Regression Analysis Report

Student Id : 2408414

Student Name : Shashank Pandey

Section : L5CG20

Module Leader : Siman Giri

Tutor : Bibek Khanal

Submitted on : 02-10-2025

Table of Contents

Regression on Energy Consumption

Abstract

The major purpose of this research is to form a predictive model of energy consumption, using regression techniques to this end. More so, the paper is intended to identify the significant factors that influence energy consumption and develop a strong regression model for forecasting. As energies tend to become increasingly important toward the use of sustainable development applications, this analysis could possibly turn research toward the identification of important parameters for energy consumption and powerful construction of regression models for their predictions. The data set used for this study is the Energy Consumption Dataset that was drawn from Kaggle, which contains various attributes that contribute to energy use.

The research methodology includes exploratory data analysis (EDA) statistical models of decision tree regression, random forest regression, linear regression on optimized hyper-parameters and feature selection for better model performance.

The model developed will be finalized by evaluating using important performance metrics such as the R-squared (R^2) value, along with Mean Squared Error (MSE), all resulting in the best model's R^2 value and an MSE values. It is evident that these results predict important independent variables in energy consumption, along with suggestions for further improvements.

1. Introductions

1.1 Problems Statements

The effective use of energy resources is a key issue in relation to sustainable development. In attempting to develop a model for forecasting that can forecast consumption in terms of factors, such a model can make efficient choices for policymakers, professionals, and providers of energy in relation to efficiency and loss in terms of consumption of energy. Energy demand forecasting is most important in urban planning, infrastructure development, and conservation of the environment.

1.2 Dataset

This analysis makes use of the Energy Consumption Dataset from Kaggle and was developed by Jinil Patel in 2024. A variety of factors such as temperature, humidity, and several other environmental or operational-related determinants can be useful in deciding energy consumption. The dataset perfectly aligns itself with the objective of the United Nations Sustainable Development Goal (UNSDG) 7: Affordable and Clean Energy, which holds for improving energy efficiency and sustainable consumption patterns. Through analyzing the dataset for this study, the research contributes to broader purpose of energy sustainability.

1.3 Objectives

The vision that shapes this research study is to build a predictive regression model that predicts energy consumption on the basis of certain given independent variables. This is done by exploring the data set, finding important and relevant predictors, building and testing the regression models, and fine-tuning performance via hyper-parameter optimization and feature selection. Ultimately, such a model should prove accurate and interpretable, aiding energy management and perhaps also energy policies.

2. Methodology

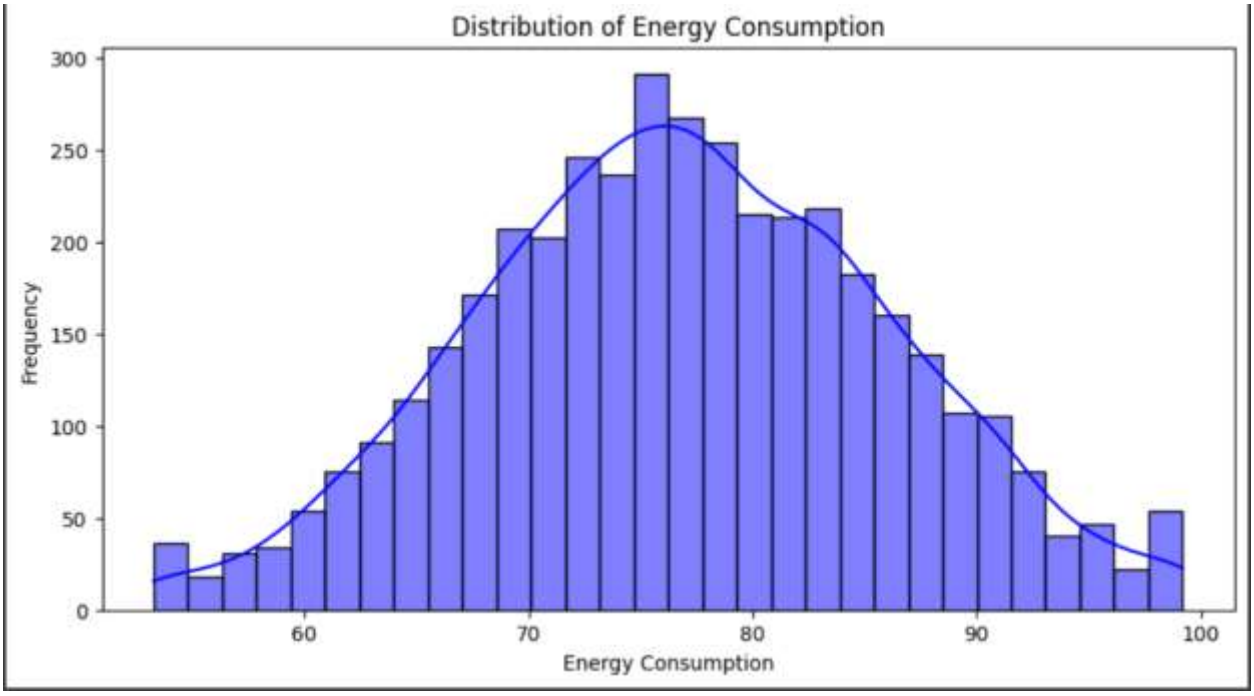
2.1 Data Pre-processing

Data preprocessing is a critical step for model accuracy and reliability. Missing values have first been examined in the data and, when in need, have been removed and replaced with respective statistics. Duplicate observations have been identified and removed for integrity in data. Numerical values have been normalized and scaled wherever in need for value uniformity in the input. Outlier techniques for value determination and extreme value suppression, with a tendency to affect model performance, have been employed.

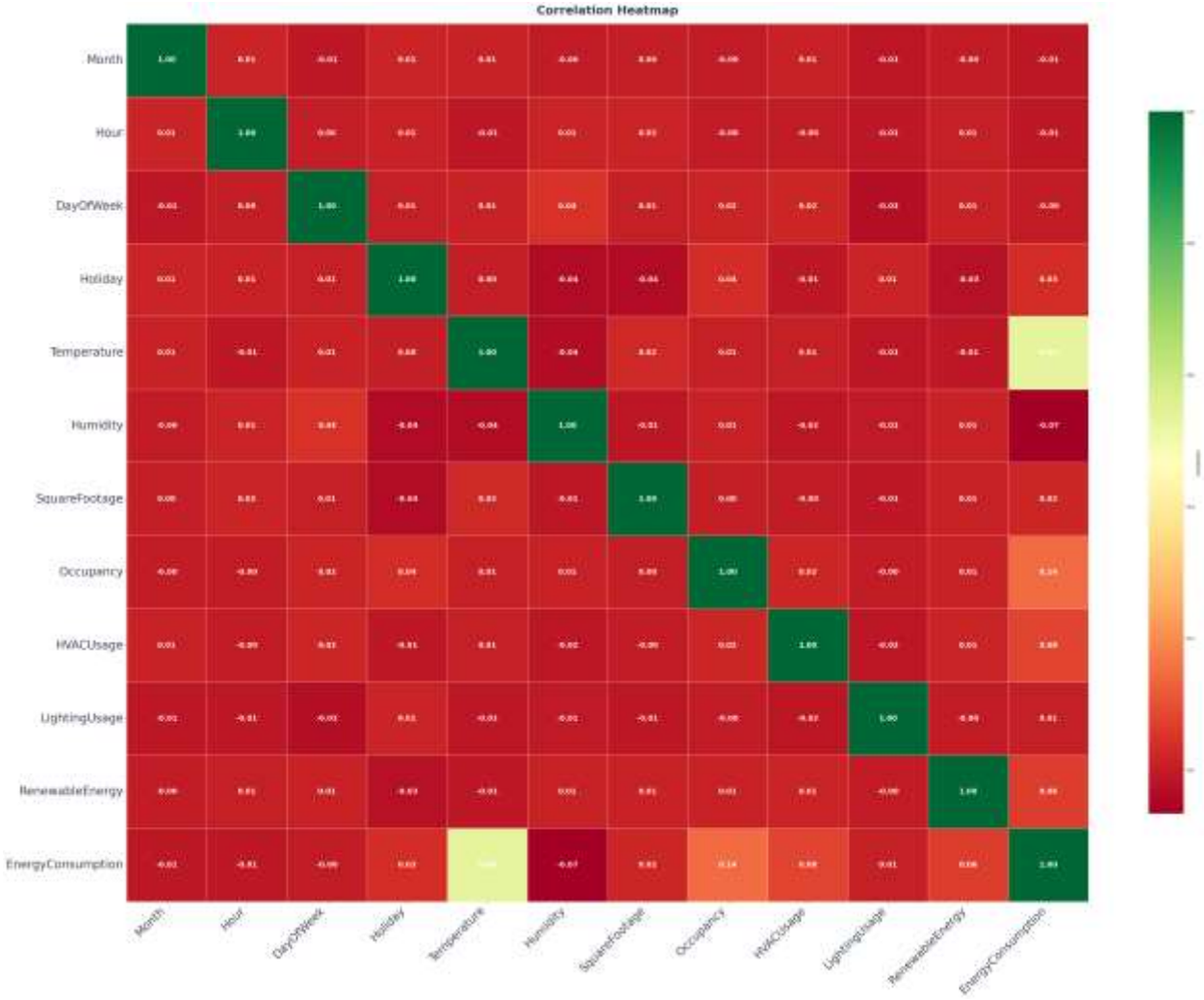
2.2 Exploratory Data Analysis - EDA

Exploratory Data Analysis (EDA) was conducted to gain insight into the distribution and relationships among variables. Various statistical and visualization techniques were employed, including:

- Histograms and kernel density estimation (kde) plots to assess the distribution of numeric features.

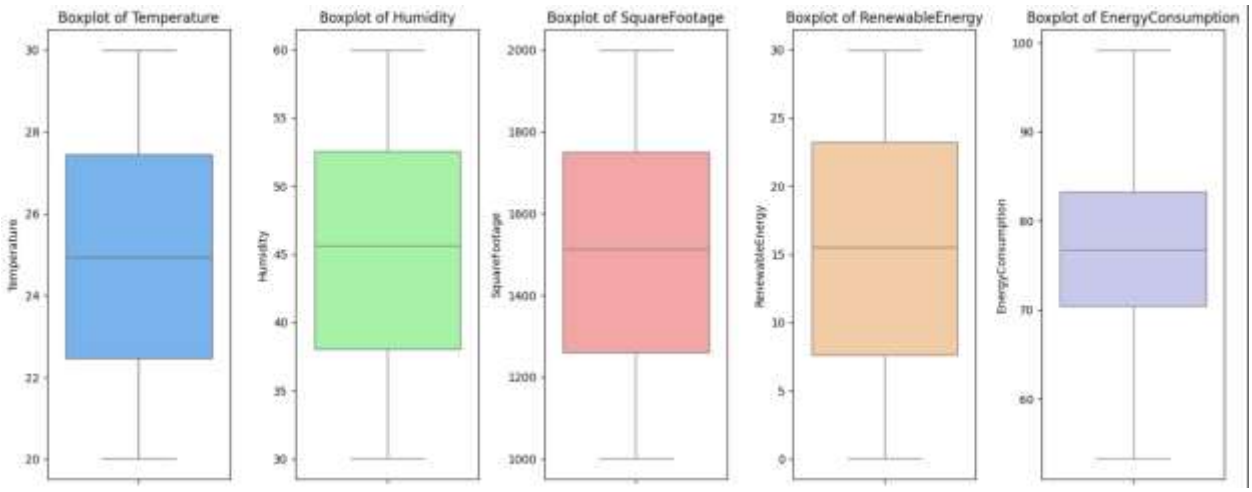


- **Correlation heatmaps** to examine relationships between independent and dependent variables.

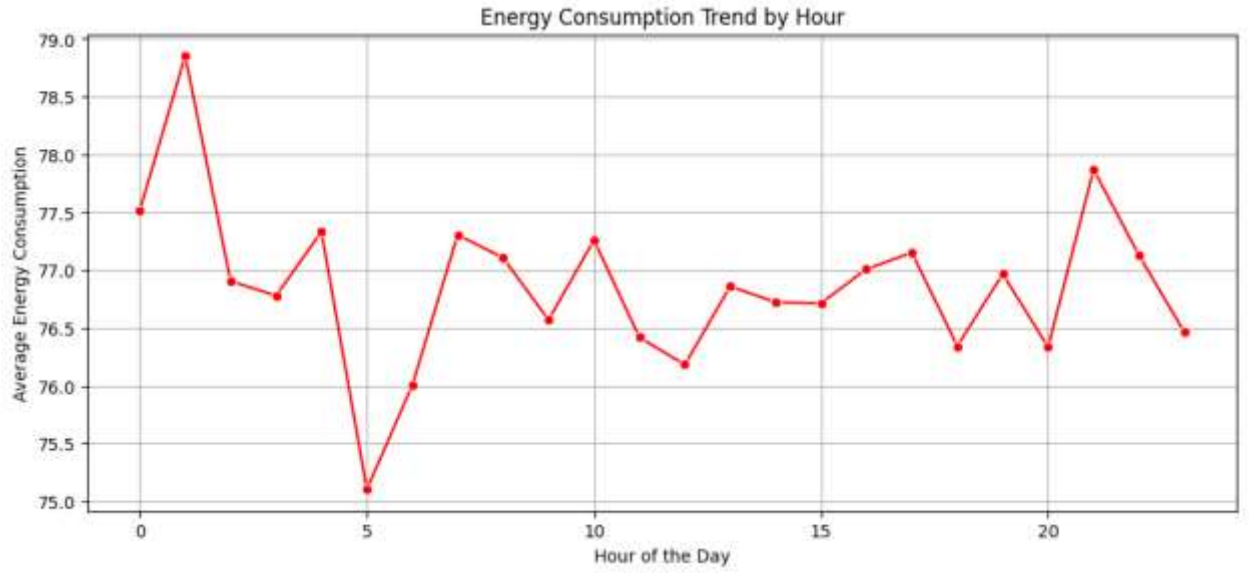


- **Boxplots and line plots** to detect potential outliers and trends.

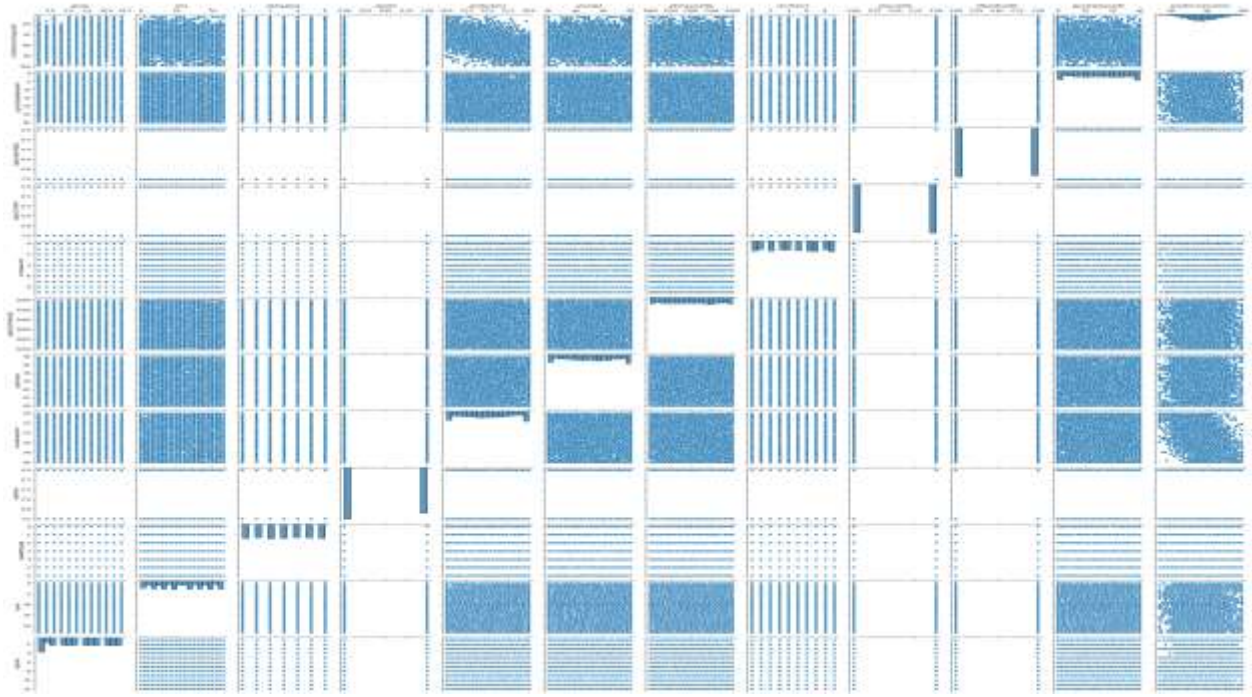
  Boxplots:



  Lineplot:



- **Pair plots** to explore interactions between different features.



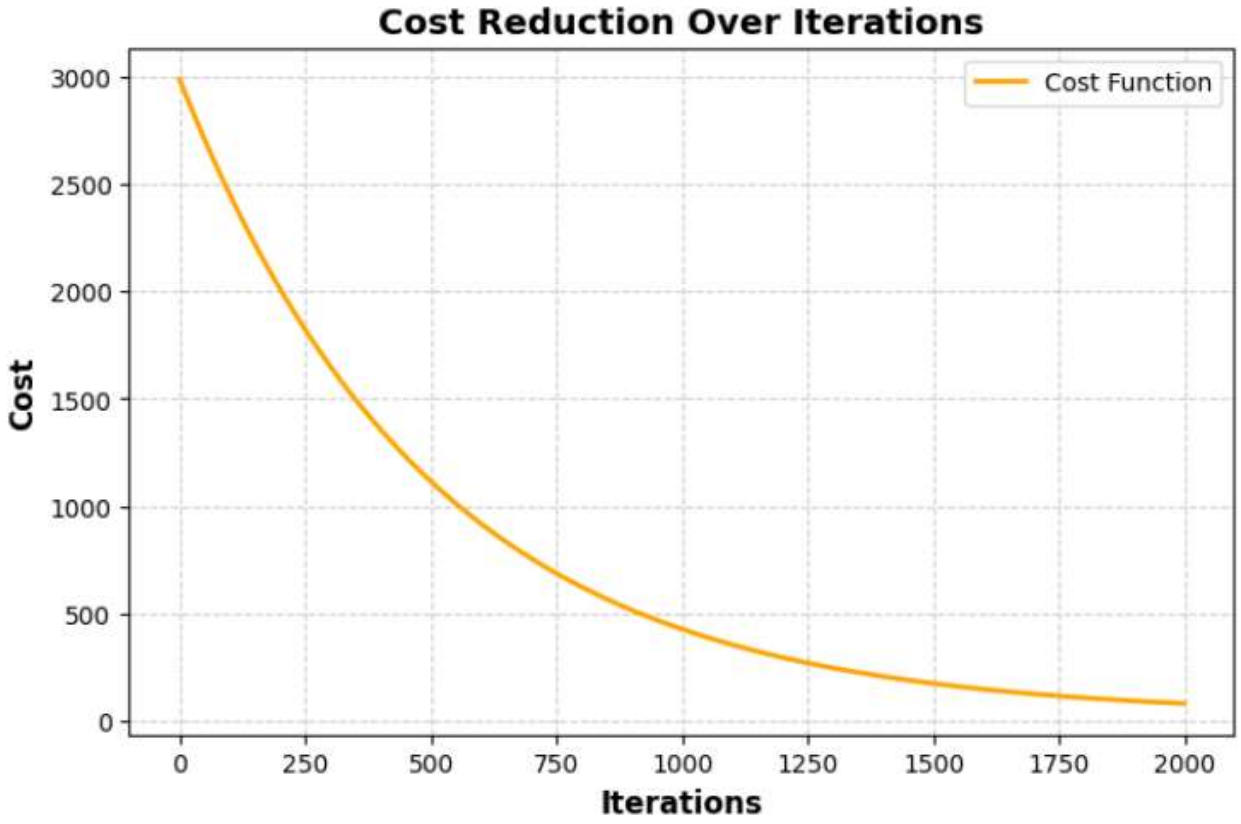
The analysis revealed that energy consumption exhibits a near-normal distribution, with minimal skewness. Additionally, certain independent variables displayed strong correlations with energy usage, highlighting their importance in model building.

2.3 Model Build

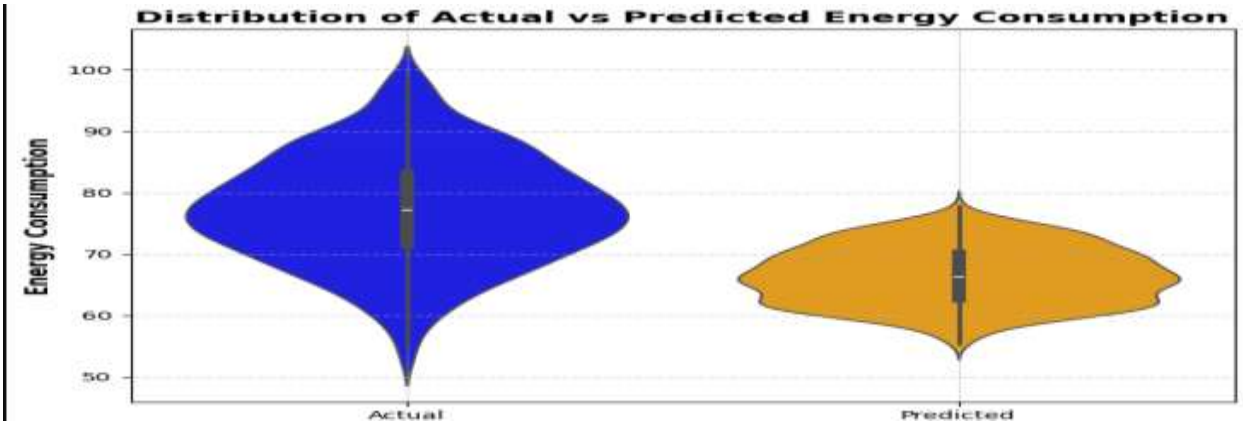
To construct a robust predictive framework, three regression models were implemented:

1. **Linear Regression (from scratch)** – A basic statistical method that presupposes that independent and dependent variables have a linear relationship.

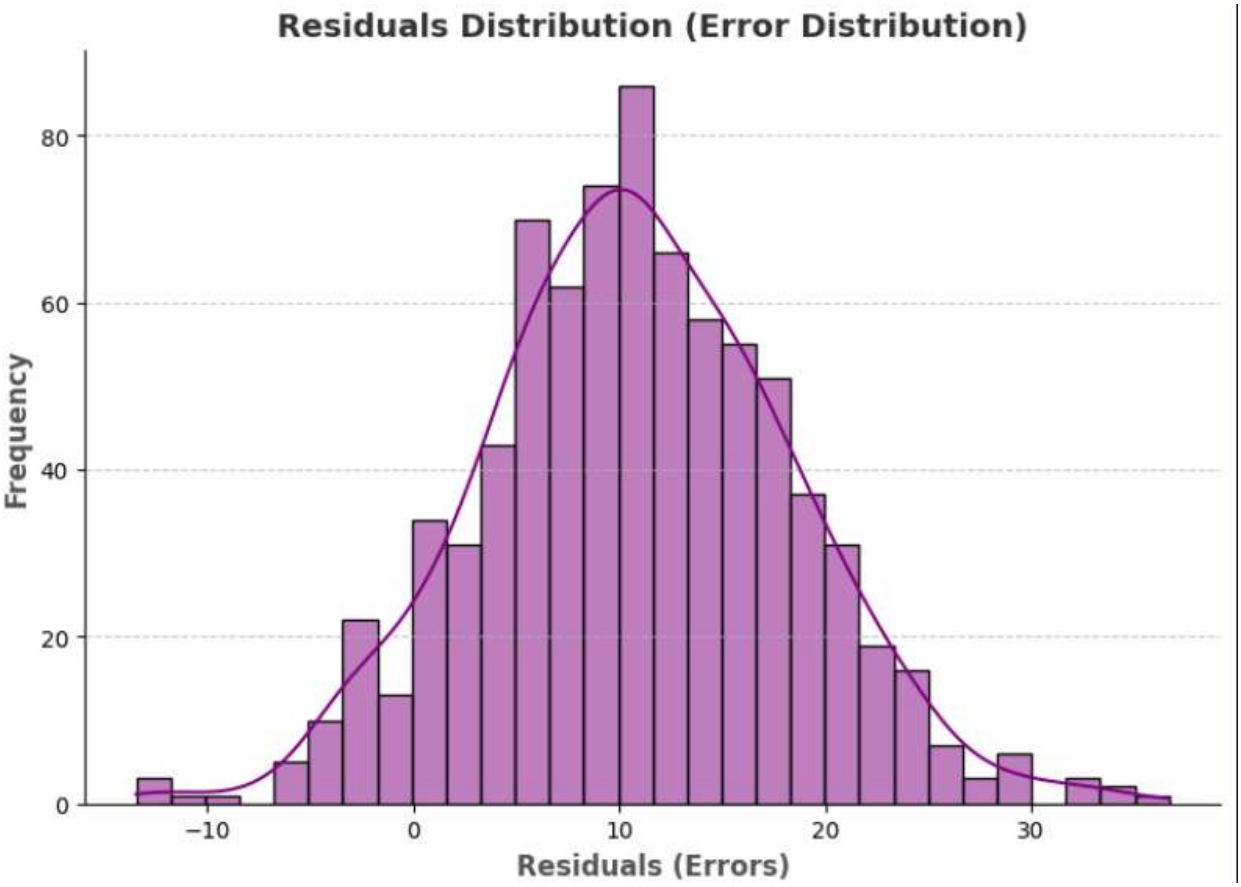
    - Plot which shows cost reductions over iterations:



    - Violin Plot showing distribution of Actual , Predicted Energy Consumption:

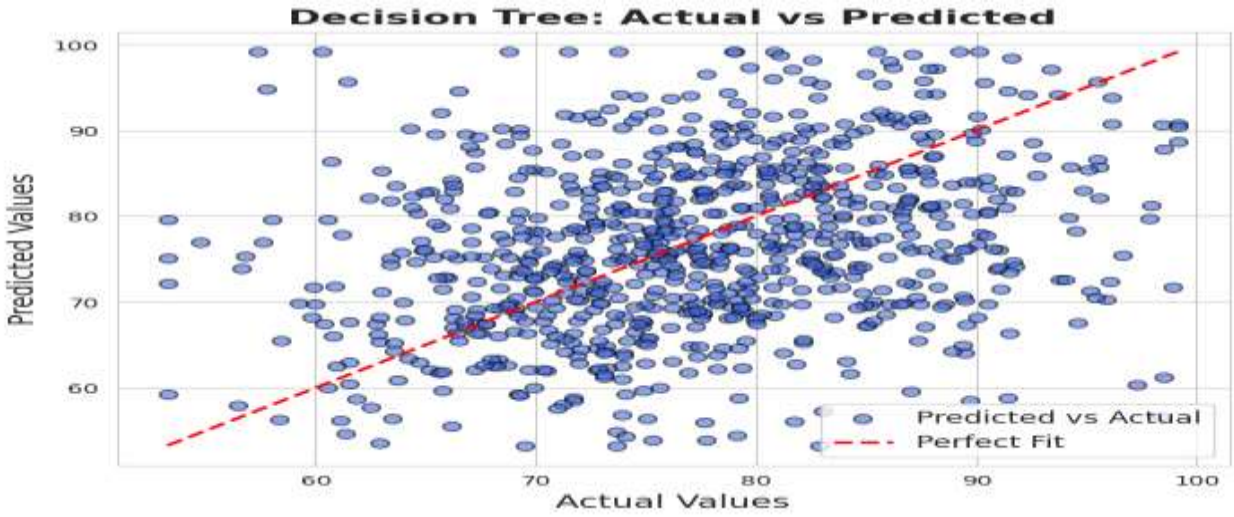


- Histogram Plot showing the Residual Distribution:

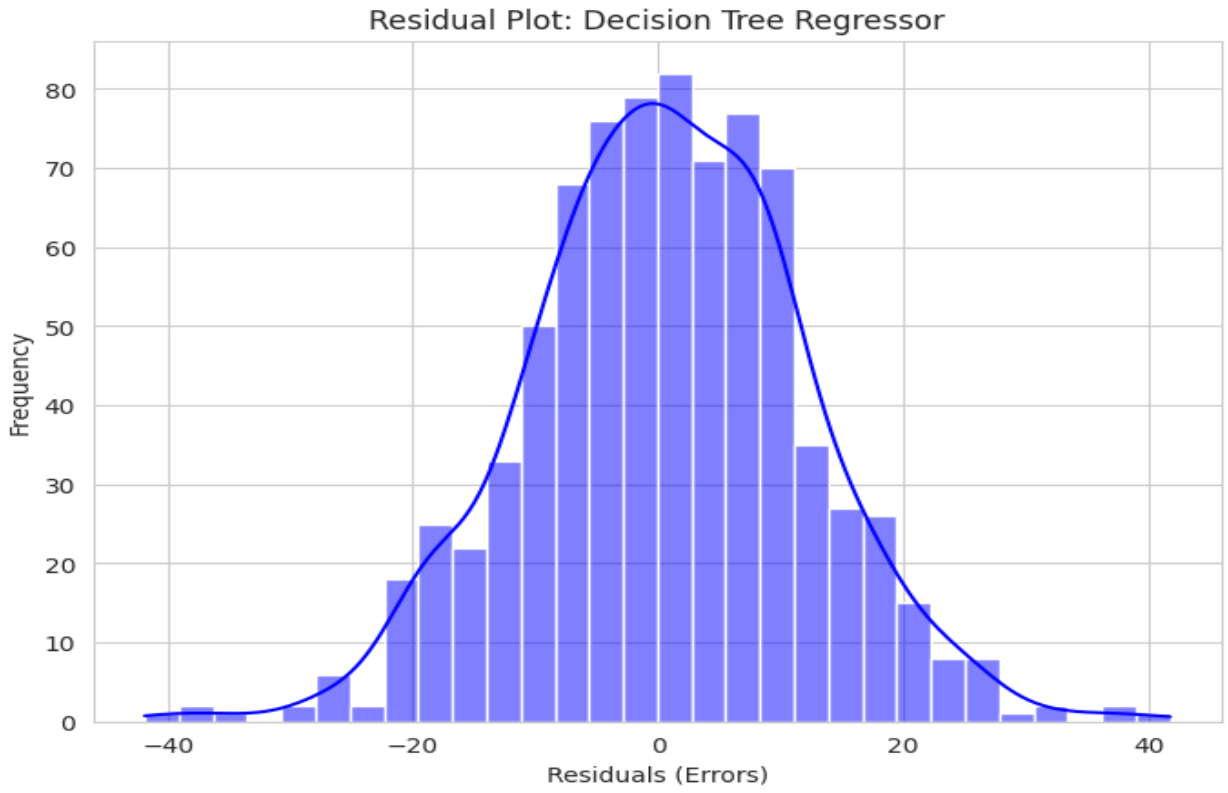


2. **Decision Tree Regression** – A non-linear approach that captures complex interactions between variables.

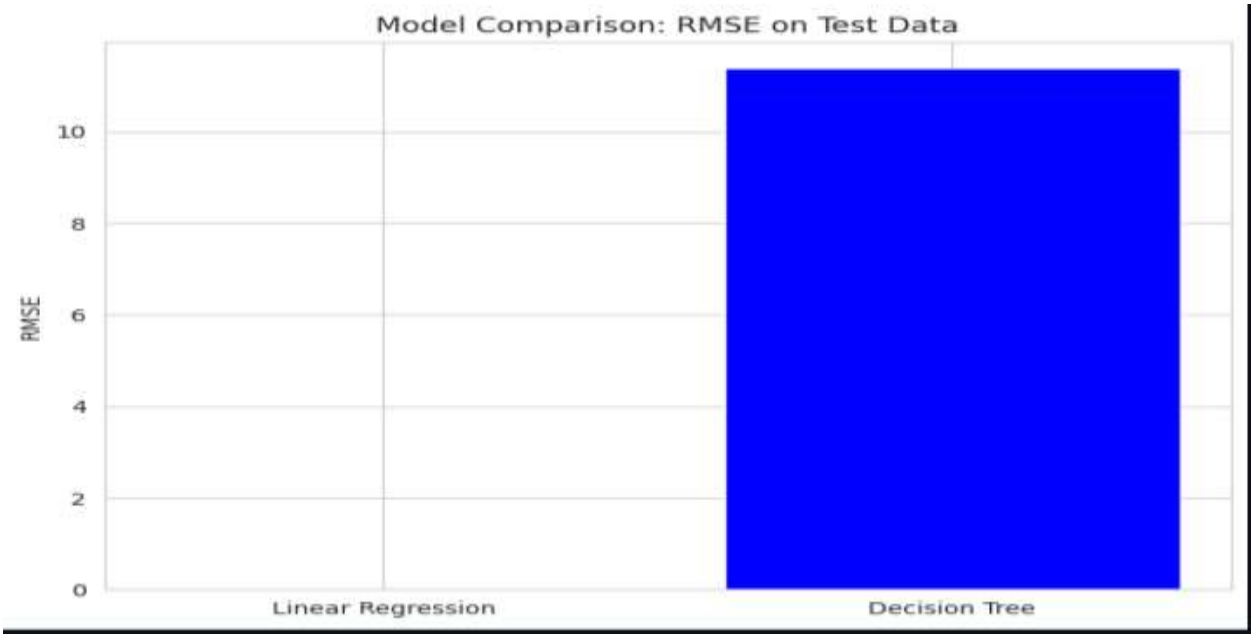
- Scatter Plot of Decision Tree: Actual and Predicted:



- Residual Plot: Decision Tree Regressor:

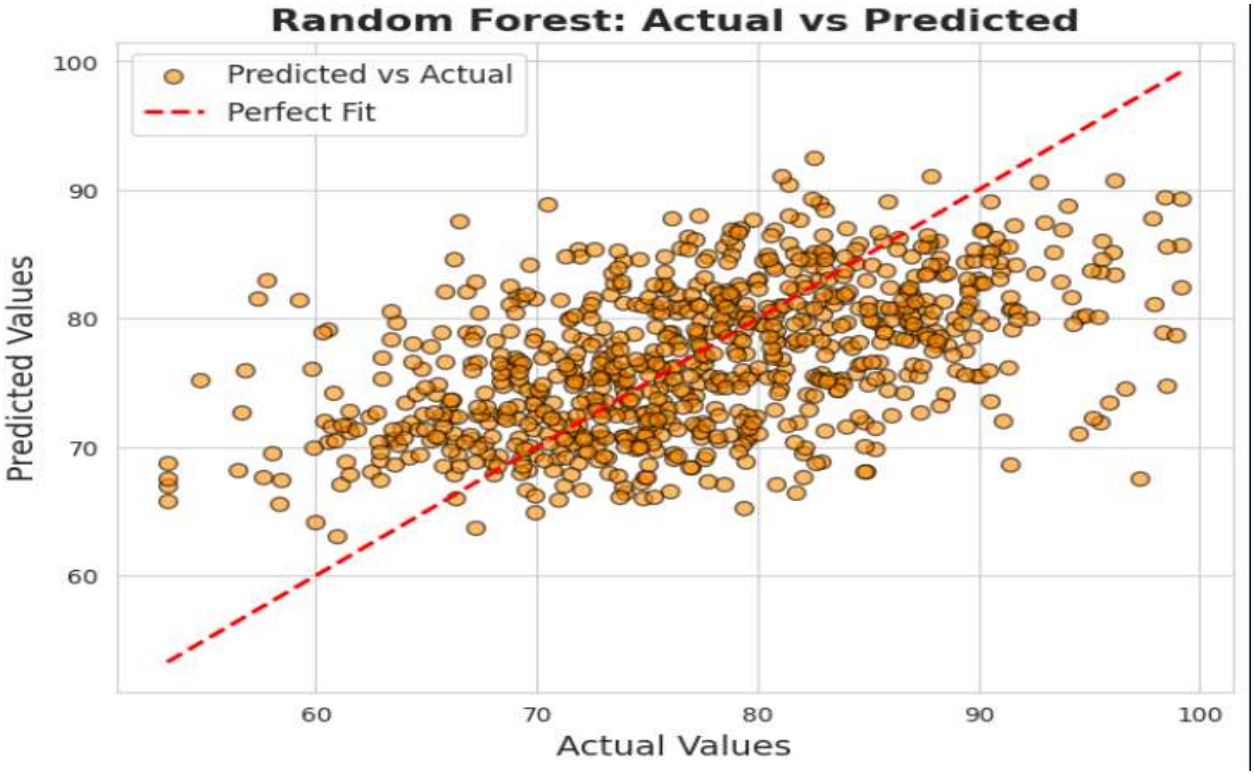


- Model Comparison: RMSE on Test Data:

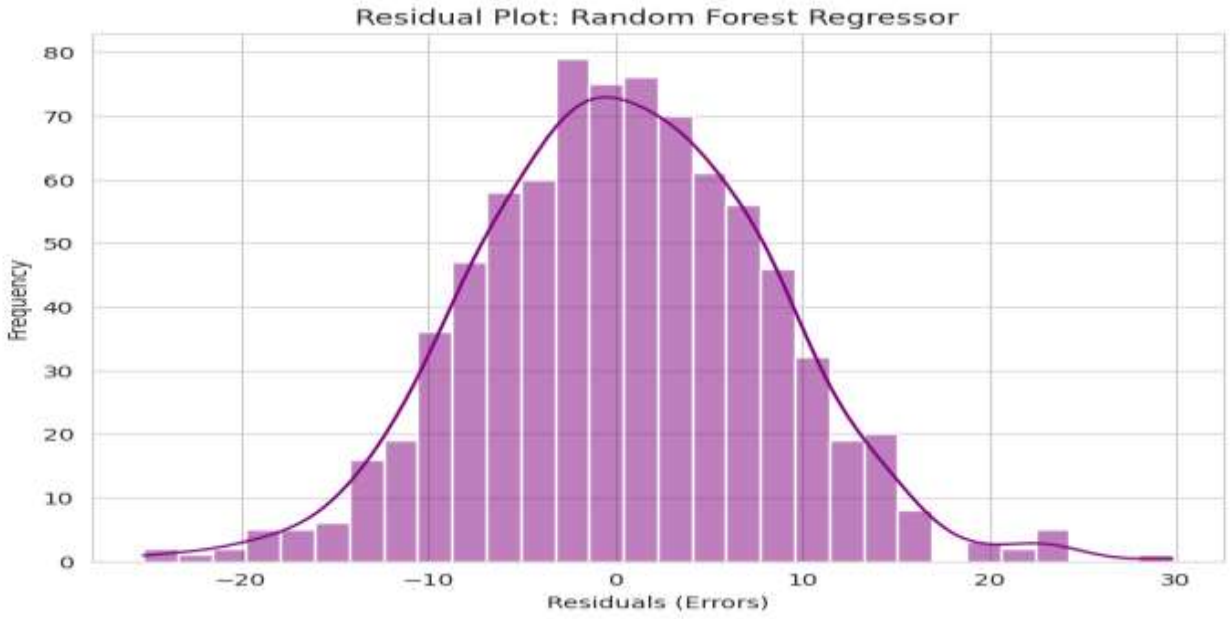


3. **Regression From Random Forest**– An ensemble learning technique that combines several decision trees to improve prediction accuracy.

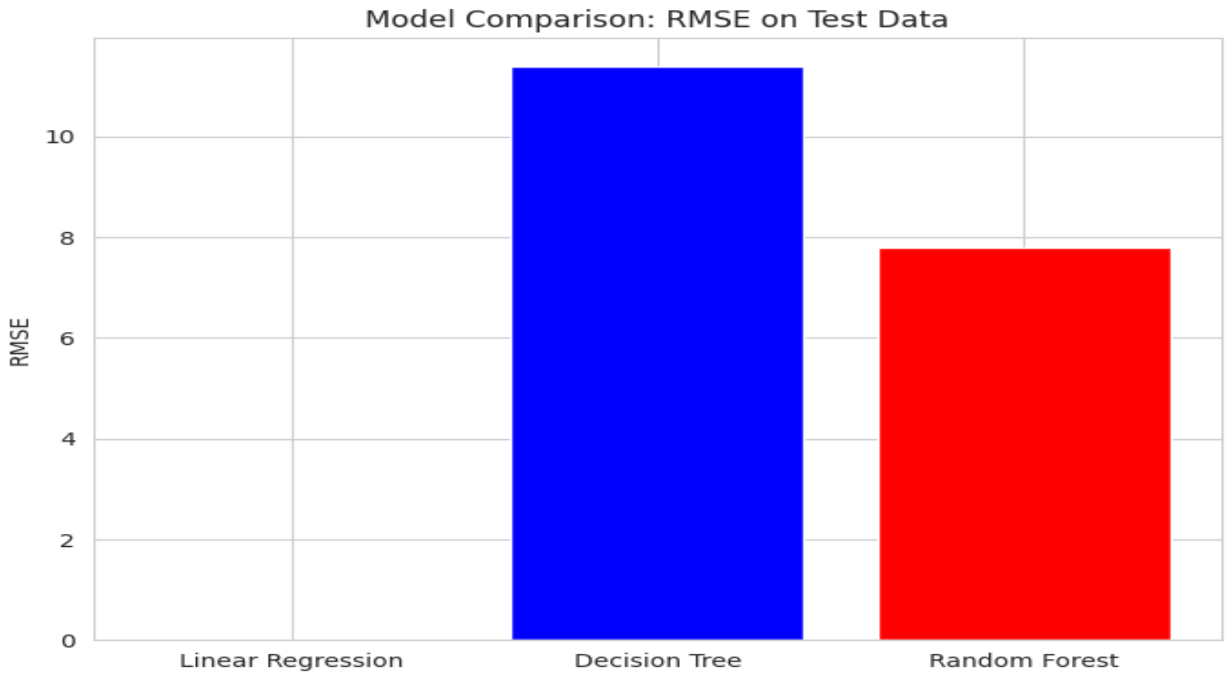
- Scatter Plot: Actual and Predicted:



- Residual Plot: Random Forest Regressor:



- Model Comparison: RMSE on Test Data:



The dataset was partitioned into training and testing subsets using an 80-20 split to ensure unbiased model evaluation.

2.4 Model Evaluations

The models were assessed using key performance metrics:

- **R-squared (R^2):** The proportion of variance in energy consumption is indicated as explained by the independent variables.
- **Mean Squared Error (MSE):** Calculates the average squared difference between expected and actual values; higher performance is indicated by lower values.

2.5 Hyper Parameter Optimizations

To enhance the efficiency of predictive models, GridSearchCV was employed to systematically search for optimal hyperparameter configurations. This process involved systematically searching for the best parameter combinations that maximize predictive accuracy. The optimal hyperparameters identified were:

- **Decision Tree Regression:**

  Best Decision Tree Hyperparameters: {'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 10}

  Decision Tree Training RMSE: 7.5188

  Decision Tree Test RMSE: 7.9769

- **Random Forest Regression:**

  Best Random Forest Hyperparameters (RandomizedSearchCV): {'n_estimators': 500, 'min_samples_split': 20, 'min_samples_leaf': 10, 'max_depth': 5}
  Random Forest Training RMSE: 7.3275
  Random Forest Test RMSE: 7.5953

- Model Performance Comparison:

Model	Best Hyperparameters	Train RMSE	Test RMSE	Conclusion
Decision Tree	max_depth=5, min_samples_leaf=1, min_samples_split=10	7.5188	7.9769	Slight overfitting, but decent performance
Random Forest	n_estimators=500, min_samples_split=20, min_samples_leaf=10, max_depth=5	7.3275	7.5953	**Best model**—lowest test RMSE, best generalization

2.6 Feature Selections

Feature selection was performed to identify the most significant predictors of energy consumption using Recursive Feature Elimination (RFE). The selected features included:

- Temperature
- Humidity
- SquareFootage
- Occupancy
- RenewableEnergy

3. Conclusions

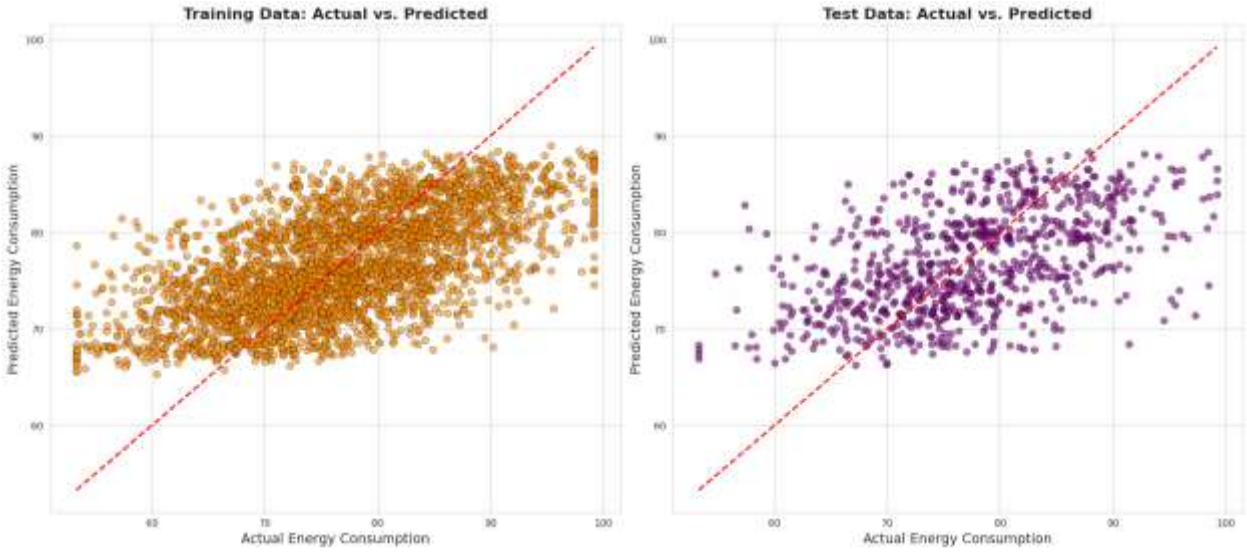
3.1 Key Finding

The most effective model was Random Forest Model Performance, achieving an R^2 value of 0.3801, MSE value of 54.3869 and RMSE value of 7.3748 on training data. Similarly, achieving an R^2 value of 0.2630, MSE value of 57.6282 and RMSE value of 7.5913 on testing data. Feature selection significantly improved model interpretability and accuracy by reducing redundant information and focusing on the most relevant predictors.

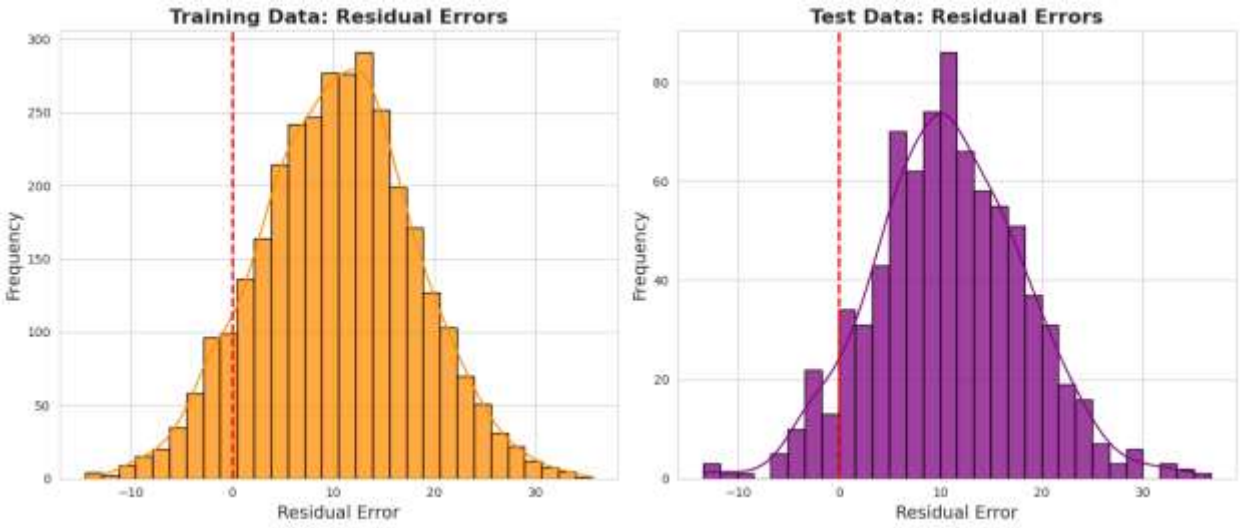
3.2 Final Models

The final model chosen was Random Forest Model Performance, as it provided the most reliable predictions while maintaining interpretability.

- Scatter Plot of Training and Testing Data: Actual vs Predicted:



- Histogram Plot of Training and Testing Data: Actual vs Predicted:



3.3 Challenges

Several challenges were encountered during the analysis, including:

- Managing missing values and ensuring data completeness.
- Balancing model complexity and interpretability.
- Computational costs associated with hyper-parameter tuning.

3.4 Future Works

To enhance the model's effectiveness, future studies could explore:

- Advanced regression techniques such as XGBoost.
- The integration of real-time energy consumption monitoring.
- The inclusion of additional external factors such as economic indicators and policy changes.

4. Discussions

4.1 Model Performances

The results suggest that Random Forest exhibited superior performance in predicting energy consumption. The findings reinforce the importance of selecting an appropriate modeling approach based on data characteristics.

Comparison of Random Forest Performance

Metric	Before Tuning	After Tuning
Training MAE	2.3168	5.8689
Test MAE	6.1846	6.0317
Training MSE	8.6644	54.3869
Test MSE	60.6401	57.6282
Training RMSE	2.9435	7.3748
Test RMSE	7.7872	7.5913
Training R^2	0.9013	0.3801
Test R^2	0.2245	0.2630

4.2 Impact of Hyper-parameter Tuning and Feature Selections

The application of hyper-parameter tuning and feature selection led to a significant improvement in model performance, reducing error rates and enhancing predictive accuracy.

4.3 Interpretation of Results

The study identified key predictors of energy consumption, offering valuable insights for optimizing energy efficiency and policy development.

4.4 Limitation

Despite its effectiveness, the model has certain limitations, including:

- Potential biases due to dataset constraints.
- Assumptions regarding feature relationships that may not fully capture real-world dynamics.

4.5 Suggestions for Future Research

Further research could involve:

- Applying deep learning methodologies to enhance predictive capabilities.
- Expanding the dataset to include broader geographical and temporal scopes.
- Investigating the impact of emerging energy-efficient technologies on consumption patterns.