YEDİTEPE UNIVERSITY
Following Atatürk's Renaissance

COMPUTER ENGINEERING DEPARTMENT

CSE 464 Case Study 2

INTRODUCTION TO DATA SCIENCE & BIG DATA ANALYTICS
(AUTUMN SEMESTER 2024)

Mahny Barazandehtar - 20210702004

YEDİTEPE UNIVERSITY
Following Atatürk's Renaissance

COMPUTER ENGINEERING DEPARTMENT

I. Executive Summary/Abstract

This case study builds upon a prior descriptive analysis of Tetouan City's power distribution network, involving three main zones influenced by temperature, humidity, wind speed, solar radiation, and time-of-day. While the initial work highlighted correlations among these factors, a more comprehensive predictive approach was needed to optimize load balancing, reduce operational inefficiencies, and guide renewable energy integration efforts.

To address these needs, multiple machine learning models—namely Linear Regression, K-Nearest Neighbors (KNN), Random Forest, XGBoost, and Support Vector Regression (SVM)—were trained and evaluated for forecasting electricity consumption at both the zone level and in aggregate. To determine the most critical drivers of consumption, two feature selection methods, Recursive Feature Elimination (RFE) and Random Forest–based importance metrics, were applied. Results indicated that ensemble methods, particularly Random Forest, achieved the highest predictive accuracy, thereby offering a valuable roadmap for more reliable and cost-effective energy management strategies in Tetouan City.

II. Background/Introduction

- **About the Client/Company:**

  Amendis oversees power distribution in Tetouan City, northern Morocco. It operates under a public-private partnership and is responsible for ensuring reliable electricity service to residential, commercial, and industrial consumers. Tetouan is divided into three main power zones (Zone 1, Zone 2, and Zone 3), each with unique consumption characteristics.

- **Industry Context:**

  The energy sector faces growing pressures from rising demand, environmental concerns, and the need for operational efficiencies. Accurate demand forecasting and load balancing are critical for minimizing waste, reducing operational costs, and maintaining grid stability. Machine learning offers advanced predictive capabilities, which help utilities like Amendis refine their resource planning and optimize distribution.

- **Goals and Objectives:**

  The primary goals of this study are to improve demand forecasting by enhancing the accuracy of consumption predictions using multiple machine learning methods. This involves evaluating various algorithms and feature selection techniques to identify the most reliable approach for predicting power consumption. Additionally, the study aims to provide actionable insights by offering evidence-based recommendations for dynamic pricing, renewable energy integration, and resource allocation strategies which enable more efficient and sustainable energy management.

III. Challenge/Problem Statement

A. The Problem:

Amendis needs to precisely forecast power consumption across three distinct zones, each influenced by environmental and temporal factors. While descriptive analytics identified important patterns, the organization lacks a robust predictive framework to plan for peaks, manage resources, and reduce operational inefficiencies.

B. Justification of the Problem (SWOT Analysis):

- **Strengths:** Amendis has access to detailed SCADA data, an established public-private partnership, and existing infrastructure for electricity distribution.
- **Weaknesses:** The current system lacks advanced predictive capabilities, leading to inefficiencies in peak-load management and resource allocation.
- **Opportunities:** There is potential for integrating renewable energy sources, leveraging advanced analytics to improve operational efficiency, and aligning with Morocco's sustainability goals.
- **Threats:** Increasing energy demands, competition from smarter grids, and regulatory pressures to reduce carbon emissions pose significant risks.

C. Why It Matters:

Unresolved inefficiencies can result in increased carbon emissions, higher operational costs, and potential power outages which undermine customer trust. Moreover, with rising energy demands, Amendis risks falling behind competitors who adopt smarter, data-driven strategies.

IV. Solution/Approach

To address these challenges, a combination of descriptive analytics and machine learning was used. Descriptive methods revealed historical trends, while machine learning models were employed for accurate future predictions. Feature selection techniques, such as Recursive Feature Elimination (RFE), ensured that the models focused on the most impactful variables.

D. Proposed Solution:

1. Conducted exploratory data analysis to uncover correlations and trends in historical energy consumption.
2. Trained predictive models (Random Forest, XGBoost) using features like temperature, humidity, wind speed, and temporal data (hour, day, month).
3. Visualized results to identify high-demand periods and the influence of environmental factors.

E. Key Features/Technologies:

The project leveraged the following tools and technologies:

1. **Python Libraries:**
    - **Pandas:** Data preprocessing and manipulation.
    - **Numpy:** Numerical and statistical analysis for correlation metrics.
    - **Matplotlib:** Visualization of consumption patterns.
    - **Scikit-learn:** For machine learning tasks.
    - **XGBoost:** For gradiant boosting models.

2. **SCADA Data:** Historical records of power consumption and environmental variables.
3. **Google Colab:** To execute and document analysis.

Access the Google Colab environment for this case study using the following link https://colab.research.google.com/drive/15OdmHpk-tMzbEvSYym6855sVBs9jvoTQ?usp=sharing

F. Implementation Process:

1. **Data Preprocessing and Feature Engineering:**

- The timestamp was dissected into Hour, Day, Month, and Weekday, allowing for fine-grained temporal analysis. Environmental data—temperature, humidity, wind speed, and solar diffuse flows were scaled using either a MinMaxScaler or StandardScaler, depending on each algorithm's sensitivity to feature magnitudes. This process ensured consistent contribution of each variable during model training.

2. **Choice of Machine Learning Models:**

- Linear Regression served as a baseline to gauge the data's linearity.
- KNN was employed for its ability to model non-linear consumption patterns through local neighbor-based comparisons.
- Random Forest used its ensemble approach to handle complex feature interactions and provide inherent feature importance metrics.
- XGBoost was integrated for its robust gradient boosting framework, which can capture subtler patterns through iterative refinement.
- SVM offered a powerful non-linear mapping if the hyperparameters were well-tuned, albeit at a higher computational cost.

3. **Feature Selection:**

- **Random Forest Feature Importance:** After training a Random Forest model, each feature's contribution to reducing variance was evaluated. This ranking guided a deeper understanding of which variables most strongly influenced consumption.

- **Recursive Feature Elimination (RFE):** Applied with Linear Regression, RFE repeatedly eliminated the least impactful predictors, refining the model down to the most consequential features. This systematic approach double-checked the outcomes of the Random Forest feature importance analysis.

4. **Model Training and Evaluation:**

- Data splits (80% training, 20% testing) were used to validate each model's ability to generalize. Metrics such as RMSE (root mean squared error), MAE (mean absolute error), R^2 (coefficient of determination), and explained variance were examined. Visualizations (e.g., predicted vs. actual scatter plots – Figures 7 - 11) helped illuminate any systematic under- or over-estimations.

V. Results/Outcomes

G. Quantifiable Results:

Random Forest generally provided the most accurate forecasts for all three zones as well as the aggregated total. In some instances, R^2 values exceeded 0.98, with low RMSE values (e.g., ~0.0140 in Zone 3) which reflect the model's ability to capture intricate relationships among features. XGBoost also exhibited strong predictive capability, occasionally approaching Random Forest's metrics but requiring more extensive hyperparameter tuning. KNN and SVM showed promise in certain scenarios but were more sensitive to parameter selection and often required significant computation time. Linear Regression, while simple and transparent, yielded moderate accuracy (R^2 around 0.60–0.65 for most zones), indicating that purely linear assumptions cannot fully encapsulate Tetouan's consumption dynamics.

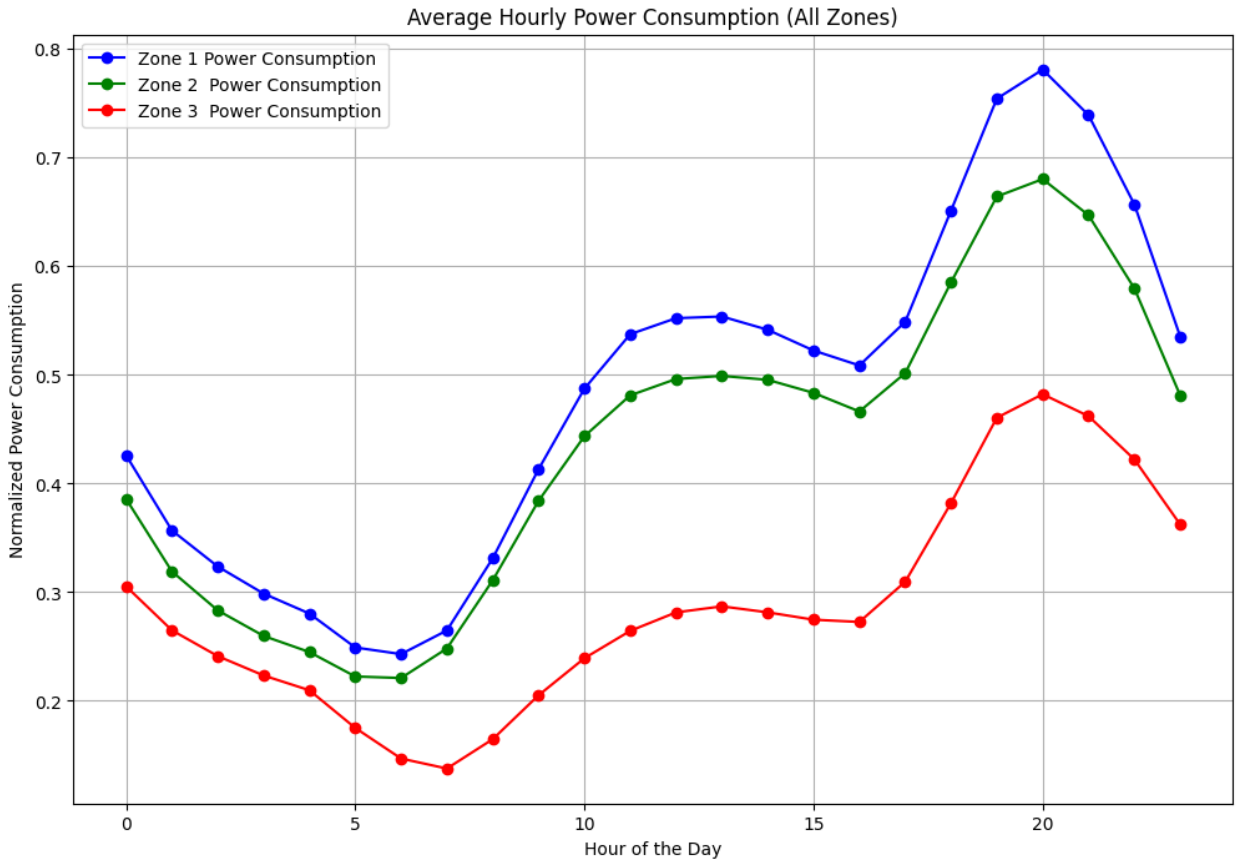

Figure 1: Average Hourly Power Consumption (All Zones)

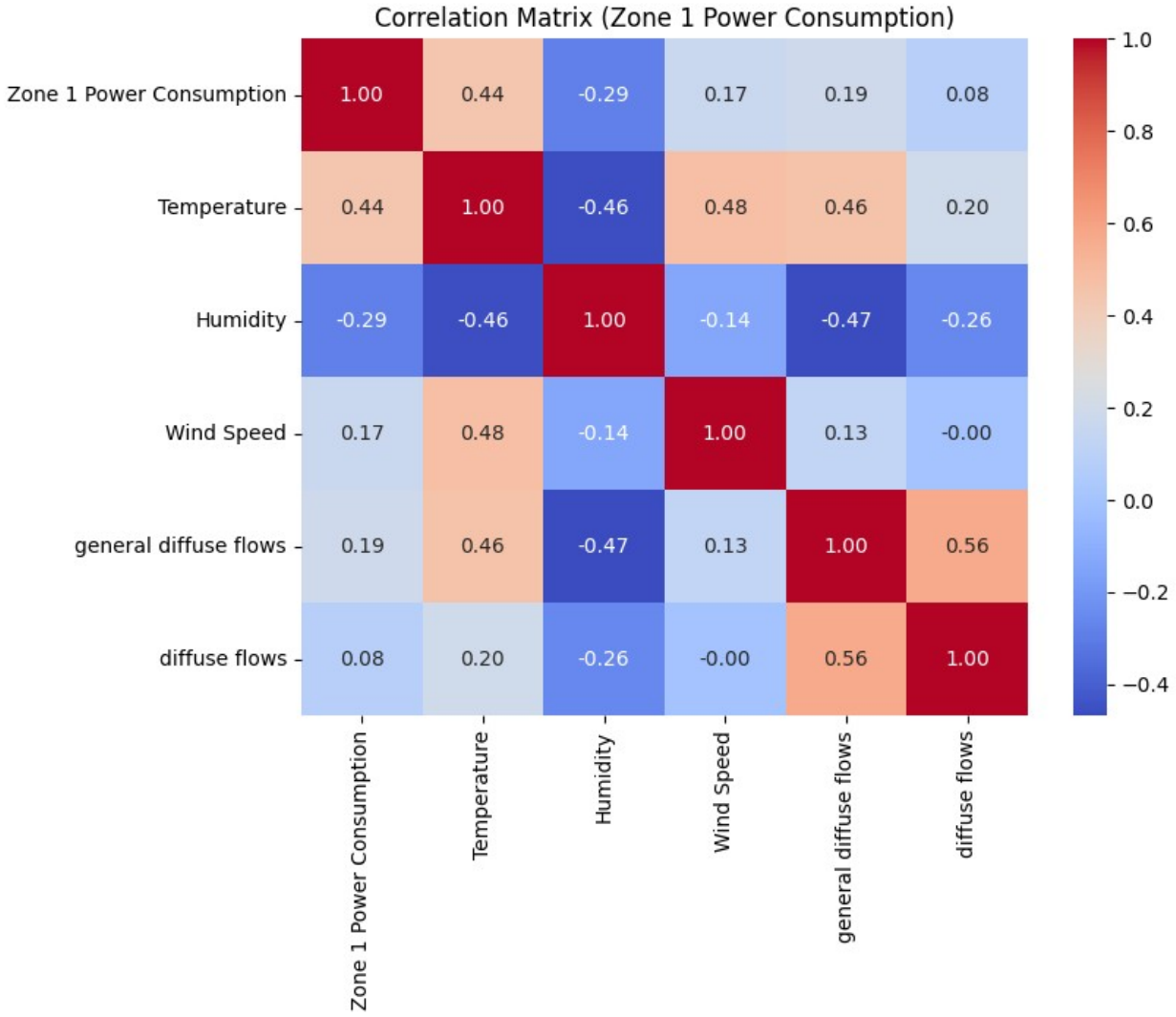

Figure 2: Correlation Matrix (Zone 1 Power Consumption)

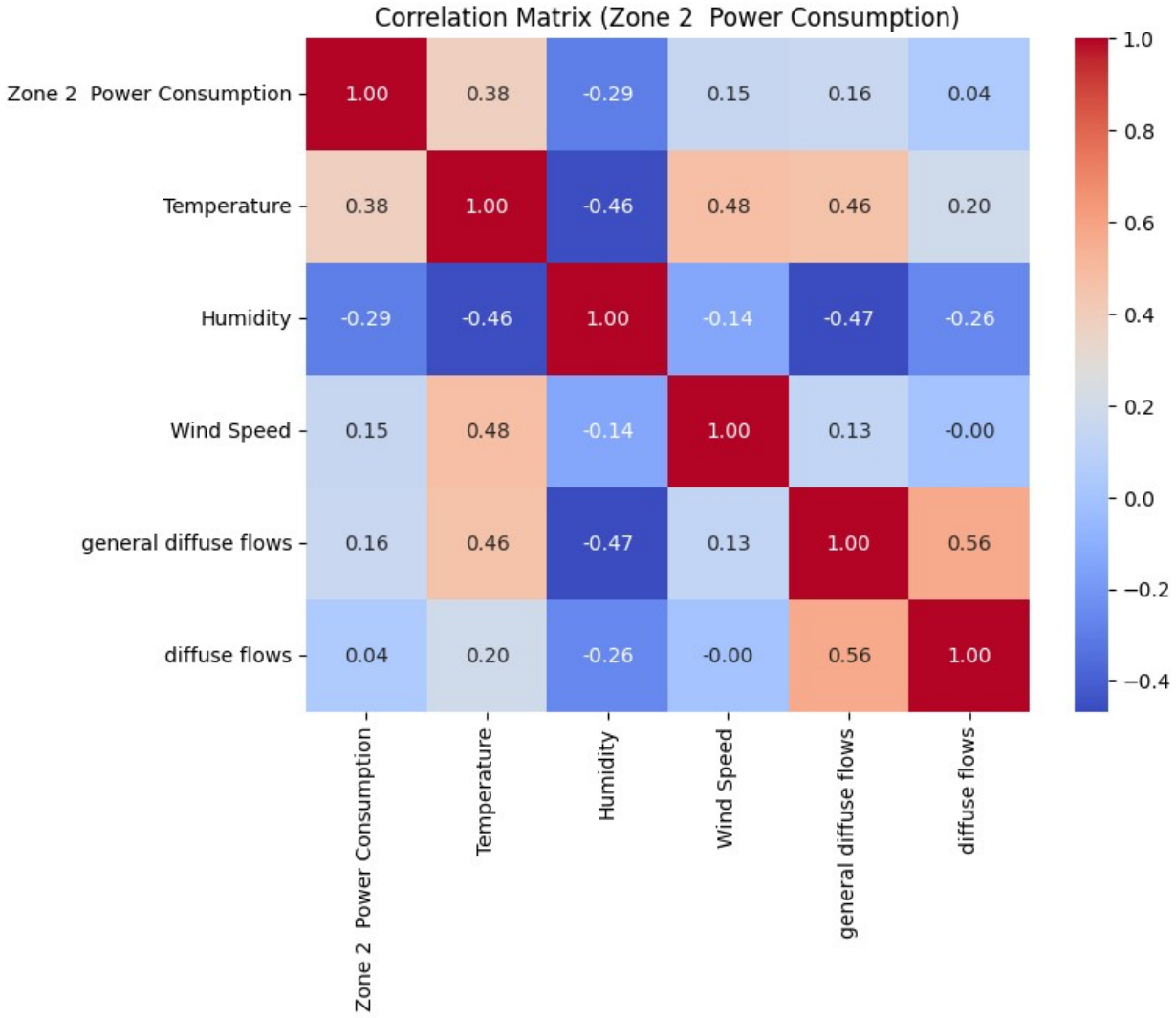

Figure 3: Correlation Matrix (Zone 2 Power Consumption)

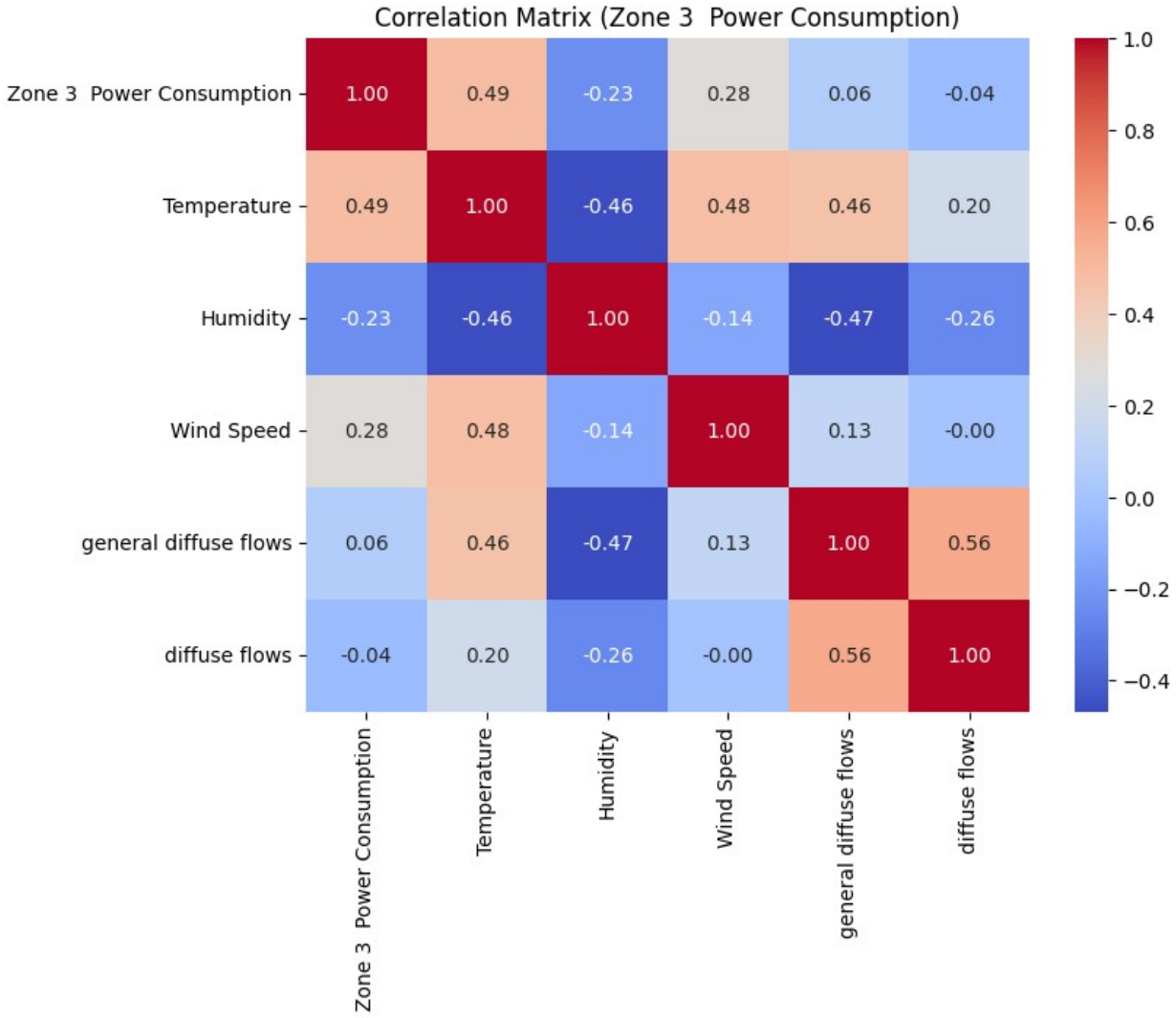

Figure 4: Correlation Matrix (Zone 3 Power Consumption)

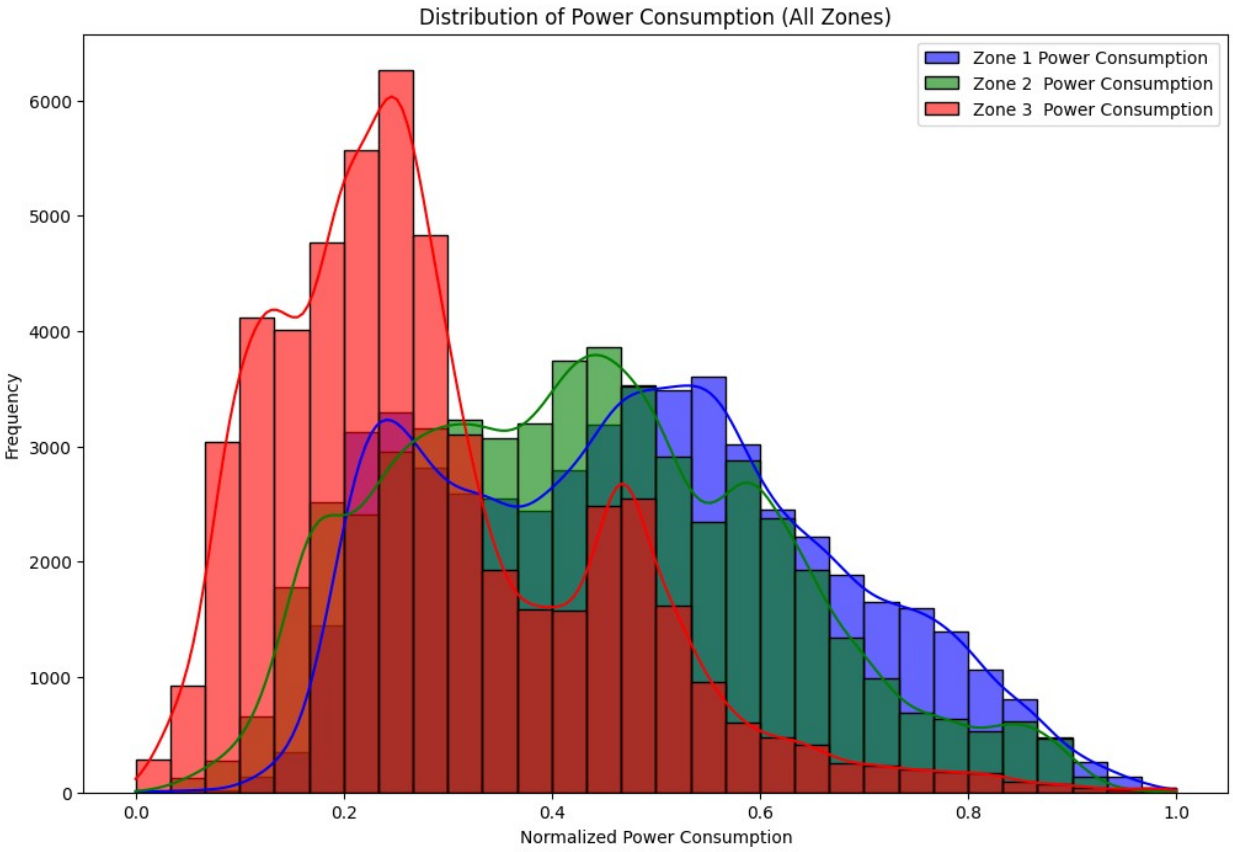

Figure 5: Distribution of Power Consumption (All Zones)

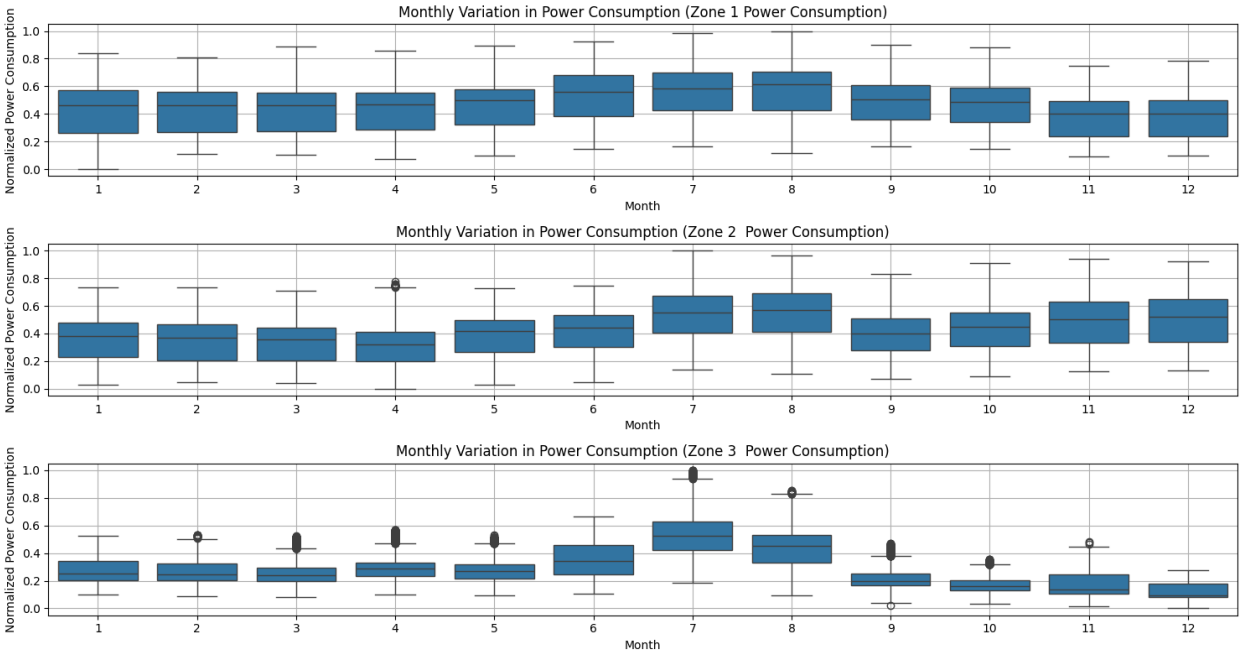

Figure 6: Monthly Variation in Power Consumption (All Zones)

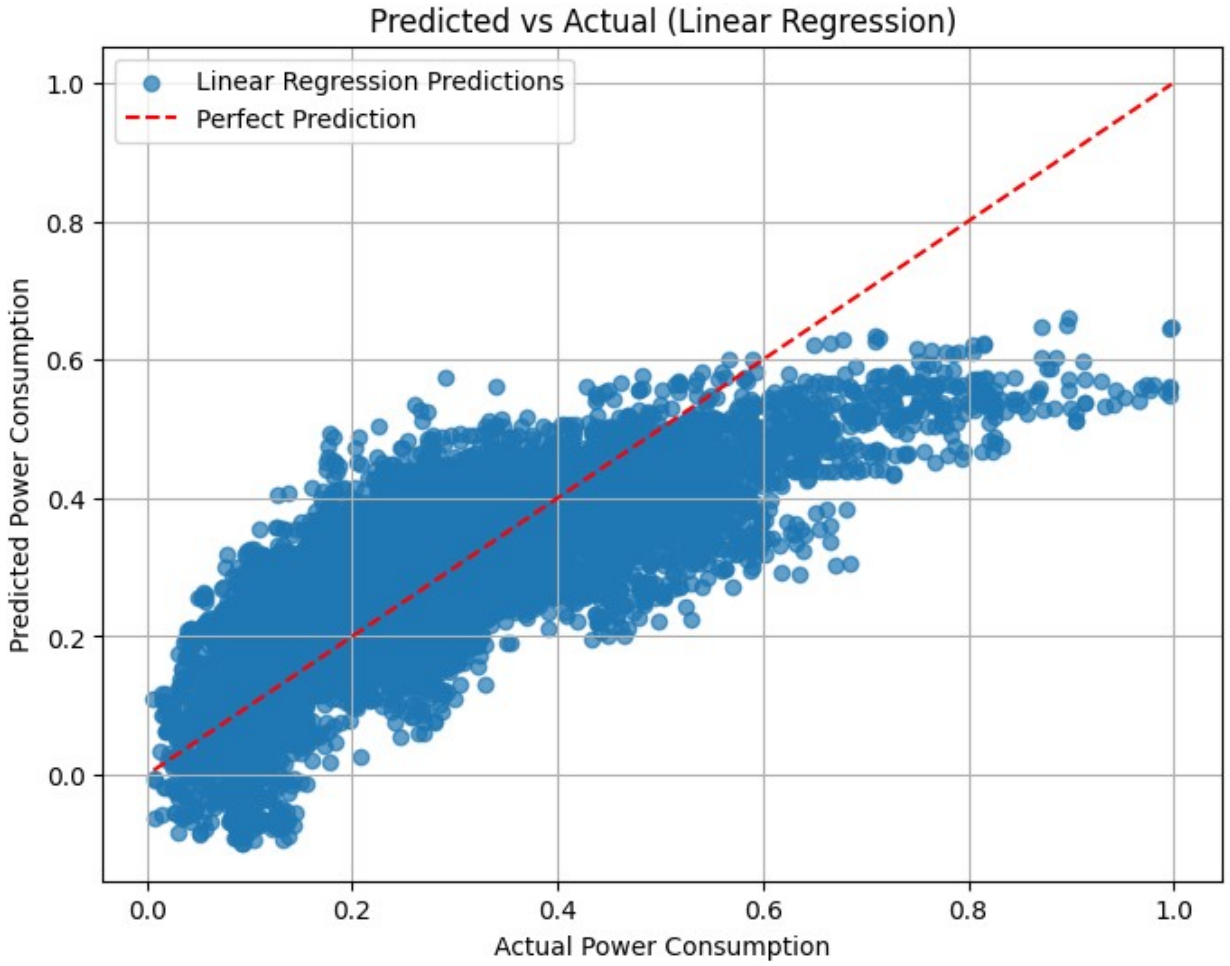

Figure 7: Predicted vs Actual (Linear Regression)

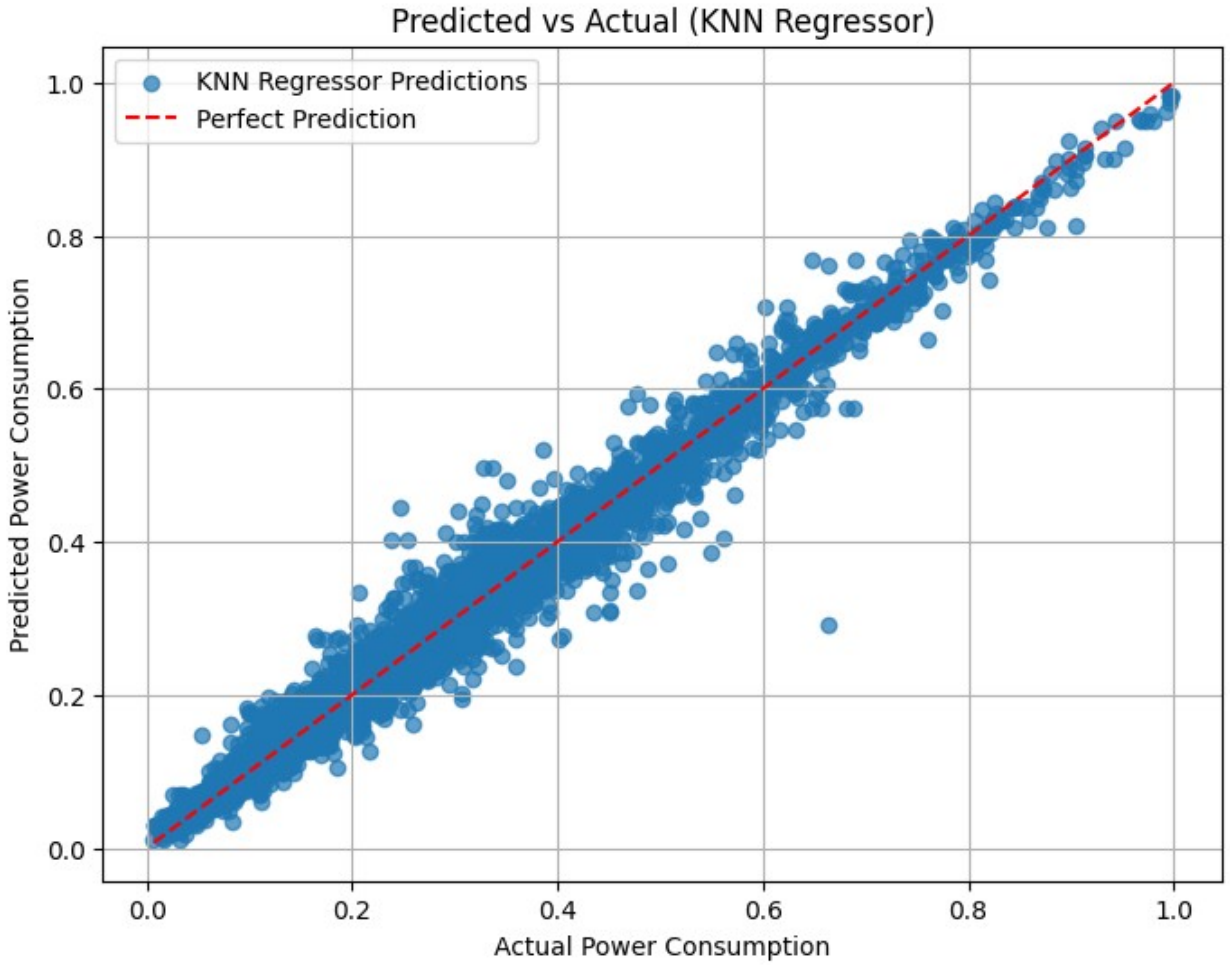

Figure 8: Predicted vs Actual (KNN Regression)

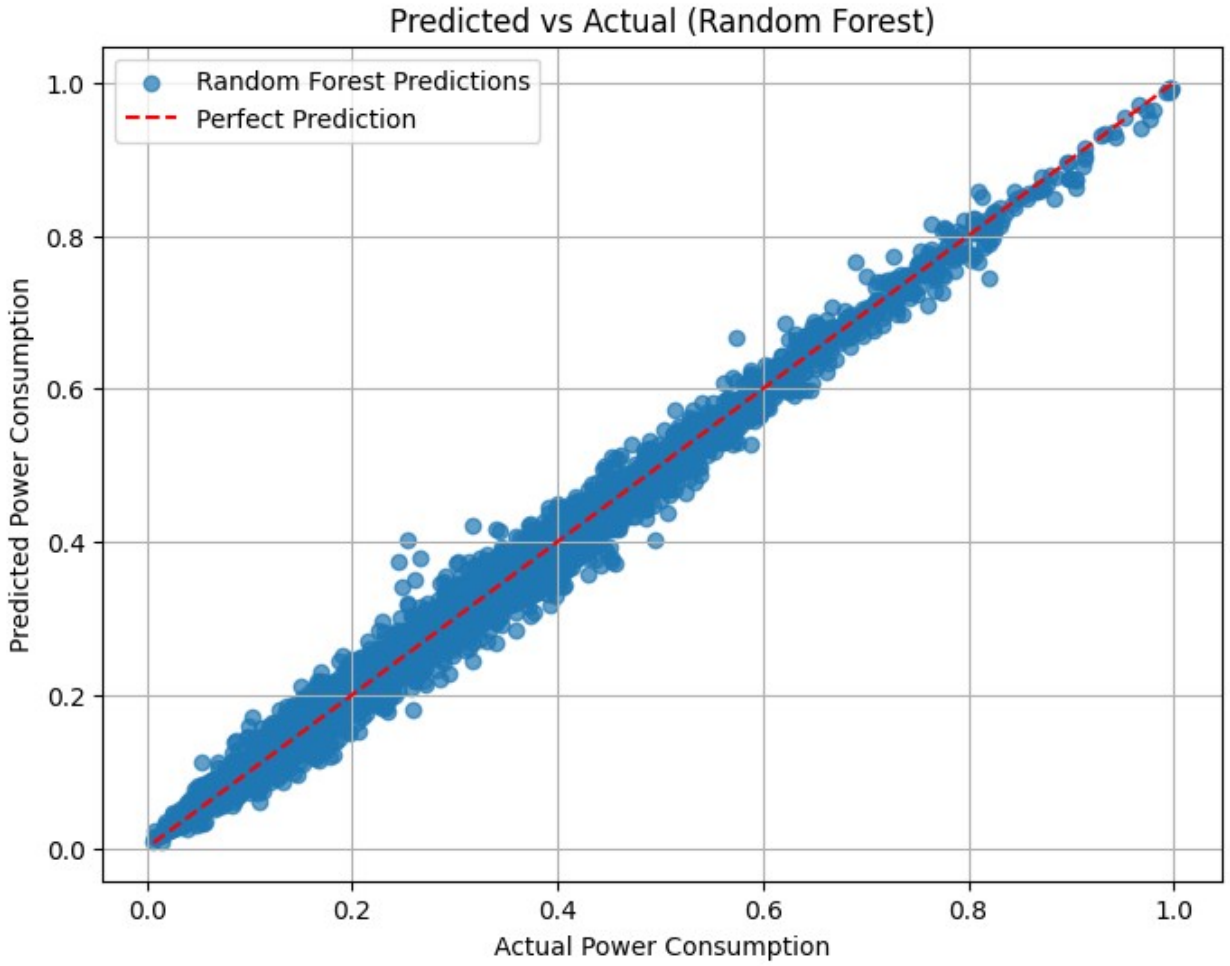

Figure 9: Predicted vs Actual (Random Forest)

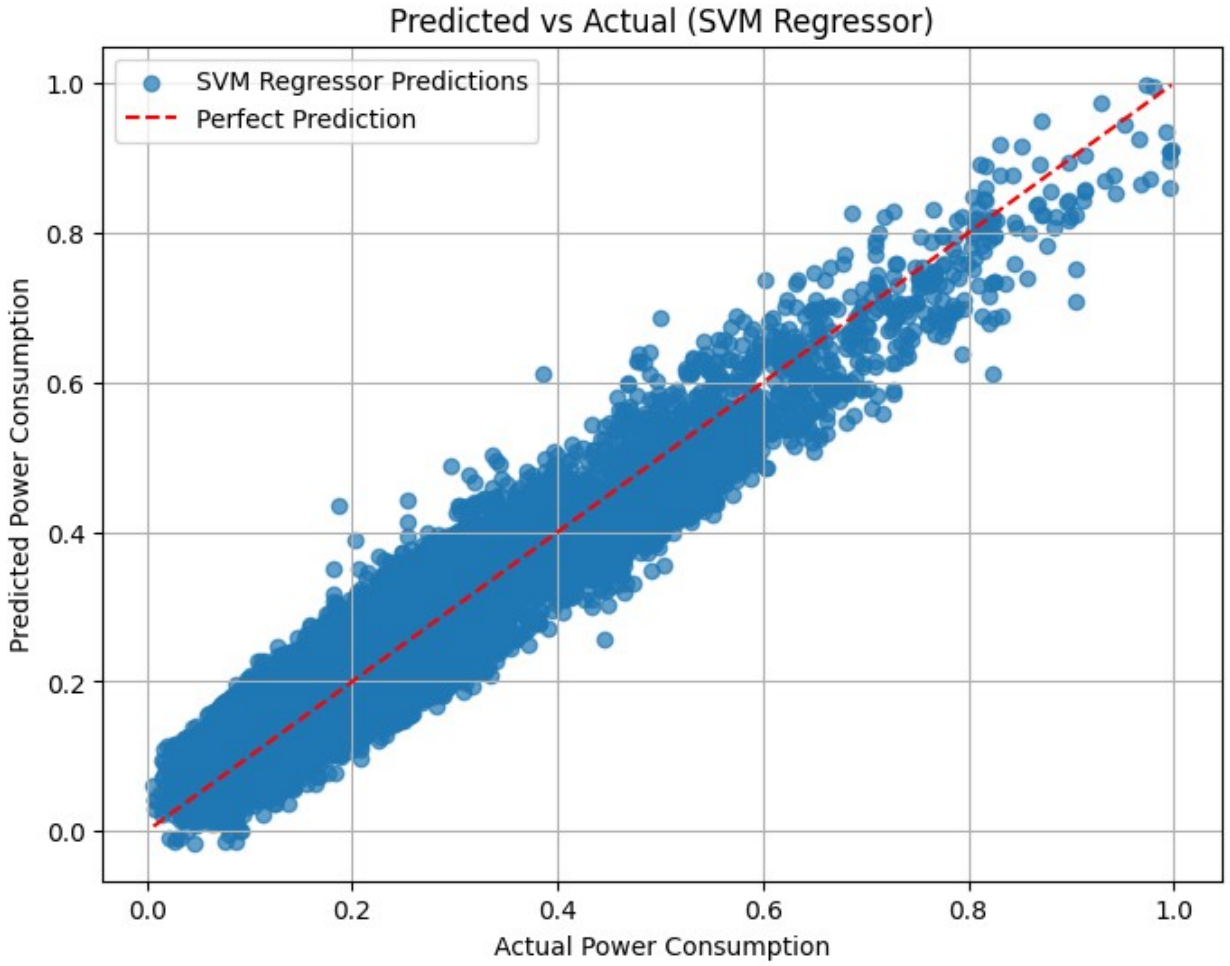

Figure 10: Predicted vs Actual (SVM Regressor)

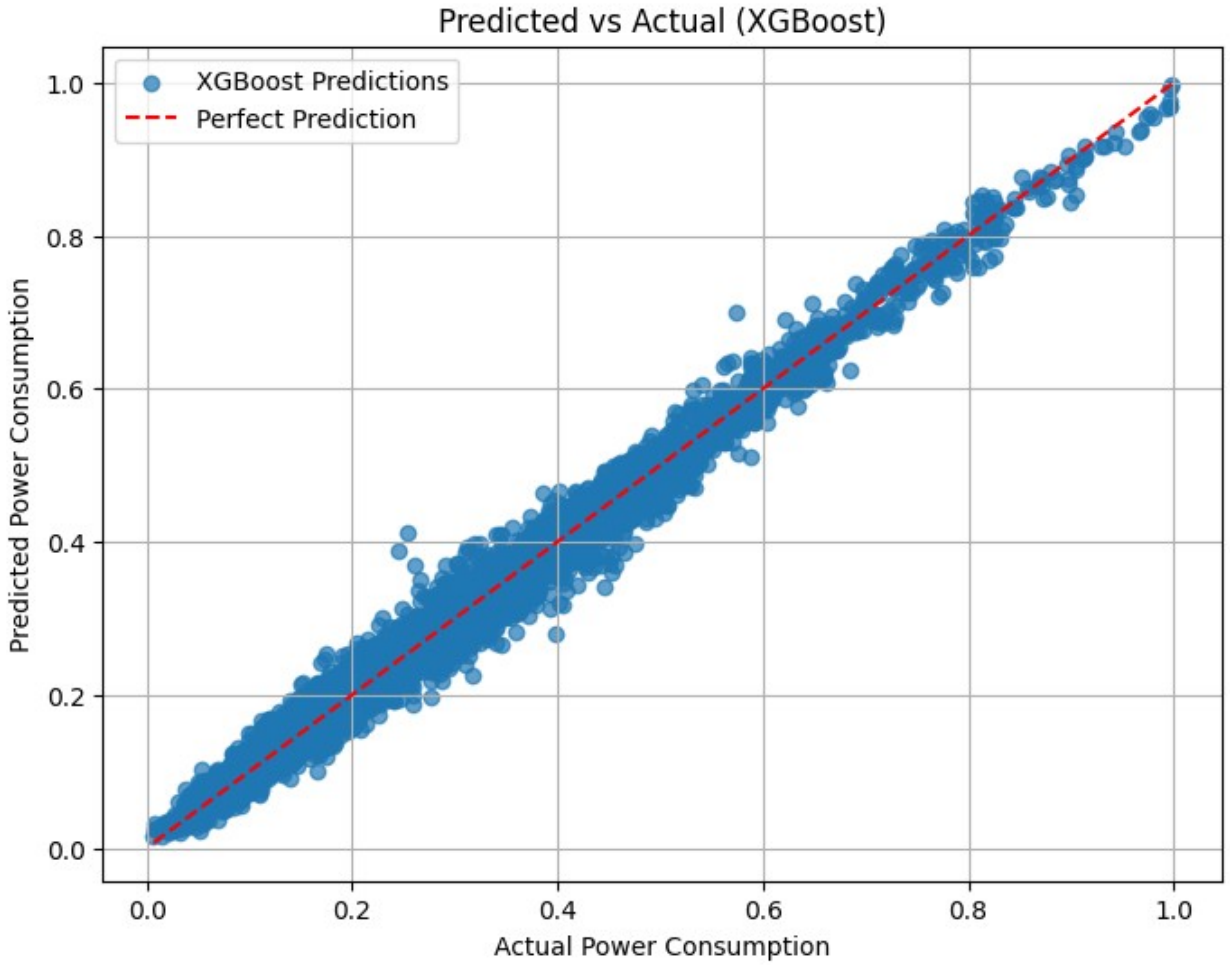

Figure 11: Predicted vs Actual (XGBoost)

H. Qualitative Impact:

The evaluation of machine learning models not only demonstrated clear performance hierarchies but also provided actionable insights for operational and strategic improvements. Random Forest and XGBoost, in particular, enabled accurate forecasting of energy consumption which allow Amendis to confidently plan for peak-load periods and address potential overload risks. Zone-specific insights facilitated more effective resource allocation, minimizing downtime and optimizing load distribution. Furthermore, the models uncovered correlations between environmental factors—such as temperature and solar diffuse flows—and energy demand, guiding renewable energy integration efforts. These findings support the deployment of solar power assets in high-demand zones like Zone 3 which align with Morocco's sustainability goals. In addition to improving operational efficiency, these predictive tools enhance customer satisfaction by reducing outages and ensuring reliable power supply, while also advancing Amendis' commitment to environmental stewardship.

VI. Recommendations

I. Key Takeaways:

Insights from this project underscore the importance of leveraging machine learning to enhance energy forecasting. Ensemble methods like Random Forest and XGBoost demonstrated superior predictive accuracy, making them reliable tools for operational planning. Feature selection techniques such as Recursive Feature Elimination (RFE) ensured models focused on impactful variables which improve interpretability and performance. Zone-specific strategies, including dynamic bandwidth adjustments and renewable energy integration, can significantly improve resource management. These methods provide a scalable framework that can be expanded as Amendis integrates more data sources.

J. Challenges Overcome:

Preprocessing the SCADA data was resource-intensive, involving the conversion of timestamps and normalization of environmental variables to ensure consistency. Handling non-linear relationships in the data proved difficult for simpler models like Linear Regression which necessitated the adoption of ensemble methods. Additionally, creating clear and actionable visualizations required iterative refinement to effectively communicate findings to stakeholders. Despite these setbacks, the project delivered robust, actionable insights to optimize Tetouan City's power distribution network.

VII. Sources

1. https://scikit-learn.org/stable/supervised_learning.html
2. https://www.semanticscholar.org/paper/Comparison-of-Machine-Learning-Algorithms-for-the-%3A-Salam-Hibaoui/177512e766fe5d8624a1b3e93abd11082ac37b3f
3. https://archive.ics.uci.edu/dataset/849/power+consumption+of+tetouan+city
4. https://scikit-learn.org/stable/modules/svm.html#regression
5. Course Material