Predictive Modelling for Power Consumption in Tetouan, Morocco Using Machine Leaning Method

MSc Research Project
Data Analytics

Etinosa Eghaghe
Student ID: x23138548

School of Computing
National College of Ireland

Supervisor: Naushad Alam

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Etinosa Eghaghe
Student ID:	x23138548
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Naushad Alam
Submission Due Date:	16/09/2024
Project Title:	Predictive Modelling for Power Consumption in Tetouan, Morocco Using Machine Leaning Method
Word Count:	XXX
Page Count:	26

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Etinosa Eghaghe
Date:	16th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predictive Modelling for Power Consumption in Tetouan, Morocco Using Machine Leaning Method

Etinosa Eghaghe
x23138548

Abstract

As the world's population continues to grow, the demand for electricity consumption is on the rise, necessitating accurate prediction to meet increasing energy needs. This research uses machine learning techniques to predict power consumption across three zones in Tetouan, Morocco, a city facing fast Urbanization and increasing demands for energy. Accurate power consumption in this region are important for good energy management and planning to help reduce cost and ensure consistent power supply. Focusing on four machine learning models, Decision Trees, Random Forests, Long Short-Term Memory (LSTM) networks, and XGBoost and using historical dataset comprising DateTime, Temperature, Humidity, Wind Speed, General Diffuse Flow, Diffuse Flows, and zone-specific power consumption variables, from the UCI Machine Learning Repository, the objective of this study is to evaluate the predictive accuracy of the these models, identify main predictors of power consumption and assess their computational efficiency. The findings indicate that the XGBoost model provides the highest predictive accuracy followed by the Random Forest model. The LSTM model, effectively capture temporal dependence's, making it good for sequential predicting. The Decision Tree model serves as baseline with lower performance compare to the other models.

This research contributes to the field of energy management and demonstrates the effectiveness of advanced predictive modeling techniques narrowed to the unique characteristics of Tetouan. The knowledge gained can help in optimizing energy distribution, reducing cost and promoting sustainable development initiatives in Tetouan's region.

1 Introduction

As the world evolve and develop, population increases, forecasting power consumption becomes more and more important. It is a critical task for energy management and planning and hence knowing the specific amount of power to supply and when to supply it to customers is critical. Furthermore, accurate predictions can help in optimizing energy distribution, thereby reducing cost and ensuring a stable power supply. This study focuses on four popular machine learning techniques Decision Trees, Random Forests, LSTM networks and XGBoost to model and predict power consumption in three zones in Tetouan Morocco, a region experiencing rapid urban development and energy demands is increasing.

Prediction of power consumption is a difficult task that is influenced by many factors such as temperature humidity and wind speed. More effective models can support consumers in managing their energy usage more efficiently and help policymakers in energy sector to plan better and distribute resources properly.

1.1 Motivation

This study is driven by the growing importance of the efficient energy management in light of the increasing energy demand and the push for more sustainable energy practices. Accurate power consumption projection are more important in areas such as Tetouan, Morocco, because the region is growing in terms of urbanization and population expansion. Precise predictions can help in reducing the dangers connected to energy scarcity and wasteful energy consumption, hence promoting sustainable development objectives. Previous studies have showed potential of machine learning models, in predicting energy consumption in many regions with varying degrees of success. However, due to variation in results, there is a lack of localized studies that considers the unique characteristics and needs of Tetouan. This gap in the literature highlights the needs for tailored solutions that leverage modern predictive modeling techniques to support both energy providers consumer in a specific region.

1.2 Research Question/ Objective

The main research question guiding this study is: How well can machine learning models such as Decision Trees, Random Forests, Long Short-Term Memory(LSTM) and XGBoost models predict power consumption in Tetouan, Morocco? The objectives of this research are to evaluate predictive accuracy of these models, identify the most significant predictor of power consumption, and assess the computational efficiency.

1.3 Research Contribution

This research provides a comparative analysis of several machine learning models in aspect of power consumption prediction, using the KDD(Knowledge Discovery in Databases) methodology, this study make used of historical weather and power consumption data from the UCI machine learning repository[1]. Our study provides important insights into the effectiveness of predictive modeling techniques for energy management in a particular geographic area by customizing the model to the unique features of Tetouan energy consumption patterns and improving their parameters.

This report is structured as follow:

- **Introduction**: This section includes Background, Motivation, Research Question/ Objective,Research Contribution.
- **Related work**: This section comprehensively discusses existing literature on power consumption predicting using machine learning approaches.
- **Methodology**: Section 3 presents a detailed description of the research process, which include data collection, preprocessing, feature engineering, model training and evaluation

[1] Dataset Source: `https://archive.ics.uci.edu/dataset/849/power+consumption+of+tetouan+city`

- **Design Specification**: This section provides an explanation of the system architecture and the requirement for implementing the predictive models.

- **Implementation**: The implementation section outlines the steps involved in developing and evaluating the predictive models, including data transformation, feature extraction and model training.

- **Evaluation/Result**: A comprehensive analysis of the model performance, including visualization and statistical assessments of the results.

- **Conclusion and Future Work**: Summary of main findings, discussion of the research implications, limitations, and suggestions for future research directions

2 Related Work

Predictive power consumption modelling is essential for energy management as it guarantees sustainability and efficiency. This review of the literature critically examines many approaches that has been used to forecast electricity consumption, with an emphasis on research carried out using various modelling techniques and in a variety of geographical areas. The objective is to evaluate the efficacy of these approaches and suitability for Tetouan, Morrocco.

2.1 Ensemble Learning Techniques

Ensemble learning techniques have increase in popularity in power consumption prediction because of the model's ability to improve prediction accuracy and stability. Ves et al. (2019) proposed a stacked ensemble model combining Gradient Boosting Regression, Multi-Layer Perceptron (MLP) and Long Short-Term Memory(LSTM) networks. The model demonstrated high accuracy with mean Absolute percentage Error (MAPE) of 1.59% on aggregated data, showing the effectiveness of ensemble learning in improving stability. However, the model's performance is highly dependent on the quality and size of the training data, which may not generalize well across different datasets. Similarly, Priyadarshini et al. (2022) developed an ensemble model combining Decision Trees, Random Forests, and XGBoost, which also improved prediction performance, with coefficient of determination also known as R-squared(R2) close to 99% and reduced variance of prediction errors. Despite these benefits, the ensemble model introduced potential computational overhead and complexity in turning multiple model parameters, though it achieved superior accuracy compared to individual models. On the other hand, Blaszczyk (2022) conducted a comparative study of various ML algorithms for energy consumption prediction using the ASHRAE dataset. This research underscored the efficiency of simpler models and their competitive performance compared to more complex ones in terms of accuracy and computational efficiency. However, the study did not extensively explored the ensemble methods, which could leverage the strengths of individual models for better performance. Mystakidis et al. (2023) evaluated various ensemble techniques for time-series forecasting, finding that ensemble techniques outperformed single models in improving forecasting accuracy. However, potential overfitting with limited data necessitated careful selection of model parameters and validation techniques.

2.2 Hybrid Machine Learning Models

Hybrid models combine different algorithms to exploit their complementary strengths. Fouad et al. (2020) reviewed ML applications in smart grids, focusing on hybrid methods combining meta heuristic algorithms with traditional models like support vector machine (SVM). Their innovative combination of algorithms demonstrated enhanced performance, though it also brought increased complexity and computational requirements, potentially hindering real-time application. Khan et al. (2021) utilized a hybrid approach combining Support Vector Machine(SVM) and genetic algorithms for energy forecasting, which improved model accuracy and reduced computational time. However, this approach added complexity to the model training process, and the results showcased the potential of hybrid models for energy forecasting. Nan et al. (2022) employed a hybrid LSTM and autoregressive integrated moving average(ARIMA) model for short-term forecasting, effectively capturing both linear and nonlinear patterns in the data. Despite high computational requirements, the model demonstrated high accuracy and the ability to capture complex consumption patterns, validating the efficacy of hybrid approaches. Shin and Woo (2022) developed a hybrid model integrating decision trees and ensemble methods for energy forecasting, achieving improved robustness and accuracy of predictions. The hybrid approach is computationally intensive: however, require large amount of computational power and memory for execution, and required extensive parameter tuning.

Cao et al. (2023) proposed an integrated energy consumption prediction model for educational buildings, incorporating spatial characteristics and using cooperative game theory for feature analysis. The model significantly reduced Root Mean Square Error(RMSE) by 13.64%-34.55% and MAE by 10.25%-30.54%, showing higher prediction accuracy compared to other models. However, the need for feature engineering and data preprocessing added complexity to the model development process.

2.3 Deep Learning Techniques

Deep learning techniques have been particularly effective in capturing complex patterns in energy consumption data. Wang et al. (2020) utilized LSTM networks for long-term energy consumption prediction, focusing on periodicity. Their work demonstrated higher prediction performance compared to traditional methods such as autoregressive–moving-average(ARMA) and autoregressive fractionally integrated(ARFIMA), but faced challenges in handling missing data and required extensive tuning of the LSTM model parameters. da Silva and de Moura Meneses (2023) compared LSTM and Bi-directional LSTM (BiLSTM) models for short-term electric consumption forecasting. They found that BiLSTM outperformed LSTM with statistically significant results, although their focus on univariate time series potentially missed the complexity of multivariate dependencies. Their study provided a baseline for future studies, emphasizing BiLSTM's better accuracy and robustness. Faiq et al. (2023) integrated CNN and LSTM models for short-term consumption forecasting, achieving superior results across all validation metrics compared to individual models. Despite high prediction accuracy, the hybrid model involved high computational costs and complexity in integrating different models. Akbari-Dibavar et al. (2020) Applied LSTM and Grated Recurrent Unit(GRU) models for energy consumption prediction, demonstrating significant error reduction compared to traditional methods. However, extensive data preprocessing and model tuning were required, and the results showed substantial improvements in prediction accuracy, validating the effectiveness of these models for energy forecasting. Olu-Ajayi et al. (2022) also use deep learning techniques using large dataset of residential buildings. They compared many models which include Artificial Neural Network(ANN), Deep Neural Network(DNN), Gradient Boosting and Random Forest, for energy consumption prediction, The DNN models perform highest in their study. though DNNs showed good potential for accurate prediction, required a large amount of data for training and significant computational resources.

2.4 Tree-Based and Boosting Models

Tree-based models, particularly Random Forest and XGBoost, have shown strong performance in energy forecasting. Tan et al. (2023) used Random Forest classifiers to predict electricity consumption levels, effectively handling both numeric and categorical data. The study highlighted the robustness of Random Forest against over fitting, though performance degraded with imbalanced datasets and the method could be computationally intensive with large datasets.

Gökçe and Duman (2022) compared the performance of simple regression, Random Forest, and XGBoost algorithms for electricity demand forecasting. XGBoost provided the best performance among the tested algorithms, demonstrating its suitability for electricity demand forecasting, although the study did not explore more advanced deep learning techniques. Zhou et al. (2021) utilized ensemble learning techniques combining Random Forest,Gradient Boosted Decision Tree, Extreme Gradient Boosting, Light Gradient Boosting Machine(LightGBM) and Categorical Boosting(CatBoost) for energy forecasting, achieving high accuracy.In their study Gaussian process Regression had the highest performance, but tree-based models for example XGBoost and LightGBM also show high accuracy and stability. These models show high efficiency in nonlinear problems. However, there is need for extensive hyper-parameter tuning and had increased computational cost due to the integration of multiple models.

2.5 Adaptive and Optimization Techniques

Adaptive models and optimization techniques have also been use to improve power consumption predictions performance through fine tuning models. The Radial Basis Function (RBF),Moth-Flame Optimization(MFO),Particle Swarm Optimization (PSO), Multi-Verse Optimizer (MVO), and Artificial Bee Colony (ABC),Grey Wolf Optimizer (GWO) and its advance Grey Wolf Optimizer(AGWO), was proposed by Shanmugam and Ramana (2024) to optimize energy consumption models improving accuracy in complex optimization problems. In their evaluation, the result show that Radial Basis Function (RBF) combined with advance Grey Wolf Optimizer(AGWO) performed best. While Saglam et al. (2023) combines Radial Basis Function (RBF) neural network with metaheuristic algorithms like particle Swarm optimization (PSO) and Artificial Bee Colony (ABC). These techniques perform better than the traditional approaches in terms of accuracy Khafaga et al. (2023) introduced a hybrid method using Throated Optimization (DTO) and Stochastic Fractal Search (SFS) algorithms to optimize Long Short-Term Memory (LSTM) networks, RMSE seems lower in their approach, making it effective in capturing complex consumption patterns. However, their approach uses a complex model and involved the need of extensive data processing.

2.6 Comparative Analysis of Machine Learning Models

Salam and El Hibaoui (2018) and RoSe et al. (2023) both examined the problem of predictive modelling of power consumption using machine learning techniques, with a focus on Tetouan city, Morocco. They explore the performance of models such as feed-forward neural networks, random forest, decision trees, and support vector machines (SVM) for regression. Both studies utilize historical data from Supervisory Control and Data Acquisition (SCADA) systems. Salam and El Hibaoui (2018) specifically emphasizes the use of random forest models, showcasing smaller prediction errors compared to other models.The models were optimized using the grid search method to achieve better and accurate performance. However, their result shows the model being more accurate on the training set as compared to the test set, indicating overfitting.(RF: Train=671.7, Test=3174.7). Overfitting is indicated when a model perform significantly better in training set compared to the test set, suggesting it has learned the noise and specific patterns of the training data rather than generalizing well to unseen data. On the other hand, RoSe et al. (2023) introduces Bayesian Fine Tree (BFT) as an optimized variant for energy demand prediction, which outperforms traditional Fine Tree algorithms. The accuracy was measure by RMSE, MSE, MAE, and R-square. Both studies highlight the effectiveness of machine learning models in predicting power consumption and the importance of energy management, with Salam and El Hibaoui (2018) emphasizing the superiority of random forest, while RoSe et al. (2023) introduced a novel approach with BFT. However, there's no direct comparison between the models studied in these papers. The comparison of ML models in different context suggests the need for further investigation to enhance the accuracy of the current forecasts.

Summary and Justification for Further Research The literature review underscores the diversity of methodologies employed in predicting power consumption, using Machine Learning and Deep learning techniques.

Ensemble models, such as those combining Decision Trees, Random Forests, and XGBoost, generally perform better than single models by improving accuracy and stability. However, they often come with increased computational complexity and require extensive parameter tuning. Hybrid methods show promise by combining the strengths of different algorithms but introduce additional layers of complexity. While existing studies provide valuable insights, there remains a gap in comprehensive comparative analyses across different methodologies in the context of Tetouan, Morocco. Building upon these foundations, this research aims to address existing gaps by evaluating the effectiveness of Decision Tree, Random Forest, LSTM, AND XGBOOST-Regressor models in the context of Tetouan, Morocco. By incorporating localized data and refining model parameters, tailored to the region's unique characteristics, this research contributes to the existing body of knowledge by offering insights into the applicability of predictive modelling techniques in a specific geographic context, thereby informing energy policy decisions and fostering sustainable development initiatives in Tetouan.

3 Methodology

This study is a detail analysis and modeling development, which uses Tetouan dataset collected from UCI machine repository[2]. How well can DT, RF, LSTM and XGBoost predict power consumption in Tetouan Morocco?.To answer this question, KDD methodology is deployed. KDD methodology is a systematic process that seeks to identify valid,novel, potentially useful, and ultimately understandable patterns from large amount of data(Chumbar; 2023).Basically is the process of transforming raw data into knowledge. This involves several stages, which includes data collection, preprocessing, feature engineering, visualization, data transformation,development, model training, evaluation, and comparison of three different zones. This approach ensures a comprehensive analysis, leveraging related work in the field to inform methodology decisions. The stages are discuss in Figure 2 below.

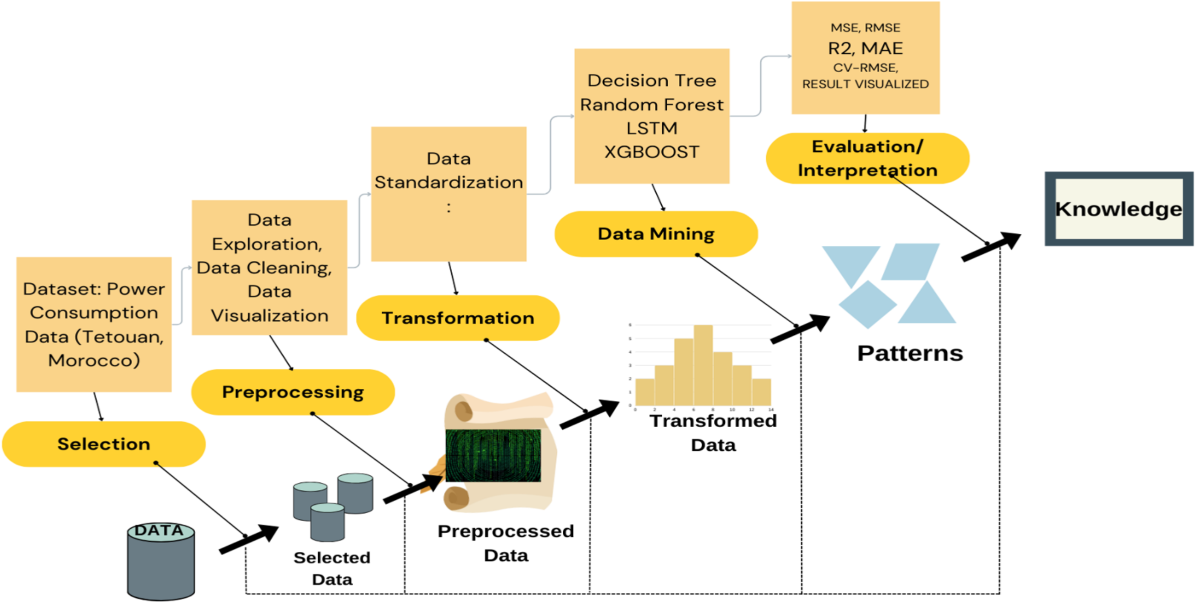


Figure 1: KDD Methodology

3.1 Data Collection and Prepossessing

3.1.1 Data Source and Description

The dataset used in this study includes historical weather and power consumption data from Tetouan, Morocco and was collected from the UCI machine learning repository. It contain 52,416 entries ranging from 01.01.2017 to 01.01.2018 with 9 features, which are DateTime, Temperature, Humidity,Wind Speed, general diffuse flows,diffuse flows, Zone1 Power Consumption, Zone2 Power Consumption, Zone3 Power Consumption.

3.1.2 Initial Data Exploration

- Data import: The dataset was imported into the python environment using pandas, a powerful data manipulation library. This step is crucial as it forms the foundation

[2] Dataset Source: `https://archive.ics.uci.edu/dataset/849/power+consumption+of+tetouan+city`

for all subsequent analysis.

- Data Summary: The datasets was examined to understand its basic statistics using the describe() function in pandas. This provided insight into the mean, median, standard deviation and range of each of the features.
- Data Dimensions: The datasets dimensions were verified using the shape function, revealing it consists of 52,416 rows and 9 columns.
- Data Types: The data types of each column were inspected to make sure compatibility with the modeling process using the 'dtypes' attribute.

3.1.3 Data Cleaning

These are essential steps to ensure that the data is consistent, accurate and ready for analysis which includes handling of missing values, renaming columns, converting data types and creating new features.

- Missing Values: The dataset was checked for missing values using 'isnull().sum()' and there was no missing values.
- Renaming of Columns: Three columns were renamed for clarity using the 'rename()' function.
- Data conversion: The DateTime column was converted to a date time type and set as the index for easier manipulation and analysis.
- Feature Engineering: Additional numerical features Hour, Day and month were extracted from the DateTime column to enrich the dataset. This was done using 'dt' accessor in pandas.

3.2 Data Visualization

Data visualization involve creating graphical representation of the data to understand the patterns, trends and relationships. Various visualizations were created which are discussed as follows,

- Line Charts: Displayed the trends of temperature, Humidity, Wind Speed and power consumption over time using Matplotlib.

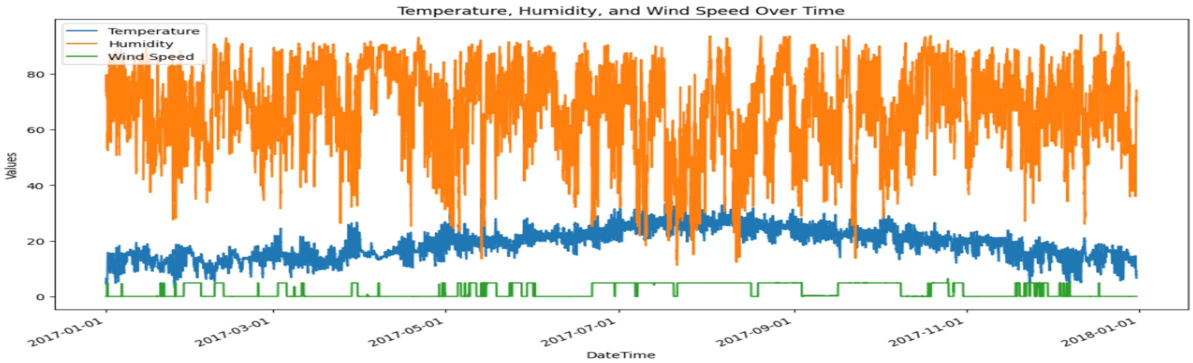


Figure 2: Temperature, Humidity and Wind speed over Time

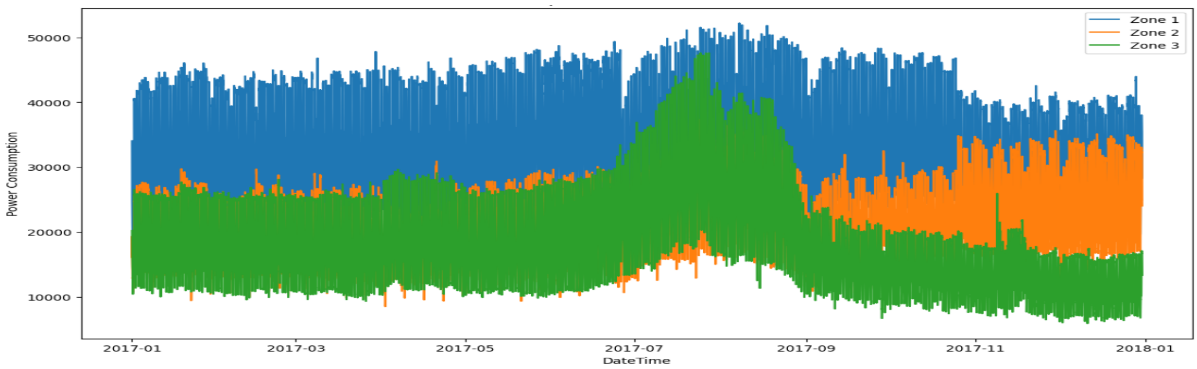


Figure 3: Power Consumption Over Time for the three Zones

- Histograms: This is to get insight into the distribution. It help in understanding the range and frequency of these variables, indicating whether the Data is skewed, normally distributed, or has any outliers.

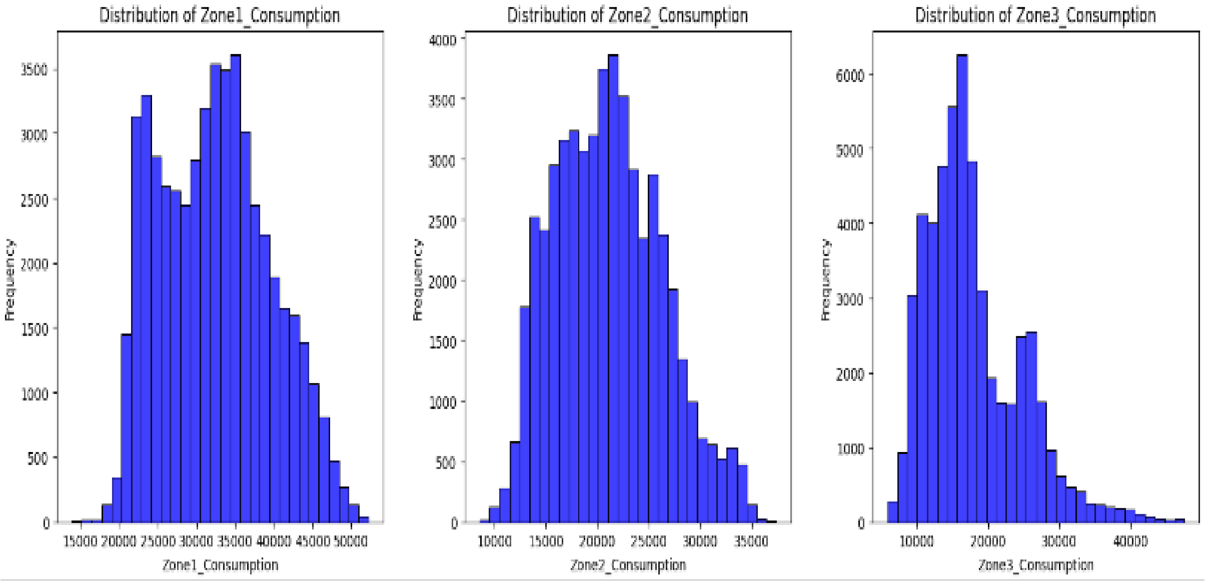


Figure 4: The histograms illustrate the distribution of power consumption across three zones, showing a multimodal pattern in Zone 1 with peaks around 180,000 and 300,000 units, a more normally distributed pattern in Zone 2 with a peak around 22,500 units, and a skewed distribution in Zone 3 with a higher frequency of lower consumption values and a peak around 17,500 units, reflecting varying levels of usage potentially due to different consumer groups, activities, and seasonal variations within each zone.

- Box Plots: This is used to showed the spread and central tendency of key variables. Also highlight any potential outliers. This is useful for quickly comparing the distributions and identifying any anomalies. This is done using seaborn, which is a data visualization package.

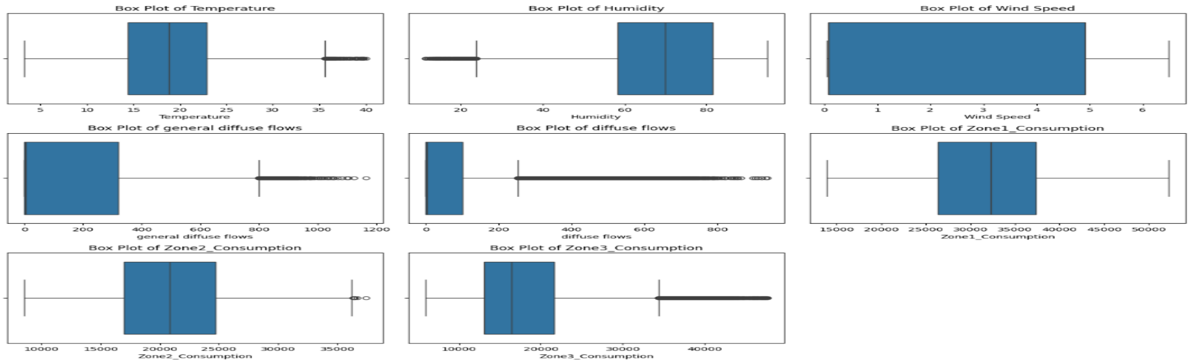


Figure 5: Box plot of Variables: It shows that there are potential outliers in the variables, this are likely due to the seasonal variations of the variable as temperature, humidity, wind speed and power consumption normally fluctuate through out the year.

- Heatmap: Displayed the correlation matrix to identify relationships between variables. using Seaborn 'heatmap()' function.

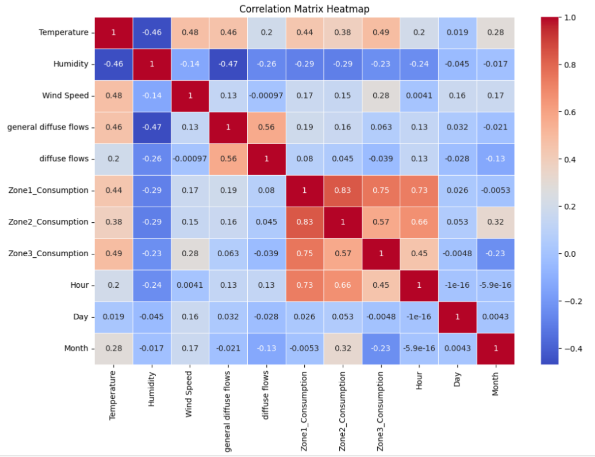


Figure 6: Correlation Plot

3.3 Model Training Methodology

A total of four machine learning and deep learning model were implemented to predict power consumption, they include Decision Tree, Random Forest, LSTM (Long Short-Term Memory) and XGBoost Regressor model. Each of the model was chosen based on it strength's and suitability for the dataset and prediction task.

Target and Feature Variables: The target variable was defined to be the three different power consumption zones respectively with the rest variables as features.

Stratified Split: At this stage, the dataset is split into two subsets: training data for model training and test data for model evaluation. The method used is StratifiedShuffleSplit from scikit-learn, which ensures that the distribution of the target variable is approximately the same in both the training and test sets. This is done by creating a new column, Stratified_Target, using quantile binning (pd.qcut) to ensure stratification. Bins with fewer than two samples are filtered out. This step is crucial as it maintains the representativeness of the splits, particularly in the presence of imbalanced datasets, thereby improving model performance and ensuring the reliability of performance metrics. This method is especially important in this research due to present of seasonality in the data, which needs to be properly represented to ensure good performance.

Standardization: This process is used to scale numeric variable in a dataset for them to have similar scale, this make sure that variable is on common scale to prevent certain variables from disproportionately influencing the model training process. It transforms each data to have a mean of 0 and a standard deviation of 1. In this analysis Features were standardized using standardscaler from 'sklearn.preprocessing' to normalize feature values.

The final stages involves developing the models and evaluating the model, detail in subsequent sections.

4 Design Specification

The predictive modeling system designed for predicting power consumption in Tetouan, Morocco, involves several critical components and requirements, The system's architecture consist of four primary stages which are data collection and storage, data preprocessing, modeling and prediction.

4.1 The System Architecture

The predictive modeling system comprises several key components. First, it involves data collection and storage, where historical data on temperature, humidity, wind speed, and power consumption are gathered and stored. Next, the system includes data preprocessing, which involves cleaning the data, engineering features, and scaling the features to ensure consistency. The modeling component follows, where Decision Tree, Random Forest, LSTM, and XGBoost models are trained using the preprocessed data. Finally, the system incorporates a prediction component that provides real-time prediction and visualization of power consumption.

4.1.1 Requirement

The system's functional requirements include the ability to ingest historical data, preprocess this data, implement and train machine learning models, evaluate the performance of these models, and provide real-time predictions of power consumption. The non-functional requirements specify that the system must achieve high prediction accuracy, efficiently handle increasing data volumes, provide clear visualizations and results, and maintain modular and well-documented code for ease of maintenance.

4.1.2 Algorithm and Model

The system employs several algorithms and models. The Decision Tree Regressor is used to split the data into subsets based on feature values, forming a tree-like structure. The Random Forest Regressor combines multiple decision trees to improve accuracy and control overfitting. The LSTM (Long Short-Term Memory) model captures temporal dependencies in time-series data, making it suitable for sequential prediction tasks. The XGBoost model is an efficient gradient boosting algorithm that builds strong predictive models by combining several weak models.

4.1.3 Data Flow

The data flow within the system begins with data ingestion, where historical data is loaded into the system. This is followed by preprocessing, which includes parsing datetime information and scaling features. The next step is model training, where the data is split into training and testing sets, and the models are trained using the training data. After training, the models are evaluated using the test data, and performance metrics are calculated. Finally, the system provides real-time prediction and visualization, comparing predicted values with actual values to ensure accuracy.

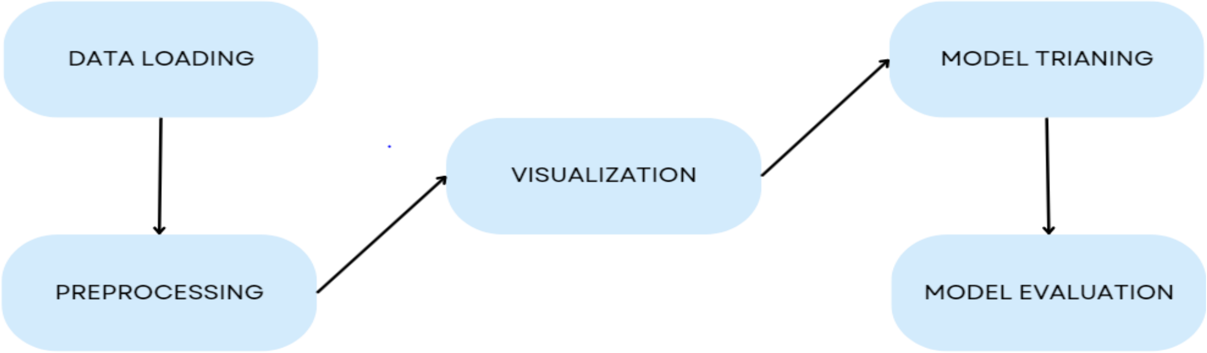


Figure 7: Data Flow

5 Implementation

This involves different stages, which includes data transformation and feature extraction, model training and evaluation, and the visualization of the results The main tools used in implementing were pandas and Numpy for data handling , Scikit-learn,TensorFlow, and Keras for model training , and Matplotlib and Seaborn for visualization

5.1 Data Transformation and Feature Extraction

The first step was to load the dataset from the local device were it was stored unto the python jupyter notebook for futher analysis. The datset was loaded into a pandas DataFrame to begin inspection. show that the dataset consist of 52,416 entries with 9 columns, which are DateTime, Temperature,Humidity,Wind Speed, general diffuse flows,diffuse flows, and power consumption in three zones. The next was to check for the missing values and handle it appropriately to ensure a clean dataset, but there was no missing values in this dataset. The DateTime column was converted to a datetime type and set as the index to facilitate time series analysis. Additional feature were engineered from the DateTime column, such as extracting the Hour, Day and Month to capture temporal patterns in power consumption. Scaling of the Feature: The numerical feature were scaled using the standarscaler from Scikit-learn to make sure they were on a similar scale, which is crucial for the performance of machine learning models.

5.2 Models Implementations

This section covers the implementation details of the four models used in our methodology. The final goal of the implementation stage was to develop and evaluate the four predictive models, output produced is transformed data, trained models and how the model performed by evaluating them.

5.2.1 Decision Tree Model

The first model which is Decision Tree Regressor was initialized using a DecisionTreeRegreesor with a maximum depth of 10 to control the complexity of the tree and avoid overfitting, also 'random state parameter was used to ensure consistency. next a dictionary to store RMSE values for each target variable. For each target, the model was initially trained using 5-fold cross validation and then prediction were done on test set. And then RMSE, MAE, R-Square were reported. Two function were defined to create scatter plots of actual vs. predicted values and to also plot the distribution of residuals.This was done for each target variable, these plots are generated to visually evaluate the model performance.

5.2.2 Random Forest Model

The Random Forest model is initialized with estimators set to 100 and a set random state of 123 for consistency. A dictionary use to store the RMSE values for each target variable. For each target variable, the model was trained using 5-fold cross validation and predictions were made on the test set. The model performance was evaluated using Mean squared Error(MSE), Root Mean Squared Error(RMSE), R-Squared(R2) and Mean Absolute Error(MAE), also cross validation obtained.A plotting function is defined to visualize the actual versus predicted values and residuals. Then the function is called for each target variable to generate the respective plots. Additionally, another function is define to plot the feature importance. which create a bar plot of the feature importance, sorted in descending order.

5.2.3 Long Sort-Term Memory(LSTM)

Developing the LSTM model in this study involve several main steps,which include scaling the features of the training and test data using 'standardscaler' to standardize the data, this is crucial for neural network models to make sure that all features contribute to the models learning process. The inputted data was reshaped into a 3D array format to meet the requirements of LSTM networks, which expects data the form [samples, timesteps,features], where each of the sample is a timestep with the given features. Several hyperparameters are then declared, including 50 LSTM units,a dropout rate of 0.2, a batch size of 32,200 epochs, and a validation split of 0.2 to monitor performance on unseen data. Using Keras 'Sequential' API, the model is constructed, which includes layer for the LSTM units and dropout for regularization, culminating in a dense output layer with the number of units equal to the number of target variables. The model is trained using the ADAM optimizer suitable for regression tasks and uses MSE as the loss function. During the training process, early stopping] is implemented with a 'patience of 10 epochs to prevent overfitting by monitoring the validation loss and halting training if no improvement in validation loss for 10 epochs, it revert to the best model weights. The process tracks training and validation losses to monitor the model's learning and convergence behavior. After successful training, the model makes Prediction on the test data, which is then inverse transformed back to their original scale using

’scaler_y.inverse_transform’. The model’s accuracy is evaluated using metrics which includes Mean squared Error(MSE), Root Mean Squared Error(RMSE), R-Squared(R2) and Mean Absolute Error(MAE). Final step, Plots are generated to visually compare actual versus predicted values and analyze the residuals for each target variable, offering insights into the accuracy and error distribution of the model’s predictions.

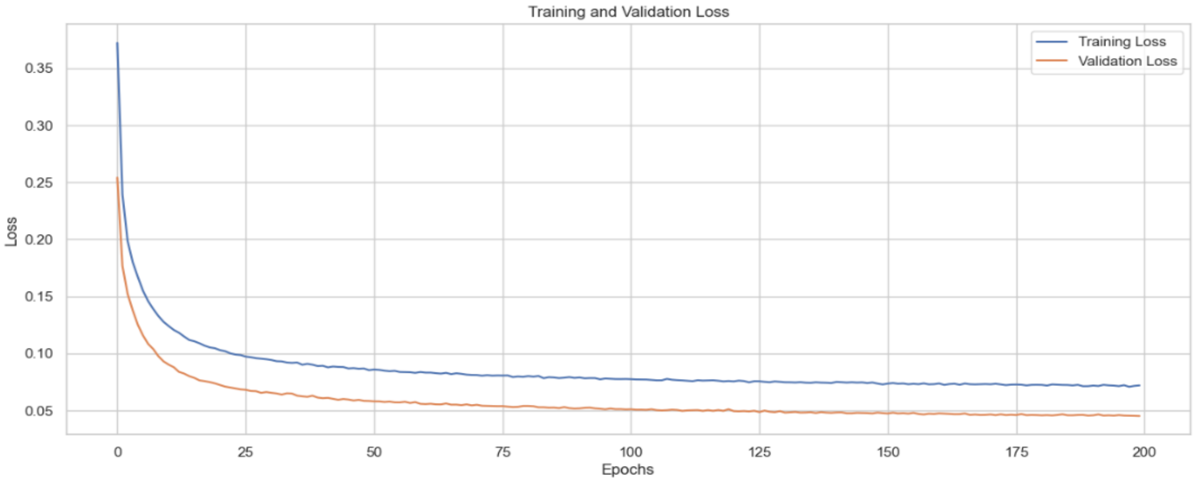


Figure 8: Trianing and Validation Loss

5.2.4 XGBoost Regressor

The data is split for regression using a stratified shuffle split aproach, with a secified test size of 20% and random state to ensure consistency. This method help to make sure that the training and test datasets have approximately the same percentage of samples of each target class as the complete set. Features are scaled using scaler, this is an important step for models like XGBoost that are sensitive to the scale of input. An XGBoost Regressor (XGBRegressor()) is intialized with default parameters and trained using the scaled training data, Then make prediction on the scaled test dataset, and these prediction are used to calculate main performance metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R2 Score, and Mean Absolute Error (MAE). Cross-validation is conducted on the training set with 5 fold using the negative root mean squared error scoring method to better assess the model performance across different subsets of training data. Metrics for each target variable are printed out, including the calculated MSE,RMSE,R2 Score and MAE, along with cross-validation RMSE score, this help to provide insight into the model’s consistency and reliability.

5.3 Tools and Languages Used

- Python: This is the primary language for all the stages of the implementation.
- Pandas and Numpy: The was used for the data manipulation, preprocessing and feature engineering.
- Scikit-learn: This was used to implement the DT and DF models, also for model evaluation.
- TensorFlow Keras: This is for developing and training the LSTM model.

- XGBoost: this is for implementing the XGBoost regression model.
- Matplotlib and seaborn: This is data visualization, helping to explore data and present model.

5.4 The Final Output

This is the four trained models ready and evaluation metrics validating their performance. these models show good solution for predicting power consumption, tailored to the unique pattern and dependencies within the dataset.

6 Evaluation/Results of the Models

The evaluation of the predictive models for power consumption in Tetouan, Morocco, involved thorough analysis of the result and main finding from each models. Each of the model's performance was assessed through the three consumption Zones, using RMSE,R2,MAE, MSE, and Cross-validation. A detail finding and insights of each model are as follow.

6.1 Decision Tree Model

The Decision Tree models show a foundational understanding of the predictive capabilities using a simple and easy to interpret structure. Different performance across the three zones. Zone3 perform better with the RMSE(1219.6042) and highest R2 score (96.64%), indicating the model predicts consumption mostly accurately in this zone. zone2 follows, with lightly higher RMSE(1462.0711) and lower R2 score (92.06%). Zone1 RMSE(1836.0731) is the highest and the lowest R2 score (93.37%) among the three zones, indicating that the model has the most difficulty predicting consumption in zone1. The cross-validated RMSE value are close to the test set RMSE value for all zones, suggesting that the models performance is reliable and consistent across different subsets of the data. This results indicated that the model had reasonable predictive accuracy. Scatter plots of actual vs predicted values and residuals were use to visualize the performances of the model in the three zone. it indicated that the model captured the general trends but struggled with finer details.

Table 1: Decision Tree Model Performance Metrics

Zone	MSE	RMSE	R2	MAE	CV-RMSE
Z1	3,371,164.5472	1,836.0731	93.37%	1,319.9929	1,817.6309
Z2	2,137,651.9169	1,462.0711	92.06%	1,075.7532	1,472.7136
Z3	1,487,434.4758	1,219.6042	96.64%	821.4085	1,215.8557

6.2 Random Forest Model

The RF model improve upon the Decision Tree model by combining multiple trees which reduced overfitting and increase accuracy. The performance are significantly better, with zone3 metrics been the best with RMSE(619.9811) and cross validation RMSE (656.5050). The Cross-validation RMSE values are close to the test set RMSE values for all the three zones as well, indicate consistent and reliable performance. Also visual comparison of the actual vs. predicted values showed a closer alignment and residuals were more evenly distributed.

Table 2: Random Forest Model Performance Metrics

Zone	**MSE**	**RMSE**	**R2**	**MAE**	**CV-RMSE**
Z1	1078350.5253	1038.4366	97.88%	685.8981	1117.3844
Z2	527578.3118	726.3459	98.04%	486.0497	805.9419
Z3	384376.5189	619.9811	99.13%	388.1489	656.5050

6.3 LSTM Model

The LSTM model leveraged the sequential nature of the data, capturing temporal dependencies effectively. Training and validation losses showed good model convergence and earlystoping prevented overfitting. The LSTM model demonstrated strong predictive performance , with lower RMSE and higher R2 scores compared to DT model. with MAE(730.5609) and R2 (97.68%)for Zone3, been the best among the three model. Plots of actual vs predicted values and residuals shows the model's ability to handle time series data well. see appendix A.

Table 3: LSTM Model Performance Metrics

Zone	**MSE**	**RMSE**	**R2**	**MAE**	**CV-RMSE**
Z1	2447823.1965	1564.5521	95.19%	1141.1768	N/A
Z2	1727951.9397	1314.5159	93.58%	970.1389	N/A
Z3	1017389.9560	1008.6575	97.70%	720.6085	N/A

6.4 XGBoost Model

The XGBoost (Extreme Gradient Boosting) is an advanced implementation of gradient boosting designed for speed and performance. It is widely use because of its potential of handling variety of data types and ability of minimizing errors through efficient training. Demonstrated the best overall performance among the models evaluated.The model across the three zones is trained using 3-fold cross-validation on each 144 candidates, totaling 720 fits, this thorough cross-validation process help to make sure the model's goodness and generalizability. The model shows strong performance across all the three zone, also zone 3 has the best performance with lowest RMSE value(622.3665) and highest R2 score (99.13%), this indicating the model predicts consumption most accurately in this zone, followed by zone2 with a little higher RMSE value(692.1694) and a little lower R2 score(98.22%) compared to zone3, Zone1 has the highest RMSE value(993.2652) and lowest R2 value(98.06%).Cross-validation RMSE values are close to the test set RMSE values for all the three zones, suggesting that the model's results is consistency and reliability. Similarly visual shows that the XGBoost predictions were very close to actual values , with minimal residuals. As showed in appendix A

Table 4: XGBoost Model Performance Metrics

Zone	MSE	RMSE	R2	MAE	CV-RMSE
Z1	986575.6893	993.2652	98.06%	705.9949	1051.3444
Z2	479098.4652	692.1694	98.22%	502.6797	738.4849
Z3	387340.0426	622.3665	99.13%	420.2049	639.1190

6.5 Discussion

The evaluation results shows that the XGBoost models consistently performance best with metrics in all three zones, this indicate it’s strength and accuracy. The Random Forest model performance also showing significant improvement over the single Decision three model. The LSTM model, though not the best performer , effectively captured the temporal dependencies in the data, highlighting it’s strength in handling sequential(time-series) data prediction.

Though the XGBoost model performance is better, handling non-linear relationship and interactions within the data efficiently, it is computationally intensive and requires careful parameter tuning. The Random Forest model also provided good result with relatively lower computational demands, it improved upon the Decision Trees by reducing overfitting and increasing accuracy through ensemble learning. The LSTM model, though little behind XGBoost in performance, effectively captured time-based patterns in power consumption.

To round it up , the evaluation shows, while the traditional machine learning models like the decision Tree and Random Forests provide good starting point, an advanced techniques such as XGBoost offer substantial improvements in predictive accuracy for power consumption in Tetouan, Morocco. These result suggest practical application in energy management and planning, providing a basis for further research and development in predictive modeling for power consumption in general.

Table 5: Compering all Model Performance Metrics

Model	Zone	MSE	RMSE	R2	MAE	CV-RMSE
Decision Tree	Z1	3371164.5472	1836.0731	93.37%	1319.9929	1817.6309
Decision Tree	Z2	2137651.9169	1462.0711	92.06%	1075.7532	1472.7136
Decision Tree	Z3	1487434.4758	1219.6042	96.64%	821.4085	1215.8557
Random Forest	Z1	1078350.5253	1038.4366	97.88%	685.8981	1117.3844
Random Forest	Z2	527578.3118	726.3459	98.04%	486.0497	805.9419
Random Forest	Z3	384376.5189	619.9811	99.13%	388.1489	656.5050
LSTM	Z1	2447823.1965	1564.5521	95.19%	1141.1768	N/A
LSTM	Z2	1727951.9397	1314.5159	93.58%	970.1389	N/A
LSTM	Z3	1017389.9560	1008.6575	97.70%	720.6085	N/A
XGBOOST	Z1	986575.6893	993.2652	98.06%	705.9949	1051.3444
XGBOOST	Z2	479098.4652	692.1694	98.22%	502.6797	738.4849
XGBOOST	Z3	387340.0426	622.3665	99.13%	420.2049	639.1190

6.6 Feature Importance and Error Analysis

Feature Importance: The analysis for both Decision Tree and Random Forest, reveals key predictors of power consumption across the three zones. The ’Hour variable was significant in zone 1,2 and 3, while the 'Month' variable was most important in zone 3 in Random Forest model. These insights guide future feature engineering and model refinement. Further details are available in the plots in appendix A.1.1 and A.2.1 respectively.

Error Analysis: Error distribution plots showed that the RF and XGBoost models had fewer and more evenly distributed errors as compared to the Decision Tree Model. this clearly indicates better model stability and reliability.

7 Conclusion and Future Work

In this comprehensive analysis, the goal was to address the existing gap in power consumption prediction by evaluating the effectiveness of Decision Tree, Random Forest, LSTM, and XGBoost models in predicting energy usage in Tetouan, Morocco. By leveraging these modelling techniques and incorporating external factors such as temperature, humidity, and wind speed, the study seeks to enhance the precision and resilience of power consumption projections in the region. given rise to the the research question, how well can Decision Tree, Random Forest, LSTM, and XGBoost models predict power consumption in Tetouan, Morocco?.

The study employ KDD (Knowledge Discovery in Databases) methodology to carry out this research approach. The data was sourced from the UCI Machine learning repository. which consist of historical weather and power consumption records. The research process, including downloading it onto pandas data frame, data exploration, data cleaning/preprocessing, Feature engineering, data visualization, model training and evaluation, was implemented using Python and it's libraries.

This research utilized four regression models, which are Decision Tree, Random Forest, LSTM and XGBoost Regressor. their performances was evaluated based on metrics such as Mean Squared Error(MSE), Root Mean Squared Error(RMSE), R-squared(R2) score and Mean Absolute Error(MAE). With main findings that XGBoost model exhibited the highest predictive accuracy across all the three zones, followed by the RF model. The LSTM model effectively captured temporal dependencies, highlighting its suitability for time-series forecasting task and Decision Tree(DT) is the baseline, perform lower than the rest model.

Although the XGBoost model achieved the best performance, it has high computation and needs careful parameter tuning. on the other hand Random Forest model provided good results with relatively lower computational demands. The LSTM model results, while slightly lower than XGBoost in performance, shows its strength in handling sequential data.

The results in this research indicate that while traditional machine learning models like Decision Tree, offer a good starting point, advance techniques such as XGBoost provide substantial improvements in predictive accuracy for power consumption.

While the current analysis provides valuable insights into predicting power consumption in Tetouan region of Morocco, More research can be build in the future on this study through exploring the following suggested areas

- Model Optimization: More optimization of model parameter, most especially for XGBoost and LSTM model, could enhance performance.

- Real-time Forecasting: By implementing real time forecasting systems to provide continuous updates on the prediction of power consumption in Tetouan, Morocco.

- Incorporating Additional Data: By including more features such as economic indicators, social data, age data and occupancy information could improve model accuracy

- Scalability and Deployment: using this study to develop more salable solution in the future for deployment in real-word energy management

All these may improve the predictive capabilities of the models providing more accurate reliable forecasting of power consumption in general, This call help in optimizing resource allocation, reduce cost and support energy management initiatives.

References

Akbari-Dibavar, A., Nojavan, S., Mohammadi-Ivatloo, B. and Zare, K. (2020). Smart home energy management using hybrid robust-stochastic optimization, *Computers & Industrial Engineering* **143**: 106425.

Blaszczyk, G. (2022). *Machine Learning Techniques for Prediction of Electricity Consumption in Buildings*, PhD thesis, Dublin, National College of Ireland.

Cao, W., Yu, J., Chao, M., Wang, J., Yang, S., Zhou, M. and Wang, M. (2023). Short-term energy consumption prediction method for educational buildings based on model integration, *Energy* **283**: 128580.

Chumbar, S. (2023). Kdd process in data science: A beginner's guide. Accessed: 2024-08-07.
URL: *https://medium.com/@shawn.chumbar/kdd-process-in-data-science-a-beginners-guide-426d1f0fc062*

da Silva, D. G. and de Moura Meneses, A. A. (2023). Comparing long short-term memory (lstm) and bidirectional lstm deep neural networks for power consumption prediction, *Energy Reports* **10**: 3315–3334.

Faiq, M., Tan, K. G., Liew, C. P., Hossain, F., Tso, C.-P., Lim, L. L., Wong, A. Y. K. and Shah, Z. M. (2023). Prediction of energy consumption in campus buildings using long short-term memory, *Alexandria Engineering Journal* **67**: 65–76.

Fouad, M., Mali, R. and Pr. Bousmah, M. (2020). Machine learning for forecasting building system energy consumption, *Innovation in Information Systems and Technologies to Support Learning Research: Proceedings of EMENA-ISTL 2019 3*, Springer, pp. 235–242.

Gökçe, M. M. and Duman, E. (2022). Performance comparison of simple regression, random forest and xgboost algorithms for forecasting electricity demand, *2022 3rd International Informatics and Software Engineering Conference (IISEC)*, IEEE, pp. 1–6.

Khafaga, D. S., El-kenawy, E.-S. M., Alhussan, A. A. and Eid, M. M. (2023). Forecasting energy consumption using a novel hybrid dipper throated optimization and stochastic fractal search algorithm, *Intelligent Automation & Soft Computing* **37**(2): 2117–2132.

Khan, P. W., Kim, Y., Byun, Y.-C. and Lee, S.-J. (2021). Influencing factors evaluation of machine learning-based energy consumption prediction, *Energies* **14**(21): 7167.

Mystakidis, A., Ntozi, E., Afentoulis, K., Koukaras, P., Gkaidatzis, P., Ioannidis, D., Tjortjis, C. and Tzovaras, D. (2023). Energy generation forecasting: elevating performance with machine and deep learning, *Computing* **105**(8): 1623–1645.

Nan, S., Tu, R., Li, T., Sun, J. and Chen, H. (2022). From driving behavior to energy consumption: A novel method to predict the energy consumption of electric bus, *Energy* **261**: 125188.

Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F. and Ajayi, S. (2022). Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques, *Journal of Building Engineering* **45**: 103406.

Priyadarshini, I., Sahu, S., Kumar, R. and Taniar, D. (2022). A machine-learning ensemble model for predicting energy consumption in smart homes, *Internet of Things* **20**: 100636.

RoSe, N., Osbourne, O., Williams, N. and Rizvi, S. S. H. (2023). A novel optimized variant of machine learning algorithm for accurate energy demand prediction for tetouan city, morocco, *International e-Conference on Advances in Computer Engineering and Communication Systems (ICACECS 2023)*, Atlantis Press, pp. 62–73.

Saglam, M., Spataru, C. and Karaman, O. A. (2023). Forecasting electricity demand in turkey using optimization and machine learning algorithms, *Energies* **16**(11): 4499.

Salam, A. and El Hibaoui, A. (2018). Comparison of machine learning algorithms for the power consumption prediction:-case study of tetouan city–, *2018 6th International Renewable and Sustainable Energy Conference (IRSEC)*, IEEE, pp. 1–5.

Shanmugam, D. and Ramana, V. V. (2024). Different meta-heuristic optimized radial basis function neural network models for short-term power consumption forecasting, *Advances in Engineering and Intelligence Systems* **3**(02): 63–82.

Shin, S.-Y. and Woo, H.-G. (2022). Energy consumption forecasting in korea using machine learning algorithms, *Energies* **15**(13): 4880.

Tan, Y.-F., Zhao, G.-Z., Cheeng, T.-H., Ooi, C.-P., Tan, W.-H. and Cheong, S.-N. (2023). Supervised machine learning techniques for power consumption usage level prediction, *2023 6th International Conference on Software Engineering and Computer Science (CSECS)*, IEEE, pp. 1–5.

Ves, A. V., Ghitescu, N., Pop, C., Antal, M., Cioara, T., Anghel, I. and Salomie, I. (2019). A stacking multi-learning ensemble model for predicting near real time energy consumption demand of residential buildings, *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, IEEE, pp. 183–189.

Wang, J. Q., Du, Y. and Wang, J. (2020). Lstm based long-term energy consumption prediction with periodicity, *energy* **197**: 117197.

Zhou, Y., Liu, Y., Wang, D. and Liu, X. (2021). Comparison of machine-learning models for predicting short-term building heating load using operational parameters, *Energy and Buildings* **253**: 111505.

A Appendix A

A.1 Decision Tree Result plot

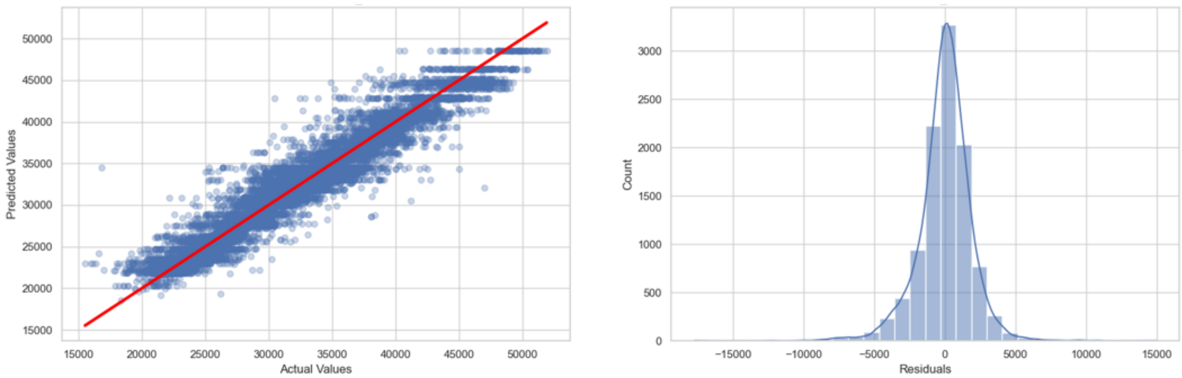


Figure 9: Decision Tree: plot for Actual vs. Predicted Value, and Residuals for Zone1

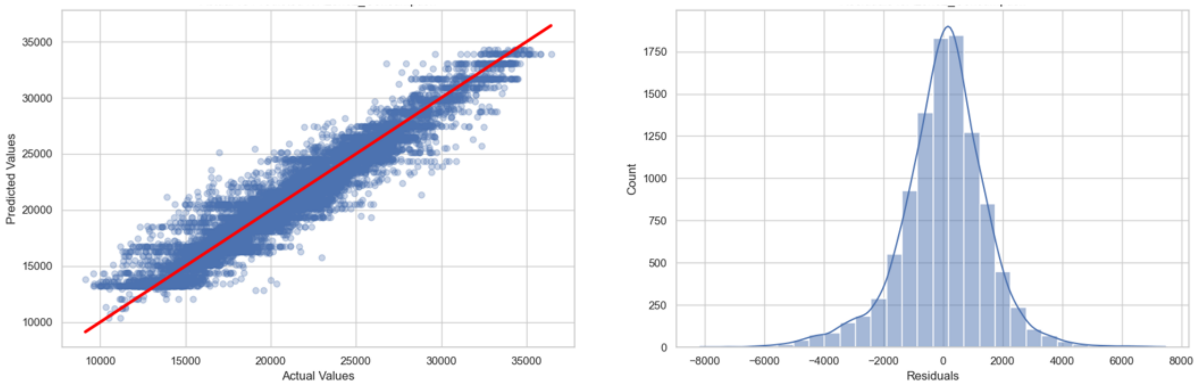


Figure 10: Decision Tree: plot for Actual vs. Predicted Value, and Residuals for Zone2

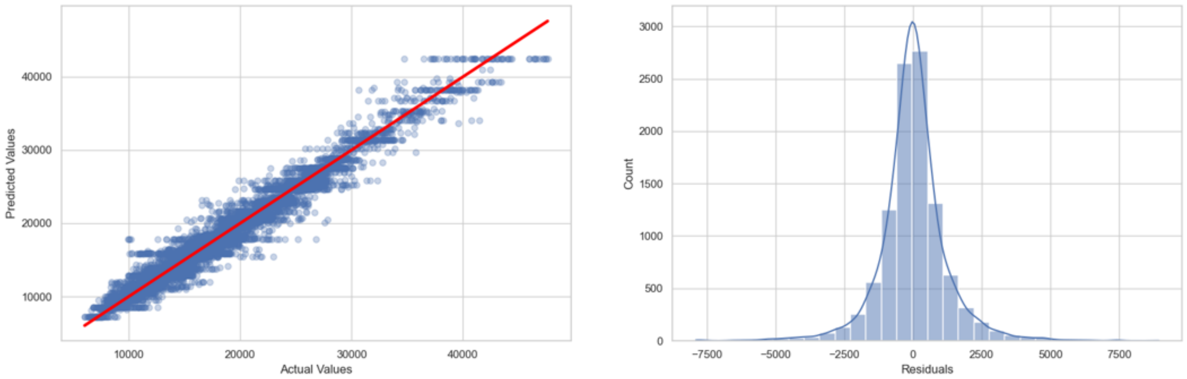


Figure 11: Decision Tree: plot for Actual vs. Predicted Value, and Residuals for Zone3

A.1.1 Decision Tree, Feature Importance Plots

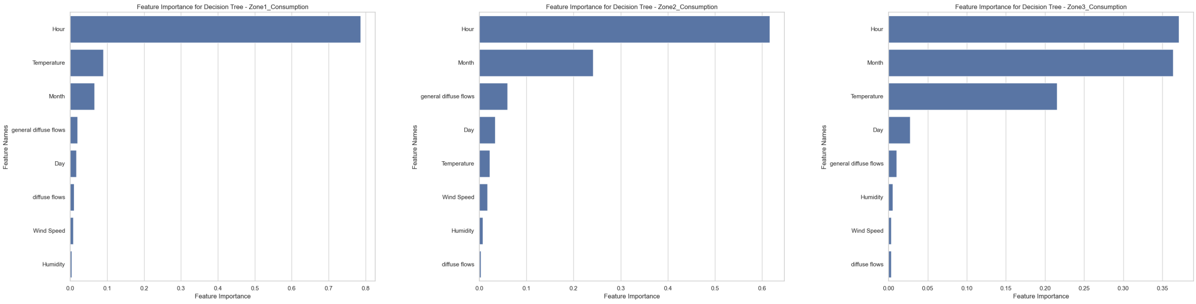


Figure 12: Decision Tree, Feature Importance Across the three Zones

A.2 Random Forest Result Plot

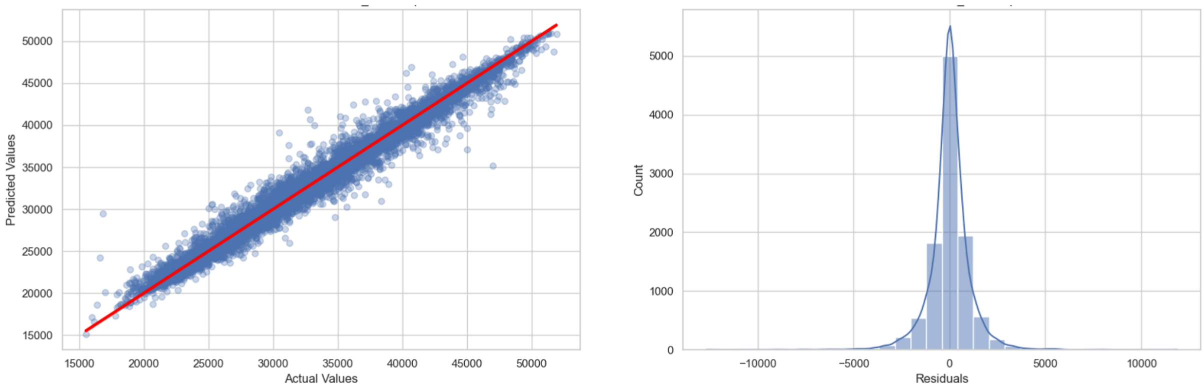


Figure 13: Random Forest: plot for Actual vs. Predicted Value, and Residuals for Zone1

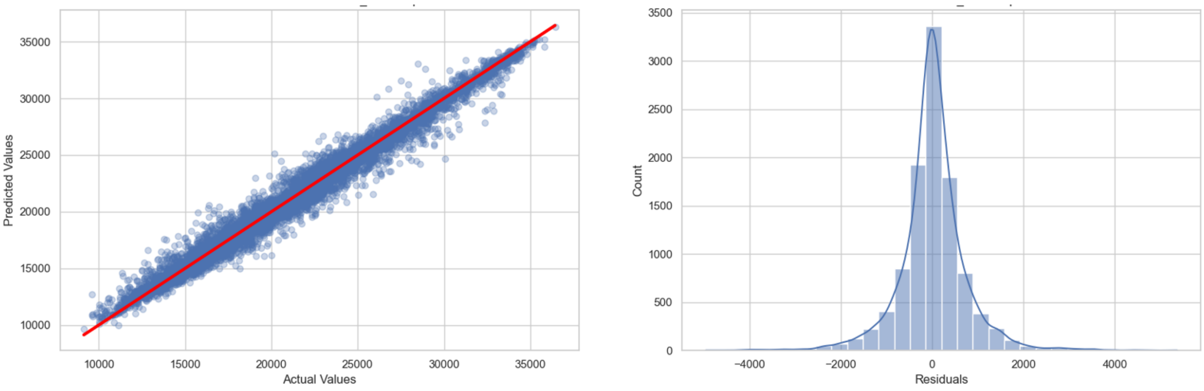


Figure 14: Random Forest: plot for Actual vs. Predicted Value, and Residuals for Zone2

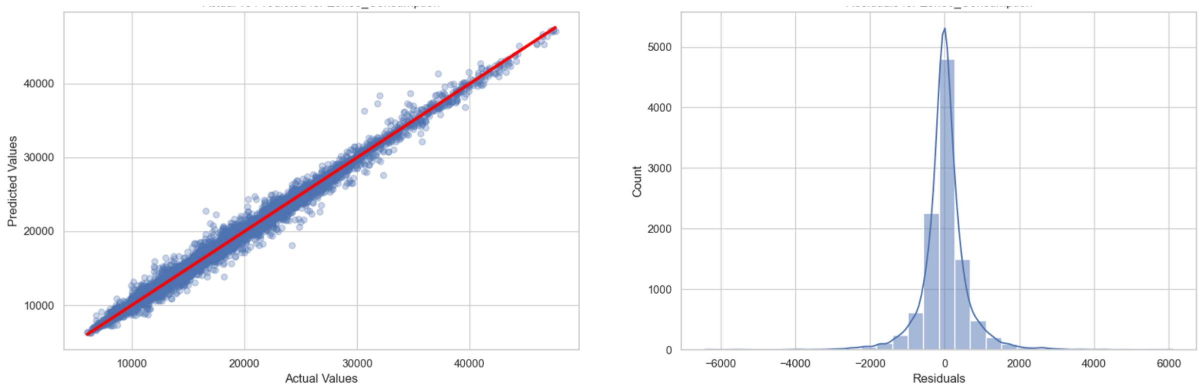


Figure 15: Random Forest: plot for Actual vs. Predicted Value, and Residuals for Zone3

A.2.1 Random Forest, Feature Importance Plots

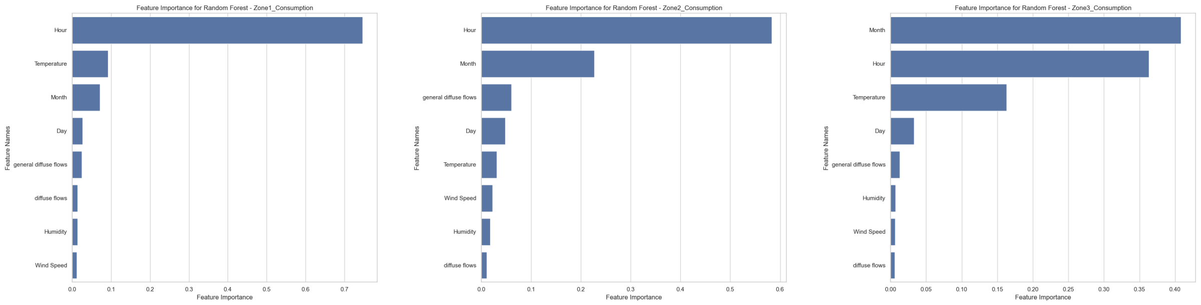
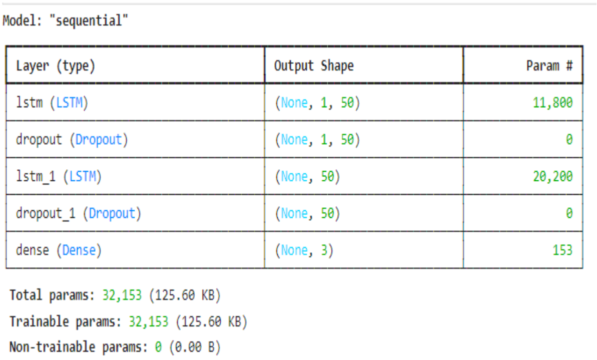


Figure 16: Random Forest, Feature Importance Across the three Zones

A.3 LSTM Result Plot



```
Model: "sequential"

┃ Layer (type)          ┃ Output Shape      ┃   Param # ┃
│ lstm (LSTM)           │ (None, 1, 50)     │    11,800 │
│ dropout (Dropout)     │ (None, 1, 50)     │         0 │
│ lstm_1 (LSTM)         │ (None, 50)        │    20,200 │
│ dropout_1 (Dropout)   │ (None, 50)        │         0 │
│ dense (Dense)         │ (None, 3)         │       153 │

 Total params: 32,153 (125.60 KB)
 Trainable params: 32,153 (125.60 KB)
 Non-trainable params: 0 (0.00 B)
```

Figure 17: LSTM Sequential

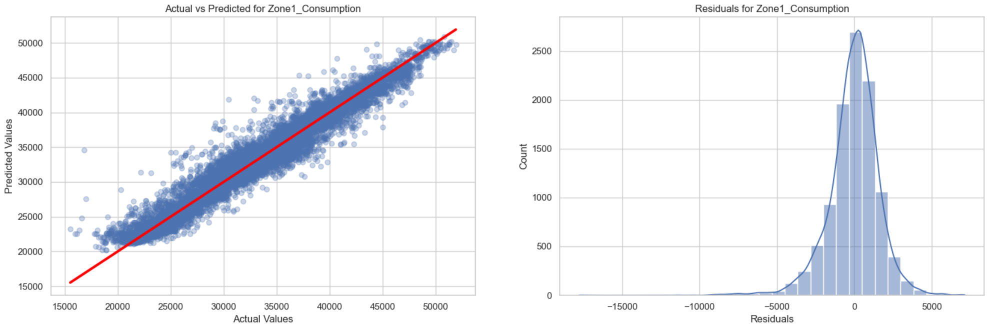


Figure 18: LSTM: plot for Actual vs. Predicted Value, and Residuals for Zone2

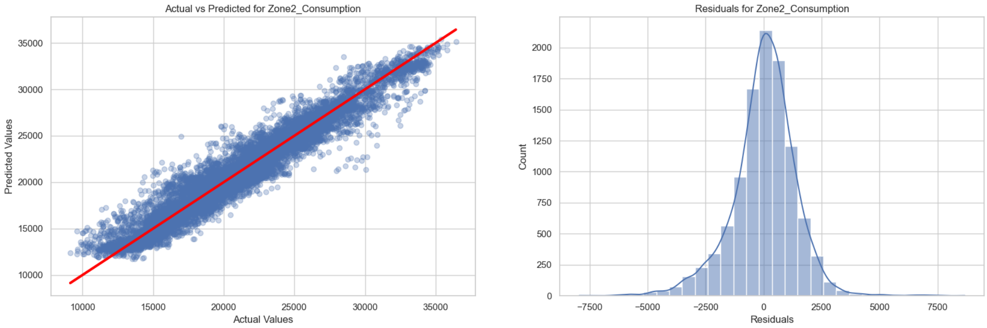


Figure 19: LSTM: plot for Actual vs. Predicted Value, and Residuals for Zone2

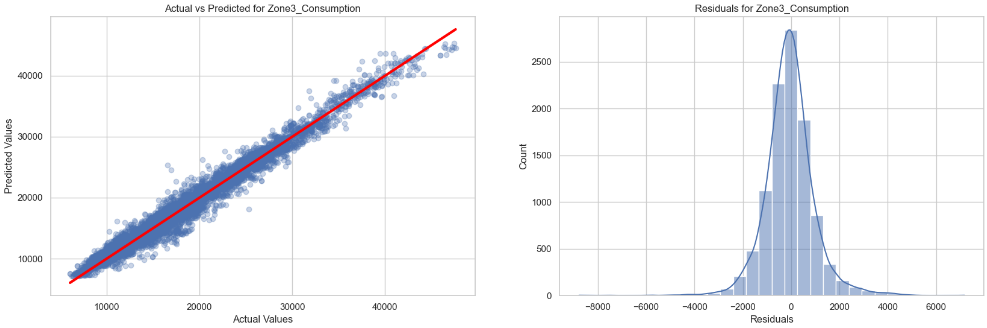


Figure 20: LSTM: plot for Actual vs. Predicted Value, and Residuals for Zone3

A.4 XGBoost Result Plot

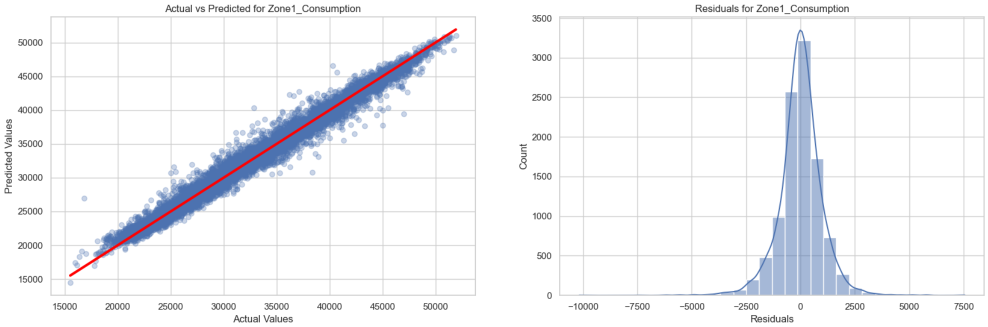


Figure 21: XGBoot: plot for Actual vs. Predicted Value, and Residuals for Zone2

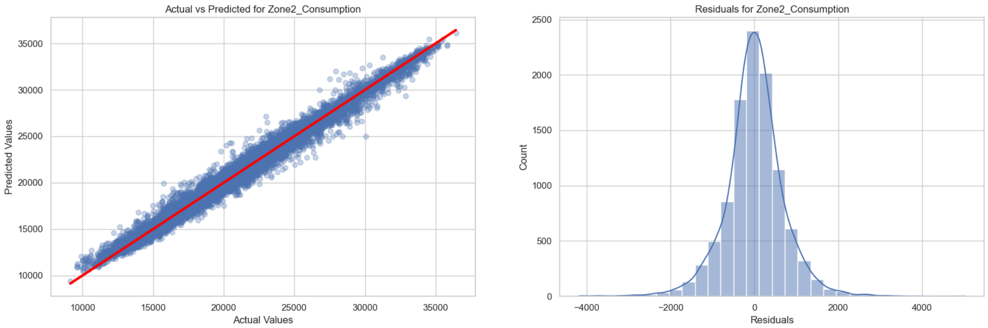


Figure 22: XGBoot: plot for Actual vs. Predicted Value, and Residuals for Zone2

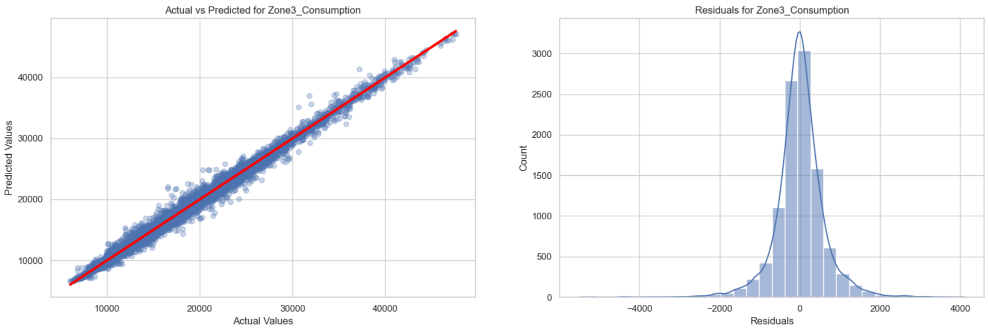


Figure 23: XGBoot: plot for Actual vs. Predicted Value, and Residuals for Zone3

A.5 Histogram of variables

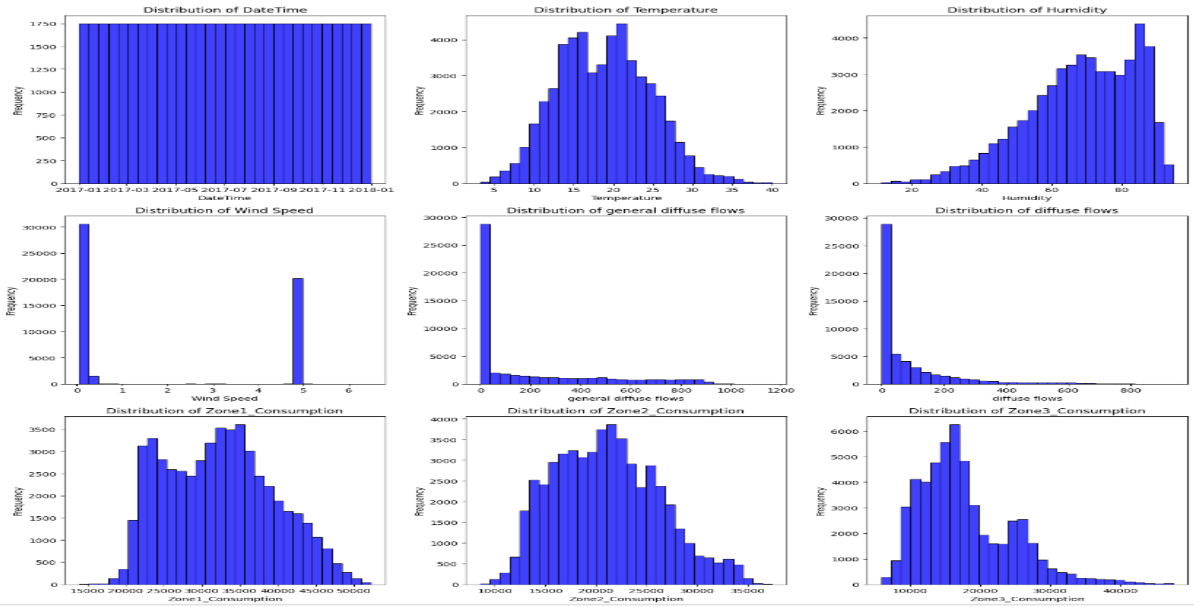


Figure 24: Histogram Showing the Distribution of the Variables: