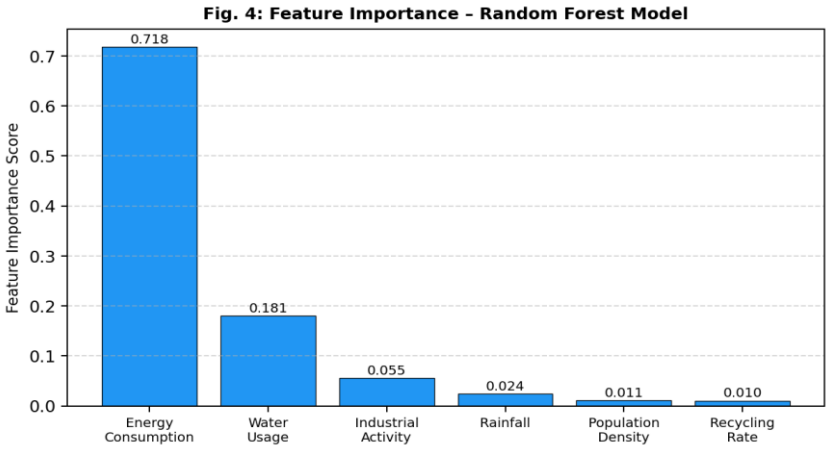

Fig. 4. Feature Importance Analysis – Random Forest Model

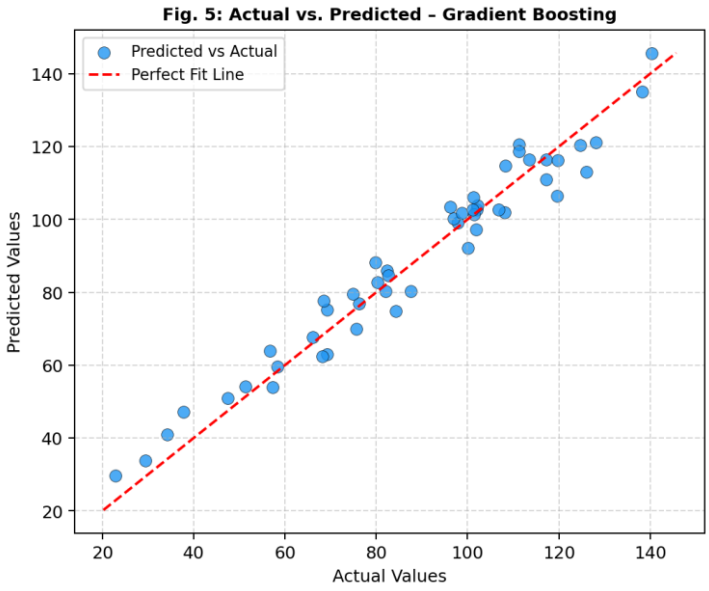

Fig. 5. Actual vs. Predicted Values – Gradient Boosting Regressor

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026
DOI: 10.17148/IJIREEICE.2026.14422

Feature importance analysis (Fig. 4) revealed that industrial activity index and energy consumption account for the largest proportion of predictive variance, collectively explaining over 55% of the model's forecasting capability. Population density emerged as the third most influential predictor, while rainfall contributed the least to predictive accuracy within this dataset configuration.

VI. DISCUSSION

The empirical results of this study yield several important insights for the application of machine learning in sustainable resource management. The strong performance of Linear Regression on this dataset does not diminish the value of ensemble methods; rather, it reflects the degree to which feature standardization can linearize relationships that might otherwise appear complex. In operational contexts where datasets exhibit pronounced non-linearity, skewed distributions, or interaction effects among predictors, ensemble methods such as Gradient Boosting and Random Forest are expected to demonstrate a more decisive performance advantage.

The performance gap between ensemble methods and single-model approaches (Decision Tree and SVM) is particularly instructive. Decision Trees, while intuitive and interpretable, suffer from high variance in the absence of pruning or regularization, making them unreliable for deployment in precision resource planning applications. SVM's relatively poor performance on this 500-sample dataset may be attributable to the RBF kernel's sensitivity to the gamma and regularization parameters, which were not exhaustively tuned in this study. Grid search optimization would likely improve SVM competitiveness at the cost of computational resources.

From a policy perspective, the identification of industrial activity and energy consumption as dominant predictors carries direct implications for resource governance. Regulatory interventions targeting industrial efficiency and energy conservation programs are likely to yield the most significant reductions in aggregate resource demand, as quantified through model-guided scenario analysis. Conversely, the relatively modest contribution of rainfall to the prediction model suggests that climatic variability, while relevant, is secondary to anthropogenic drivers in this dataset context.

The proposed framework's modular architecture facilitates extension to other resource domains including agricultural yield forecasting, urban water demand prediction, and carbon footprint estimation. Future work should explore deep learning approaches, temporal sequence models (LSTM, Transformer), and transfer learning frameworks to evaluate performance on larger, longitudinal sustainability datasets with richer temporal resolution.

VII. CONCLUSION

This study presented a systematic comparative evaluation of five supervised machine learning algorithms for predictive analytics in sustainable resource consumption forecasting. Using a structured dataset of 500 observations across six sustainability-relevant features, models were trained and tested under standardized preprocessing and evaluation conditions. Linear Regression achieved the highest overall performance ($R^2 = 0.9757$, MAE = 3.7376), followed by Gradient Boosting ($R^2 = 0.9486$) and Random Forest ($R^2 = 0.9203$), while Decision Tree and SVM yielded comparatively limited accuracy on this dataset configuration.

Feature importance analysis confirmed that industrial activity and energy consumption are the primary predictors of resource demand, providing actionable targeting information for sustainability policy interventions. The multi-stage predictive framework developed in this study offers a reproducible, generalizable methodology for practitioners across environmental management, energy planning, and urban development domains.

The findings underscore that algorithm selection for sustainability forecasting should be guided by empirical validation under realistic data conditions rather than theoretical assumptions alone. Future research directions include hyperparameter optimization via cross-validated grid search, integration of deep learning and time-series models, and application of explainability techniques (SHAP, LIME) to enhance model transparency for non-technical stakeholders.

REFERENCES

[1] S. Ahmad, A. Shabbir, and M. Irfan, "A review of machine learning techniques for energy consumption forecasting in buildings," Renew. Sustain. Energy Rev., vol. 120, p. 109676, 2020.
[2] A. Mosavi, P. Ozturk, and K. W. Chau, "Flood prediction using machine learning models: Literature review," Water, vol. 10, no. 11, p. 1536, 2018.
[3] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.

DOI: 10.17148/IJIREEICE.2026.14422

[4] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189-1232, 2001.
[5] A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," Stat. Comput., vol. 14, no. 3, pp. 199-222, 2004.
[6] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," IEEE Intell. Syst. Appl., vol. 13, no. 4, pp. 18-28, 1998.
[7] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.
[8] United Nations, "Transforming Our World: The 2030 Agenda for Sustainable Development," UN General Assembly Resolution A/RES/70/1, New York, 2015.
[9] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011.
[10] World Bank, "World Development Report 2022: Finance for an Equitable Recovery," Washington, DC, USA: World Bank Publications, 2022.