REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Electrical and Computer Engineering

ENERGY CONSUMPTION ESTIMATION USING MACHINE LEARNING WITH DATA FROM SMART METERS IN A RESIDENTIAL COMPLEX BUILDING IN IRAQ

Noor Malik Safaa AL-SHAWWAF

Master's Thesis

Supervisor

Asst. Prof. Dr. Abdullahi Abdu IBRAHIM

İstanbul, 2024

ENERGY CONSUMPTION ESTIMATION USING MACHINE LEARNING WITH DATA FROM SMART METERS IN A RESIDENTIAL COMPLEX BUILDING IN IRAQ

Noor Malik Safaa AL-SHAWWAF

Electrical and Computer Engineering

Master's Thesis

ALTINBAŞ UNIVERSITY
2024

The thesis titled “ENERGY CONSUMPTION ESTIMATION USING MACHINE LEARNING WITH DATA FROM SMART METERS IN A RESIDENTIAL COMPLEX BUILDING IN IRAQ” prepared by NOOR MALIK SAFAA AL-SHAWWAF and submitted on 00/12/2023 has been **accepted unanimously** for the the drgree of Master of Science in Electrical and Computer Engineering.

Asst. Prof. Dr. Abdullahi Abdu IBRAHIM

Supervisor

Thesis Defence Committee Members:

Asst. Prof. Dr. Abdullahi Abdu IBRAHIM	Department Computer Engineering, Altınbaş University	
Asst. Prof. Dr.	Department of Engineering, Altınbaş University	
Asst. Prof. Dr.	Department of Engineering, University	

I hereby declare that this thesis meets all format and submission requirements of a Master’s thesis.

.

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Noor Malik Safaa AL-SHAWWAF

Signature

DEDICATION

As I reach the pinnacle of my academic journey with a Master's degree, this master's thesis dedicated to the tireless support of my family, friends, and mentors who stood by me through thick and thin.

To my loving parents, to my dear father and my first teacher who instilled in me the values of hard work and determination. Thank you for being my pillars of strength and for encouraging me to pursue my dreams, my father I know you are happy at afterlife world.

To my beloved husband, who has always stood by my side, providing me with love and unwavering support. Your patience and understanding during the times when I needed it the most appreciated more than words can express. You provided me with constant encouragement, and your faith in my abilities inspired me to keep pushing. Your sacrifices, late-night study sessions, and unwavering belief in my capabilities helped me stay focused.

To my thesis adviser, thank you for being patient with me and for always being available to guide me in the right direction. Your advice, wisdom, and mentorship helped me stay on track, and your unwavering belief in my capabilities encouraged me to aim higher.

To my inspiring professors and mentors, whose guidance, expertise, and encouragement have shaped my academic journey and who have challenged me to reach my full potential. Your passion for your fields of study and your willingness to share your knowledge have been instrumental in my academic growth.

I stand before you today, proud and grateful, as I express my heartfelt appreciation to the person who played a pivotal role in shaping my academic journey to my previous advisor Dr. Hasan Abd Al-Kader and my new advisor Dr. Abdullahi Abdu Ibrahim, you have been my mentor, my guide, and my support throughout the pursuit of my Master's degree, and for that, I am deeply grateful.

I am immensely grateful to my supervisor, Assistant Dr. Abdullahi Abdu Ibrahim, for his invaluable guidance, mentorship and encouragement throughout my journey towards obtaining a Master's degree. Dr. Abdullahi Abdu Ibrahim expertise in the field of engineering and his commitment to the success of his students has been an inspiration to me. His insightful feedback, constructive criticism, and unwavering support has truly made a difference in my academic and personal development.

I am also grateful to Dr. Hakan Koyuncu, the teaching staff, and my colleagues, who have been an essential part of this academic journey. The collaborative effort and collective knowledge shared have undoubtedly made this journey much more exciting, enriching, and fulfilling. You all have made a positive impact on my personal and professional growth.

To my dear friends, who have made this journey enjoyable and memorable, thank you for providing me with your love and laughter during the times when I needed them the most.

Thank you all for being part of my life and for supporting me on this journey towards my Master's degree. This achievement is as much yours as it is mine, and I will forever cherish the memories and the invaluable lessons learned.

Finally, I dedicate this thesis to all those who aspire to pursue higher education but are afraid of taking the first step. May my journey serve as an inspiration, a reminder that with hard work and the support of loved ones, anything is possible.

I close with gratitude and appreciation to all those who played a part in making my academic dreams come true. Without each and every one of you, I would not be where I am today.

Lastly, I extend my thanks to the engineering department staff for providing an excellent infrastructure and resources necessary to make my academic pursuits possible. Your dedication and support to our needs have not gone unnoticed and have played a significant role in shaping our academic journey.

In conclusion, I express my profound gratitude to everyone who contributed to making my Master's degree a success. I am privileged to have had Dr. Abdullahi Abdu Ibrahim as my supervisor, and I thank him from the bottom of my heart for his invaluable support and guidance. My academic journey would not have been complete without all of you, and for that, I am forever grateful.

Thank you once again to everyone who has played a role in my journey towards earning my Master's degree. Your support and encouragement has truly meant the world to me and will always be remembered with heartfelt gratitude.

Finally, to myself, for staying true to my goals and pushing through the obstacles to achieve this milestone. This is a testament to the strength and determination that resides within me.

ABSTRACT

ENERGY CONSUMPTION ESTIMATION USING MACHINE LEARNING WITH DATA FROM SMART METERS IN A RESIDENTIAL COMPLEX BUILDING IN IRAQ

AL-SHAWWAF, Noor Malik Safaa

M.Sc., Electrical and Computer Engineering , Altınbaş University,

Supervisor: Asst. Prof. Dr. Abdullahi Abdu IBRAHIM

Date: January / 2024

Pages: 126

The exponential growth in population and their overall reliance on the usage of electrical and electronic devices have increased the demand for energy production , energy saving schemes and nowadays a major worldwide concern. As the building sector is major energy consumer.

This research project aims to determine the energy usage of a residential complex building in Iraq by using machine learning strategies to the data collected by smart meters in that building. The critical focus is producing accurate projections of future loads in the residential complex over two or three years to surpass any potential artificial intelligence barriers. Data gathering, pre-processing, algorithm selection, model training, model assessment, and energy consumption estimation are the steps involved in the technique. The dataset has been corrected, and the three algorithms used are linear regression, decision trees, random forests ,gradient boosting, bagging regressor ,extra trees egressor

,SVR , lasso , ridge , elastic net , K-nearest nighbors and neural network. The protection of users' privacy and the safety of their data are among the ethical issues. An accurate calculation of the energy consumed is required to manage energy and reduce costs effectively. The findings of this study can be used to address the problem of inefficient management of energy resources in residential structures in Iraq and other countries with comparable conditions.

Keywords: Energy Consumption, Machine Learning, Smart Meters, Residential complex, Energy Management, Iraq.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

LIST OF CHARTS

Pages

ABBREVIATIONS

IEA	:	International Energy Agency
MOE	:	Ministry of Electricity
RMSE	:	Root Mean Square Error
MAE	:	Mean Absolute Error
R^2	:	R-squared
ARIMA	:	Auto Regressive Integrated Moving Average
LSTM	:	Long Short-Term Memory
LSTMs	:	Long Short-Term Memory networks
ANN	:	Artificial Neural Networks
RNNs	:	Recurrent Neural Networks
IoT	:	Internet of Thing
IQR	:	Interquartile Range
LOOCV	:	Leave-One-out Cross-Validation
COD	:	Coefficient of Determination
GDPR	:	General Data Protection Regulation
RAM	:	Random-Access Memory
CPU	:	Central Processing Unit
GPU	:	Graphics Processing Unit
Ltd	:	Limited
E.g.	:	For example
MLPRegressor	:	Multi-Perceptron Regressor
SVR	:	Support Vector Regression

LIST OF SYMBOLS

Y : It is Dependent Variable (Energy Consumption)

X : It is Independent Variable (E.G. Temperature, Time of Day)

0β0 : The Coefficients of the Linear Model 1β1

: The Coefficients of the Linear Model E :

Represents the Error Term

N : It is the Number of Samples in the Testing Set

Yi : It is the Actual Energy Consumption Value for Sample I

Yi^ : It is the Predicted Energy Consumption Value for Sample I

R2 : Coefficient of Determination

RSS : Sum of Squares of Residuals

TSS : Total Sum of Squares

N : It is the Number of Data Points in the Test Set

Yi : It is Actual Energy Consumption Value for I-Th Data Point

^Yi^ : It is Predicted Energy Consumption Value for I-Th Data Point

¯y¯ : It is the Mean of the Actual Energy Consumption Values

1. INTRODUCTION

Energy needs have skyrocketed in recent years as a result of factors like rising populations, increased reliance on electricity and other modern conveniences, expanding economies, and dramatic climatic shifts.and It is predicted that residential structures' energy consumption may rise over the next few years as a result of population expansion, urbanisation, and monetary development, making them one of the world's largest consumers of power [1]. The residential zone accounts for about 70% of Iraq's average electricity consumption , making it Iraq's largest electricity consumer [2], chart 1.1explain the increase in Iraqs energy consumption from 2017 to 2022.

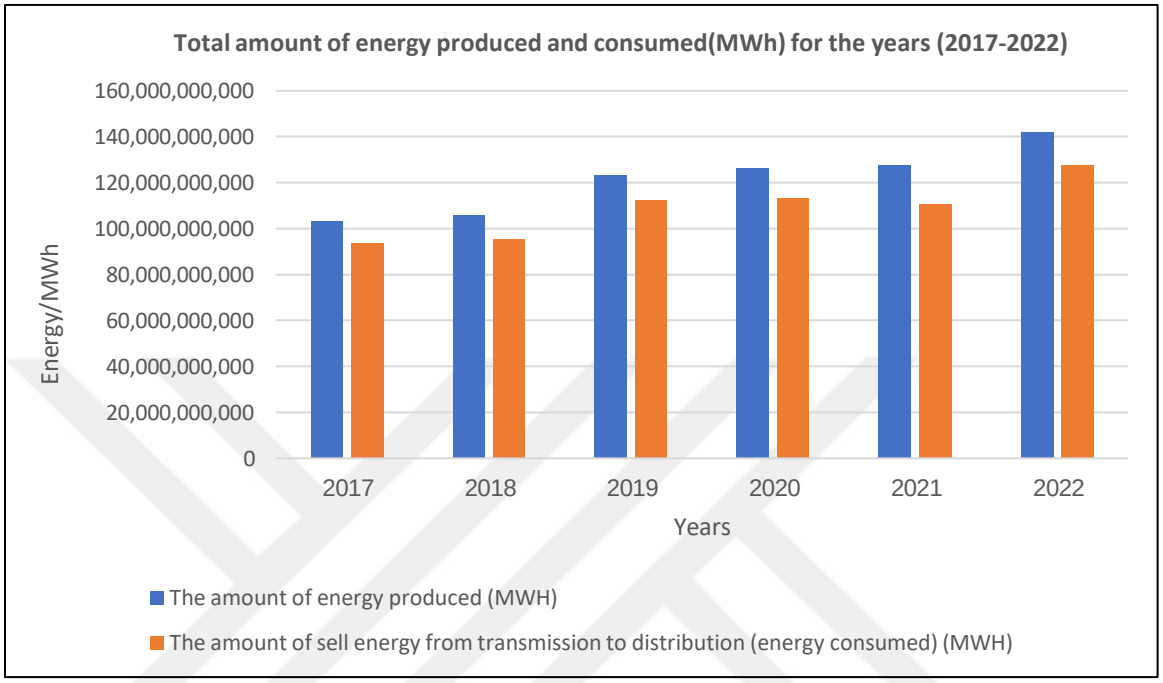
As a result, it's essential to encourage robust energy control in residential buildings to reduce wasted strength, convey down month-to-month electricity costs, and lessen the harmful impact of strength use on the environment.

Estimating the quantity of energy used in residential homes has historically been finished with statistical models and methodological strategies based on guidelines. However, these strategies have limits in phrases of accuracy and reliability, and they do now not think about the dynamic and complicated nature of energy intake in residential structures. Moreover, these strategies do not recall the character of the records being amassed. The improvement of intelligent meters and algorithms for system learning has made it feasible to nicely forecast the energy used in residential homes, establishing new possibilities. Machine-gaining knowledge of algorithms can analyze tremendous volumes of information and find patterns and traits, while intelligent meters supply real-time and complete records of the quantity of electricity consumed.

The purpose of this study is to employ methods similar to these to estimate the amount of energy consumed by a residential complex building in Iraq using the data collected from smart meters. Data collection, pre-processing, algorithm selection, model training, model assessment, and energy consumption estimation are the components that make up the methodology of the research. The dataset has been corrected, and the algorithm to be used are linear regression, decision tree, random forest, gradient boosting, bagging regressor, extra trees regressor, SVR, lasso, ridge, elastic net, K-nearest nighbors and neural networks.

By giving precise estimates of energy usage, the purpose of this research project is to, in the broadest sense, contribute to the effective control of energy consumption in residential structures in Iraq and other nations like it.

Chart 1.1: Diagram of Energy Consumption at Iraq from 2017 Until 2022.



1.1 BACKGROUND

The ever-increasing demand for energy in residential complexes has become a significant challenge for countries like Iraq, where population growth, urbanization, and economic development contribute to the surge in electricity consumption. Previous studies have shown that applying machine learning techniques in energy management, particularly for forecasting how much energy would be used, has significant promise. For example, Atanassov et al. (2019) conducted research in which they used machine learning to forecast the amount of energy used in residential buildings in Bulgaria [1]. The findings revealed that the strategy could potentially minimize prediction error by as much as 30 percent. Similarly, Hussain et al. (2020) [2] conducted another study in which they predicted the amount of energy used in commercial buildings using machine learning algorithms. They were successful in their forecast, obtaining an accuracy of over 90 percent.

This is a significant challenge because residential complexes account for a significant portion of a country's overall energy consumption (International Energy Agency, 2019) [3]. The residential sector is Iraq's most significant energy user, since it is responsible for a

sizeable share of the country's overall average energy consumption and ranks among its most energy-intensive sectors (Ministry of Electricity, 2018) [4]. It is vital for residential structures to have effective energy management in order to minimize the negative effects of energy consumption on the surrounding environment, reduce the amount of energy that is wasted, and lower the monthly cost of energy.

Historically, the estimation of energy consumption in residential buildings relied on statistical models and rule-based methodologies, both of which frequently lacked accuracy and reliability due to their limited consideration of dynamic and complex energy consumption patterns. In recent years, however, there has been a shift toward more modern methods of estimating energy consumption in residential buildings. In addition, these old systems did not take use of the real-time and comprehensive data offered by smart meters, which reduced their capacity to accurately predict future energy usage.

The development of intelligent meters and progress made in machine learning algorithms provide exciting new possibilities to transform energy management and increase the accuracy of energy consumption forecasting. While machine learning algorithms can analyze enormous amounts of data, recognize trends, and generate accurate forecasts, smart meters can give information on real-time energy use at a granular level.

The history of smart meters in Iraq is relatively recent, and their usage has been driven by the need to improve energy management and address various challenges faced by the country's electricity sector. Smart meters were introduced in Iraq as part of the government's efforts to modernize the electricity infrastructure and enhance energy efficiency in residential and commercial buildings.

Historically, Iraq has faced significant challenges in its electricity sector, including electricity shortages, unreliable energy supply, and high energy wastage. The demand for electricity has been steadily increasing due to population growth, urbanization, and economic development, leading to an increasing strain on the energy grid. Additionally, inefficient energy management practices and outdated metering systems contributed to significant energy losses and revenue leakage for the government.

To address these issues, the Iraqi government initiated a modernization plan for its electricity sector, which included the deployment of smart meters. Smart meters are digital devices that can measure and record electricity consumption in real-time or at regular intervals. Unlike traditional meters, smart meters provide two-way communication

capabilities, allowing both the utility and consumers to monitor and manage energy usage more effectively.

1.2 PROBLEM STATEMENT

The old methods for calculating the amount of energy consumed in residential buildings do not have the level of precision necessary for efficient energy management and cost reduction. Because of the dynamic and complicated nature of energy consumption patterns, as well as the restricted consideration of real-time data, these traditional methodologies have limited accuracy. In order to solve this problem, it is necessary to apply the principles of machine learning and to utilize the information that is gathered from smart meters in order to provide precise forecasts of the energy that was consumed in residential complexes in the near future.

1.3 OBJECTIVE

This research project's principal purpose is to apply machine learning algorithms to anticipate the quantity of energy consumption in a residential complex building in Iraq. The building in question is (Iraq gate) located in Baghdad and its the largest and modern project in all over Iraq. Particular goals consist of the following:

a. Collect information on the amount of energy used by the building using the smart meters that have been placed in the residential complex.
b. The data should be preprocessed so that noise and inconsistencies may be removed, which was ensure that the data are accurate and reliable.
c. Determine which machine learning algorithm was provide the best accurate prediction of future energy use by comparing the effectiveness of a number of different methods.
d. In order to identify the patterns and trends in energy use, train the machine learning model by utilizing the pre-processed data.
e. Evaluate the usefulness of the machine learning model in properly estimating energy usage by measuring its performance using measures such as root mean square error and mean absolute error.
f. Utilizing the trained machine learning model, formulate an estimate for the amount of electricity that was utilized by the residential complex over the following two to three years.

1.4 RESEARCH QUESTIONS

a. How does the energy consumption pattern vary in a residential complex building in Iraq, and what are the factors influencing these variations?
b. How can machine learning algorithms be applied to predict future energy usage in the residential complex using data from smart meters, and which algorithm demonstrates the highest accuracy in energy consumption estimation?
c. What are the performance metrics (such as Root Mean Square Error and Mean Absolute Error) for different machine learning algorithms in predicting energy consumption, and which algorithm yields the most reliable results?
d. How well does the machine learning model estimate the amount of energy consumed in the residential complex over the next two to three years, and how does the predicted consumption align with the actual energy usage data?
e. What are the potential implications of accurate energy consumption estimation for effective energy management, cost reduction, and sustainable energy utilization in residential structures in Iraq and similar regions?

1.5 SIGNIFICANCE

The findings of this study project have major significance for successful energy management in residential buildings in Iraq and other nations that are confronting issues that are comparable to those that Iraq faces. An assessment of energy consumption that is both accurate and exact can result in a reduction in energy waste, an improvement in energy efficiency, and the most effective exploitation of any available energy resources. This project seeks to discover consumption patterns and trends that may not be visible using standard techniques of data analysis by utilizing smart meters and machine learning algorithms. This was accomplished by leveraging technology.

The use of techniques from machine learning in energy management can permit fast reactions to fluctuations in energy demand, which in turn enables building managers to make educated decisions regarding energy use. The findings of this study have the potential to offer policymakers, energy managers, and building operators significant insights that can be used to improve energy management methods and reduce the negative impact that energy use has on the environment.

1.6 SCOPE AND LIMITATION

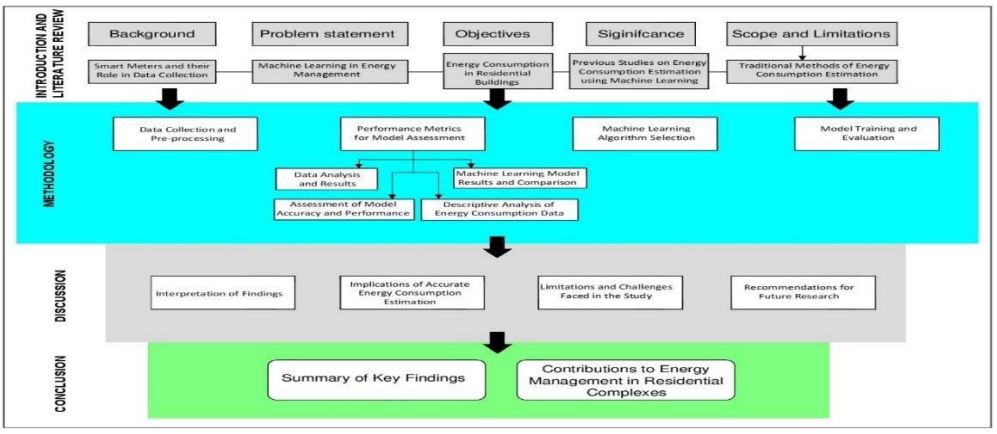
Using information gathered from smart meters, the scope of this study project has been narrowed down to explicitly estimate the amount of electricity used in a residential complex building located in Iraq. The scope of the study include activities such as data collection and pre-processing, method selection, model training and assessment, as well as calculation of energy usage. Nevertheless, it is necessary to recognize that there are certain restrictions:

a. The quality and granularity of the data obtained from smart meters was determine how accurate the energy consumption estimate was. If there are any problems with the data collecting, it might affect how accurate the model is.
b. The outcomes of the study and the performance of the model might be different depending on the machine learning technique that was chosen and the size of the dataset that was made available for training.
c. Although machine-learning algorithms can increase estimates of energy use, it is important to keep in mind that other external factors, such as shifts in weather and user behavior, can also have an impact on energy consumption and may not be fully accounted for by the model.

This research project aims to provide valuable insights into accurate energy consumption estimation, which can contribute to more efficient energy management in residential structures in Iraq and other regions facing similar challenges. Despite these limitations, the project's goal is to provide valuable insights into accurate energy consumption estimation.

1.7 THESIS STRUCTURE

Chart 1.2: Thesis Structure Project.



2. LITERATURE REVIEW

2.1 INTRODUCTION

This chapter provides a comprehensive literature analysis to investigate the status of estimating energy consumption in residential complex buildings using machine learning using data obtained from smart meters. The purpose of this literature review is to investigate the value of data from smart meters and the role that machine learning algorithms play in providing an accurate forecast of energy usage. This chapter was providing insights into the effectiveness of machine learning techniques and their potential impact on energy management in residential structures by analysing previous research and case studies. These studies were used to provide insights into the effectiveness of machine learning techniques. The study project is broken down into several different categories, and each of those categories is addressed in a separate portion of the literature review.

2.2 ENERGY CONSUMPTION ESTIMATION IN RESIDENTIAL COMPLEX

It is necessary for efficient energy management to have an accurate calculation of the amount of electricity used in residential complexes. Traditional techniques of estimating, such as statistical models and rule-based approaches, have limits in terms of accuracy and flexibility to dynamic energy consumption patterns. Traditional methods of estimation, such as statistical models and rule-based approaches, have drawbacks (Atanassov et al., 2019) [5]. On the other hand, the advancement of machine learning algorithms and the availability of real-time data from smart meters have opened new opportunities for projections of energy usage that are more precise and dependable. Machine learning has been shown to have great promise in the field of energy management, and previous research has shown that it can surpass traditional approaches in terms of the accuracy of its predictions.

2.2.1 Challenges in Energy Consumption Estimation

Due to the dynamic and varied nature of energy usage, estimating the amount of electricity consumed in residential complex buildings involves several issues. The following are some of the most significant difficulties:

a. Variability In Energy Usage: Energy consumption patterns in residential buildings can vary significantly based on factors such as weather conditions, occupancy, and individual user behaviors. Traditional methods often struggle to capture these variations accurately.
b. Non-Linear Relationships: The relationship between energy consumption and influencing factors may not always be linear. Non-linearities make it challenging conventional statistical models to capture the complexity of energy usage patterns.
c. High-Dimensional Data: Smart meters and other sensing technologies in modern buildings generate vast amounts of data. Analyzing and extracting meaningful insights from this high-dimensional data require advanced data processing and modeling techniques.
d. Data Sparsity And Noise: In some cases, data may be missing or incomplete due to meter failures or other issues. Additionally, the data collected from smart meters may contain noise or outliers that can impact the accuracy of the estimation.
e. Real-Time Prediction: Many applications, such as demand response and load balancing, require real-time energy consumption prediction. Traditional methods may not be able to provide timely and accurate forecasts in such scenarios.

2.2.2 Advancements with Machine Learning

The use of machine learning as a potential solution to the difficulties associated with estimating Energy consumption in residential complex buildings is becoming increasingly popular. Machine learning algorithms, in contrast to traditional approaches, which rely on predetermined rules or assumptions, can adapt to new situations and learn from data in order to produce predictions.

a. Regression Models: Linear regression, polynomial regression, and support vector regression are commonly used machine learning algorithms for energy consumption estimation. These models can capture the relationships between energy usage and influencing factors, even in non-linear scenarios.
b. Time Series Forecasting: Time series forecasting algorithms, such as ARIMA (Autoregressive Integrated Moving Average) and LSTM (Long Short-Term Memory), are well-suited for capturing temporal dependencies in energy consumption data. They can handle seasonality and trends in long-term energy usage patterns.

c. Ensemble Methods: Ensemble methods, like Random Forest and Gradient Boosting, combine multiple models to improve prediction accuracy. They can mitigate the impact of noisy data and handle high-dimensional feature spaces effectively.
d. Deep Learning: Deep learning techniques, including neural networks, have shown promise in energy consumption estimation tasks. They can automatically learn hierarchical representations from data and capture complex relationships between variables.
e. Anomaly Detection: Machine learning models can also be utilized for anomaly detection in energy consumption data. Identifying unusual energy usage patterns can help detect faults or inefficient energy usage in the building.

2.3 TRADITIONAL METHODS OF ENERGY CONSUMPTION ESTIMATION

For several decades, conventional approaches of estimating energy consumption have been utilized in order to approximate the amount of energy that is used in residential structures. These techniques include of statistical models and rule-based approaches, both of which derive their predictions from prior data and are governed by a set of predetermined procedures. Although these methods have seen widespread use, they suffer from several drawbacks that prevent them from being accurate and prevent them from adapting to changing patterns of energy consumption.

Estimation of energy usage has often been performed with the use of statistical methods like time-series analysis and moving averages (Khan & Wu, 2018) [6]. These models are constructed using historical data on energy usage and assume that patterns of consumption from the past was continue into the foreseeable future. In spite of the fact that they are able to provide accurate forecasts for consistent energy use, they frequently fail to catch unexpected shifts in consumption patterns or abnormalities in that behavior. In addition, these approaches may not take external variables into consideration, such as changes in the behavior of inhabitants or variations in the weather, which can have a considerable impact on the amount of energy that is consumed.

Rule-based techniques make use of predetermined algorithms that estimate energy usage in accordance with a set of parameters based on a set of specified criteria (Sutharshan & Jirutitijaroen, 2018) [7]. To get an approximation of energy use, a rule-based method may, for instance, consider the number of inhabitants, the time of day, and the weather

conditions. While it may not be difficult to put these strategies into practice, they frequently lack the adaptability necessary to accommodate a variety of building types or the behaviors of occupants. As a consequence of this, their precision might be reduced, which is especially problematic in intricate residential constructions that have fluctuating energy requirements. In addition, in order to be trained, statistical models and rule-based techniques both need access to historical data, and the accuracy of their predictions is highly dependent on both the availability of the dataset and the quality of the information included within it. These approaches might not produce accurate estimates in circumstances in which there is a lack of data or data that cannot be relied upon.

Researchers and energy industry specialists have begun to implement machine learning strategies in order to circumvent the constraints imposed by conventional approaches. Algorithms that learn through machine learning are able to examine huge volumes of data, including information gathered in real time by smart meters, in order to uncover intricate patterns and correlations that have an impact on how much electricity is consumed. Machine learning models are able to constantly learn and improve their predictions over time since they make use of sophisticated algorithms. This makes machine learning models more suited for dynamic and developing energy consumption scenarios.

In the context of estimating energy consumption, machine learning models may take into consideration not just previous patterns of energy usage but also environmental elements like temperature, humidity, and the time of day to produce more accurate forecasts. Because machine-learning algorithms are able to handle various variables and non-linear interactions, they are attractive candidates for improving the prediction of the amount of electricity consumed in residential complex buildings.

2.4 SMART METERS AND REAL –TIME DATA COLLECTION

The collection of data in real time regarding residential buildings' energy use is made significantly easier by the use of smart meters. These cutting-edge gadgets deliver comprehensive reports on the consumption of electrical energy at regular intervals, often once every 15 minutes or once every hour (Khan & Wu, 2018) [6]. The availability of data in real time offers a more detailed knowledge of patterns of energy usage, which in turn enables improved decision-making in the context of energy management. Consumers are

able to better manage their energy use and associated expenses with the assistance of demand response programs, which are supported by smart meters: the type of smart meter, which is using right now (LUNA LSM45), See figure 2.1 below:

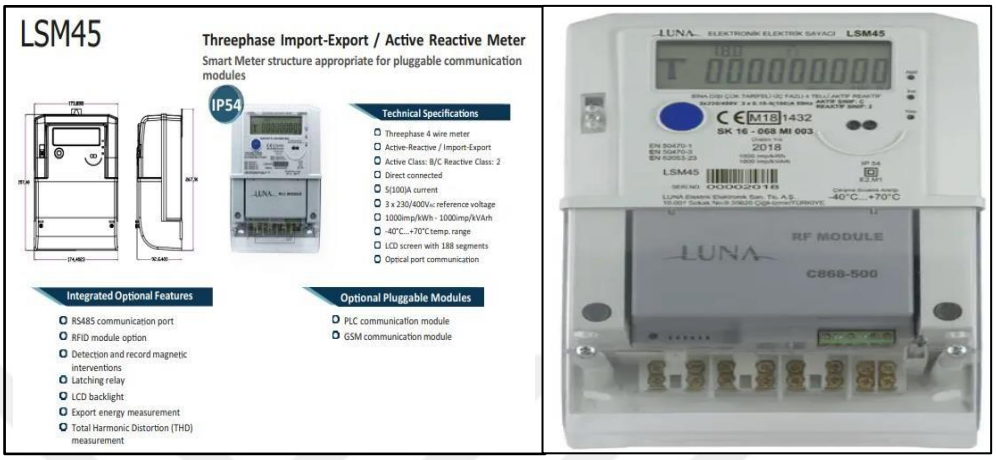


Figure 2.1: Smart Meter (LUNA LSM45).

2.4.1 Advantages of Smart Meters

Smart meters offer several advantages over traditional analog meters, making them indispensable tools in the quest for efficient energy management in residential complex buildings [8]:

a. Real-Time Monitoring: Smart meters provide real-time data on electricity consumption, allowing users and energy managers to track energy usage patterns as they happen. This immediate feedback empowers residents to make informed decisions about their energy usage and identify opportunities for energy savings.
b. Granular Data: Unlike conventional meters that provide monthly readings, smart meters offer granular data, capturing energy consumption at shorter intervals. This fine-grained data allows for a deeper understanding of energy usage patterns and facilitates the detection of irregularities or anomalies.
c. Automated Meter Reading: Smart meters eliminate the need for manual meter readings, reducing human errors and the associated costs of manual data collection. The automated process also enhances the efficiency of billing and energy management operations.
d. Two-Way Communication: Smart meters can communicate with energy utilities in a bidirectional manner. This communication capability enables utilities to remotely

monitor energy consumption, manage energy distribution, and implement demand-response strategies.

e. Load Profiling: With real-time data, smart meters enable load profiling, which involves analyzing and characterizing energy consumption patterns over specific time periods. Load profiling is valuable for forecasting energy demands and optimizing energy distribution.

2.4.2 Use of Real-Time Data in Energy Management

Real-time data collected by smart meters can be leveraged in various energy management strategies for residential complex buildings [9]:

a. Demand Response Programs: Energy utilities can use real-time data to implement demand response programs. By incentivizing residents to adjust their energy usage during peak hours, utilities can reduce strain on the grid and prevent blackouts.
b. Energy Conservation: Real-time data empowers residents to monitor their energy consumption and identify areas of energy wastage. By actively managing their electricity usage, residents can reduce their overall energy consumption and contribute to sustainability goals.
c. Fault Detection: Smart meters' real-time data can be used for fault detection and rapid response to energy-related issues. For example, sudden spikes in energy usage may indicate equipment malfunctions or leaks, allowing prompt rectification.
d. Load Balancing: Real-time data assists energy managers in optimizing load balancing across the residential complex. By distributing the energy load evenly, energy managers can avoid overloading specific circuits and enhance overall system efficiency.
e. Energy Billing: With real-time data, energy billing can be more accurate and transparent. Residents can access detailed energy usage information and track their consumption, leading to better awareness and control of energy costs.

2.5 MACHINE LEARNING

Machine learning is a game-changing area of AI that is essential for reliable estimates of building energy use. Without being explicitly programmed, machine learning allows

computers to learn patterns, make predictions, and steadily improve their performance over time.

In our research project, machine learning serves as a powerful tool for modeling and forecasting energy usage trends. This process involves several key components:

a. Data Collection And Pre-Processing: We gather historical energy consumption data, including temperature, current, voltage, and other relevant variables, to create a comprehensive dataset. Data preprocessing techniques applied to clean, normalize, and prepare the data for analysis.
b. Machine Learning Algorithm Selection: We choose suitable machine learning algorithms that align with our research objectives. Common algorithms for regression tasks, such as Linear Regression, Random Forest, and Neural Networks, considered for predicting energy consumption.
c. Model Training And Evaluation: The selected algorithms trained on a portion of the dataset, learning from historical patterns and relationships within the data. Another part of the dataset reserved for testing and evaluating the model's performance. The model iteratively refines its predictions to minimize errors.
d. Performance Metrics For Model Assessment: The precision and consistency of machine learning models evaluated using many performance measures. These include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R2).

Machine learning provides the capability to capture complex temporal dependencies, identify non-linear relationships, and adapt to changing patterns in energy consumption. By harnessing the energy of these algorithms, our research aims to enhance the precision of energy consumption estimation, contributing to more efficient energy management in residential complexes.

Through the application of machine learning techniques, we strive to unlock valuable insights from data, ultimately leading to informed decisions and improved energy efficiency in the residential context.

2.6 PROPOSED ALGORITHM

In order to make an accurate estimate of the amount of electricity used in a residential complex building, the suitable machine learning algorithms must be use. These algorithms must be able to capture the numerous correlations that exist within the data. In this part, we

show the algorithms that selected for this research project and highlight the relevance of these algorithms with regard to the estimation of energy consumption.

2.6.1 Algorithm Implementation

Python and the necessary machine learning libraries used to implement the selected algorithms. For example, scikit-learn was used to implement Random Forest and Gradient Boosting, while Tensor Flow or PyTorch was used to create LSTM networks. The following steps make up the implementation process:

a. Data Preparation: The pre-processed information, which was include readings from smart meters as well as auxiliary data, was divided into input characteristics, also known as independent variables, and the target variable (energy consumption).
b. Algorithm Configuration: Techniques such as grid search and random search was utilized in order to fine-tune the hyper parameters of each algorithm. Some examples of these hyper parameters are the number of trees in Random Forest and the learning rate in Gradient Boosting.
c. Model Training: The training data used to train the algorithms, which was enable the algorithms to understand the fundamental correlations that exist between the input characteristics and the patterns of energy usage.
d. Model Validation: The models' overall effectiveness evaluated, and any necessary adjustments to their hyper parameters made, using the validation data. This stage avoids the models from being over fit and ensures that they generalize effectively to data that has not yet been observe.
e. Model Evaluation: Finally, the results of the tests was use to evaluate the performance of the models. Accuracy and performance of each method may be evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R- squared value.

2.6.2 Ensemble Method

An ensemble technique may be used if one wanted to further improve the accuracy of their forecasts regarding energy usage. The final forecast that is produced by using ensemble techniques is more reliable and accurate than those produced by using individual models alone. Combining the results of several machine learning algorithms, such as Random

Forest, Gradient Boosting, and LSTM models, can be accomplished in a number of ways. For instance, one could use a weighted average or a voting method.

The energy consumption can be more accurately estimated using ensemble approaches since these methods can compensate for the shortcomings of individual models while also capitalizing on their strengths.

This research project attempts to properly estimate energy consumption within the building of a residential complex by making use of a mixture of the algorithms known as Random Forest, Gradient Boosting, and LSTM. The goal of the study is to capitalize on the strengths that each method possesses. The subsequent parts are going to go into the training, validation, and assessment of these models, and then the subsequent section is going to estimate the energy usage based on their insights.

2.7 MACHINE LEARNING ALGORITHMS FOR ENERGY CONSUMPTION PREDICTION

First, we must define what machine learning (ML) is it is important to understand to what each word refers. Machine learning is composed of two notions: machine and learning. The machine is the subject of the process, it must learn and develop itself, and that is from where learning comes. Actually, learning means changing the behavior with experience. In machines, learning involves the amelioration of prediction based on changes in data or the program. Tom Mitchell gave the following definition when asked about machine learning: A computer program said to learn from experience Ewith respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with Experience E .and there is five prominent languages in the field of machine learning through analysis and comparison they are Java script, Lisp9, Java9R, Python , Figure 2.2 below explains why we chose Python for this research.

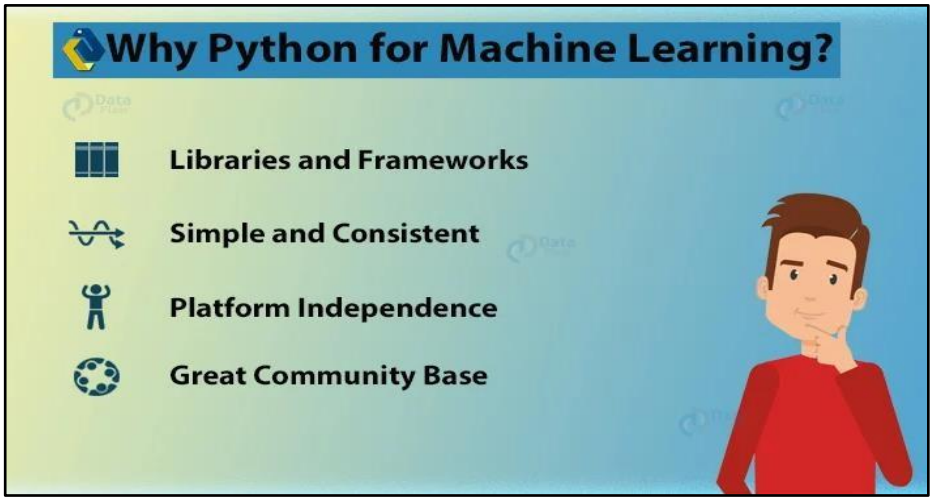


Figure 2.2: Reason of Using Python at Machine Learning.

The prediction of energy consumption may be accomplished using a wide variety of methods offered by machine learning. In the process of energy estimation, supervised learning methods such as regression, support vector machines (SVM), and artificial neural networks (ANN) have seen widespread use (Sutharshan & Jirutitijaroen, 2018) [7]. These algorithms train their models with historical data containing values for previously measured energy consumption, and then utilize those models to produce forecasts about future energy use. On the other hand, unsupervised learning methods like as clustering and anomaly detection have been utilized for energy analysis. These algorithms have been used to group similar patterns of energy use and to discover anomalous patterns of energy usage, respectively (Jain et al., 1999) [8].

Figure 2.3 , Describe the 8 python machine learning algorithms that we chose some of them and apply at my thesis.

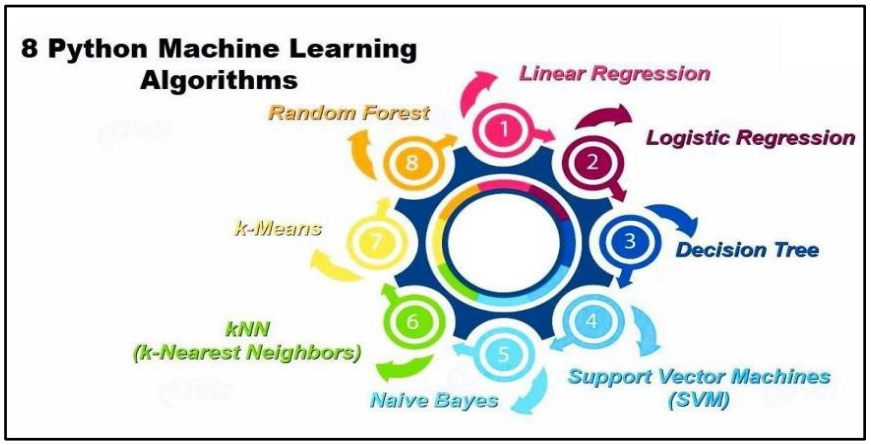


Figure 2.3: 8 Python Machine Learning Algorithms.

2.7.1 Linear Regression

When it comes to machine learning techniques, linear regression is one of the most straightforward and popular choices for estimating future energy usage [9]. It is a supervised learning approach that builds a linear connection between the dependent variable (energy consumption) and one or more independent variables. In other words, the energy consumption serves as the dependent variable (e.g., temperature, time of day, occupancy). The algorithm draws a line that best matches the data points, so reducing the amount of variance that exists between the measured and the anticipated levels of energy use [10].

The following is an example of a straightforward linear regression model's formula showing in equation 2.7.1:

$$Y=\beta 0+\beta 1X+\varepsilon \tag{2.1}$$

Where:

a. Y is the dependent variable (energy consumption).

b. X is the independent variable (e.g., temperature, time of day).

c. $0\beta 0$ and $1\beta 1$ are the coefficients of the linear model.

d. ε represents the error term.

Linear regression is computationally efficient and interpretable, making it a popular choice for initial energy consumption prediction tasks. However, its accuracy may be limited in capturing complex non-linear relationships in energy consumption data [11].

2.7.2 Decision Trees

Energy consumption data has complicated linkages, but decision trees, a non-linear machine learning technique, are up to the task. Decision trees create a tree-like structure by repeatedly subdividing the data into subgroups depending on the most important features. Attribute-based decisions are represented by the core nodes, while estimated energy consumption values are shown by the leaf nodes. [12].

The advantages of decision trees include their ability to handle both numerical and categorical data, easy interpretability, and resistance to outliers. However, decision trees may suffer from overfitting, especially when the tree depth is not appropriately controlled [13].

2.7.3 Random Forest

The goal of Random Forest, an ensemble learning approach, is to increase prediction accuracy while decreasing overfitting by combining numerous decision trees. The training procedure involves building many decision trees and then using the average of their predictions to make a final judgement. [14].

By including a random element into the tree-building process, Random Forest is able to overcome the constraints of using individual decision trees. Each decision tree learns from a different subset of the data, and at each node, only some of the characteristics are taken into account for making a split. By using a wide variety of building blocks, decision trees may decrease error and increase precision. [15].

2.7.4 Ridge Regression

Overfitting may be avoided with the use of the penalty element in the cost function that Ridge Regression introduces by adding L2 regularization to the linear regression cost function.

The algorithm aims to minimize the sum of squared differences between the predicted and actual values, along with the regularization term, Ridge Regression can be solved using techniques like closed-form solutions or optimization algorithms like gradient descent.
The algorithm aims to minimize the sum of squared differences between the predicted and actual values, along with the regularization term, Ridge Regression can be solved using techniques like closed-form solutions or optimization algorithms like gradient descent.

2.7.5 Lasso Regression

Lasso Regression is another linear regression technique, but it adds L1 regularization to the cost function, Similar to Ridge, Lasso aims to prevent overfitting by adding a penalty term to the linear regression cost. However, Lasso uses L1 regularization, which encourages sparsity in the model by driving, the errors made by the previous ones, at each step, a new tree trained to predict the residual errors of the previous ensemble. This process continues until a stopping some coefficients to exactly zero. This makes it useful for feature selection, Lasso can be solved using optimization techniques like coordinate descent.

2.7.6 Support Vector Regression (SVR)

Based on the success of support vector machines in classification, SVR applies those same ideas to regression. It seeks to find a hyperplane that has the highest margin while also minimising the discrepancy between the anticipated and actual values.
To discover a nonlinear connection between input and output variables, SVR maps the data into a higher-dimensional space using a kernel function. The effectiveness of the SVR model is affect by the selection of the kernel function (which might be linear, polynomial, or radial basis function, among others).

2.7.7 Gradient Boosting

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, i.e., models that make very few assumptions about the data, which are typically simple decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods,

but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

Advantages of Gradient Boosting are often provides predictive accuracy that cannot be trumped. Lots of flexibility - can optimize on different loss functions and provides several hyper parameter tuning options that make the function very fit.

2.7.8 Elastic Net

Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models.

Elastic Net Regression is a versatile tool in the data scientist's toolkit, offering a robust way to handle complex datasets with multi-collinearity and high dimensionality. Its ability to perform feature selection and manage the bias-variance trade-off makes it a valuable algorithm for many machine-learning tasks.

2.7.9 Multi-Perceptron Regressor (MLPRegressor)

An MLPRegressor is a convolutional neural network designed specifically for regression analysis. It built from an input layer, one or more hidden layers, and an output layer, all of which are composed of neurons (nodes).

Each neuron in the network applies a weighted sum of its inputs, passes the result through an activation function, and forwards it to the next layer. The model learns the optimal weights during training using techniques like backpropagation and gradient descent.

MLPRegressor can approximate complex, nonlinear relationships between input features and target values, making it suitable for a wide range of regression problems.

2.7.10 Extra Trees Regression

This class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over the best for feature selections.Fisher score is one of the most widely used supervised feature selection methods. The algorithm we will use

returns the ranks of the variables based on the fisher's score in descending order. We can then select the variables as per the case.

2.7.11 K-Neighbors Regressor

An instance-based and non-parametric regression method is the K-Nearest Neighbours Regressor. In order to make a prediction; it locates the K-nearest neighbours of a particular data point in the training dataset and uses a weighted average of the predicted values from those neighbours to make the final prediction. How many neighbours are included into the forecast is determine by the value of K, and those neighbours' weights might be uniform or distance based.

K-Neighbours Because of its ease of use and implementation, Repressor is a great option for datasets of any size. However, it may not do well with high-dimensional data and is prone to being picky about the value of K.

2.7.12 Neural Networks

Neural networks, particularly deep learning models, have shown promising results in energy consumption prediction tasks due to their ability to capture intricate relationships in data. Deep learning models consist of multiple layers of interconnected neurons that can learn complex representations from raw data [16].

For energy consumption prediction, recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) are commonly used. RNNs are well-suited for sequential data, making them ideal for time series prediction tasks. LSTMs, a type of RNN, address the vanishing gradient problem and can efficiently capture long-term dependencies in time series data [17].

2.7.13 Model Evaluation Metrics

When evaluating the performance of machine learning algorithms for energy consumption prediction, various metrics are used. Commonly used evaluation metrics include [18]:

a. Mean Absolute Error (MAE): MAE measures the average absolute difference between the actual and predicted energy consumption values. It provides a straightforward interpretation of the prediction error.

b. Root Mean Square Error (RMSE): RMSE is the square root of the average of the squared differences between the actual and predicted energy consumption values. It penalizes larger prediction errors more than MAE and is widely used in regression tasks.
c. R-Squared (R^2): R-squared represents the proportion of variance in the energy consumption data explained by the machine-learning model. It ranges from 0 to 1, with higher values indicating better model fit.

2.8 CASE STUDIES OF MACHINE LEARNING IN ENERGY MANAGEMENT

Several case studies have been conducted to evaluate the effectiveness of machine learning in energy consumption estimation. These studies often involve the implementation of machine learning algorithms on real-world datasets obtained from smart meters. Results show that machine learning models can achieve high accuracy in predicting energy consumption, thus facilitating better energy management in residential complexes [19]. These case studies provide valuable insights into the performance and practicality of machine learning techniques in real-life scenarios.

2.8.1 Case Study 1: Smart Grid Energy Optimization Using Machine Learning

In a case study conducted by Kim et al. (2018) [20], machine learning algorithms were applied to optimize energy consumption in a smart grid environment. The researchers utilized data from smart meters installed in residential buildings to predict energy demand patterns accurately. They employed a combination of LSTM neural networks and random forests to forecast energy consumption over different time horizons. The LSTM neural networks were capable of capturing temporal dependencies in the data, while the random forests helped reduce overfitting and improve generalization.

The results of the case study showed that the machine learning models achieved a high level of accuracy in predicting energy consumption, outperforming traditional statistical methods. The optimized energy consumption based on the machine learning predictions led to a significant reduction in overall energy costs for the residential complex. This study demonstrated the potential of machine learning algorithms in achieving energy-efficient operations in smart grid environments.

2.8.2 Case Study 2: Energy Consumption Forecasting in Smart Homes

In a study by Matijasevic et al. (2019) [21], machine learning techniques were employed to forecast energy consumption in smart homes equipped with smart meters and IoT devices. The researchers used a combination of decision trees and SVM to predict energy consumption patterns based on real-time data from smart meters and environmental sensors. The decision trees were effective in handling non-linear relationships between energy consumption and various factors such as temperature, occupancy, and time of day. SVM, on the other hand, was instrumental in capturing complex patterns in the data.

The results of the case study demonstrated that the machine learning models could accurately predict energy consumption in smart homes, facilitating proactive energy management and cost reduction. By leveraging real-time data from smart meters and IoT devices, the residents could make informed decisions to optimize energy usage and reduce their carbon footprint. This study highlighted the significance of machine learning in enabling energy-efficient practices at the individual household level.

2.8.3 Case Study 3: Predictive Energy Management in Residential Complex Buildings

A case study conducted by Olu-Ajayi et al. (2020) [22] focused on predictive energy management in a large residential complex building in an urban setting. The researchers used historical energy consumption data from smart meters and integrated weather forecasts to predict future energy demand accurately. They employed an ensemble approach, combining multiple machine learning algorithms such as random forests, gradient boosting, and neural networks to enhance prediction accuracy.

The findings of the case study indicated that the machine learning models significantly outperformed traditional statistical methods in energy consumption prediction. The predictive energy management system allowed the building management to optimize energy distribution, reduce peak demand, and implement demand response strategies effectively. As a result, the residential complex achieved substantial energy cost savings and contributed to a more sustainable energy consumption pattern in the urban area.

2.8.4 Case Study 4: Deep Learning for Energy Load Forecasting

A case study by Oprea & Bara (2019) [23] explored the application of deep learning models in energy load forecasting for a mixed-use residential and commercial building complex. The researchers used LSTM neural networks and attention mechanisms to capture temporal dependencies and improve the models' interpretability. The attention mechanisms helped identify the most influential factors affecting energy load, such as occupancy, weather conditions, and time of day.

The results of the case study demonstrated that the deep learning models provided accurate load forecasting, even during periods of high variability. The improved load forecasting enabled the building complex to optimize energy usage, reduce peak demand, and minimize energy costs. Additionally, the attention mechanisms allowed the building management to identify potential areas for energy efficiency improvements, leading to more sustainable operations.

2.9 CHALLENGES AND LIMITATION

Despite the promising results, the implementation of machine learning for energy consumption estimation in residential buildings also faces several challenges and limitations. One of the primary challenges is the need for a large and high-quality dataset for training machine learning models. Gathering and preprocessing such data can be time- consuming and resource-intensive. Additionally, the dynamic nature of energy consumption patterns in residential buildings can pose difficulties in accurately capturing fluctuations and seasonal variations [24].

Moreover, the interpretability of machine learning models is another concern, especially in critical applications like energy management. Black-box models, such as deep neural networks, may provide accurate predictions but lack transparency in explaining their decision-making process. Interpretable machine learning approaches, like decision trees or linear regression, might be preferred in scenarios where model explain ability is crucial.

Furthermore, the deployment of machine learning algorithms in real-world environments requires careful consideration of the computational resources and processing energy needed. Implementing complex algorithms on edge devices, such as smart meters, may be challenging due to hardware constraints [25].

Lastly, the ethical implications of utilizing smart meter data also need to be addressed. Ensuring data privacy and protecting residents' personal information are crucial aspects that researchers and policymakers must consider when utilizing smart meter data for energy management.

2.10 SUMMARY

This literature review provides a comprehensive overview of the current research and developments in energy consumption estimation using machine learning with data from smart meters in residential complex buildings in Iraq. The studies reviewed demonstrate the potential of machine learning algorithms in accurately predicting energy consumption and improving energy management practices. Despite the challenges and limitations, the integration of machine learning with smart meter data offers a promising approach to address the inefficiencies in energy consumption and contribute to sustainable energy management practices in Iraq and beyond. The subsequent chapters was built upon this literature review to develop a methodology and conduct a data-driven analysis to estimate energy consumption in the target residential complex building in Iraq.

3. METHODOLOGY

The methodology that was used for the research project entitled “Energy Consumption Estimation Using Machine Learning with Data from Smart Meters in a Residential Complex Building in Iraq" is presented in this chapter. This chapter provides an overview of the study design, including the techniques of data collecting, data pre-processing, algorithm selection, model training, model assessment, and energy consumption calculation. In order to accomplish the goals of the research and acquire accurate forecasts of the amount of energy consumed in the residential complex, it is essential to follow the stages and processes outlined in this technique.

3.1 DATA COLLECTION

The collection of data for this study endeavor took place in the building that is part of the residential complex that is located in Iraq. The intelligent meters that have been set up all around the structure was serve as the major data source. Smart meters are sophisticated electronic devices that monitor and record the amount of energy used at regular intervals, often once every an hour but the reading every 6 hours according to setup by operation company. The information that these meters offer, which is both precise and thorough, concerns the quantity of energy that is utilized by inhabitants of the complex [26].

In order to successfully gather the data, authorization and approved got from ministry of electricity and operation company as per letters in appendix A.5, page #127 and #128, as mentioned figure 3.1 below:

Figure 3.1: Authorization Letter.

The information that is gathered from the smart meters was in the form of time-stamped readings, which was illustrate the levels of energy consumption at a variety of time intervals. Each data entry was including a timestamp in addition to the number that corresponds to the amount of energy usage that was captured by the smart meter [27].

The information that is gathered from the smart meters [28] was saved in a safe database, to which only authorized staff will have access. This is done to protect the data's privacy and maintain its confidentiality. In order to preserve the residents' privacy, procedures for data anonymization were used to scrub the records of any information that may be used to identify individual inhabitants.

It is of the utmost importance to make certain that the data obtained is accurate and comprehensive. The smart meters subjected to routine maintenance and calibration in order to guarantee that they was work appropriately and that accurate data collected. In addition, methods of data validation used so that any outliers or discrepancies in the data may be located and corrected.

In addition, the data from the smart meters should be supplemented with information from other relevant data sources, like as data on the weather [29], patterns of occupancy, and the features of the building. These new data sources have the potential to give insightful information on the elements that impact the residential complex's overall energy use [30].

In general, the process of gathering resident data would be carried out in an honest and open manner, with the residents' consent and privacy being given the utmost importance.

When designing a strong and trustworthy machine learning model for estimating the

amount of electricity consumed in a residential complex building, it was essential to have access to data from smart meters that is both accurate and real-time.

3.2 DATASET

The standard of the datasets that are utilized for the purpose of instructing and assessing the performance of the machine learning models is what lays the groundwork for an accurate calculation of the energy consumption. In the next part, we were talk about the datasets that was used for this research project, which was take place between the years 2020 and 2023. These records are extremely helpful in capturing the temporal patterns, changes, and trends in energy usage that occur within the residential complex building.

3.2.1 Dataset Description

The dataset utilized in this research project encompasses a range of variables, each providing valuable insights into energy consumption. Below, we discuss which columns from the dataset are deemed useful and which ones are considered for exclusion:

3.2.1.1 Useful columns for energy consumption prediction

a. Subscriber No: This column could potentially be used to identify individual subscribers, which might be relevant if we wish to analyze the energy consumption patterns of specific customers.
b. Read Date: This is a critical column as it contains the date and time of the energy consumption readings, essential for time-series analysis.
c. T (Total Energy consumption / Watt. hour): This column represents the actual energy consumption data, serving as the target variable for our prediction model.
d. T1: Energy consumption of National grid / Watt. hour.
e. T4: Generator Energy consumption / Watt. hour.

3.2.1.2 Columns that shall be ignored

a. Customer Name: Unless we require customer-specific analysis, this column can often be ignored.
b. T2, T3: Backup Energy consumption.

c. Current R, Current S, Current T: These columns likely represent current readings, which are essential for energy analysis.

d. Voltage R, Voltage S, Voltage T: Voltage readings can also be relevant for energy analysis.

e. Demand Date, Inductive, Capacitive, Meter Serial Nr, Meter Last Connection, Ri Export, Rc Export, T Export, T1Export, T2Export, T3Export, Demand Delivery, Demand Delivery Date, Current R Output, Current S Output, Current T Output, Voltage R Access, Voltage S Access, Voltage T Access, Reading Date, Reading Time, Reading Day, IsletmeKodu, Modem Last Connection Date: These columns seem to contain additional details and metadata that may not be directly related to energy consumption prediction and are likely candidates for exclusion.

Our dataset selection process is guided by the need for relevant and informative features that can aid in building accurate machine learning models for energy consumption prediction.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Subscriber No	CustomerName	ReadDate	T	Consumption	T1	T2	T3	T4	CurrentR	CurrentS	CurrentT	VoltageR	VoltageS	VoltageT	Demand	Demand Date	Inductive	Capacitive	MeterSerialNr	MeterLastConnect
16	b11.f12.f61	YENİ ABONE	7/31/2022 11:59:59 PM	10806.650	00000.000	00000.000	00000.000	00000.000	10806.650	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00010.716	7/19/2022 9:15:00 AM	00697.453	00080.387	41003511	03s:10dk:39sn (8/13/2023 1
17	b11.f12.f62	YENİ ABONE	7/31/2022 11:59:59 PM	02786.026	00000.000	00000.000	00000.000	00000.000	02786.026	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00008.464	7/1/2022 6:00:00 PM	00113.931	00301.932	41003509	02s:27dk:34sn (8/13/2023 1
18	b11.f13.f64	YENİ ABONE	7/31/2022 11:59:59 PM	05241.689	00000.000	00000.000	00000.000	00000.000	05241.689	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00010.792	7/24/2022 4:45:00 PM	00298.536	00035.899	41004096	02s:19dk:52sn (8/13/2023 1
19	b11.f13.f65	YENİ ABONE	7/31/2022 11:59:59 PM	00000.068	00000.000	00000.000	00000.000	00000.000	00000.068	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	7/1/2022 12:00:00 AM	00000.000	00000.013	41004119	01s:47dk:42sn (8/13/2023 1
20	b11.f13.f66	YENİ ABONE	7/31/2022 11:59:59 PM	04556.644	00000.000	00000.000	00000.000	00000.000	04556.644	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00014.048	7/20/2022 9:00:00 PM	00305.648	00045.191	41004095	02s:12dk:36sn (8/13/2023 1
21	b11.f13.f67	YENİ ABONE	7/31/2022 11:59:59 PM	02136.362	00000.000	00000.000	00000.000	00000.000	02136.362	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00003.932	7/6/2022 2:45:00 PM	00035.202	00190.588	41004094	02s:22dk:01sn (8/13/2023 1
22	b11.f14.f68	YENİ ABONE	7/31/2022 11:59:59 PM	00011.101	00000.000	00000.000	00000.000	00000.000	00011.101	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	7/1/2022 12:00:00 AM	00000.639	00000.075	41004101	02s:10dk:42sn (8/13/2023 1
23	b11.f14.f69	YENİ ABONE	7/31/2022 11:59:59 PM	08120.566	00000.000	00000.000	00000.000	00000.000	08120.566	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00007.528	7/31/2022 6:30:00 PM	00565.264	00059.140	41004100	02s:26dk:41sn (8/13/2023 1
24	b11.f14.f70	YENİ ABONE	7/31/2022 11:59:59 PM	02282.563	00000.000	00000.000	00000.000	00000.000	02282.563	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	7/1/2022 12:00:00 AM	00176.521	00064.459	41004075	01s:57dk:55sn (8/13/2023 1
25	b11.f14.f71	YENİ ABONE	7/31/2022 11:59:59 PM	05083.284	00000.000	00000.000	00000.000	00000.000	05083.284	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00014.648	7/24/2022 1:15:00 PM	00303.673	00033.814	41004102	02s:21dk:08sn (8/13/2023 1
26	b11.f14.f72	YENİ ABONE	7/31/2022 11:59:59 PM	03745.892	00000.000	00000.000	00000.000	00000.000	03745.892	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00009.704	7/19/2022 11:30:00 PM	00201.054	00324.406	41004103	02s:09dk:50sn (8/13/2023 1
27	b11.f16.f78	YENİ ABONE	7/31/2022 11:59:59 PM	06724.969	00000.000	00000.000	00000.000	00000.000	06724.969	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00015.140	7/21/2022 8:45:00 PM	00370.645	00172.655	41004034	03s:14dk:48sn (8/13/2023 1
28	b11.f16.f79	YENİ ABONE	7/31/2022 11:59:59 PM	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	7/1/2022 12:00:00 AM	00000.000	00000.000	41004040	01s:46dk:29sn (8/13/2023 1
29	b11.f16.f81	YENİ ABONE	7/31/2022 11:59:59 PM	01779.848	00000.000	00000.000	00000.000	00000.000	01779.848	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00014.076	7/21/2022 8:45:00 PM	00120.204	00007.600	41004038	03s:19dk:02sn (8/13/2023 1
30	b11.f16.f82	YENİ ABONE	7/31/2022 11:59:59 PM	02619.394	00000.000	00000.000	00000.000	00000.000	02619.394	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00012.100	7/10/2022 5:15:00 PM	00137.101	00383.577	41004035	03s:07dk:51sn (8/13/2023 1
31	b11.f17.f83	YENİ ABONE	7/31/2022 11:59:59 PM	03288.879	00000.000	00000.000	00000.000	00000.000	03288.879	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00004.128	7/20/2022 4:00:00 PM	00146.407	00267.413	41004053	03s:03dk:46sn (8/13/2023 1
32	b11.f17.f84	YENİ ABONE	7/31/2022 11:59:59 PM	04410.640	00000.000	00000.000	00000.000	00000.000	04410.640	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00014.384	7/23/2022 1:45:00 PM	00262.151	00332.795	41004050	03s:17dk:34sn (8/13/2023 1
33	b11.f17.f87	YENİ ABONE	7/31/2022 11:59:59 PM	03245.614	00000.000	00000.000	00000.000	00000.000	03245.614	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00012.884	7/2/2022 1:00:00 PM	00188.715	00183.274	41002745	02s:04dk:24sn (8/13/2023 1
34	b11.f3.f13	YENİ ABONE	7/31/2022 11:59:59 PM	00203.090	00000.000	00000.000	00000.000	00000.000	00203.090	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00011.236	7/31/2022 11:30:00 AM	00012.569	00008.225	41003720	03s:02dk:47sn (8/13/2023 1
35	b11.f3.f14	YENİ ABONE	7/31/2022 11:59:59 PM	06430.743	00000.000	00000.000	00000.000	00000.000	06430.743	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00014.244	7/16/2022 5:45:00 PM	00340.728	00168.611	41003721	02s:49dk:05sn (8/13/2023 1
36	b11.f3.f15	YENİ ABONE	7/31/2022 11:59:59 PM	00060.194	00000.000	00000.000	00000.000	00000.000	00060.194	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00000.272	7/16/2022 2:00:00 PM	00000.391	00206.008	41003777	02s:15dk:28sn (8/13/2023 1
37	b11.f3.f16	YENİ ABONE	7/31/2022 11:59:59 PM	03584.098	00000.000	00000.000	00000.000	00000.000	03584.098	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00007.784	7/31/2022 3:15:00 PM	00203.848	00049.702	41003774	02s:08dk:42sn (8/13/2023 1
38	b11.f3.f17	YENİ ABONE	7/31/2022 11:59:59 PM	00117.151	00000.000	00000.000	00000.000	00000.000	00117.151	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00001.964	7/24/2022 3:30:00 PM	00005.271	00011.001	41003718	02s:58dk:12sn (8/13/2023 1
39	b11.f4.f18	YENİ ABONE	7/31/2022 11:59:59 PM	08052.463	00000.000	00000.000	00000.000	00000.000	08052.463	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00009.096	7/19/2022 3:00:00 PM	00433.811	00086.310	41003693	01s:55dk:40sn (8/13/2023 1
40	b11.f4.f20	خديجة محمود علي	7/31/2022 11:59:59 PM	02168.231	00000.000	00000.000	00000.000	00000.000	02168.231	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00015.568	7/12/2022 11:00:00 PM	00159.564	00243.074	41003684	02s:06dk:57sn (8/13/2023 1
41	b11.f4.f21	YENİ ABONE	7/31/2022 11:59:59 PM	01847.249	00000.000	00000.000	00000.000	00000.000	01847.249	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00014.892	7/23/2022 6:30:00 PM	00092.392	00064.006	41003688	01s:57dk:11sn (8/13/2023 1
42	b11.f4.f22	المستجر/جوان صلاح الدين	7/31/2022 11:59:59 PM	06631.641	00000.000	00000.000	00000.000	00000.000	06631.641	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00013.740	7/19/2022 7:15:00 PM	00426.866	00117.092	41003614	02s:46dk:52sn (8/13/2023 1
43	b11.f5.f23	YENİ ABONE	7/31/2022 11:59:59 PM	02155.169	00000.000	00000.000	00000.000	00000.000	02155.169	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00013.796	7/17/2022 7:45:00 PM	00104.855	00201.096	41003823	02s:50dk:02sn (8/13/2023 1
44	b11.f5.f24	YENİ ABONE	7/31/2022 11:59:59 PM	00404.632	00000.000	00000.000	00000.000	00000.000	00404.632	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	7/1/2022 12:00:00 AM	00000.877	00159.366	41003724	03s:13dk:01sn (8/13/2023 1
45	b11.f5.f25	YENİ ABONE	7/31/2022 11:59:59 PM	05794.691	00000.000	00000.000	00000.000	00000.000	05794.691	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00013.332	7/21/2022 10:45:00 PM	00165.852	00402.370	41003821	03s:00dk:23sn (8/13/2023 1
46	b11.f5.f26	YENİ ABONE	7/31/2022 11:59:59 PM	08879.717	00000.000	00000.000	00000.000	00000.000	08879.717	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00010.456	7/20/2022 5:30:00 PM	00522.723	00169.679	41003815	02s:51dk:17sn (8/13/2023 1
47	b11.f6.f28	YENİ ABONE	7/31/2022 11:59:59 PM	00538.943	00000.000	00000.000	00000.000	00000.000	00538.943	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00014.600	7/22/2022 7:15:00 PM	00024.792	00123.577	41003517	03s:11dk:42sn (8/13/2023 1
48	b11.f6.f29	ليث يحيى	7/31/2022 11:59:59 PM	03259.423	00000.000	00000.000	00000.000	00000.000	03259.423	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00015.364	7/7/2022 7:30:00 PM	00118.351	00123.999	41003522	02s:55dk:05sn (8/13/2023 1
49	b11.f7.f33	مصطفى عصام حسين	7/31/2022 11:59:59 PM	05651.764	00000.000	00000.000	00000.000	00000.000	05651.764	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00015.768	7/19/2022 11:00:00 PM	00273.999	00124.024	41003523	03s:16dk:11sn (8/13/2023 1
50	b11.f7.f34	YENİ ABONE	7/31/2022 11:59:59 PM	03343.667	00000.001	00000.000	00000.000	00000.000	03343.666	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00011.680	7/23/2022 10:30:00 PM	00173.650	00263.689	41003514	02s:56dk:05sn (8/13/2023 1
51	b11.f7.f35	YENİ ABONE	7/31/2022 11:59:59 PM	00002.043	00000.000	00000.000	00000.000	00000.000	00002.043	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	7/1/2022 12:00:00 AM	00000.004	00000.410	41003443	02s:13dk:34sn (8/13/2023 1
52	b11.f7.f37	YENİ ABONE	7/31/2022 11:59:59 PM	00001.968	00000.000	00000.000	00000.000	00000.000	00001.968	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	7/1/2022 12:00:00 AM	00000.005	00000.292	41003502	01s:48dk:47sn (8/13/2023 1
53	b11.f8.f38	YENİ ABONE	7/31/2022 11:59:59 PM	01245.204	00000.000	00000.000	00000.000	00000.000	01245.204	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00014.256	7/7/2022 6:45:00 PM	00075.246	00006.467	41003519	02s:47dk:48sn (8/13/2023 1
54	b11.f8.f42	YENİ ABONE	7/31/2022 11:59:59 PM	04363.461	00000.000	00000.000	00000.000	00000.000	04363.461	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00013.112	7/1/2022 5:45:00 PM	00344.424	00140.666	41003515	02s:52dk:14sn (8/13/2023 1
55	b11.f9.f43	YENİ ABONE	7/31/2022 11:59:59 PM	05531.106	00000.000	00000.011	00000.000	00000.000	05531.095	00000.000	00000.000	00000.000	00000.000	00000.000	00000.000	00015.012	7/16/2022 1:45:00 AM	00242.748	00311.766	41003404	02s:30dk:22sn (8/13/2023 1

Figure 3.2: Example of Raw Dataset for Building 11 at July.2022.

3.2.2 Data Pre-Processing

The pre-processing of the data is an important phase in the research technique that was used for this project. It entails cleaning, converting, and preparing the data gathered from smart meters to make it acceptable for training machine learning models effectively. This is done in order to ensure that the models can learn from the data in an accurate manner. In order to guarantee that the data are of a high quality, free from noise and inconsistencies, and suitable for analysis and modeling, pre-processing is performed on the data [31].

The performance of machine learning algorithms may be adversely affected by a variety of problems that may be present in the data that was acquired from the smart meters. Some of these problems include missing values, outliers, and noise. In light of this, the following strategies for the pre-processing of data was implemented:

3.2.2.1 Data cleaning

At this point, we were address any data that is missing or incomplete that we may have. It's possible that a broken meter or a misunderstanding in the communication chain led to the loss of data. To guarantee that the dataset was not be corrupted in any way, the missing information was either be deleted or filled up using appropriate methods, such as interpolation or mean imputation, respectively.

3.2.2.2 Outlier detection

Extreme numbers that are considerably different from the rest of the data are referred to as outliers. The accuracy of the model's predictions and its performance can both be impacted by outliers. The identification and management of outliers was accomplished with the assistance of powerful statistical techniques like the Interquartile Range (IQR) and the Z-score. Iraq gate, its big project for construct residential complex buildings at Baghdad, it content 48 buildings with variety of floors, our study we collect dataset for 10 building (1,2,3,5,6,7.8,9,10 and 11) which are started up as functional operation with different times (started from 2020 until 2023) and the rest still under construction, See the layout of Iraq gate project at figure 3.3 below.

Figure 3.3: Top View of Layout of Iraq Gate Project.

3.2.2.3 Data collection

We collected useful columns for Energy Consumption Prediction for each month at each flat in one sheet for each year, See figure 3.4 below.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	No#	Subscriber No#	Jan-22	Feb-22	Mar-22	Apr-22	May-22	Jun-22	Jul-22	Aug-22	Sep-22	Oct-22	Nov-22	Dec-22
2			T / Wh	T / Wh	T / Wh	T / Wh	T / Wh	T / Wh	T / Wh	T / Wh	T / Wh	T / Wh	T / Wh	T / Wh
3	1	b1 colling system	41326.998	41501.323	42141.757	44468.606	47315.796	50754.218	54521.806	58921.215	62091.531	64085.970	64761.649	65233.746
4	2	b1 elevator L1	08790.978	09299.296	09824.746	10266.168	10723.982	11231.438		12162.904	12626.887	13077.745	13521.198	14009.586
5	3	b1 fan	80945.661	82654.039	83662.719	87845.144	91650.957	94835.211		101583.052			115771.865	120730.400
6	4	b1.f1.f3	25130.451	25930.214	26863.683	27651.721	28651.650	29932.303	30967.697	32101.604	33123.106	34132.691	34801.334	35890.149
7	5	b1.f1.f4	14984.229	15640.279	16489.036	17065.223	18192.815	19885.517	21620.820	23567.944	24905.414	25801.821	26256.707	27251.025
8	6	b1.f1.f5	39098.750	39145.690	39145.816	39146.299	39146.891	39174.008	39362.358	39428.144	39496.808	39826.355	39845.561	40107.928
9	7	b1.f1.f6	42482.746	42550.542	42550.731	42697.214	43665.285	44021.892	44112.051	44212.056	44764.727	44976.728	45519.057	46912.031
10	8	b1.f1.f7	28874.296	29089.935	29391.305	30193.425	31940.383	34733.607	37196.885	39179.734	40971.000	42610.206	43274.109	43944.576
11	9	b1.f10.f48	32820.708	33479.429	34178.899	35152.928	36441.205	38516.766	40675.326	43375.963	45398.877	46304.656	46902.952	47713.002
12	10	b1.f10.f49	10034.414	10789.717	11184.531	11341.304	11767.790	11984.922	12657.054	13113.477	13339.838	13563.864	13765.192	14012.016
13	11	b1.f10.f50	10302.243	11012.244	11680.115	13021.657	14646.411	16907.629	19263.432	22379.624	24760.608	26570.965	27440.815	28438.534
14	12	b1.f10.f51	14356.872	14434.127	14522.049	14623.325	14734.526	14954.182	15152.852	15338.518	15421.064	15547.630	15708.426	15806.220
15	13	b1.f10.f52	13935.583	14391.692	14718.226	15078.790	15557.920	16349.377	16950.841	17878.731	18605.332	19063.775	19326.306	19666.679
16	14	b1.f11.f53	14539.791	14965.870	15416.348	16154.586	16791.327	17124.996	17292.286	17390.852	17501.869	17875.308	18040.161	18242.351
17	15	b1.f11.f54	23852.192	24602.878	25166.606	26121.433	27552.241	29504.596	31496.785	32475.348	33967.023	35117.443		36336.723
18	16	b1.f11.f55	38943.697	40190.871	41123.651	42116.971	43969.058	46643.765	49364.442	51773.080	54163.958	55424.371	56078.737	57160.917
19	17	b1.f11.f56	03574.434	04049.891	04305.053	04527.852	04902.084	05742.426	07626.966	08986.314	10276.868	11102.933	11357.969	11637.896
20	18	b1.f11.f57	04363.763	04885.142	05370.469	05929.971	06821.032	08190.949	09581.489	11424.496	12829.318	13738.754	14311.685	14902.861
21	19	b1.f12.f58	20088.853	20512.063	21052.639	21643.181	22436.407	24427.332	25444.460	27797.830	29473.854	30317.421	30669.950	31047.795
22	20	b1.f12.f59	20212.786	20868.373	21319.222	21774.582	22418.018	23602.340	24917.662	26219.775	27319.745	27668.295	27969.752	28601.760
23	21	b1.f12.f60	16514.861	18147.133	19382.762	20834.098	22658.417	24902.760	27179.684	29780.630	31669.255	32845.140	33771.256	35031.631
24	22	b1.f12.f61	30632.484	31917.254	33039.907	34323.479	36206.928	38898.416	41220.206	44032.539	46205.884	47702.166	48398.793	49164.151
25	23	b1.f12.f62	35823.028	37174.425	38532.578	39901.972	41706.135	44317.331	46471.944	49477.831	51671.369	53189.328	54268.982	56037.084
26	24	b1.f13.f63	08162.623	08162.623	08162.623	08162.843	08162.843	08162.843	08163.884	08163.884	08163.884	08682.012	08775.672	08816.630
27	25	b1.f13.f64	36788.454	37972.202	39392.032	41130.310	43343.510	46045.978	48946.474	52747.751	55008.285	56808.604	56891.914	57022.096
28	26	b1.f13.f65	23925.880	24377.527	24957.028	25853.349	27021.193	29191.309	30687.393	31380.306	33051.782	33997.738	34478.197	34829.869

Figure 3.4: Example of Dataset for Building -1 at 2022 Before Reshape.

3.2.2.4 Data encoding

In order for the machine learning algorithms to be able to work with categorical variables, it is necessary to transform them into a numerical format first. To do that we make those steps:

First, Rename the subscriber into customer ID.

Second, the months will be in one column instead of every month in columns.

Third, the year will be fixed for each row.

We implemented reshape by using Python code as shown in figure 3.5 and the outcomes shown in appendix A.4, Page #126 and mentioned in figure 3.6 below :

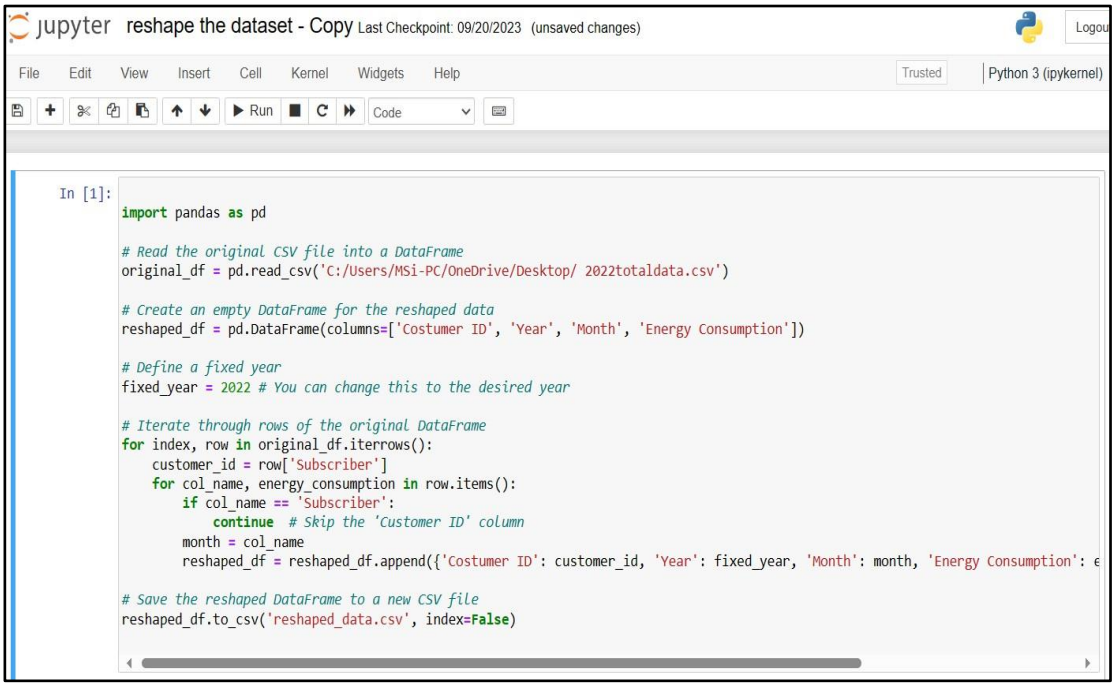


Figure 3.5: Code of Python for Reshape.

	A	B	C	D
1	Costumer ID	Year	Month	Energy Consumption
2	b1 colling system	2022	22-Jan	41326.998
3	b1 colling system	2022	22-Feb	41501.323
4	b1 colling system	2022	22-Mar	42141.757
5	b1 colling system	2022	22-Apr	44468.606
6	b1 colling system	2022	22-May	47315.796
7	b1 colling system	2022	22-Jun	50754.218
8	b1 colling system	2022	22-Jul	54521.806
9	b1 colling system	2022	22-Aug	58921.215
10	b1 colling system	2022	22-Sep	62091.531
11	b1 colling system	2022	22-Oct	64085.97
12	b1 colling system	2022	22-Nov	64761.649
13	b1 colling system	2022	22-Dec	65233.746
14	b1 elevator L1	2022	22-Jan	8790.978
15	b1 elevator L1	2022	22-Feb	9299.296
16	b1 elevator L1	2022	22-Mar	9824.746
17	b1 elevator L1	2022	22-Apr	10266.168
18	b1 elevator L1	2022	22-May	10723.982
19	b1 elevator L1	2022	22-Jun	11231.438
20	b1 elevator L1	2022	22-Jul	
21	b1 elevator L1	2022	22-Aug	12162.904
22	b1 elevator L1	2022	22-Sep	12626.887
23	b1 elevator L1	2022	22-Oct	13077.745
24	b1 elevator L1	2022	22-Nov	13521.198
25	b1 elevator L1	2022	22-Dec	14009.586
26	b1 fan	2022	22-Jan	80945.661
27	b1 fan	2022	22-Feb	82654.039
28	b1 fan	2022	22-Mar	83662.719

Figure 3.6: Example of Dataset for Building -1 at 2022 After Reshape.

Due to machine learning algorithm cannot recognized words that's why we change months name to numerical and we do that by using Python code shown in figure 3.7 and the outcome as shown in appendix A.4 , Page #126 as mentioned in figure 3.8 below:

The data, after being preprocessed, was split up into a training set and a testing set. In order to train the machine learning model, the training set was utilized, while the testing set was used to evaluate the model's overall effectiveness.

```
In [2]: import pandas as pd

        # Read your CSV file into a DataFrame (assuming the CSV file is named 'your_data.csv')
        df = pd.read_csv('C:/Users/MSi-PC/OneDrive/Desktop/reshape dataset at 2023.csv')

        # Define a dictionary to map month names to numbers
        month_name_to_number = {
            'Jan': 1, 'Feb': 2, 'Mar': 3, 'Apr': 4, 'May': 5, 'Jun': 6,
            'Jul': 7, 'Aug': 8, 'Sep': 9, 'Oct': 10, 'Nov': 11, 'Dec': 12
            # You can add more months as needed
        }

        # Extract the month name part from the "Month" column and map month names to numbers
        df['Month'] = df['Month'].str.extract('([A-Za-z]+)', expand=False).map(month_name_to_number)

        # Save the modified2 DataFrame to a new CSV file
        df.to_csv('C:/users/MSi-PC/OneDrive/Desktop/1/modified6_data.csv', index=False)
```

Figure 3.7: Code Python for Modify.

	A	B	C	D
1	Costumer ID	Year	Month	Energy Consumption
2	b1 colling system	2022	1	41326.998
3	b1 colling system	2022	2	41501.323
4	b1 colling system	2022	3	42141.757
5	b1 colling system	2022	4	44468.606
6	b1 colling system	2022	5	47315.796
7	b1 colling system	2022	6	50754.218
8	b1 colling system	2022	7	54521.806
9	b1 colling system	2022	8	58921.215
10	b1 colling system	2022	9	62091.531
11	b1 colling system	2022	10	64085.97
12	b1 colling system	2022	11	64761.649
13	b1 colling system	2022	12	65233.746
14	b1 elevator L1	2022	1	8790.978
15	b1 elevator L1	2022	2	9299.296
16	b1 elevator L1	2022	3	9824.746
17	b1 elevator L1	2022	4	10266.168
18	b1 elevator L1	2022	5	10723.982
19	b1 elevator L1	2022	6	11231.438
20	b1 elevator L1	2022	7	
21	b1 elevator L1	2022	8	12162.904
22	b1 elevator L1	2022	9	12626.887
23	b1 elevator L1	2022	10	13077.745
24	b1 elevator L1	2022	11	13521.198
25	b1 elevator L1	2022	12	14009.586
26	b1 fan	2022	1	80945.661
27	b1 fan	2022	2	82654.039
28	b1 fan	2022	3	83662.719

Figure 3.8: Example for Dataset Building 1 at 2022 After Modify.

The generalization performance of a model may be evaluated with the use of a method called cross-validation. It entails breaking the data up into several different subsets, or folds, and then training the model on a variety of different permutations of these folds. This technique contributes to the estimation of the model's performance on data that has not yet been observed.

The dataset was improved as a result of employing these strategies for preprocessing the data, and any problems that are caused by the data was fixed. The machine learning model

was trained and validated using the pre-processed data in order to get an accurate estimate of the amount of electricity that is consumed within the residential complex building.

3.3 ALGORITHM SELECTION

The choice of algorithm is an important phase in the methodology of this research project's approach because it sets the machine learning model that was used to forecast the amount of electricity consumed in the building that houses the residential complex. Several different machine learning algorithms was investigated and assessed depending on how well they can provide an accurate estimate of the amount of energy that was consumed [32].

In order to make accurate projections of energy usage, we were investigating the following machine learning algorithms:

a. Linear Regression: When it comes to regression problems, a basic approach that's also quite successful is called linear regression. A linear equation is used to describe the relationship that exists between the input characteristics and the goal variable, which in this case is energy consumption. When there is a linear connection between the input data and the target variable, linear regression is an appropriate statistical method to use.
b. Decision Trees: The input attributes are used as the basis for decision trees, which are non-linear models that divide the data into subsets. They are able to capture non-linear correlations as well as the interactions that occur between characteristics. The results of using decision trees may be interpreted, and they are able to process both category and numerical data.
c. Random Forest: To improve accuracy and reduce overfitting, ensemble learning methods like Random Forest combine many decision trees into a single model. To do this, it constructs several decision trees on different subsets of the data and then combines the combined forecasts from these trees. The Random Forest algorithm has gained widespread popularity because to its dependability and its ability to handle large and complex datasets.
d. Gradient Boosting: Another strategy for ensemble learning is called gradient boosting, and it involves building numerous weak learners (often decision trees) in a sequential fashion, with each one attempting to fix the mistakes made by its predecessor. Gradient

boosting is an effective technique that frequently delivers high levels of accuracy when used to prediction problems.

e. Support Vector Regression (SVR): The Support Vector Machine algorithm is used for regression, and SVR is a variant on that technique. The goal is to locate a hyperplane that provides the greatest fit for the data while yet allowing for some degree of inaccuracy. When working with non-linear data, SVR proves to be extremely helpful.
f. Neural Networks: Deep learning models, and more especially neural networks, have demonstrated impressive performance across a variety of areas, one of which is the prediction of energy use. Because they are able to recognize intricate patterns and interconnected relationships in the data, neural networks are well-suited for high-dimensional and non-linear data sets.
g. Long Short-Term Memory (LSTM): A kind of recurrent neural network known as LSTM is very effective when used to time series data. LSTM is able to successfully capture both short-term and long-term patterns and trends since energy consumption data frequently displays temporal interdependence.

A comparative analysis of the performance of the candidate algorithms was carried out with the assistance of pertinent metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared value. This was allowed for the selection of the algorithm that is best suited for this project (R2). Utilizing these criteria, one was able to evaluate the degree to which each algorithm is accurate and capable of generalization. In addition, strategies based on cross-validation was implemented in order to validate the models and prevent overfitting [33].

Based on the information that has been gathered and preprocessed, the final algorithm that is chosen was the one that exhibits the best overall performance in terms of properly calculating the amount of energy that is consumed by the building that houses a residential complex.

3.4 MODEL TRAINING

The chosen machine learning algorithm was trained on the pre-processed data in order to develop a predictive model for estimating energy usage as part of this research project's approach [34], which makes model training an essential step [35]. Finding the best

parameters and patterns in the data that was enable the model to generate correct predictions is one of the tasks that are performed throughout the training phase [36].

3.4.1 Data Splitting

The procedure of data splitting is an essential aspect of the machine learning process [37], particularly when it comes to the training of models. In order to do this, the dataset is first partitioned into two subsets: the training set and the testing set. The primary goals of data splitting are to assess the trained model's ability to generalize to data it has not previously seen and to avoid the model from becoming overly accurate by the split using library called (train test split) in python to split that data to a training data 80% and testing data 20%.

3.4.1.1 Training set

In most cases, anywhere between 70 and 80 percent of the total data is contained inside the training set. This subset is what the machine learning algorithm is trained on in order to discover the underlying patterns and connections that are present in the data. The prediction error is reduced by a process called iterative parameter adjustment, which is driven by the training data and performed by the algorithm [38].

3.4.1.2 Testing set

The remaining fraction of the data, which is somewhere between 20 and 30 percent, is comprised of the testing set. It functions as an independent dataset that the trained model did not view when it was being trained. After the training phase, the model is examined using the testing set to determine how well it does on data for which it has not been previously prepared. The results of this assessment assist offer an estimate of how effectively the model can generalize to new data and provide an estimate of how well it performs in the actual world [39].

The purpose of data splitting is to replicate the performance of the model in a real-world scenario in which it meets fresh data that it has not previously been exposed to. We may acquire insights into possible concerns such as overfitting by evaluating the accuracy of the model on data that it has never seen before. This is a situation in which the model performs well on the data that it was trained on, but it performs badly on fresh data.

Creating an extra subset known as the validation set is possible in some circumstances, particularly in situations in which the dataset is restricted. The validation set is utilized during the process of hyper parameter tuning, which involves testing several possible configurations of the model's hyper parameters to see which one produces the best results. The model is next trained on the combined training and validation sets, and then it is assessed using the testing set. This process is repeated once the optimal hyper parameters have been selected based on their performance on the validation set.

The procedure of data splitting is absolutely necessary in order to acquire a model that is dependable and generalizable. It gives us the ability to evaluate how well the machine learning model was perform in real-world scenarios and gives us a measure of its effectiveness in predicting energy consumption in the residential complex building based on data from the smart meters. This is made possible by the fact that it is possible for us to use real-world data.

3.4.2 Model Initialization

Model initialization is a critical stage in the training of machine learning algorithms, such as those used for energy consumption estimates in this research project. In this research project, it was used to estimate the amount of energy consumed. Before beginning the training process, it requires specifying starting values for each of the parameters that make up the model. Model initialization has the purpose of providing a beginning point for the optimization method so that it may repeatedly update the parameters and discover the optimal values that minimize the prediction error. This is accomplished by providing a starting point for the model.

The particular approach of model initialization must be tailored to the particular machine learning algorithm that is being applied. Take, for example:

In the context of neural networks, "model initialization" refers to the process of assigning initial weights and biases to each neuron that makes up the network. These weights and biases influence the degree to which neurons are connected to one another as well as the level of activity required to trigger neuronal activity. Common methods for initialization include selecting random values from a Gaussian distribution, employing predefined schemes such as Xavier/Glorot initialization or He initialization, both of which aim to keep

the signal's magnitude consistent across all network layers, and using random initialization from a Gaussian distribution [40].

Model initialization for linear regression requires specifying starting values for the coefficients and the intercept term. This is done so that the model may be fit to the data. These settings define the slope of the regression line as well as its location inside the feature space.

Support Vector Machines (SVM): In SVM, model initialization entails picking an initial hyperplane that separates the data points of distinct classes by a margin that is the maximum possible.

Decision Trees: Model initialization in decision trees normally entails picking a root node and establishing the first splits based on the feature values. This is done in decision trees.

It is extremely important to select a suitable model initialization, since this can have an effect on both the speed at which the model converges and the quality of the model overall. If the initialization is not done correctly, the model may converge slowly or become trapped in a local optimum, which was prevent it from arriving at the best possible solution.

In addition, different machine learning algorithms have different levels of sensitivity to initialization. Because of the complexity of their structures, deep neural networks, for instance, might be very sensitive to the starting choices that are made. Techniques of appropriate initialization are required to enable the proper training of the model and its capacity to capture the underlying patterns in the data efficiently.

It is important to keep in mind that the initialization of the model is just the beginning of the training process. During training, the model's parameters was updated based on the data in order to enhance the model's overall performance. During the iterative optimization process, the starting parameter values was modified until the model converges on a solution that reduces the amount of variance that exists between the energy consumption that was predicted and the actual values.

3.4.3 Temporal Split

The dataset was temporally segmented into training, validation, and testing sets in order to confirm that the models can generalize effectively to data that they have not previously encountered. The years 2020 to 2022 was included in the data used for training, whereas

the years 2022 and after was included in the data used for validating the model. The most current measurements, which were taken in 2023, was make up the data for the testing.
This temporal split reflects the real-world scenario in which models are trained on historical data and tested on more recent data, simulating their ability to accurately forecast energy consumption. In this scenario, models are split into two groups: one group is trained on historical data, and the other group is tested on more recent data.

3.4.4 Model Training

The training of models is an essential and iterative part of the technique that was used for this research effort. Creating a predictive model for estimating the amount of electricity consumed in the residential complex building requires inputting the data after it has been pre-processed into the machine learning algorithm that has been chosen and modifying the algorithm's parameters. The primary purpose of model training is to improve the accuracy of the model's predictions by reducing the amount of error that exists between the values that are predicted by the model and the actual amount of energy that is used [41].
The stages that make up the process of training models may be broken down into the following categories:

a. Input Data Preparation: The data, after having been preprocessed, are separated into two subsets: the training set and the testing set. The model is trained with the data from the training set, which accounts for around 70–80 percent of the total, while the testing set, which accounts for the remaining 20–30 percent, is used only to assess how well the model is doing.
b. Training Data Feeding: The machine learning algorithm is given the training data, and it uses the input characteristics (such as past energy use, weather data, occupancy levels, and so on) to understand the underlying patterns and correlations that are present within the data.
c. Optimization Algorithm: During training, the algorithm iteratively updates the model's parameters to minimize the prediction error. This process involves utilizing an optimization algorithm, such as gradient descent, to adjust the model's parameters in the direction that reduces the error.
d. Loss Function: To quantify the prediction error, a loss function (also known as a cost function) is used. The choice of the loss function depends on the specific machine

learning algorithm and the nature of the problem. For regression tasks, Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) are commonly used loss functions.

e. Backpropagation (Neural Networks): In the case of neural networks, the backpropagation algorithm is employed to calculate the gradients of the loss function with respect to the model's parameters. These gradients indicate the direction and magnitude of the parameter updates needed to minimize the error.
f. Epochs and Batch Size: Model training is typically performed over multiple epochs, where each epoch represents one complete pass through the training data. For large datasets, the data is divided into batches, and the model is updated after processing each batch. The batch size is another hyper parameter that can impact training efficiency and convergence.
g. Hyper parameter Tuning: Some machine learning algorithms have hyper parameters that need to be set before training. These hyper parameters significantly influence the model's performance and may need to be tuned using techniques like grid search or random search to find the optimal values.
h. Early Stopping: To avoid overfitting, early stopping may be employed during training. During early stopping, the training process is halted if the model's performance begins to deteriorate as measured on a validation set.
i. Cross-Validation: Cross-validation is used to ensure that the model's performance is not biased by the data splitting. It involves dividing the data into multiple subsets (folds), training the model on different combinations of these folds, and then averaging the performance results to obtain a more robust estimate of the model's accuracy.
j. Model Evaluation: Throughout the training process, the model's performance is continually evaluated on the testing set, using the evaluation metrics determined earlier, such as MAE, RMSE, and R2. This evaluation helps monitor the model's progress and identify potential issues, such as overfitting or under fitting.
k. Final Model: Once the training process is complete, and the model's performance is satisfactory, the final model is obtained. This model is ready to be used for energy consumption estimation on new, unseen data from the smart meters in the residential complex building.

It is important to note that the success of the energy consumption estimation heavily relies on the effectiveness of the model training process. The trained model's accuracy and

generalization capabilities was playing a significant role in achieving the research project's objectives of optimizing energy management in the building [42].

3.4.5 Evaluation

Evaluation is a critical phase in the methodology of this research project. After training the machine learning model on the training dataset, it is essential to assess its performance on unseen data to determine its accuracy and generalization capabilities. The evaluation process provides valuable insights into how well the model can estimate energy consumption based on data from the smart meters in the residential complex building [43]. The evaluation of the trained model involves the following steps:

a. Testing Data Prediction: The testing dataset, which was set aside earlier and not used during training, is now fed into the trained model. The model uses the input features from the testing data to make predictions of energy consumption for each data point.
b. Evaluation Metrics: To quantify the performance of the model, various evaluation metrics are used. The choice of metrics depends on the specific nature of the problem and the goals of the research project. Commonly used evaluation metrics for regression tasks include:

a. Mean Absolute Error (MAE): “It measures the average absolute difference between the predicted values and the actual values of energy consumption. It provides a straightforward interpretation of prediction accuracy” [44].
b. Root Mean Squared Error (RMSE): “RMSE is similar to MAE but penalizes large prediction errors more. It is useful when large errors are particularly undesirable.” [44]
c. R-squared (R2): R-squared measures the proportion of variance in the target variable (energy consumption) that is predictable from the input features. It gives an indication of how well the model fits the data. [44]
c. Comparison to Baseline Models: In some cases, it is essential to compare the performance of the machine learning model to baseline models or existing traditional methods for energy consumption estimation. This comparison helps to demonstrate the effectiveness and superiority of the proposed machine learning approach.
d. Overfitting and Under fitting Analysis: Evaluation also involves analyzing the model for signs of overfitting or under fitting. Overfitting occurs when the model performs well on the training data but poorly on unseen data, indicating that it has memorized the training

examples without generalizing well. Under fitting, on the other hand, happens when the model is too simplistic and fails to capture the underlying patterns in the data.

e. Interpretability of Results: It is crucial to interpret the results of the evaluation to gain meaningful insights into the factors influencing energy consumption in the residential complex building. Understanding which features have the most significant impact on consumption can help building managers and occupants optimize energy usage.
f. Validation and Robustness: To ensure the reliability and robustness of the evaluation results, cross-validation can be applied. Cross-validation involves dividing the data into multiple subsets and performing evaluation on different combinations of these subsets. This helps to obtain a more reliable estimate of the model's performance and minimizes the potential bias introduced by the data splitting.
g. Iterative Improvement: If the model does not meet the desired performance criteria, further iterations of data pre-processing, algorithm selection, and model training may be performed to improve the results.

The evaluation outcomes were determining the success of the research project in developing an accurate and efficient predictive model for energy consumption estimation in the residential complex building. The model's ability to provide reliable estimations was have practical implications for optimizing energy management, reducing costs, and promoting sustainable practices [45].

3.4.6 Hyper Parameter Tuning

The machine learning approach used in this study relies heavily on hyper parameter optimization. The hyper parameters of the chosen machine learning algorithm must be adjusted to their optimum values before training can begin; these values cannot be learnt during the training process. The proper selection of hyper parameters significantly impacts the model's performance and its ability to accurately estimate energy consumption in the residential complex building.

The following are the steps involved in hyper parameter tuning:

a. Hyper parameter Selection: Depending on the chosen machine learning algorithm, there are specific hyper parameters that need to be tuned. For instance, in a neural network, the learning rate, number of hidden layers, number of neurons per layer, and dropout

rate are some of the hyper parameters that require tuning. Each hyper parameter plays a critical role in the behavior and performance of the model.

b. Grid Search: One common approach to hyper parameter tuning is grid search. Grid search involves defining a range of values for each hyper parameter and then exhaustively searching all possible combinations of these values. For example, if the learning rate can take values [0.001, 0.01, 0.1], and the number of hidden layers can be [1, 2, 3], grid search was evaluating the model's performance for combinations like (0.001, 1), (0.001, 2), (0.001, 3), (0.01, 1), and so on.
c. Random Search: In cases where the hyper parameter search space is vast, random search can be a more efficient approach. Instead of trying all possible combinations like in grid search, random search samples random combinations of hyper parameter values over a predefined number of iterations. This can lead to better results in less computational time when compared to grid search.
d. Cross-Validation for Hyper parameter Tuning: To prevent overfitting during hyper parameter tuning, cross-validation is used. The dataset is divided into multiple subsets, and different combinations of hyper parameters are evaluated on these subsets. This process helps in estimating the model's performance more reliably and ensures that the hyper parameter choices are not biased by the specific data split.
e. Scoring Metric for Tuning: The evaluation metric used during hyper parameter tuning is crucial. It guides the search for the best hyper parameters based on the desired performance criteria. For example, if the goal is to minimize prediction errors, the RMSE might be used as the scoring metric during tuning.
f. Automated Tuning Techniques: In addition to manual tuning using grid search or random search, automated hyper parameter tuning techniques like Bayesian optimization and genetic algorithms can also be employed. These techniques use mathematical optimization methods to search for the best hyper parameters more efficiently.
g. Best Hyper parameter Combination: After evaluating various hyper parameter combinations, the set of hyper parameters that results in the best-performing model is selected. This combination is then used to train the final model on the entire training dataset.

Hyper parameter tuning ensures that the machine learning model is fine-tuned to achieve its best possible performance. By finding the optimal hyper parameters, the model's ability to accurately estimate energy consumption in the residential complex building is enhanced, making it a valuable tool for energy management and conservation efforts.

3.4.7 Cross-Validation

Cross-validation is a vital method used throughout the model training process to examine the performance and generalization potential of the machine learning model. It's a more thorough assessment of the model's efficacy than just using a train/test split to see how well it did on new data. To guarantee the precision and stability of the energy consumption estimate model [46], cross-validation was used in this study.

Cross-validation relies on the following principles:

a. K-Fold Cross-Validation: K-Fold Cross-Validation is one of the most commonly used techniques. The data is folded into 'K' roughly equal-sized chunks. This procedure is repeated 'K' times, with a new fold serving as the test set each time and the other folds serving as the training set. The final assessment criteria are the weighted average of the K best iterations' performances. By doing so, we eliminate the possibility of selecting just certain data points for training and testing.
b. Stratified Cross-Validation: In cases where the dataset is imbalanced or has specific patterns that need to be preserved during cross-validation, stratified cross-validation is used. This technique ensures that each fold has a similar distribution of data points from different classes or categories, which is especially important when dealing with skewed datasets.
c. Leave-One-Out Cross-Validation (LOOCV): The amount of data points in a dataset determines the value of K in K-Fold Cross-Validation, of which LOOCV is a particular instance. In each cycle, just one data point is utilized for testing, while the others are used for training. LOOCV employs virtually all available data for training and testing, which results in a more precise evaluation of the model's performance. However, for huge datasets, it may be computationally costly.
d. Hold-Out Validation: In this approach, a portion of the dataset is held out as a validation set, and the rest is used for training the model. This validation set is then used to evaluate the model's performance. While hold-out validation is simple and

computationally efficient, it may lead to higher variance in the performance estimate compared to K-Fold Cross-Validation, especially for smaller validation sets.

e. Benefits Of Cross-Validation: Cross-validation helps to identify potential issues such as overfitting or under fitting. It provides a more accurate assessment of how well the model was perform on unseen data, which is crucial for the practical deployment of the energy consumption estimation model. By using cross-validation, the model's hyper parameters can be fine-tuned to achieve better generalization and improved accuracy on new data.
f. Cross-Validation And Hyper parameter Tuning: During hyper parameter tuning (as discussed in section 3.4.5), cross-validation is commonly used to evaluate different hyper parameter combinations. This helps to ensure that the best hyper parameters are chosen based on their performance on multiple subsets of the data, rather than being biased by a single train-test split.

By employing cross-validation techniques, this research project can assess the performance of the machine learning model for energy consumption estimation accurately. It provides a more reliable estimate of the model's accuracy, robustness, and generalization ability, contributing to the overall success of the project.

3.4.8 Early Stopping

To avoid overfitting and enhance machine learning models' capacity to generalize, early halting is a regularization strategy employed often during model training. To prevent the energy consumption, estimate model from continuing to train after its performance on the validation set has dropped, early stopping was employed in this study [47].

The process of early stopping is as follows:

a. Training And Validation Sets: As discussed earlier, the dataset is divided into training and testing sets. During model training, a portion of the training data is set aside as the validation set.
b. Monitoring The Validation Loss: During the training process, after each epoch or a certain number of iterations, the model's performance is evaluated on the validation set using an evaluation metric, such as Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE). The validation loss is calculated based on these metrics.

c. Stopping Criteria: Early stopping requires defining a stopping criterion, which is usually based on the performance of the model on the validation set. Commonly, the training is stopped if the validation loss does not improve or starts to increase for a certain number of consecutive epochs. This indicates that the model has started to over fit the training data and is not generalizing well to new, unseen data.
d. Model Reversion: Once the stopping criterion is met, the model's weights and parameters are reverted to the point where the validation loss was at its lowest. This ensures that the model retains the best performance achieved during the training process and prevents it from overfitting.
e. Benefits Of Early Stopping: Early stopping helps prevent the model from memorizing noise in the training data and encourages it to learn more general patterns that can be applied to new data. It improves the model's ability to generalize and make accurate predictions on unseen data, which is crucial for the successful application of the energy consumption estimation model in a real-world setting.
f. Early Stopping And Training Time: Implementing early stopping can also lead to a reduction in training time, especially if the training process is computationally expensive. Instead of training for a fixed number of epochs, early stopping allows the model to stop as soon as it reaches the optimal performance on the validation set, saving computational resources.
g. Choosing The Right Patience: The number of epochs to wait before stopping the training process is a hyper parameter known as "patience." Setting the patience too low may lead to premature stopping, resulting in a suboptimal model. On the other hand, setting it too high may lead to prolonged training, negating the benefits of early stopping. The patience value needs to be chosen carefully based on the characteristics of the dataset and the complexity of the model.

By incorporating early stopping in the model training process, this research project aims to develop a energy consumption estimation model that achieves better generalization, improved accuracy, and reliability in predicting energy usage in the residential complex building.

3.4.9 Model Finalization

Model finalization is the last step in the model training process, where the trained machine learning model is prepared for deployment and future use in energy consumption estimation. Once the model has been trained and its performance on the validation set is satisfactory, it undergoes finalization to ensure its readiness for practical applications. This step involves several key activities [48]:

a. Retraining on Full Dataset (Optional): Depending on the approach used during training, it might be beneficial to retrain the model using the entire pre-processed dataset, including both the training and testing sets. This can lead to a more generalized model that has seen more data and, in some cases, can improve overall performance.
b. Hyper parameter Persistence: The optimal hyper parameters determined during the hyper parameter tuning process are saved and recorded for future reference. These hyper parameters are crucial as they define the configuration of the model that achieved the best performance during training.
c. Saving Model Parameters: The model's learned parameters, such as weights and biases in neural networks or coefficients in linear regression, are saved for future use. These parameters encode the patterns and relationships learned during the training process and was used for energy consumption estimation on new data.
d. Serialization And Format: The model is serialized and saved in a specific format, such as JSON or HDF5, which allows for easy storage and portability. Serialization ensures that the entire model structure and its parameters are saved as a single file, making it convenient to transport and load the model on different platforms.
e. Documentation: Comprehensive documentation of the model is created, describing the architecture, hyper parameters, and any specific considerations made during the training process. This documentation helps other researchers and practitioners understand and replicate the model.
f. Model Versioning: To manage the evolution of the model and track changes, model versioning is maintained. This is particularly useful if the model needs to be updated or retrained in the future to incorporate new data or adapt to changing conditions.
g. Integration And Deployment: The finalized model is integrated into the energy management system of the residential complex building, ready to receive real-time data from the smart meters and provide energy consumption estimations. The deployment

process ensures that the model is seamlessly incorporated into the existing infrastructure.

h. Monitoring And Maintenance: After deployment, continuous monitoring and maintenance are performed to ensure that the model's performance remains optimal. Regular updates and retraining might be necessary if the building's energy consumption patterns change over time.

By finalizing the trained model, this research project aims to create a reliable and accurate energy consumption estimation tool that can aid in optimizing energy usage, identifying inefficiencies, and making informed decisions for sustainable energy management in the residential complex building.

Throughout the training process, the main goal is to develop a predictive model that can accurately estimate energy consumption in the residential complex building based on the data from the smart meters. The effectiveness and accuracy of the trained model was crucial for achieving the research project's objectives and providing valuable insights for energy management in the building.

3.5 MODEL EVALUATION

Model evaluation is a critical step in the machine learning process that assesses the performance of the trained model on unseen data. It allows us to understand how well the model generalizes to new instances and how accurately it can predict energy consumption in the residential complex building. The evaluation provides insights into the model's strengths and weaknesses, helping us identify potential issues and determine whether the model meets the research project's objectives.

3.5.1 Metrics for Evaluation

Several evaluation metrics was used to measure the model's performance in estimating energy consumption. The following metrics are commonly employed in regression tasks:

3.5.1.1 Mean absolute error (MAE)

The performance of regression models, such as the machine learning model utilized in this study, is often measured by their Mean Absolute Error (MAE). It measures the typical absolute disparity between forecasted and observed values. The mean absolute error

(MAE) quantifies how much the model's predictions deviate from the truth on average [49].

The formula for calculating MAE is as showing in equation 3.5.1.1:

$$MAE = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j| \quad (3.1)$$

Where:

n is the number of samples in the testing set.

yi is the actual energy consumption value for sample *i*.

*yi^* is the predicted energy consumption value for sample *i*.

A lower MAE value indicates better performance, as it means the model's predictions are closer to the true values. Conversely, a higher MAE indicates that the model's predictions have larger errors, suggesting less accurate estimations.

In the context of this research project, the MAE was used alongside other evaluation metrics, such as RMSE and R2, to comprehensively assess the model's performance in estimating energy consumption. The combination of these metrics was providing a comprehensive understanding of the model's accuracy and its suitability for energy management applications in the residential complex building.

3.5.1.2 Root mean square error (RMSE)

Another common statistic used to evaluate the efficacy of regression models, such as the machine learning model employed here, is the root mean square error (RMSE). RMSE measures the square root of the average of the squared discrepancies between the anticipated values and the actual values. It is favored because it more accurately reflects the model's accuracy by penalizing greater prediction mistakes than smaller ones.
The formula for calculating RMSE is as showing in equation 3.5.1.2 [49]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{j-1}^{n} (y_j - \hat{y}_j)^2} \quad (3.2)$$

Smaller root-mean-squared errors (RMSE) between predicted and observed values imply higher model performance. However, a larger RMSE indicates that the model's predictions are less close to the true values, and so are less reliable.

Interpretation of RMSE is similar to MAE, but with the added benefit of penalizing large errors more.

In the context of this research project, the RMSE, along with other evaluation metrics like MAE and R2, was provide a comprehensive assessment of the model's accuracy. By considering multiple metrics, the research can gain a deeper understanding of the model's performance in energy consumption estimation, leading to more informed decisions regarding energy management strategies.

3.5.1.3 R-squared (R2) score

To assess how well a regression model fits the data, statisticians calculate a value called the R-squared (R2) score. It offers an indicator of how effectively the independent variables (features) explain the variation in the dependent variable (target). The R2 score is useful for evaluating the accuracy with which a machine learning model predicts energy consumption based on historical data.

A perfect R2 score would be 1. If the value is 1, then the model is a perfect fit for the data and accounts for all of the variation in the dependent variable. A score of 0 for R2 indicates that the model does not provide any explanation for the variation in the data, and that its predictions are identical to the target variable's mean. When R2 is negative, the model is less accurate than just forecasting the mean.

The formula for calculating the R2 score is as showing in equation 3.5.1.3 [49]:

$$R^2 = 1 - \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \quad (3.3)$$

Where:

a. *n* is the number of data points in the test set.

b. *yi* is the actual energy consumption value for the i-th data point.

c. ^*yi*^ is the predicted energy consumption value for the i-th data point.

d. ¯*y*¯ is the mean of the actual energy consumption values.

The R2 score quantifies how much the variance of the predicted values differs from the variance of the actual values. A higher R2 score indicates a better fit of the model to the data, suggesting that the model's predictions are more accurate compared to using the mean as a predictor.

3.5.1.4 Mean square error (MSE)

Mean squared error (MSE), also called mean squared deviation (MSD), the average squared difference between the value observed in a statistical study and the values predicted from a model. When comparing observations with predicted values, it is necessary to square the differences as some data values will be greater than the prediction (and so their differences will be positive) and others will be less (and so their differences will be negative). Given that observations are as likely to be greater than the predicted values as they are to be less, the differences would add to zero. Squaring these differences eliminates this situation.
The formula for the mean squared error as showing in equation 3.5.1.4:

$$MSE = \frac{1}{n}\sqrt{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2} \quad (3.4)$$

Where:

n is the number of samples in the testing set.

yi is the actual energy consumption value for sample *i.*

*yi^* is the predicted energy consumption value for sample *i.*

The Σ indicates that a summation is performed over all values of i.

If the prediction passes through all data points, the mean squared error is zero. As the distance between the data points and the associated values from the model increase, the mean squared error increases. Thus, a model with a lower mean squared error more accurately predicts dependent values for independent variable values.

In the context of this research project, the R2 score, along with other evaluation metrics like MAE, MSE and RMSE, was provide a comprehensive assessment of the model's accuracy and performance in energy consumption estimation. By considering multiple

metrics, the research can gain a more complete understanding of the model's strengths and limitations, supporting data-driven decisions for energy management in the residential complex building.

The evaluation results were compared to the project's objectives and the performance of traditional methods used for energy consumption estimation. The main goal is to develop a model that provides accurate and reliable energy consumption estimates for the residential complex building in Iraq. If the model meets the desired accuracy and outperforms traditional methods, it can be applied to forecast future energy consumption and contribute to effective energy management in the building.

3.6 ENERGY CONSUMPTION ESTIMATION

The process of energy consumption estimation involves using the trained machine learning model to forecast the amount of electricity that was consumed in the residential complex building over the next two to three years. This estimation is crucial for effective energy management and cost reduction. By utilizing the model's predictions, building managers and energy providers can make informed decisions on resource allocation and demand planning [50].

To perform the energy consumption estimation, the following steps was taken:

a. Data Preparation: The historical data used for training and testing the machine learning model was organized and pre-processed to ensure consistency and accuracy. Any new data collected during the estimation period was also be incorporated into the dataset.
b. Feature Selection: The relevant features or input variables used during the training phase, such as weather conditions, occupancy patterns, and building characteristics, was selected for the energy consumption estimation.
c. Data Prediction: The selected features fed into the trained machine-learning model to generate energy consumption predictions for each time interval in the estimation period. The model was using its learned patterns and relationships to forecast electricity usage.
d. Performance Assessment: The accuracy of the energy consumption predictions was evaluated using the previously mentioned performance metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), R-squared (R2) Score, and Coefficient of Determination (COD). The model's performance was compared to the historical data and traditional estimation methods.

e. Future Energy Consumption Projection: Based on the model's predictions, the total energy consumption for the residential complex building was projected for the next two to three years. This projection was provided valuable insights into future energy demands and help in devising effective energy management strategies.
f. Decision Making: The estimated energy consumption values was used by building managers, energy providers, and policymakers to make informed decisions about energy allocation, pricing, and resource planning. Effective energy management based on accurate predictions can lead to reduced energy waste and cost savings.

It is important to note that energy consumption estimation is not a static process and should be periodically updated with new data to ensure continued accuracy. As the model receives additional data over time, it can be retrained to improve its performance and provide more precise estimates. Through energy consumption estimation, this research project aims to contribute to the efficient management of energy resources in residential complexes in Iraq. By harnessing the capabilities of machine learning and smart meters, this approach can lead to a more sustainable and environmentally responsible energy consumption pattern. Furthermore, the findings from this study can serve as a blueprint for other residential buildings facing similar energy management challenges, both in Iraq and other countries with comparable conditions.

3.7 ETHICAL CONSIDERATIONS

Ethical considerations are of paramount importance in any research project involving human participants and sensitive data. This chapter outlines the ethical principles and safeguards that was implemented throughout the energy consumption estimation project using machine learning with data from smart meters in a residential complex building in Iraq.

3.7.1 Informed Consent

Obtaining informed consent from the operation company of the residential complex is crucial before collecting any data. The research team was explaining the purpose of the study, the data collection process, and how the data was used. Participants was informed that their participation is voluntary, and they have the right to withdraw their consent at any time without consequences.

3.7.2 Data Privacy and Security

Ensuring the privacy and security of participants' data is a primary ethical concern. All data collected from smart meters and other sources was stored securely using encrypted databases and access controls. Only authorized personnel were having access to the data, and data sharing was limited to anonymized and aggregated datasets to protect individual privacy.
To minimize potential risks, the research team was implemented strict data anonymization and follow data protection regulations, such as the General Data Protection Regulation (GDPR) and relevant local laws.

3.7.3 Confidentiality

Confidentiality of participants' data was strictly maintained. The research team was not disclosing any personally identifiable information of the participants in any publication or presentation. Data referred to by unique identifiers, and the researchers was avoiding any information that may lead to the identification of individual participants.

3.7.4 Minimizing Harm

The research team was taking all necessary precautions to minimize harm to participants. The energy consumption estimation process was not causing any harm or disruption to the residents' daily lives. Participants' data was used solely for research purposes and was not be shared with any third parties.

3.7.5 Transparency and Accountability

The research team was transparent about the study's purpose, methods, and findings. Any potential conflicts of interest were disclosed, and all research procedures was adhered to ethical guidelines set by relevant institutions and regulatory bodies.

3.7.6 Data Retention and Disposal

Data collected for the energy consumption estimation project was retained for the duration necessary for analysis and validation. Once the research is completed, the data was securely archived or deleted in compliance with data protection regulations.

3.7.7 Ethical Review

Ethical considerations are fundamental to the successful and responsible conduct of the energy consumption estimation project. By upholding the principles of informed consent, data privacy, confidentiality, minimizing harm, transparency, and accountability, the research team was ensuring the protection of participants' rights and welfare. Adhering to ethical guidelines was contribute to the credibility and validity of the study's findings and foster trust among participants and stakeholders.

3.8 SUMMARY

In this chapter, the methodology for the energy consumption estimation project using machine learning with data from smart meters in a residential complex building in Iraq is presented. The chapter outlines the steps that was followed to achieve the objectives of the research and addresses the ethical considerations to ensure the protection of participants' rights and privacy.

The methodology begins with data collection, where information on energy consumption was gathered from smart meters installed in the residential complex. The data was then undergoing a pre-processing phase, which includes cleaning, normalization, and feature selection, to ensure data accuracy and reliability.

Next, the appropriate machine learning algorithm was selected for energy consumption prediction. Various algorithms, such as regression models, neural networks, and decision trees, was evaluated based on their performance indicators, such as Root Mean Square Error and Mean Absolute Error.

The chosen algorithm was then be trained using the pre-processed data to develop a model for energy consumption prediction. The model's performance was assessed using various metrics to ensure its accuracy and effectiveness in estimating energy consumption.

The energy consumption estimation was conducted based on the trained model, allowing for projections of energy usage over the next two to three years. The estimated energy consumption was compared with actual consumption to evaluate the model's precision and reliability.

The chapter also addresses ethical considerations, emphasizing the importance of informed consent, data privacy, confidentiality, minimizing harm, transparency, accountability, and ethical review. Participants' informed consent was obtained, and their data was anonymized

and securely stored to protect their privacy. The research team was transparent about the study's purpose and methods, and potential conflicts of interest was disclosed. An ethics committee was review and approve the study protocol to ensure adherence to ethical standards.

Overall, this chapter provides a comprehensive and systematic approach to conducting the energy consumption estimation project while upholding ethical principles and ensuring the protection of participants' rights and welfare. The methodology is designed to generate accurate and reliable predictions of energy consumption in the residential complex, contributing to effective energy management and sustainability efforts in Iraq and beyond.

4. PROPOSED METHOD

This section details the suggested approach for predicting energy usage in an Iraqi apartment complex using machine learning methods. Methods for estimating future electricity usage using data from smart meters are discussed here. These procedures involve a number of steps, from gathering raw data through testing and tuning the models. The suggested approach takes into account the difficulties provided by fluctuating energy consumption patterns and makes use of the strengths of state-of-the-art machine learning algorithms to arrive at accurate predictions [50].

The acquired data is verified for accuracy and reliability before being used to train machine learning models in the first step of the technique. The steps include filling in any blanks, finding any anomalies, extracting useful characteristics from timestamps, and normalizing the data to a common scale. The dataset is further separated into training and testing sets for model assessment, and categorical variables are encoded for interoperability with machine learning techniques. Models are validated using cross-validation techniques to avoid overfitting.

Following this critical stage, a variety of machine learning techniques are investigated for estimating energy usage. Linear regression, decision trees, random forest, gradient boosting, support vector regression (SVR), neural networks, lasso regression, ridge regression, elastic net and k -nearest neighbor algorithms are selected based on their ability to anticipate the complex's energy usage based on its specific features, we

After an appropriate method has been chosen, the focus shifts to model training. This phase incorporates data partitioning, model initialization, iterative parameter optimization, hyper parameter tweaking, and validation set additions (if desired). In order to reduce prediction error and accurately reflect underlying data trends, the iterative optimization process involves adjusting model parameters.

The effectiveness of trained models is evaluated in the last phase, which is known as model evaluation. This process makes use of metrics such as Mean Absolute Error (MSE), Root Mean Square Error (RMSE), and R-squared value. Methods of cross-validation are absolutely necessary for determining whether or not the models are accurate and whether or not they may be utilized with new data. We select the model that provides the most

accurate results in relation to these metrics so that we can provide trustworthy estimates of the amount of energy consumed by the building that houses a residential complex.

The objective of the approach that has been suggested is to lay a reliable groundwork for the application of machine learning algorithms and the data collected by smart meters in order to arrive at energy consumption estimates that are reliable and accurate. This method lays the groundwork for precise energy management in residential buildings by integrating efficient data preprocessing, careful algorithm selection, painstaking model training, we implemented several algorithms for daily, weekly and annual prediction.

4.1 SYSTEM SETUP

Before digging into the implementation of the suggested technique, it is vital to create the system setup and environment in which the research project was carried out. This must be done before moving on to the implementation of the method. The configuration of the system includes the software tools, programming languages, and hardware resources that was used to carry out the various steps of the technique, beginning with the pre-processing of the data and continuing on through the training and assessment of the model [51].

4.1.1 Software Tools and Libraries

Data manipulation, the creation and assessment of machine learning models, and evaluation are all made easier by a collection of specialized software tools and libraries, which are essential to the effective application of the approach that has been suggested. It is planned to make use of the following applications and libraries:

a. Python Programming Language: Python provides a versatile and comprehensive environment for data analysis, machine learning, and scientific computing. Its extensive ecosystem of libraries makes it well-suited for this research project.

b. Pandas: Pandas is a powerful data manipulation and analysis library that offers data structures and functions essential for cleaning, transforming, and exploring datasets.

Figure 4.1: Screenshot Showing Pandas Library.

c. NumPy: NumPy provides support for large, multi-dimensional arrays and matrices, along with a wide range of mathematical functions to operate on these arrays.

d. Scikit-Learn: Scikit-Learn is a robust machine learning library that includes various algorithms for classification, regression, clustering, and more [52]. It also offers tools for model evaluation and selection.

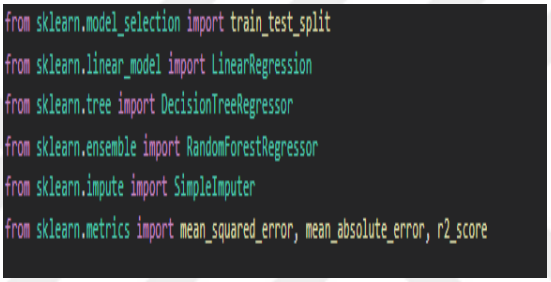


Figure 4.2: Screenshot Showing Imported Libraries.

e. Tensor Flow and Keras: Tensor Flow is an open-source machine learning framework developed by Google. Keras is an API that runs on top of Tensor Flow, simplifying the process of building, training, and evaluating neural networks.

f. Matplotlib And Seaborn: These libraries are essential for creating visualizations and plots to analyze the data and model performance.

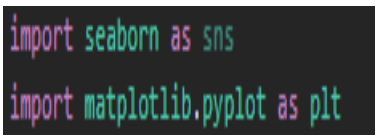


Figure 4.3: Screenshot Showing Matplotlib and Seaborn.

4.1.2 Hardware Resources

The suggested approach may be carried out on a regular home computer; however, making use of hardware resources that have a greater processing capacity can greatly speed up the training process. This is especially true when working with complicated algorithms or extensive datasets. Therefore, having access to a computer that has a sufficient amount of RAM, a powerful CPU, and ideally a GPU that is solely dedicated to the machine can increase the effectiveness of the research endeavor.

4.1.3 Data Storage and Management

In light of the sensitive nature of data pertaining to energy use as well as the requirement to preserve both privacy and security, a data storage and management system that is both secure and manageable was developed. The data that is gathered from smart meters was saved in a database that is organized and was only allow access to those who are allowed. In order to avoid any loss of data, adequate data backup measures were put into place.

4.2 TESTING AND RESULTS

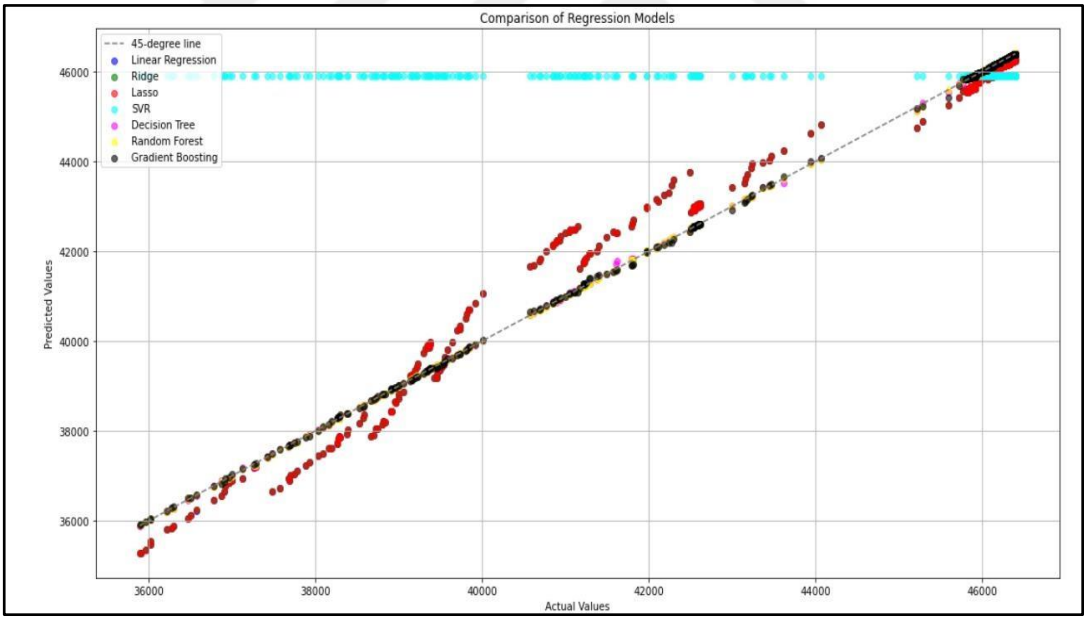
4.2.1 First Test and Result for Daily Predict

We selected an example of the dataset for building one for the customer b1f1f3 and take one day for different hours, we executed daily predict by using seven algorithms (Linear Regression, Ridge Regressor, lasso , SVR , Decision tree Regressor, Random forest Regressor and Gradient Boosting Regressor).

We calculate the avrage of R^2 and MSE results by using algorithm above comparison it and select the best one model using python code in appendix A.1, page #118, #119 and #120 .

	A	B	C	D	E	F	G
1	Subscriber No	Year	Month	Day	Hour	Minute	Energy consumption
2	b1.f1.f3	2023	8	30	17	7	46405.577
3	b1.f1.f3	2023	8	30	17	6	46405.569
4	b1.f1.f3	2023	8	30	17	5	46405.561
5	b1.f1.f3	2023	8	30	17	3	46405.552
6	b1.f1.f3	2023	8	30	16	59	46405.524
7	b1.f1.f3	2023	8	30	16	58	46405.515
8	b1.f1.f3	2023	8	30	16	55	46405.493
9	b1.f1.f3	2023	8	30	16	53	46405.484
10	b1.f1.f3	2023	8	30	16	51	46405.467
11	b1.f1.f3	2023	8	30	16	49	46405.458
12	b1.f1.f3	2023	8	30	16	6	46405.257
13	b1.f1.f3	2023	8	30	16	5	46405.213
14	b1.f1.f3	2023	8	30	14	50	46400.481
15	b1.f1.f3	2023	8	30	14	49	46400.414
16	b1.f1.f3	2023	8	30	14	48	46400.295
17	b1.f1.f3	2023	8	30	13	57	46396.441
18	b1.f1.f3	2023	8	30	13	56	46396.372
19	b1.f1.f3	2023	8	30	13	53	46396.212
20	b1.f1.f3	2023	8	30	13	52	46396.139
21	b1.f1.f3	2023	8	30	13	51	46396.06
22	b1.f1.f3	2023	8	30	13	47	46395.858
23	b1.f1.f3	2023	8	30	13	46	46395.796
24	b1.f1.f3	2023	8	30	13	44	46395.737
25	b1.f1.f3	2023	8	30	13	43	46395.675
26	b1.f1.f3	2023	8	30	13	40	46395.647
27	b1.f1.f3	2023	8	30	13	39	46395.638
28	b1.f1.f3	2023	8	30	13	37	46395.628

Figure 4.4: Screenshot for Sample of Daily Dataset for B1 F1 F3.



` **Figure 4.5:** Screenshot for Comparison of Regression Models with Actual for Daily Dataset of B1 F1 F3.

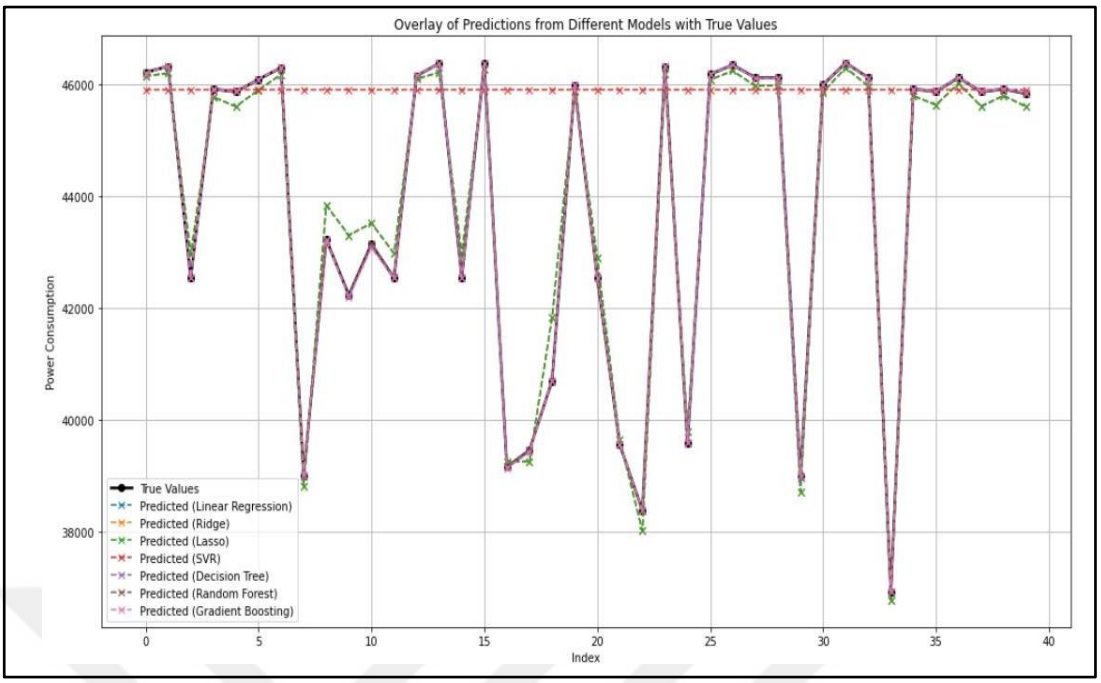


Figure 4.6: Screenshot for Comparison of ML Algorithms for Daily Energy Consumption for B1 F1 F3.

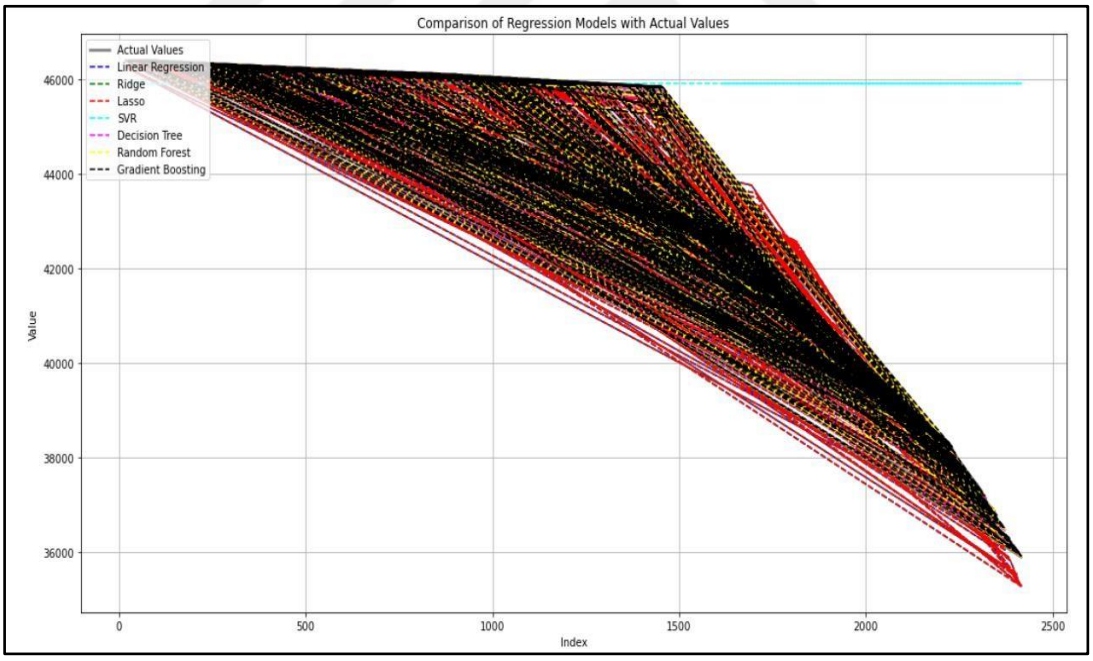


Figure 4.7: Screenshot of Comparison at Regression Models for Daily with Actual of B1 F1 F3.

Table 4.1: Table Summary of Comparison Algorithms for Daily B1 F1 F3.

Algorithm	R^2	MSE	Predict
Linear Regression	0.981	182767	41434
Ridge	0.981	182778	41435
Lasso	0.981	182776	41435
SVR	-0.382	137400	45911
Decision Tree Regressor	0.99	1042.46	40596.261
Random Forest Regressor	0.99	121.48	40588
Gradient Boosting Regressor	0.99	534.1	40566

Result For Daily Predict:

In this test we chose building one, first floor, flat three (b1. f1. f3) and used the dataset shown in screenshot at figure 4.4 (total dataset 2420 started from date 01-Jan-2023 until 30-Aug-2023) and we selected random date 15-May-2023 at hour 14 the predict are 40566 Watt. hour and the real value are 40537 Watt. Hour.

From figure 4.6 and 4.7, showing that all models almost fitted between real and predict expect SVR model.

From table 4.1 showing the value of R^2 is higher than 0.98 for all models expect SVR model. This result lets us to elaborate another test.

4.2.2 Second Test and Result for Weekly Predict

We selected an example of the dataset for building one for the customer b1f1f3 and take one week for different days, we executed weekly predict by using ten algorithms (Linear Regression, Ridge Regressor, lasso, SVR, Decision tree Regressor, Random forest

Regressor, Gradient Boosting Regressor, Elastic Net, K-Nearest neighbour and Neural Network).

We calculate the avrage of R^2 , MSE , RMSR and Gross vailated results by using algorithm above comparison it and select the best one model using python code in appendix A.2, page #121 and #122.

	A	B	C	D
1	Subscriber No	Year	Week Number	Energy Consumption
2	b1.f1.f3	2023	1	36171.275
3	b1.f1.f3	2023	2	36602.014
4	b1.f1.f3	2023	3	36918.87
5	b1.f1.f3	2023	4	37350.187
6	b1.f1.f3	2023	5	37600.462
7	b1.f1.f3	2023	6	37924.702
8	b1.f1.f3	2023	7	38273.892
9	b1.f1.f3	2023	8	38562.974
10	b1.f1.f3	2023	9	38750.352
11	b1.f1.f3	2023	10	38941.715
12	b1.f1.f3	2023	11	39093.176
13	b1.f1.f3	2023	12	39242.43
14	b1.f1.f3	2023	13	39399.826
15	b1.f1.f3	2023	14	39551.877
16	b1.f1.f3	2023	15	39680.767
17	b1.f1.f3	2023	16	39832.345
18	b1.f1.f3	2023	17	39849.345
19	b1.f1.f3	2023	18	40000.923
20	b1.f1.f3	2023	19	40152.501
21	b1.f1.f3	2023	20	40669.028
22	b1.f1.f3	2023	21	40915.784
23	b1.f1.f3	2023	22	41230.587
24	b1.f1.f3	2023	23	41509.401
25	b1.f1.f3	2023	24	41951.518
26	b1.f1.f3	2023	25	42236.153
27	b1.f1.f3	2023	26	42542.354
28	b1.f1.f3	2023	27	43054.686

Figure 4.8: Screenshot for Sample of Weekly Dataset for B1 F1 F3.

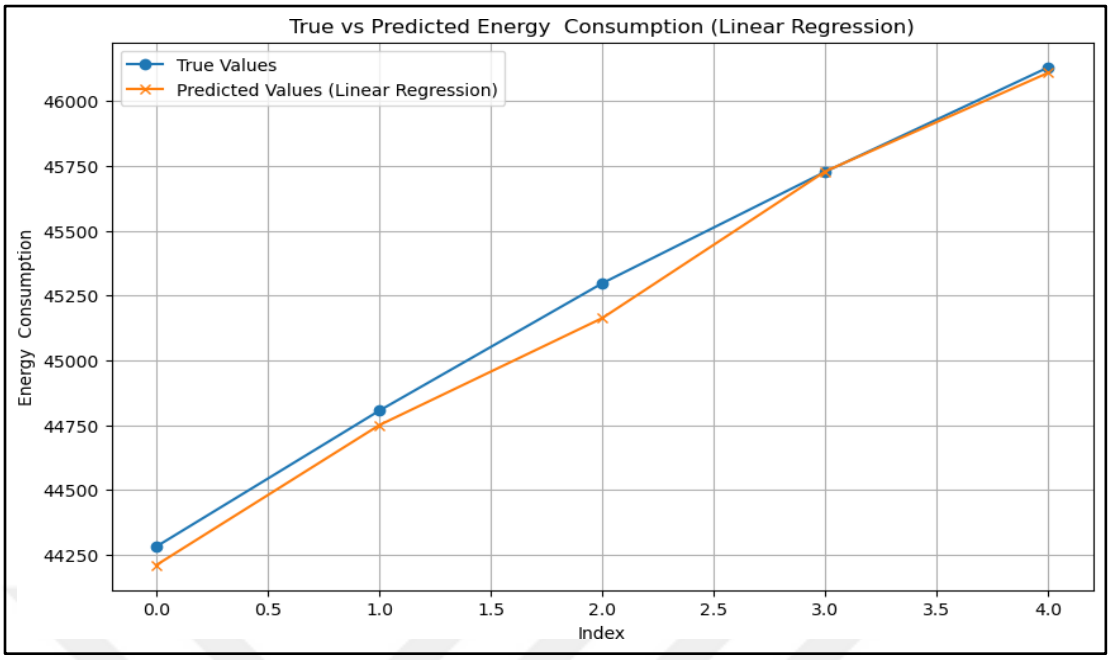


Figure 4.9: Screenshot for Weekly Energy Consumption of Linear Regression.

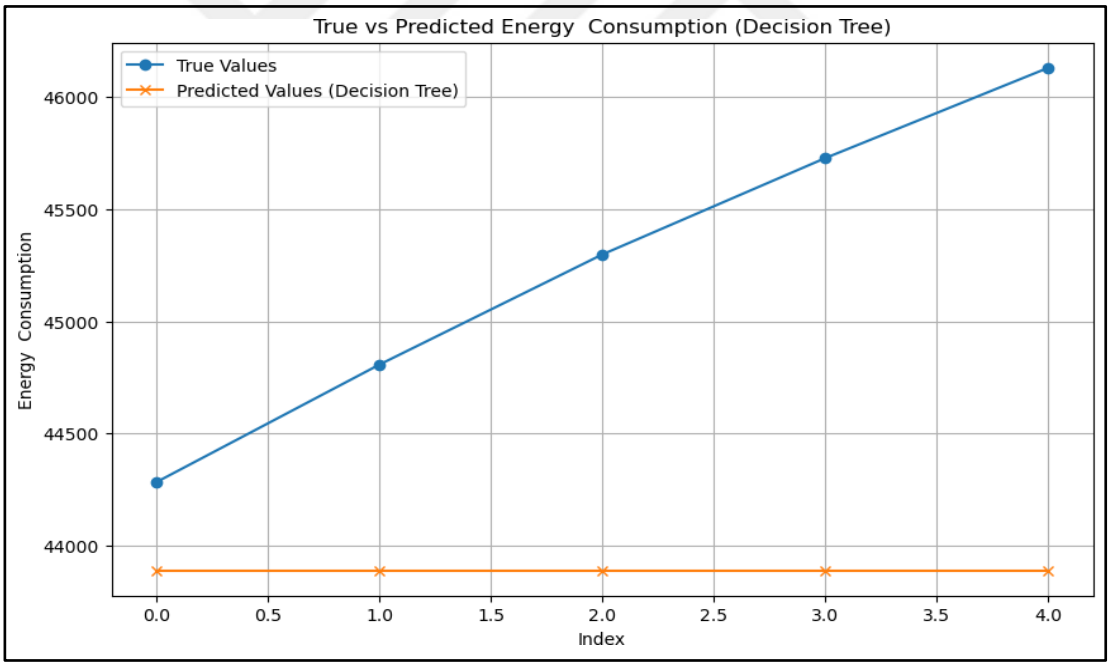


Figure 4.10: Screenshot for Weekly Energy Consumption of Decision Tree.

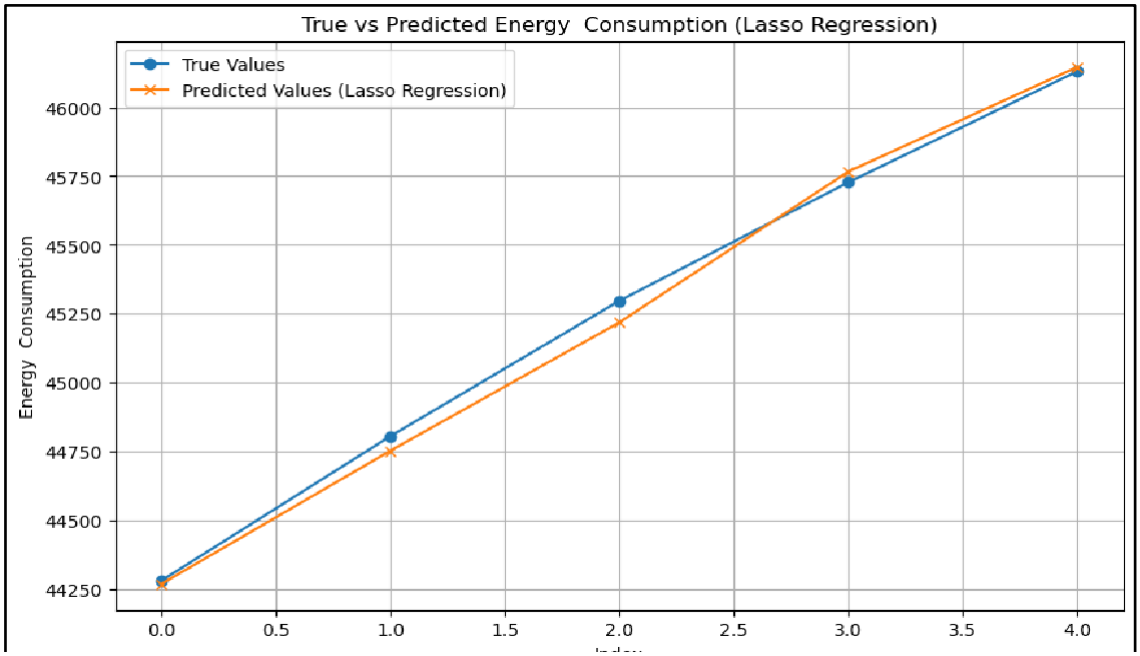


Figure 4.11: Screenshot for Weekly Energy Consumption of Gradient Boosting.

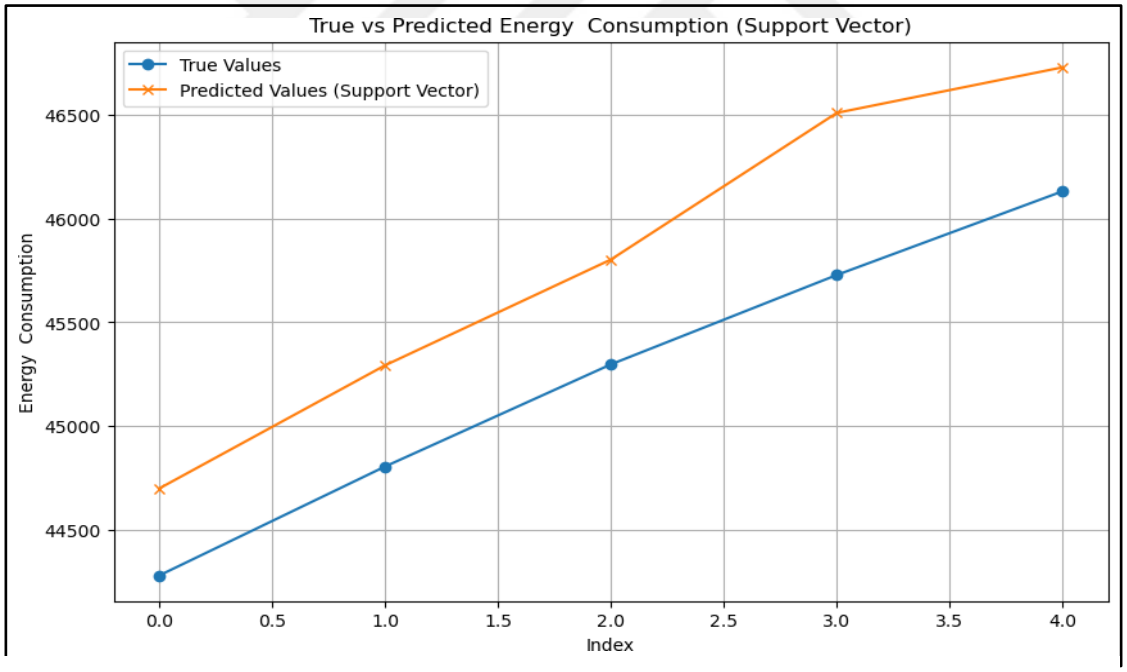


Figure 4.12: Screenshot for Weekly Energy Consumption of Support Vector.

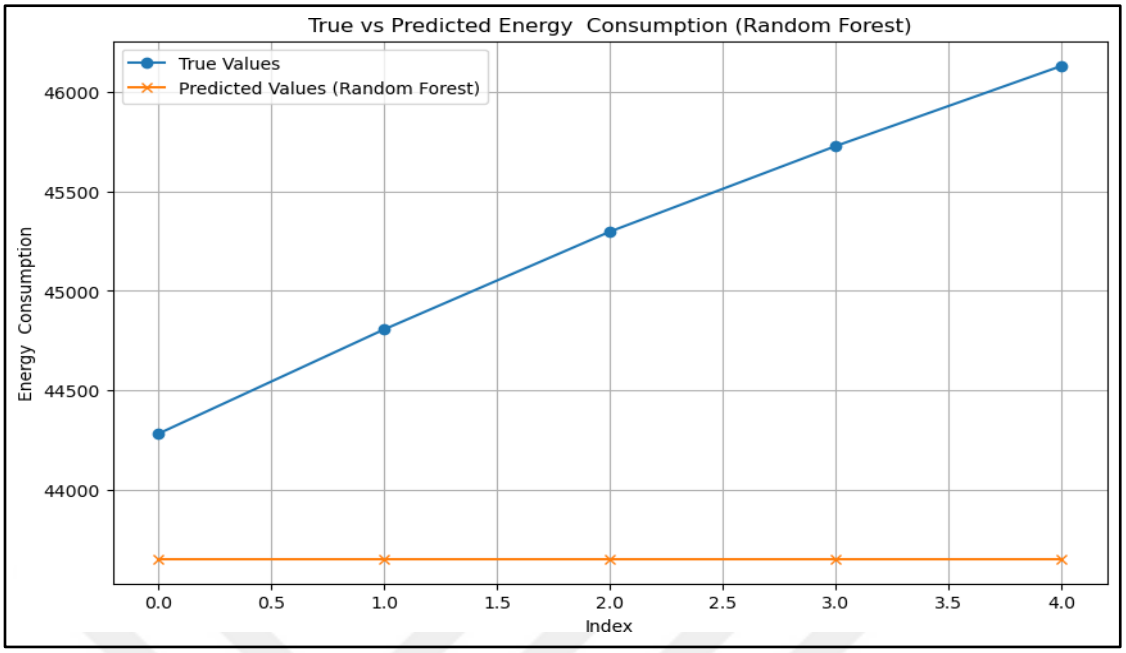


Figure 4.13: Screenshot for Weekly Energy Consumption of Lasso Regression.

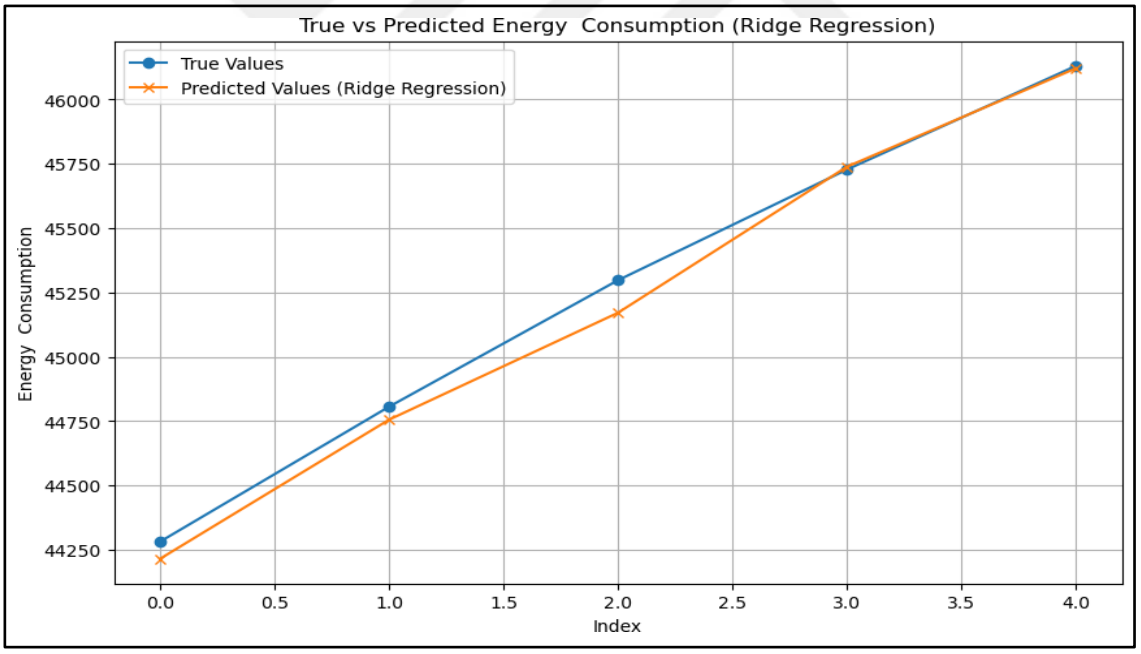


Figure 4.14: Screenshot for Weekly Energy Consumption of Ridge Regression.

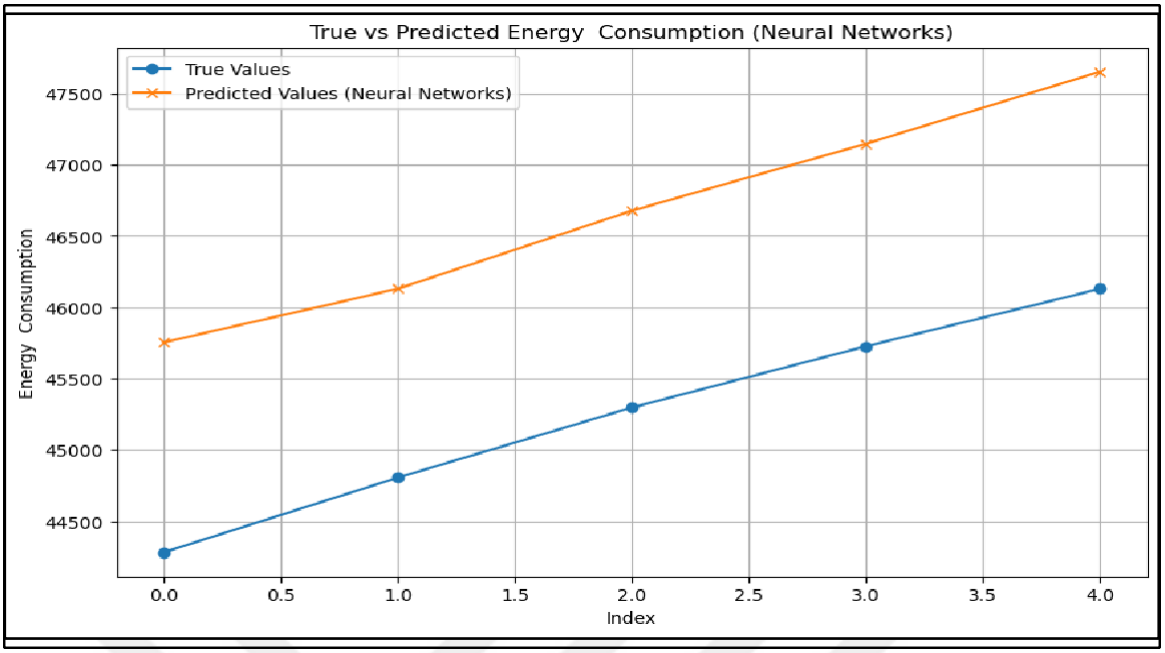


Figure 4.15: Screenshot for Weekly Energy Consumption of Random Forest.

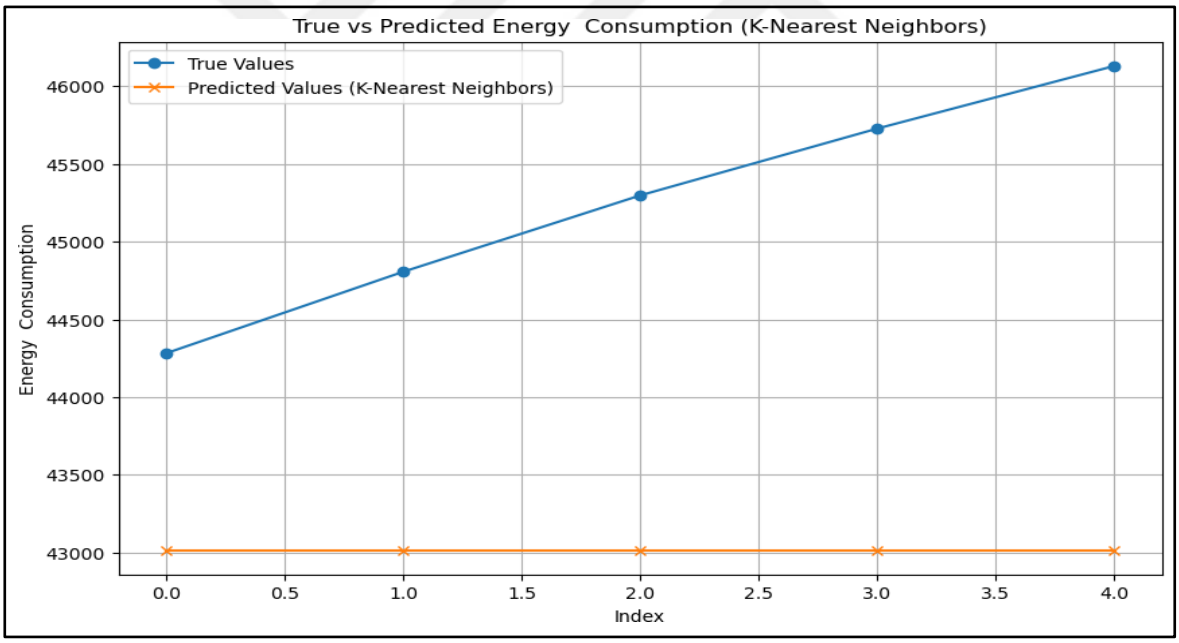


Figure 4.16: Screenshot for Weekly Energy Consumption of K-Nearest Neighbors.

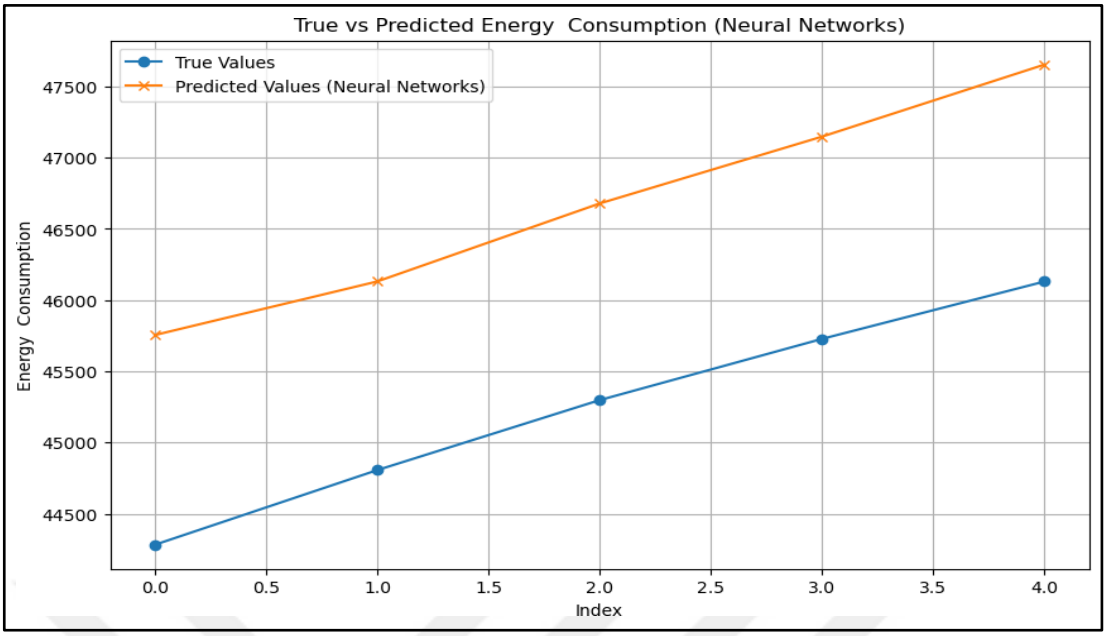


Figure 4.17: Screenshot for Weekly Energy Consumption Neural Networks.

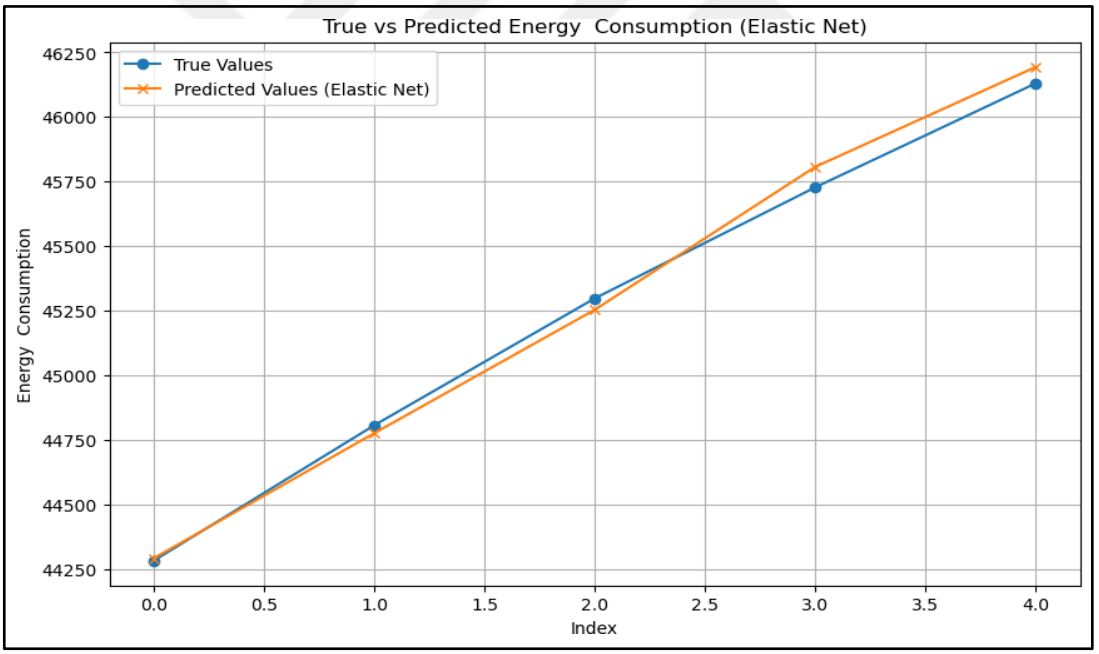


Figure 4.18: Screenshot for Weekly Energy Consumption of Elastic Net.

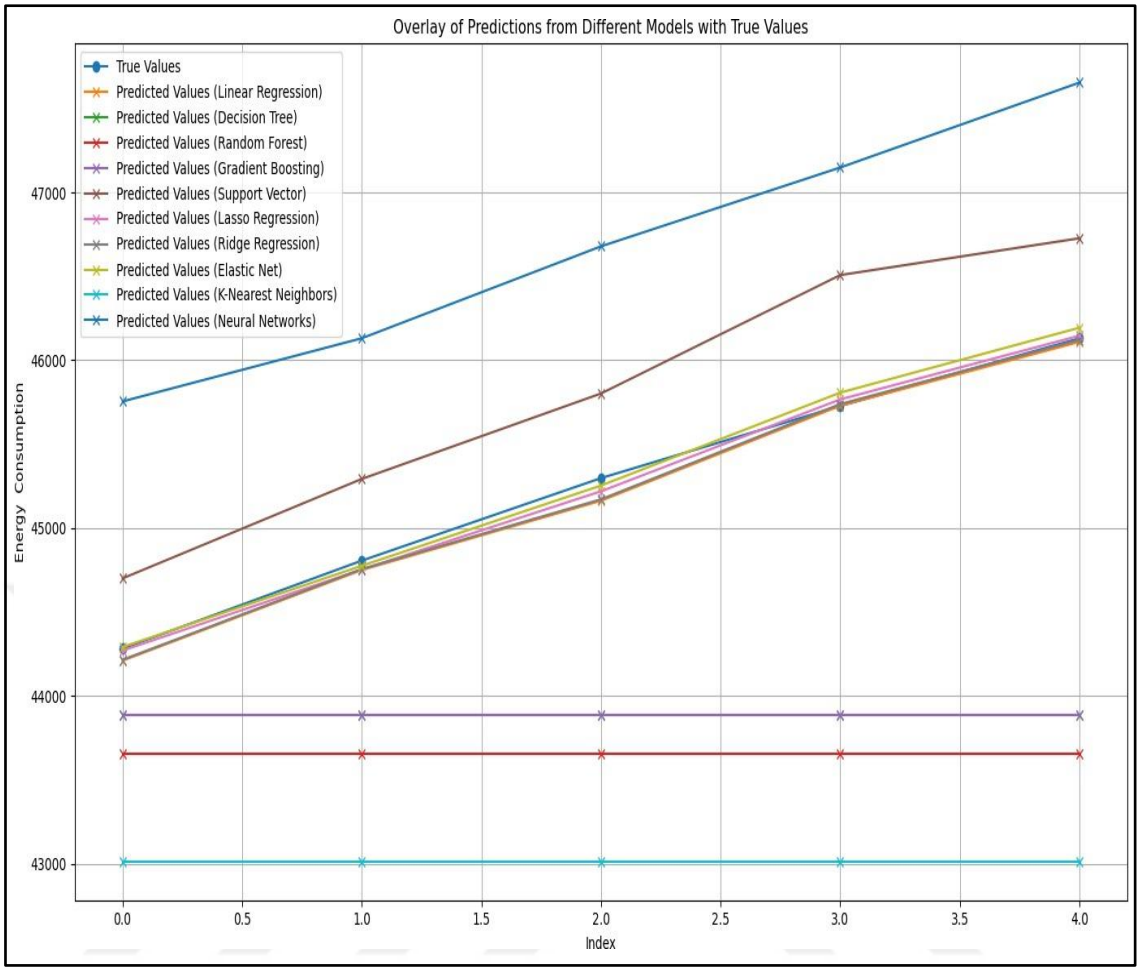


Figure 4.19: Screenshot for Comparisons of ML Algorithms for Weekly Energy Consumption for B1 F1 F3.

Table 4.2: Table Summary of Comparisons Algorithms for Weekly B1 F1 F3.

Algorithm	R^2	MAE	RMSE	Gross Validated
Linear Regression	0.99	56.77	73.10	-1.57
Ridge Regressor	-0.19	52.75	68.04	0.55
Lasso Regressor	-1.53	39.52	46.57	-0.19
SVR	-5.89	577.06	570.90	-1.53
Decision Tree Regressor	-4.32	1360.50	1509.7	-5.89
Random Forest Regressor	-6.13	1619.65	1746.8	-12.52
Gradient Boosting Regressor	-4.32	1360.63	1509.8	-5.89
Elastic Net	0.55	45.35	0.99	0.58
K-Nearest Neighbors	0.99	2233.78	51.53	-38.85
Neural Network	-3.75	1424.70	1426.38	-104.97

Result For Weekly Predict:

In this test we chose building one, first floor, flat three (b1. f1. f3) and used the dataset shown in screenshot at figure 4.8 (total dataset 34 weeks started from date 01-Jan-2023 until 30-Aug-2023)

We implemented ten different models as shown in figure 4.19, showing that the linear Regression and K-Nearest neighbour are the best models.

From table 4.2 showing the value of R^2 are 0.99 for two models are okay and the other failed. From result of weekly predict, seems datasets insufficient for the several of algorithms, motivate us to start looking forward with annuals prediction.

4.2.3 Third Test and Result for Monthly Predict 2020

We took all buildings (at beginning of project only three buildings operation at 2020), we executed monthly predict by using five algorithms (Decision tree Regressor, Random forest Regressor, Gradient Boosting Regressor, Bagging Regressor and Extra Trees).

We calculate the avarge of R^2 and MSE results by using algorithm above comparison it and select the best one model using python code in appendix A.3, page #123,#124 and #125.

	A	B	C	D
1	Costumer ID	Year	Month	Energy Consumption
2	b1 colling system	2020	9	11473.997
3	b1 colling system	2020	10	14686.046
4	b1 colling system	2020	11	15566.829
5	b1 colling system	2020	12	16036.007
6	b1 elevator L1	2020	9	1258.379
7	b1 elevator L1	2020	10	1660.231
8	b1 elevator L1	2020	11	2152.338
9	b1 elevator L1	2020	12	2668.76
10	b1 fan	2020	9	13350.482
11	b1 fan	2020	10	17421.993
12	b1 fan	2020	11	21393.275
13	b1 fan	2020	12	25482.299
14	b1.f1.f3	2020	9	5915.996
15	b1.f1.f3	2020	10	6823.894
16	b1.f1.f3	2020	11	7453.591
17	b1.f1.f3	2020	12	8647.367
18	b1.f1.f4	2020	9	1337.097
19	b1.f1.f4	2020	10	1480.61
20	b1.f1.f4	2020	11	1656.278
21	b1.f1.f4	2020	12	1858.546
22	b1.f1.f5	2020	9	5654.197
23	b1.f1.f5	2020	10	6887.893
24	b1.f1.f5	2020	11	7774.272
25	b1.f1.f5	2020	12	9777.464
26	b1.f1.f6	2020	9	10476.226
27	b1.f1.f6	2020	10	12609.886
28	b1.f1.f6	2020	11	13965.942

Figure 4.20: Screenshot for Sample of Dataset Forall Three Buildings at 2020.

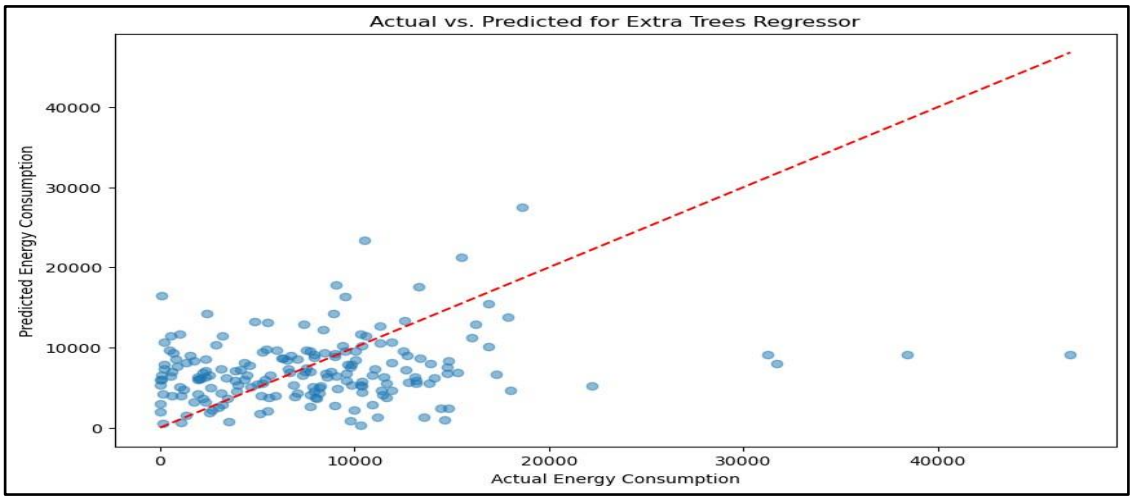


Figure 4.21: Screenshot for Gradient Boosting Regressor with Actual at 2020.

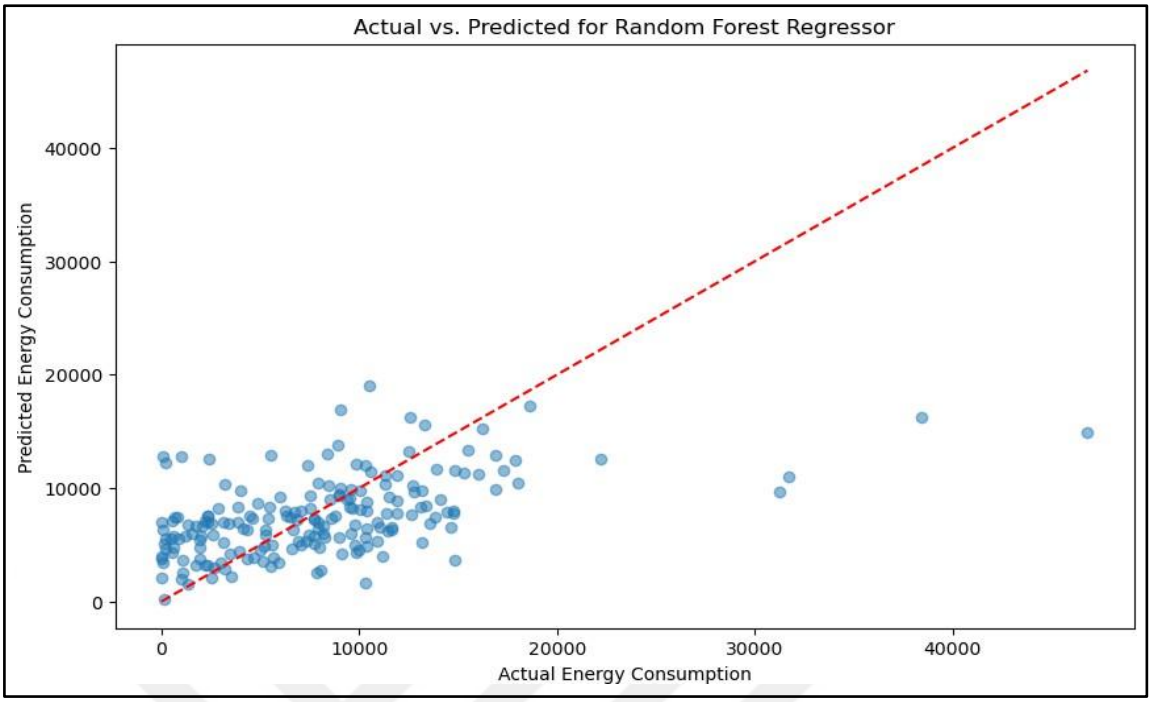


Figure 4.22: Screenshot for Random Forest Model with Actual at 2020.

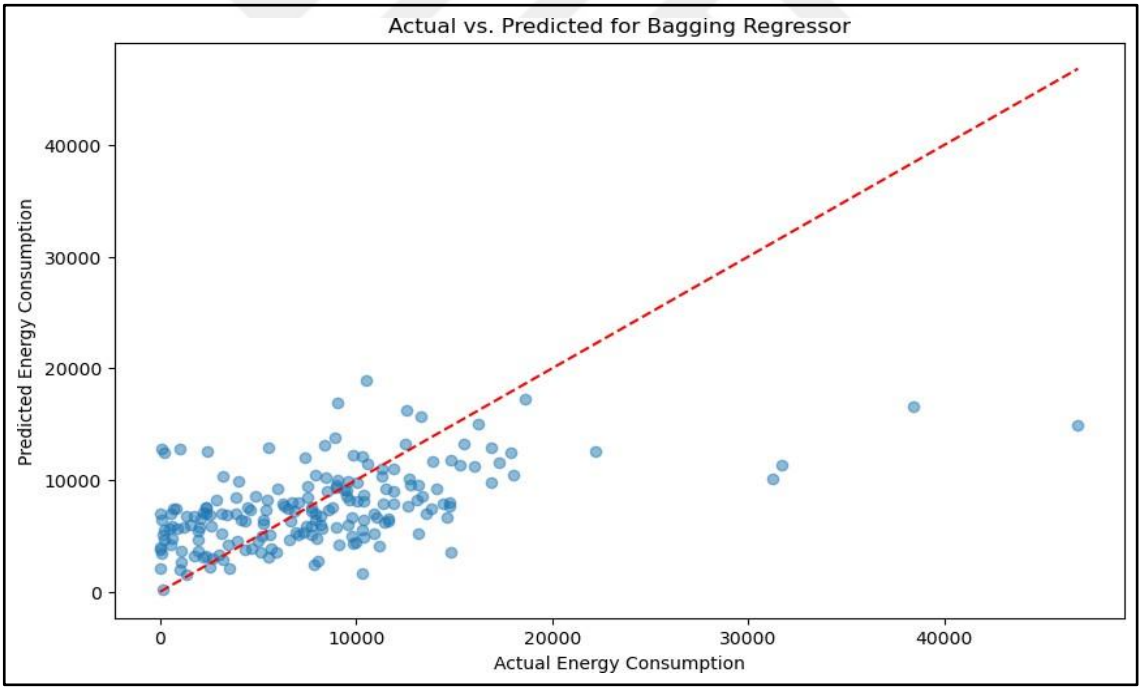


Figure 4.23: Screenshot for Bagging Regressor with Actual at 2020.

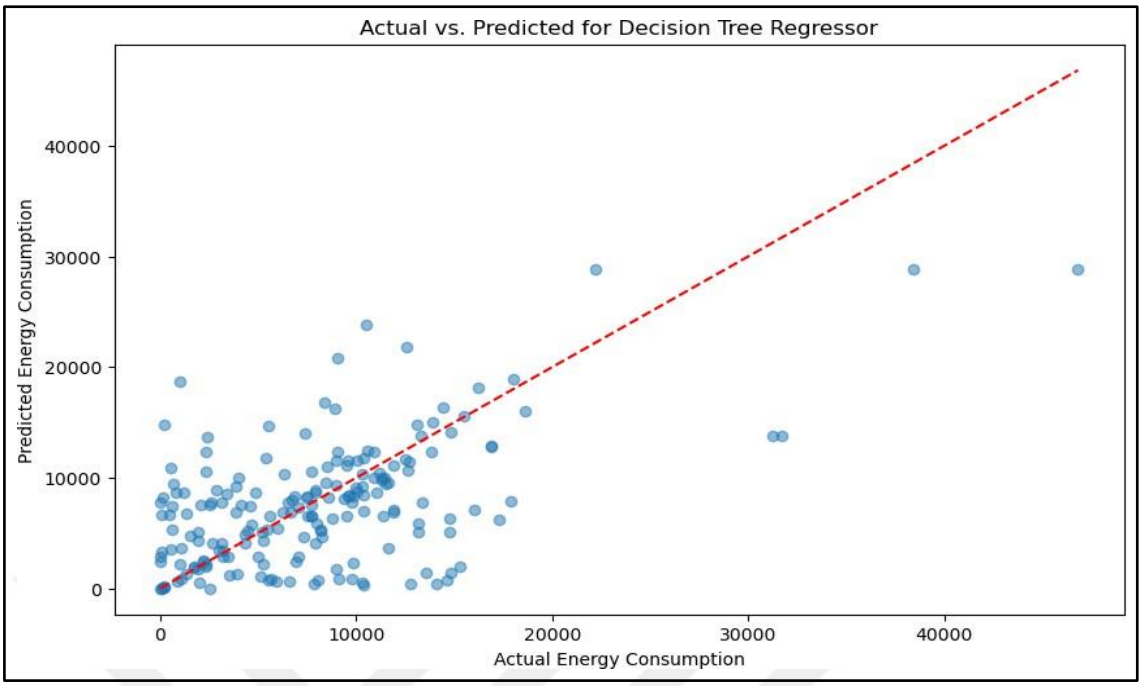


Figure 4.24: Screenshot for Decision Tree Regressor with Actual at 2020.

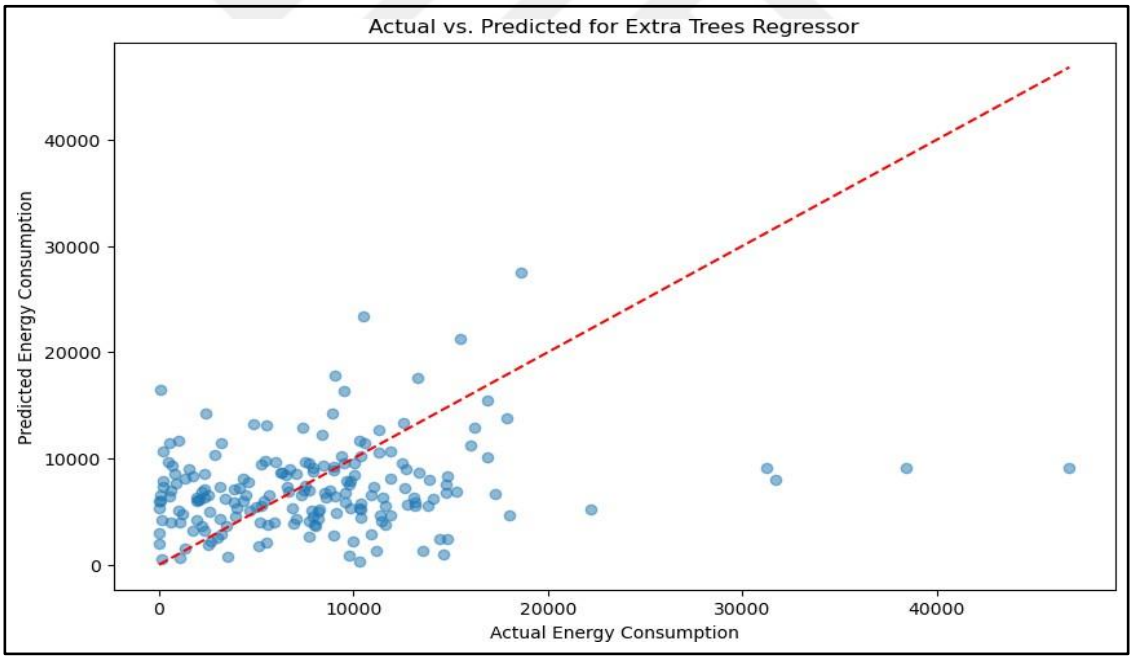


Figure 4.25: Screenshot for Extra Trees Regressor with Actual at 2020.

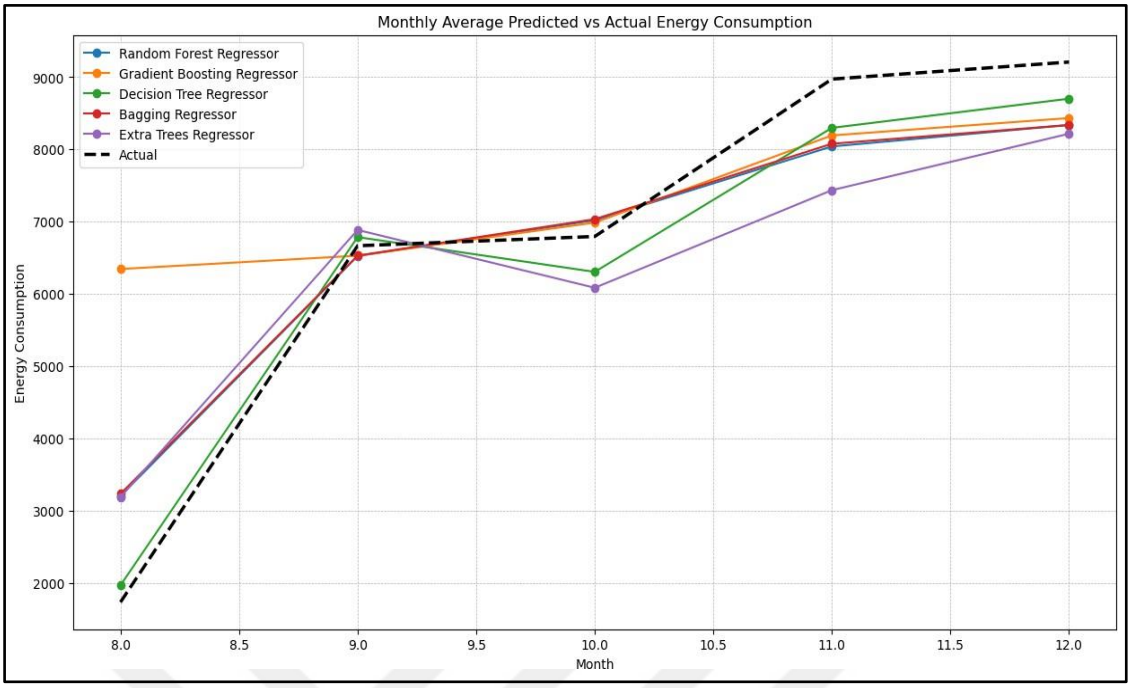


Figure 4.26: Screenshot for Comparisons of ML Algorithms for Monthly Energy Consumption at 2020.

Table 4.3: Table Summary of Comparisons Algorithms for Monthly at 2020.

Algorithm	R^2	MSE	Predict
Random Forest Regressor	0.27	30802509.78	3550.41
Gradient Boosting	0.36	26776834.26	4125.21
Decision Tree Regressor	0.21	33464574.74	1258.37
Bagging Regressor	0.27	30626762.15	3643.14
Extra Trees Regressor	-0.13	48200125.30	1258.37

Result For Monthly Predict 2020:

In this test we implemented for three building and used the dataset shown in screenshot at figure 4.20 (total dataset 950 started from date Sept-2020 until Dec-2020) and we selected

random costumer b1 fan at date Sept-2020 the predict are 1258.37 Watt. hour and the real value are 13350.48 Watt. Hour.

From figure 4.26, showing that all models are failed.

From table 4.3 showing the value of R^2 are less than 0.36 for all models and the predicts values far away from the real value.

These results are failed due to insufficient dataset.

4.2.4 Forth Test and Result for Monthly Predict 2021

We took all buildings (Seven buildings operation at 2021), we executed monthly predict by using five algorithms (Decision tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, Bagging Regressor and Extra Tree).

We calculate the avrage of R^2 and MSE results by using algorithm above comparison it and select the best one model using python code in appendix A.3, page #123,#124 and #125.

	A	B	C	D
1	Costumer ID	Year	Month	Energy Consumption
2	b1 colling system	2021	1	16299.255
3	b1 colling system	2021	2	16782.82
4	b1 colling system	2021	3	16787.919
5	b1 colling system	2021	4	17609.837
6	b1 colling system	2021	5	20897.569
7	b1 colling system	2021	6	24493.515
8	b1 colling system	2021	7	29133.784
9	b1 colling system	2021	8	33430.464
10	b1 colling system	2021	9	37325.22
11	b1 colling system	2021	10	39700.255
12	b1 colling system	2021	11	40848.973
13	b1 colling system	2021	12	41233.37
14	b1 elevator L1	2021	1	3190.965
15	b1 elevator L1	2021	2	3571.454
16	b1 elevator L1	2021	3	4035.986
17	b1 elevator L1	2021	4	4448.213
18	b1 elevator L1	2021	5	4834.649
19	b1 elevator L1	2021	6	5330.49
20	b1 elevator L1	2021	7	5796.686
21	b1 elevator L1	2021	8	6259.125
22	b1 elevator L1	2021	9	6778.838
23	b1 elevator L1	2021	10	7300.862
24	b1 elevator L1	2021	11	7794.399
25	b1 elevator L1	2021	12	8298.945
26	b1 fan	2021	1	29569.293
27	b1 fan	2021	2	33709.453
28	b1 fan	2021	3	38468.944

Figure 4.27: Screenshot for Sample of Dataset for all Seven Buildings at 2021.

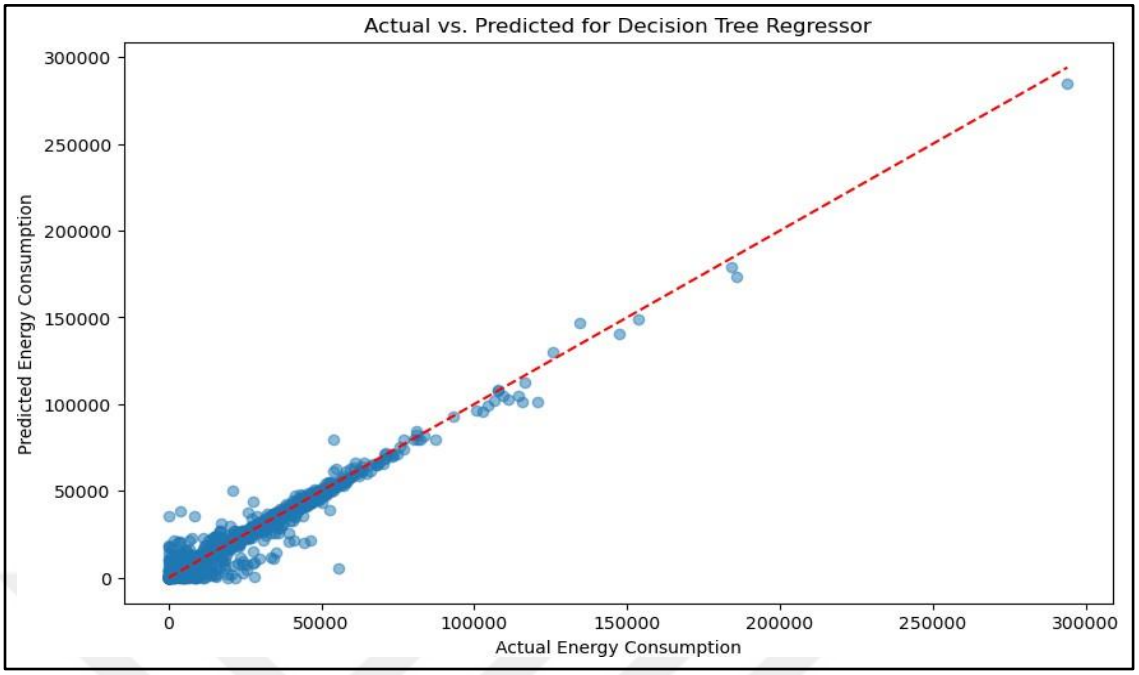


Figure 4.28: Screenshot for Random Forest Regressor with Actual at 2021.

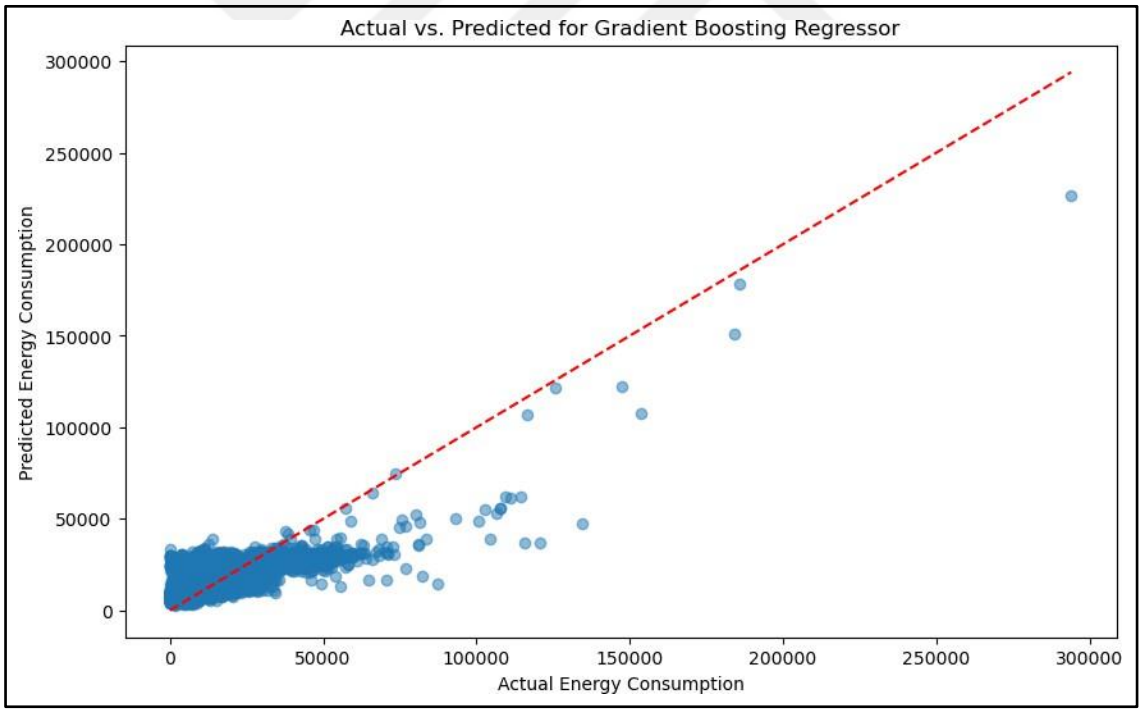


Figure 4.29: Screenshot for Gradient Boosting Regressor with Actual at 2021.

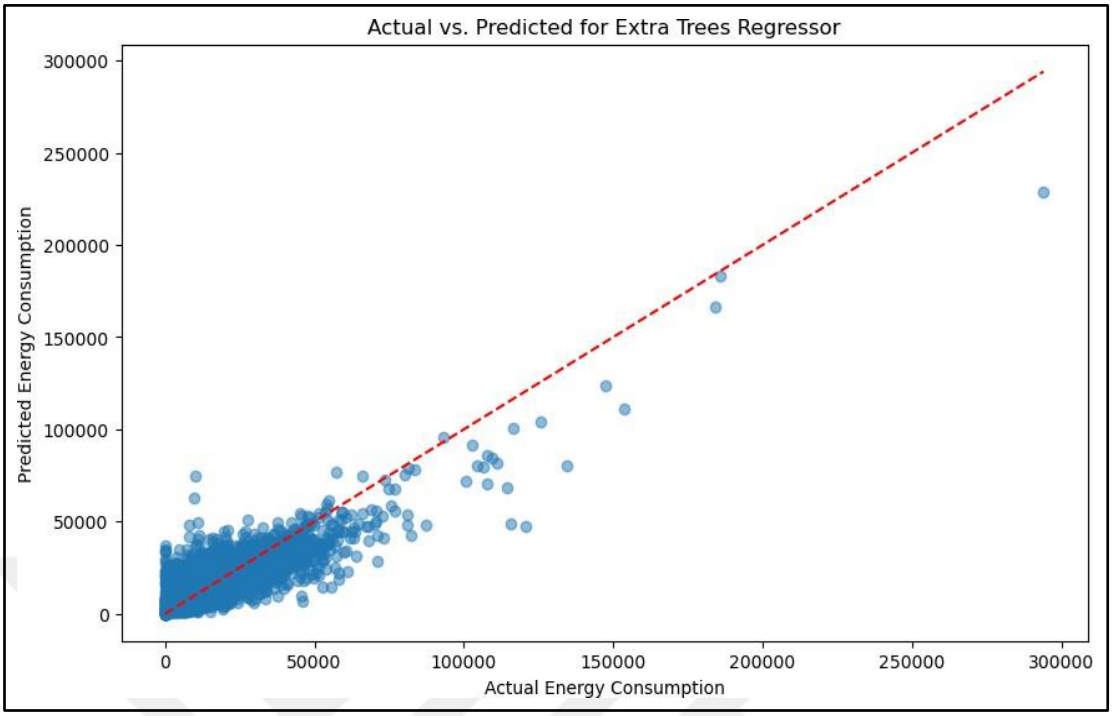


Figure 4.30: Screenshot for Decision Tree Regressor with Actual at 2021.

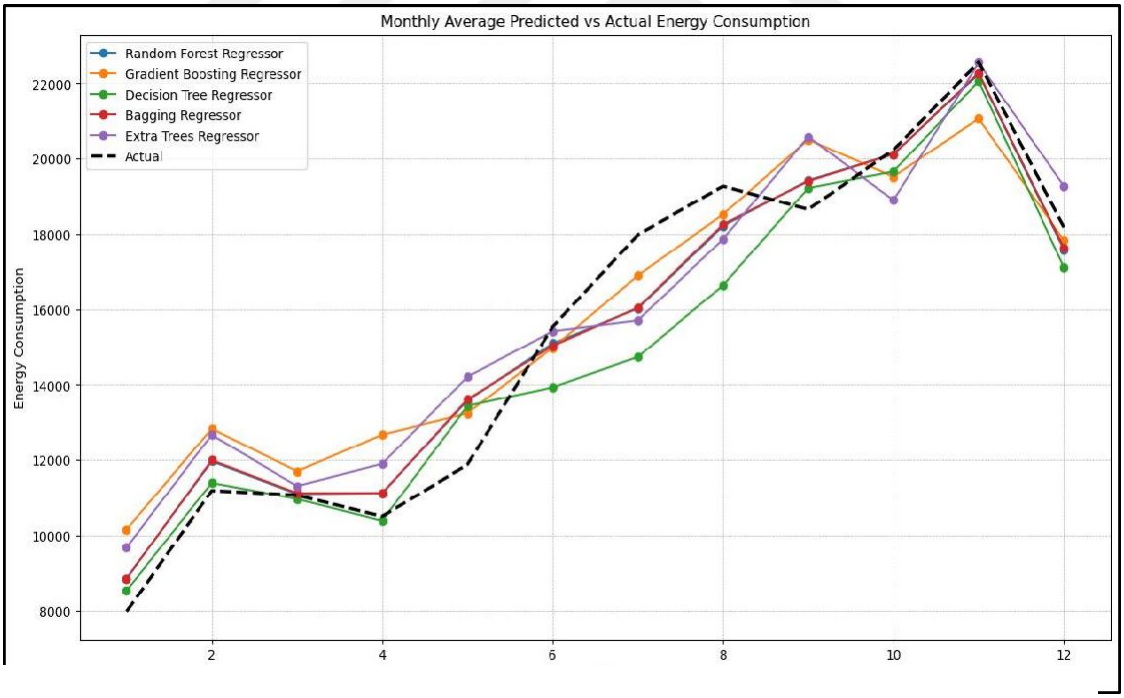


Figure 4.31: Screenshot For Bagging Regressor with Actual at 2021.

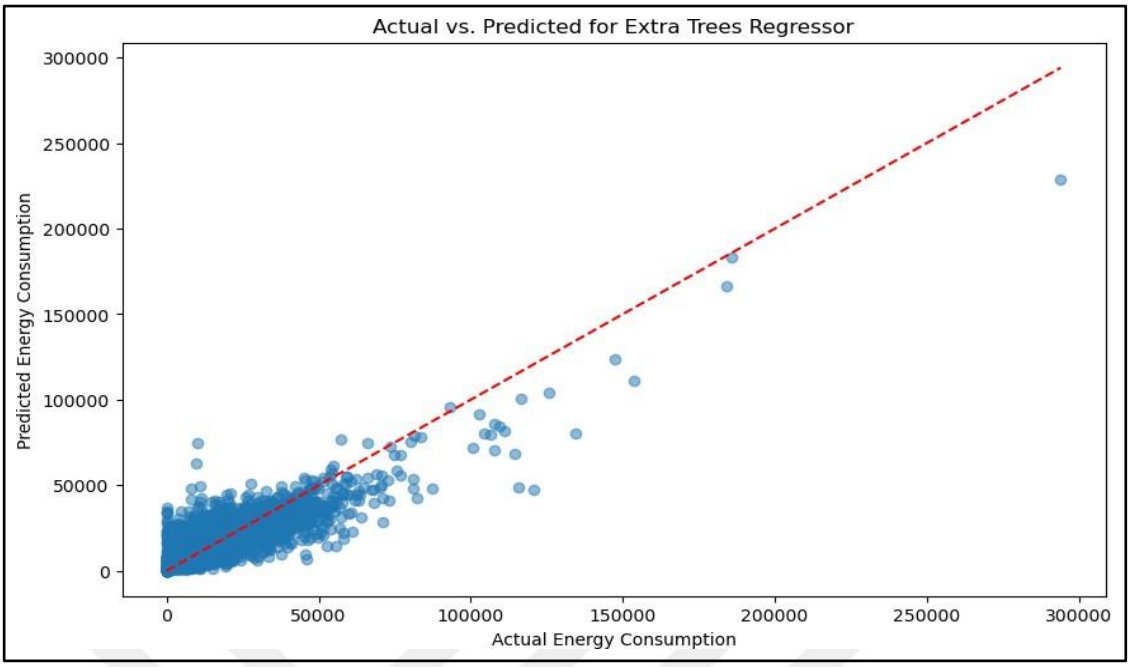


Figure 4.32: Screenshot for Extra Trees Regressor with Actual at 2021.

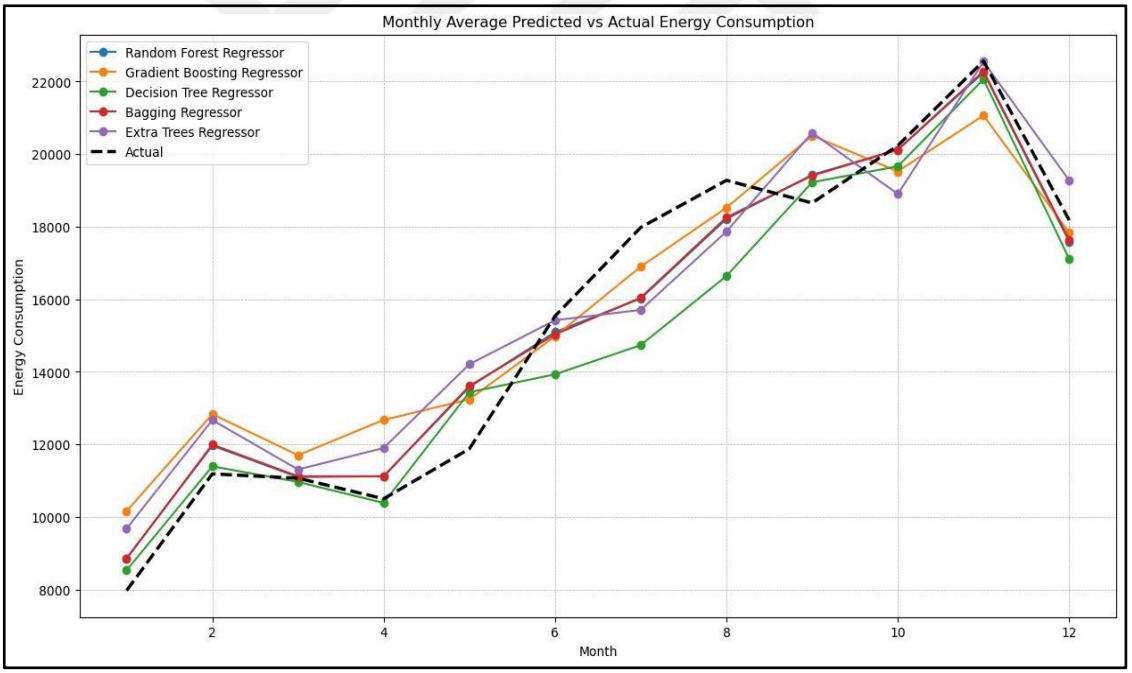


Figure 4.33 Screenshot for Comparisons of ML Algorithms for Monthly Energy Consumption at 2021.

Table 4.4: Table Summary of Comparison Algorithms for Monthly at 2021.

Algorithm	R^2	MSE	Predict
Random Forest Regressor	0.75	43108754.05	33568.48
Gradient Boosting	0.50	86114762.89	8027.91
Decision Tree Regressor	0.77	40306878.8	21925.2
Bagging Regressor	0.75	42342942.20	33700.6
Extra Trees Regressor	0.79	161645243.7	25125.7

Result For Monthly Predict 2021:

In this test we implemented for seven buildings and used the dataset shown in screenshot at figure 4.27 (total dataset 3255 started from date Jan-2021 until Dec-2021) and we selected random costumer b1 fan at date Jan-2021 the predict are 25125.7 Watt. hour and the real value are 29569 Watt. Hour.

We implemented five different models as shown in figure 4.33, showing that the predict in extra trees regressor are nearest model to actual compared with other.

From table 4.4 showing the value of R^2 are less than 0.79 for four model are suitable except one failed.

From result of monthly predict of 2021, seems datasets encouraged us to implement same algorithms for next prediction year.

4.2.5 Fifth Test and Result for Monthly Predict 2022

We took all buildings (Ten buildings operation at 2022), we executed monthly predict by using five algorithms (Decision tree Regressor, Random forest Regressor, Gradient Boosting Regressor, Bagging Regressor and Extra Tree).

We calculate the average of R2 and MSE results by using algorithm above comparison it and select the best one model using python code in appendix A.3, page #123, #124 and #125.

	A	B	C	D
1	Costumer ID	Year	Month	Energy Consumption
2	b1 colling system	2022	1	41326.998
3	b1 colling system	2022	2	41501.323
4	b1 colling system	2022	3	42141.757
5	b1 colling system	2022	4	44468.606
6	b1 colling system	2022	5	47315.796
7	b1 colling system	2022	6	50754.218
8	b1 colling system	2022	7	54521.806
9	b1 colling system	2022	8	58921.215
10	b1 colling system	2022	9	62091.531
11	b1 colling system	2022	10	64085.97
12	b1 colling system	2022	11	64761.649
13	b1 colling system	2022	12	65233.746
14	b1 elevator L1	2022	1	8790.978
15	b1 elevator L1	2022	2	9299.296
16	b1 elevator L1	2022	3	9824.746
17	b1 elevator L1	2022	4	10266.168
18	b1 elevator L1	2022	5	10723.982
19	b1 elevator L1	2022	6	11231.438
20	b1 elevator L1	2022	8	12162.904
21	b1 elevator L1	2022	9	12626.887
22	b1 elevator L1	2022	10	13077.745
23	b1 elevator L1	2022	11	13521.198
24	b1 elevator L1	2022	12	14009.586
25	b1 fan	2022	1	80945.661
26	b1 fan	2022	2	82654.039
27	b1 fan	2022	3	83662.719
28	b1 fan	2022	4	87845.144

Figure 4.34: Screenshot for Sample of Dataset for all Ten Buildings at 2022.

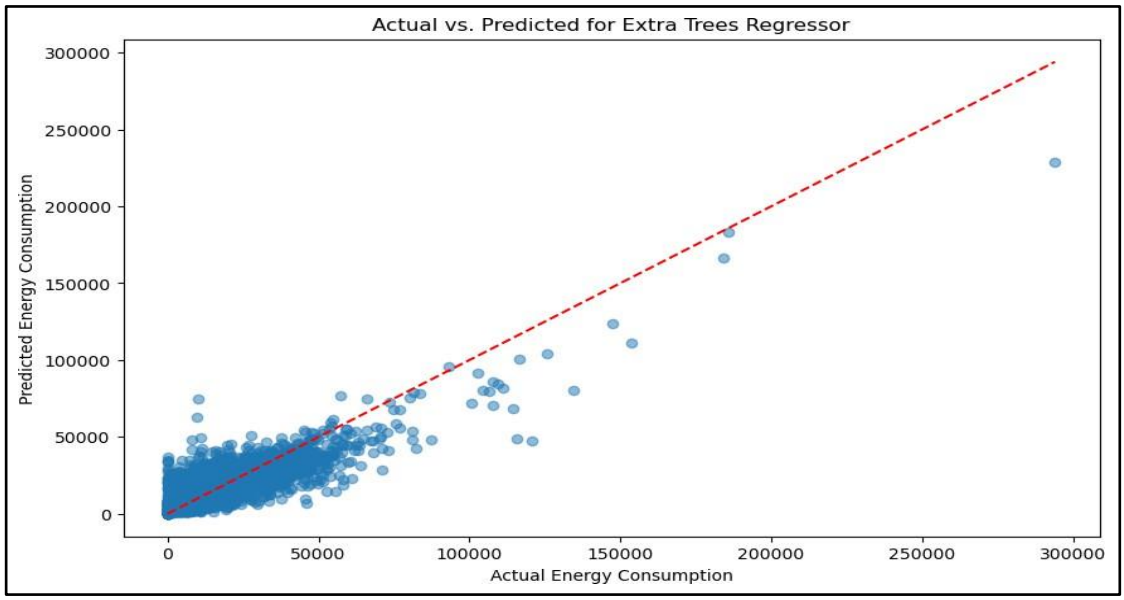


Figure 4.35: Screenshot for Random Forest Regressor with Actual at 2022.

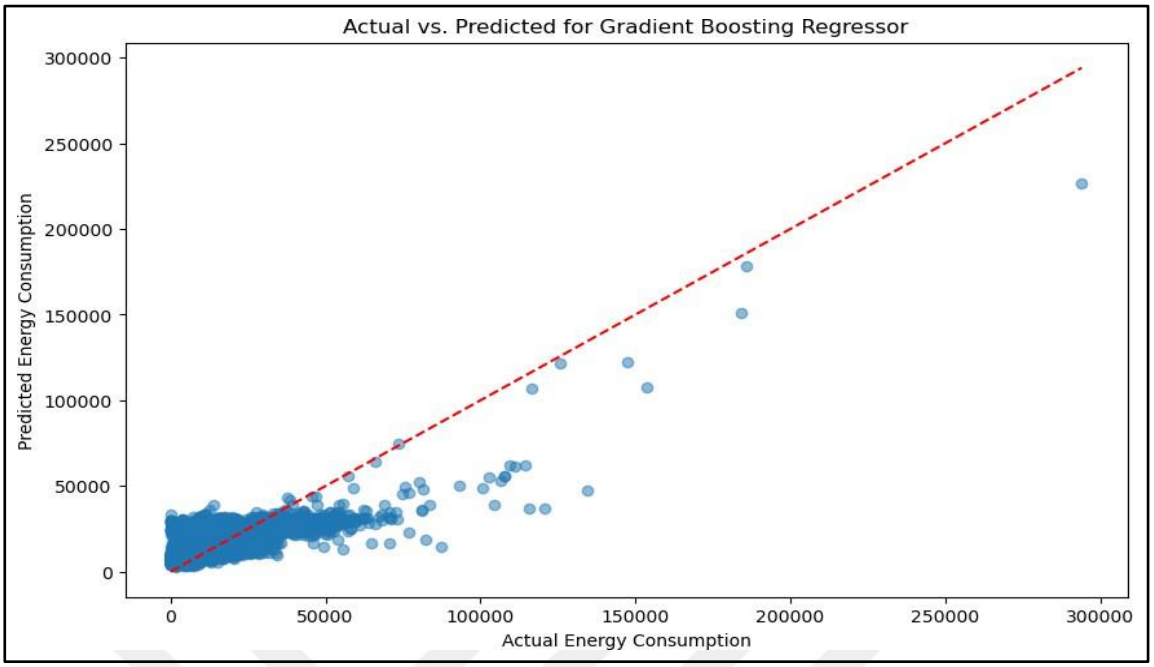


Figure 4.36: Screenshot for Gradient Boosting Regressor with Actual at 2022.

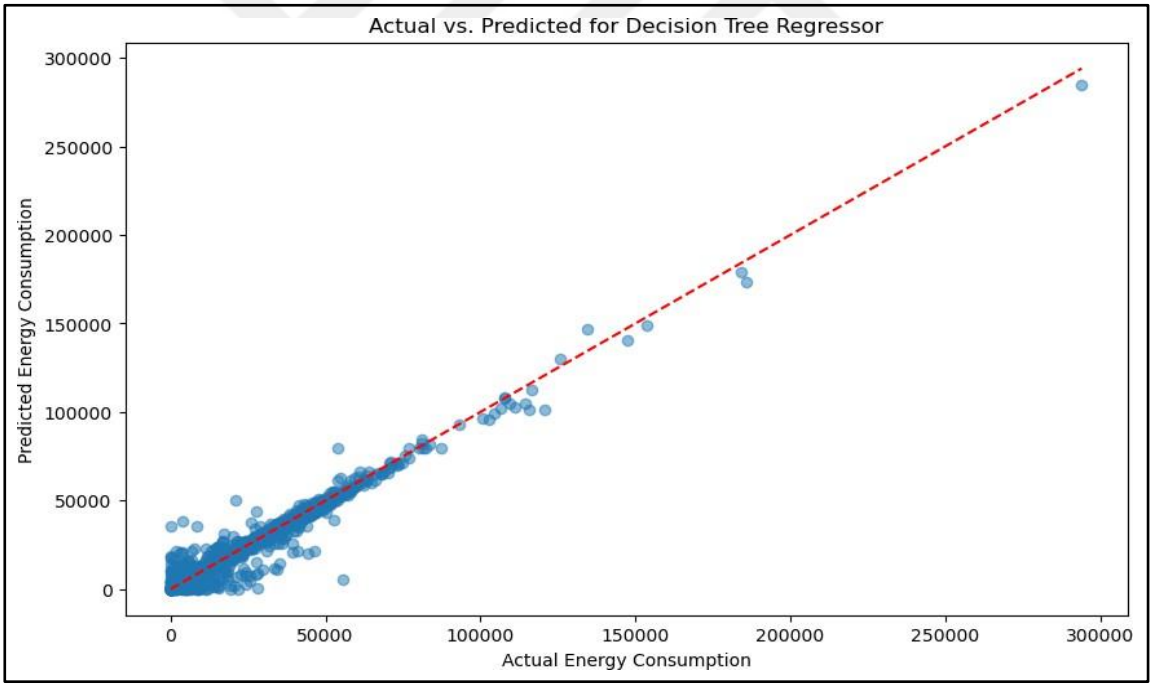


Figure 4.37: Screenshot for Decision Tree Regressor with Actual at 2022.

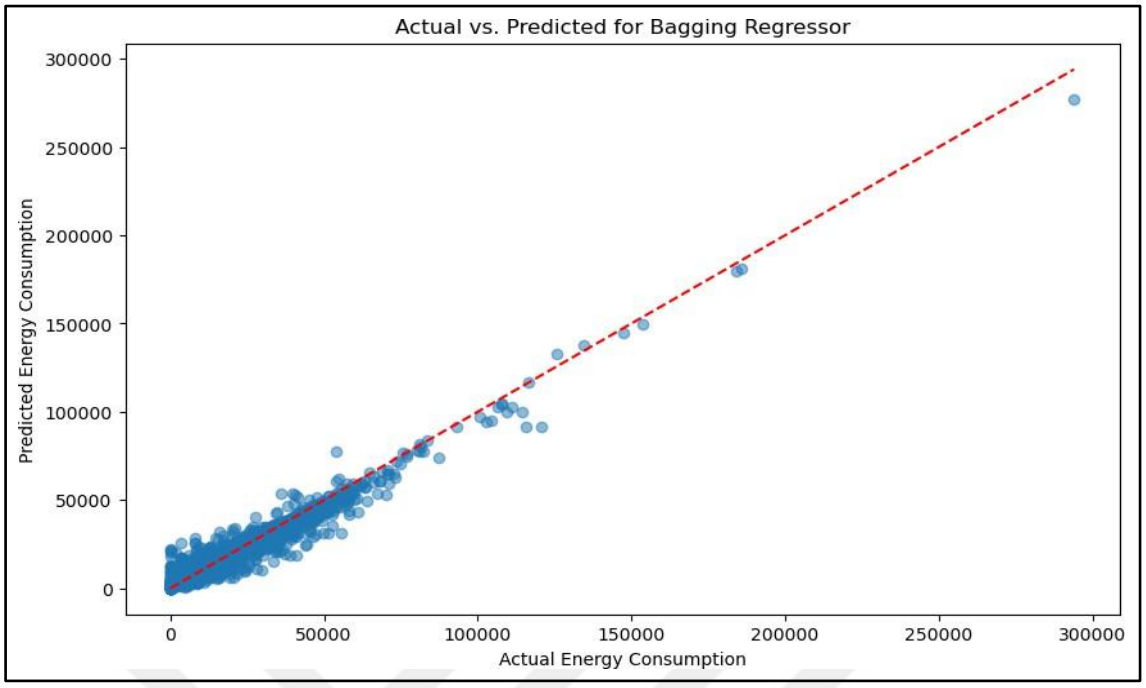


Figure 4.38: Screenshot for Bagging Regressor with Actual at 2022.

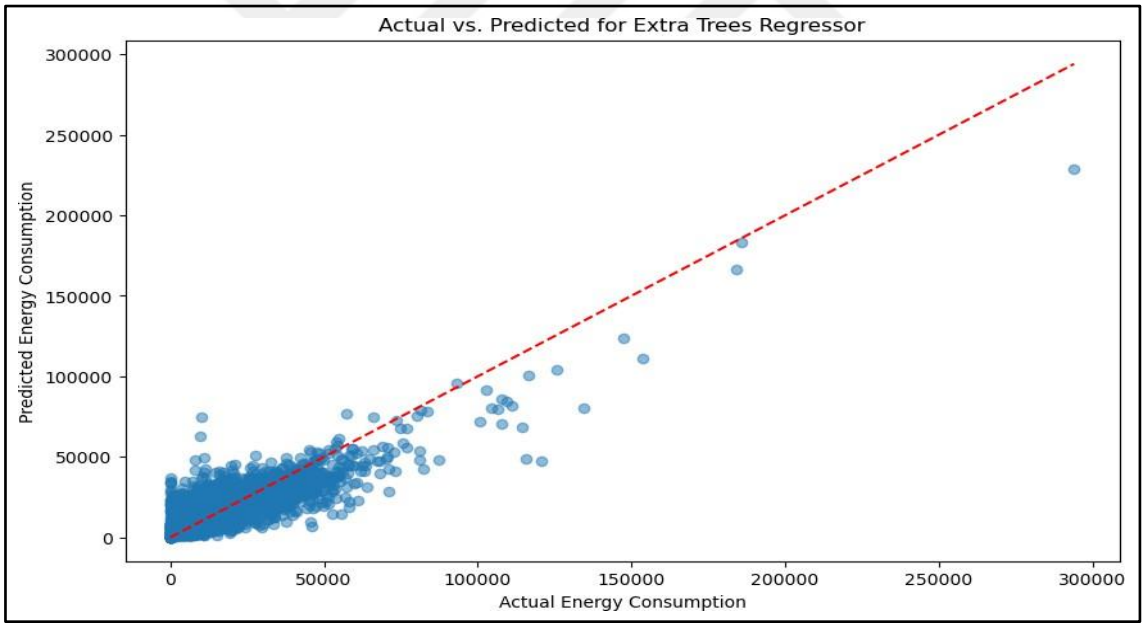


Figure 4.39: Screenshot for Extra Trees Regressor with Actual at 2022.

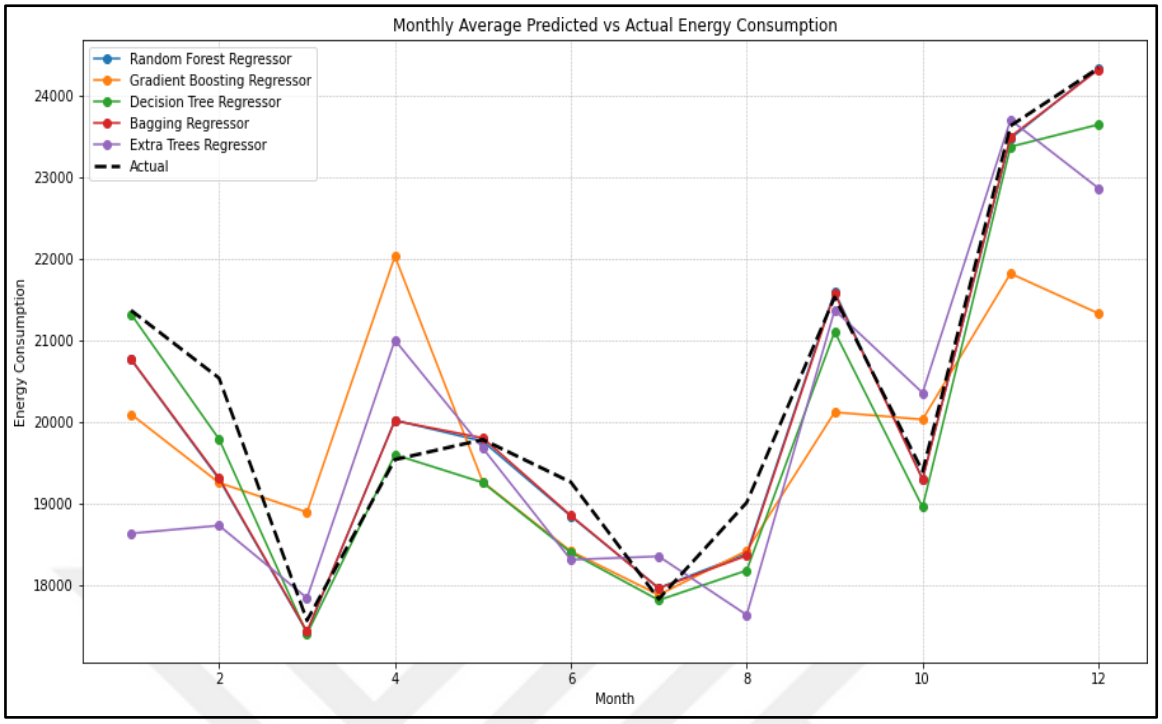


Figure 4.40: Screenshot for Comparisons of ML Algorithms for Monthly Energy Consumption at 2022.

Table 4.5: Table Summary of comparisons algorithms for monthly at 2022.

Algorithm	R^2	MSE	Predict
Random Forest Regressor	0.93	25580195.4	117023.78
Gradient Boosting	0.58	17376335.2	106927.7
Decision Tree Regressor	0.95	19591039.6	112838.25
Bagging Regressor	0.94	25578542.5	117023.78
Extra Trees Regressor	0.70	125279781.5	112838.25

Result For Monthly Predict 2022:

In this test we implemented for ten buildings and used the dataset shown in screenshot at figure 4.34 (total dataset 9401 started from date Jan-2022 until Dec-2022) and we selected

random b1 fan at date Jan-2022 the predict are 112838.25 Watt. hour and the real value are 112311.4 Watt. Hour.

From figure 4.40, showing that four models almost fitted between real and predict expect Gradient Boosting model.

From table 4.5 showing the value of R^2 is higher than 0.70 for four models expect Gradient Boosting model.

From result of monthly predict of 2022, seems datasets encouraged us to implement same algorithms for year 2023 prediction.

4.2.6 Sixth Test and Results for Monthly Predict 2023

We took all buildings (Ten buildings operation at 2023), we executed monthly predict by using five algorithms (Decision tree Regressor, Random forest Regressor, Gradient Boosting Regressor, Bagging Regressor and Extra Tree).

We calculate the average of R2 and MSE results by using algorithm above comparison it and select the best one model using python code in appendix A.3, page #123, #124 and #125.

	A	B	C	D
1	Costumer ID	Year	Month	Energy Consumption
2	b1 colling system	2023	1	65681.201
3	b1 colling system	2023	2	66128.14
4	b1 colling system	2023	3	66862.062
5	b1 colling system	2023	4	67679.108
6	b1 colling system	2023	5	69072.408
7	b1 colling system	2023	6	71639.193
8	b1 colling system	2023	7	74481.965
9	b1 colling system	2023	8	78692.933
10	b1 elevator L1	2023	1	14469.283
11	b1 elevator L1	2023	2	14856.376
12	b1 elevator L1	2023	3	15290.829
13	b1 elevator L1	2023	5	16156.153
14	b1 elevator L1	2023	7	17034.122
15	b1 elevator L1	2023	8	17487.619
16	b1 fan	2023	1	125643.073
17	b1 fan	2023	2	129768.345
18	b1 fan	2023	3	133554.47
19	b1 fan	2023	4	137170.737
20	b1 fan	2023	5	141008.519
21	b1 fan	2023	7	148386.216
22	b1 fan	2023	8	152111.82
23	b1.f1.f3	2023	1	37434.773
24	b1.f1.f3	2023	2	38637.831
25	b1.f1.f3	2023	3	39385.774
26	b1.f1.f3	2023	4	40053.429
27	b1.f1.f3	2023	5	41151.561
28	b1.f1.f3	2023	6	42486.139

Figure 4.41: Screenshot for Sample of Dataset for all Ten Buildings at 2023.

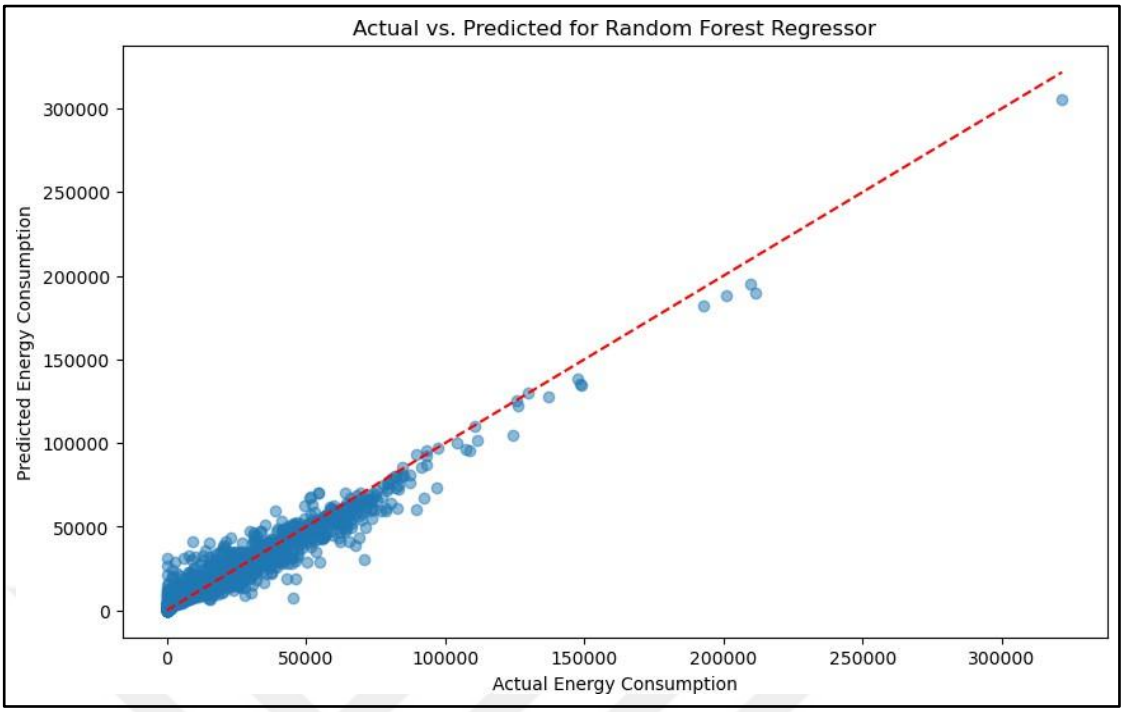


Figure 4.42: Screenshot for Random Forest Regressor with Actual at 2023.

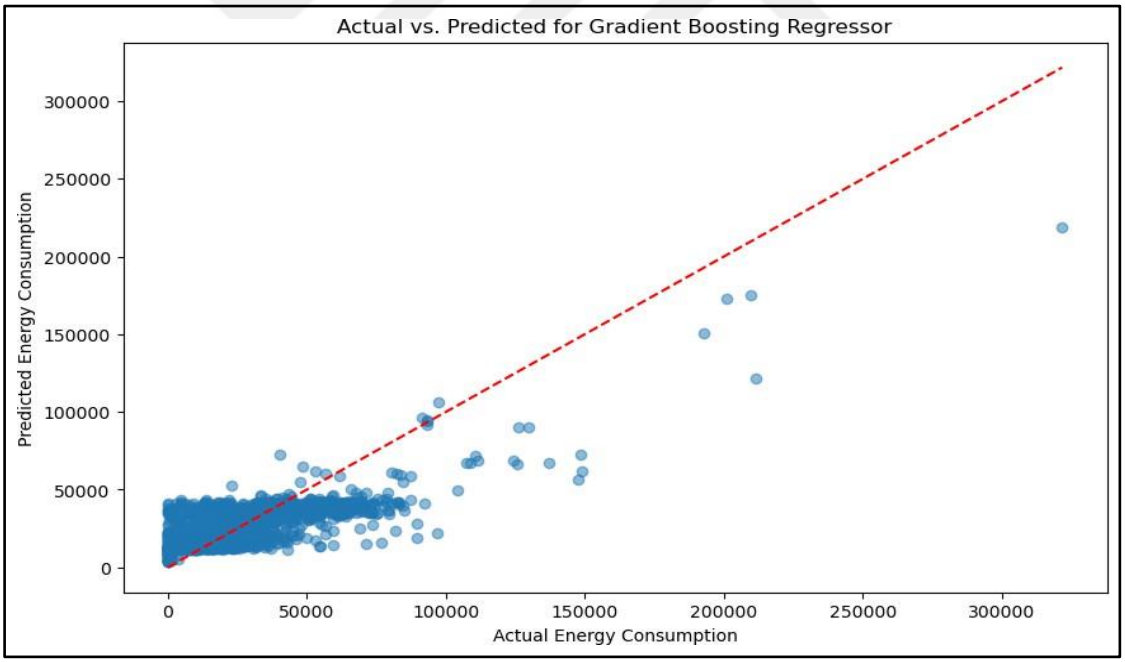


Figure 4.43: Screenshot for Gradient Boosting Regressor with Actual at 2023.

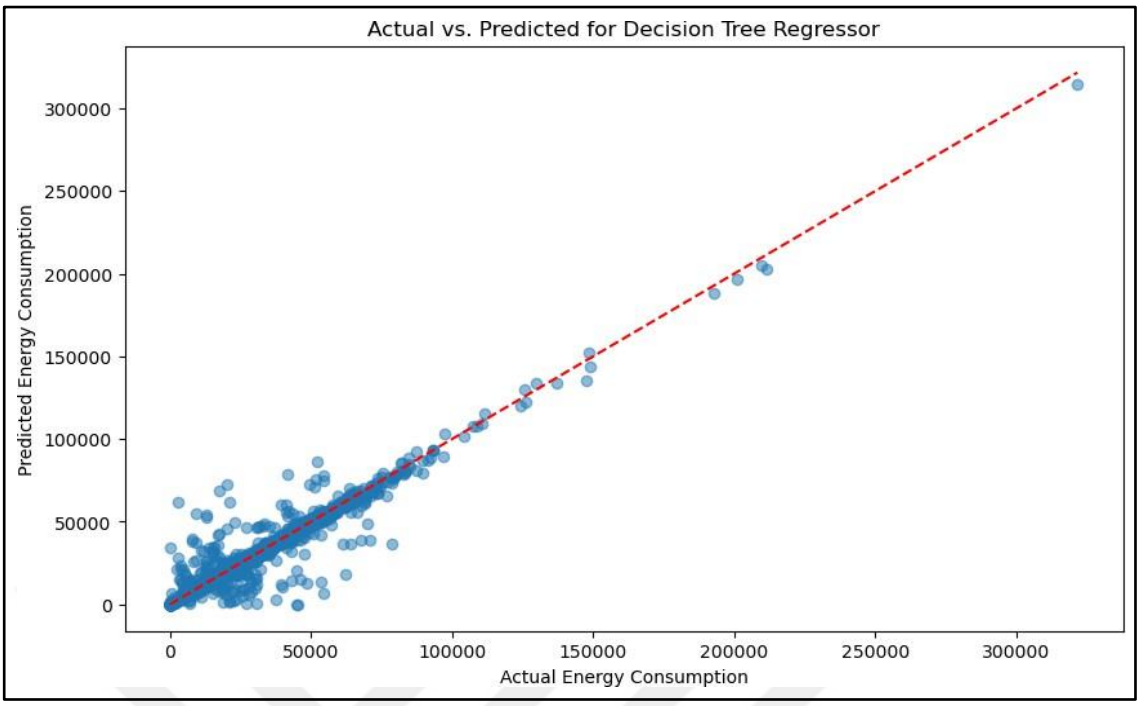


Figure 4.44: Screenshot for Decision Tree Regressor with Actual at 2023.

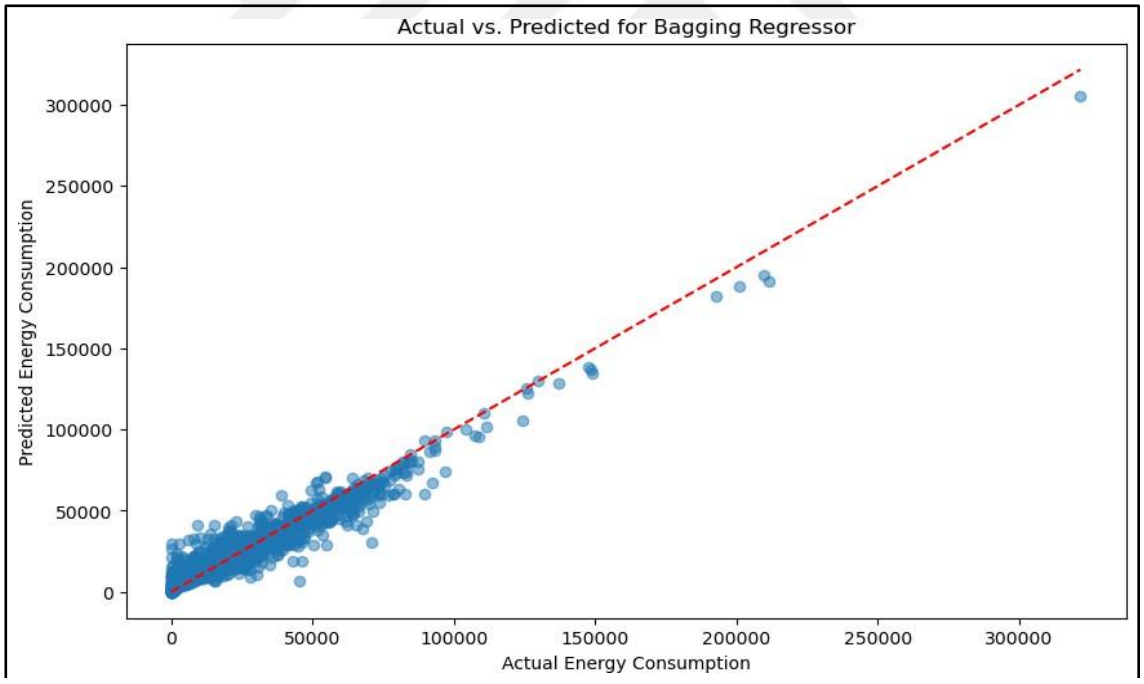


Figure 4.45: Screenshot for Bagging Regressor with Actual at 2023.

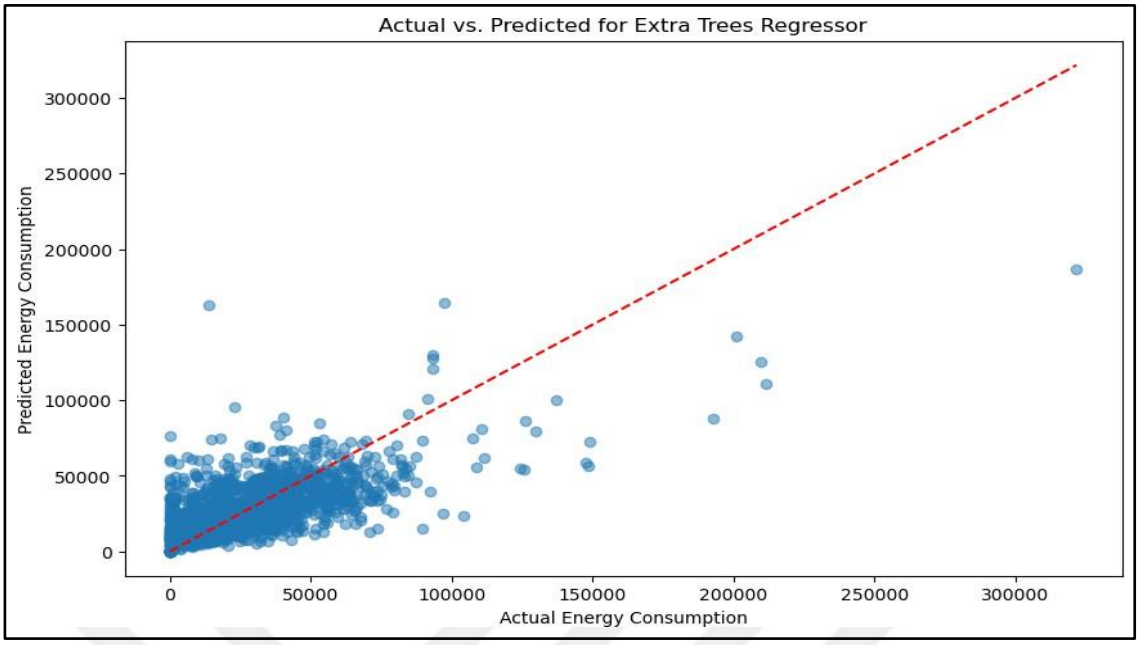


Figure 4.46: Screenshot for Extra Trees Regressor with Actual at 2023.

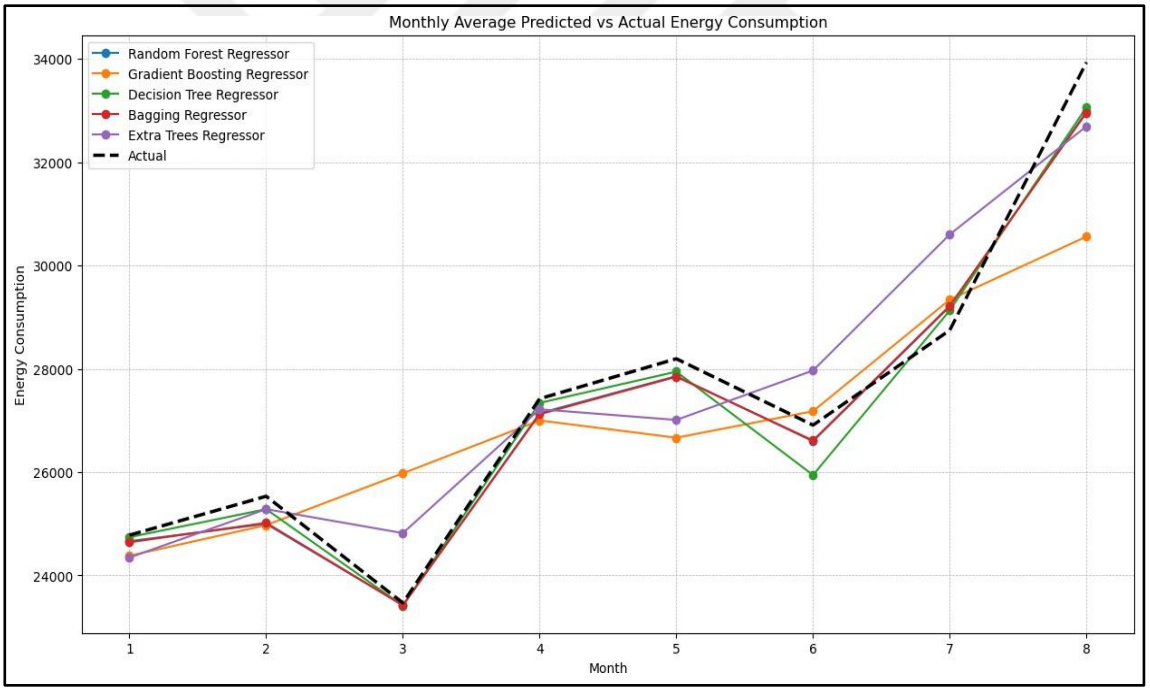


Figure 4.47: Screenshot for Comparisons of ML Algorithms for Monthly Energy Consumption at 2023.

Table 4.6: Table Summary of Comparisons Algorithms for Monthly at 2023.

Algorithm	R^2	MSE	Predict
Random Forest Regressor	0.92	45925294.6	166112.9
Gradient Boosting	0.58	281107172.3	146389.5
Decision Tree Regressor	0.92	49634563.5	166027.2
Bagging Regressor	0.93	45882398.1	166112.9
Extra Trees Regressor	0.43	354108271.3	166027.2

Result For Monthly Predict 2023:

In this test we implemented for ten buildings and used the dataset shown in screenshot at figure 4.41 (total dataset 7775 started from date Jan-2023 until Sept-2023) and we selected random b1 fan at date Jan-2023 the predict are 166027.2 Watt. hour and the real value are 166561.15 Watt. Hour.

From figure 4.47, showing that three models almost fitted between real and predict expect Gradient Boosting and Extra trees models.

From table 4.6 showing the value of R^2 is higher than 0.92 for three models expect Gradient Boosting and Extra trees models.

After those six of tests and results, we have clear idea about the best models and we will be mentioned that in comparison.

4.3 TOTAL RESULTS AND COMPARISION

In this section, we present the findings and comparisons of energy consumption projections using various machine learning algorithms for the years 2020, 2021, 2022, and 2023. These projections cover different timeframes, including daily, weekly, monthly, and yearly patterns. The primary aim of these predictions is to verify the algorithm's performance by forecasting the next status, ensuring that it aligns closely with the existing dataset. To evaluate the precision of these predictions, we employ two essential performance measures: R^2 (R-Square error), Root Mean Square Error (RMSE) , Mean Square Error (MSE) [53] and Gross Vailated .

4.3.1 RESULTS

In this part of the article, we were discussed the outcomes of our experiment in which we predicted future energy usage by employing a variety of machine learning algorithms for

the years 2020, 2021, 2022, and 2023. R^2 (R-Square error), Root Mean Square Error (RMSE), Mean Square Error and Gross Vailated are the four major metrics that are used in the process of evaluating the performance of any method. The accuracy and precision of the predictions that the models provide may be gleaned from the information provided by these measures.

The findings shed light on how each algorithm has fared throughout the years. In terms of both of R^2, RMSE, MSE and Gross Vailated it is clear that the Linear Regression method produces respectable outcomes on a regular basis [54]. This shows that the Linear Regression model is capable of accurately predicting future energy use and identifying underlying trends in the data for daily and weekly [55].

The Random Forest's performance and Bagging Regressor method produces respectable outcomes on a regular basis. This shows that The Random Forest's performance and Bagging Regressor models are capable of accurately predicting future energy use and identifying underlying trends in the data for monthly.

The Extra trees Regressor algorithm's fluctuating performance over time may be an indication that it is highly sensitive to variations in input data [56]. Similar to other algorithms, SVR and Gradient Boosting Regressors varies from year to year.

In sum, we may use these findings to better pick algorithms, develop our models, and identify promising avenues for future research. In the next part, we'll analyze the data and offer conclusions based on our findings.

4.3.2 COMPARISON

4.3.2.1 Daily comparison

a. Linear Regression exhibited an MSE of 182767. The R-squared (R2) value was 0.98, indicating optimal fit.
b. Decision Tree had an MSE of 142.46. The R2 value was 0.99, indicating optimal fit.
c. Random Forest showed an MSE of 121.48. The R2 value was 0.99, indicating optimal fit.
d. Gradient Boosting showed an MSE of 534.1. The R2 value was 0.99, indicating optimal fit.
e. Ridge showed an MSE of 182778. The R2 value was 0.98, indicating optimal fit.
f. Lasso showed an MSE of 182776. The R2 value was 0.98, indicating optimal fit.

g. SVR showed an MSE of 13740. The R2 value was -0.382, indicating a relatively weak fit.

4.3.2.2 Weekly comparison

a. Linear Regression had MSE of 56.77. The R2 value was 0.99, indicating optimal fit.
b. Decision Tree showed MSE of 1360. The R2 value was -0.43, indicating a relatively weak fit.
c. Random Forest exhibited MSE of 1619.6. The R2 value was -0.61, indicating a relatively weak fit.
d. Gradient Boosting showed MSE of 1360. The R2 value was -0.43, indicating a relatively weak fit.
e. SVR showed MSE of 577. The R2 value was -0.58, indicating a relatively weak fit.
f. Lasso showed MSE of 93.5. The R2 value was -0.15, indicating a relatively weak fit.
g. Ridge showed MSE of 250. The R2 value was -0.19, indicating a relatively weak fit.
h. Elastic Net showed MSE of 45.35. The R2 value was 0.55, indicating a relatively meddle fit.
i. K-Nearest Neighbors showed MSE of 2233.7. The R2 value was 0.99, indicating optimal fit.
j. Neural Network showed MSE of 1424.7. The R2 value was -0.37, indicating a relatively weak fit.

4.3.2.3 Monthly comparison

For 2020:

a. Bagging yielded MSE of 3062672. The R2 value was 0.27, suggesting a poor fit.
b. Decision Tree had MSE of 33464574. The R2 value was 0.21, suggesting a poor fit.
c. Random Forest showed MSE of 30892509. The R2 value was 0.27, suggesting a poor fit.
d. Gradient Boosting showed MSE of 26776834. The R2 value was 0.36, suggesting a poor fit.
e. Extra Trees showed MSE of 48820125. The R2 value was -0.13, suggesting a poor fit.

For 2021:

f. Bagging yielded MSE of 4234294.2. The R2 value was 0.75, indicating a relatively meddle fit.
g. Decision Tree had MSE of 40306878. The R2 value was 0.77, indicating a relatively meddle fit.
h. Random Forest showed MSE of 43108754.05. The R2 value was 0.75, indicating a relatively meddle fit.
i. Gradient Boosting showed MSE of 86114762. The R2 value was 0.50, indicating a relatively meddle fit.
j. Extra Trees showed MSE of 1616452. The R2 value was 0.79, indicating a relatively meddle fit.

For 2022:

k. Bagging yielded MSE of 25578542. The R2 value was 0.93, indicating optimal fit.
l. Decision Tree had MSE of 19591039. The R2 value was 0.95, indicating optimal fit.
m. Random Forest showed MSE of 25580195. The R2 value was 0.93, indicating optimal fit.
k. Gradient Boosting showed MSE of 175376335. The R2 value was 0.58, suggesting a poor fit.
l. Extra Trees showed MSE of 12527978. The R2 value was 0.70, indicating a relatively meddle fit.

For 2023:

n. Bagging yielded MSE of 45882398. The R2 value was 0.93, indicating optimal fit.
o. Decision Tree had MSE of 49634563. The R2 value was 0.92, indicating optimal fit.
p. Random Forest showed MSE of 45925294.6. The R2 value was 0.92, indicating optimal fit.
m. Gradient Boosting showed MSE of 281107172.3. The R2 value was 0.58, suggesting a relatively meddle fit.
n. Extra Trees showed MSE of 354108271.3. The R2 value was 0.43, suggesting a poor fit.

4.3.3 OVERALL COMPARISON

Over the course of the past four years, Decision Tree Regressor has demonstrated consistently competitive performance with MSE values that are on average rather low and R2 highest values. Both the Bagging and Random Forest algorithms experienced various

degrees of success throughout the years, with the Bagging method sometimes achieving greater success than Random Forest did [57]. The variance in performance might be ascribed to a number of factors, including the qualities of the dataset, the importance of the features, and the sensitivity of the algorithm.

While in daily and weekly Linear Regressor has demonstrated consistently competitive performance with MSE values that are on average rather low and R2 highest values.

In conclusion, in light of the findings and the comparison, Decision Tree Regressor demonstrated a performance that was generally consistent and accurate in estimating energy consumption over the different years. As a consequence, it is an excellent contender for this particular application. It is essential to keep in mind, however, that the selection of the algorithm may also be influenced by other considerations, such as the interpretability, computing efficiency, and scalability of the solution.

Here is a comparison of percentage predictions of the successful tests results at table 4.7 below.

Overall, the findings suggest that the Decision Tree model consistently outperformed the other models, showcasing its effectiveness in predicting both daily and monthly energy consumption patterns with costumers. The Bagging model also demonstrated competitive performance, while the Linear Regression model displayed moderate accuracy across both prediction intervals for daily and weekly [58].

In future stages might comprise further refining of the selected algorithm and the investigation of new methods that have the potential to improve the forecast accuracy for the energy consumption estimation similar to others residential complex buildings.

Table 4.7: Comparisons of Percentage Predictions of the Successful Tests Results.

Test Type	Linear Regressor	Random Forest	Decision Tree	Bagging Regressor
Daily	98%	99%	99%	99%
Weekly	99%	N/A	N/A	N/A
Monthly 2021	N/A	75%	77%	75%
Monthly 2022	N/A	93%	95%	94%
Monthly 2023	N/A	92%	92%	93%

5. CONCLUSION

Using machine learning methods, the purpose of this research project was to produce an accurate and trustworthy model for calculating the amount of electricity consumed by a residential complex building in Iraq. The technique consisted of collecting the data, preparing the data, selecting the algorithm to use, training the model, and analyzing the results. In the course of this in-depth procedure, useful insights on the efficacy of several algorithms in estimating energy usage were gathered.

During the period of data collection, time-stamped values of energy use were gathered from smart meters positioned at various locations around the facility. Concerns about privacy have been addressed by the anonymization of data and the storing of that data securely. To improve the quality of the dataset and better understand the elements that influence energy use, other data sources, such as weather information and building characteristics, were integrated.

The preparation of the data was an extremely important step in assuring the dataset's high quality. In order to get the data ready for training, several different methods, including data cleaning, outlier identification, data transformation, feature engineering, and data encoding, were utilized. Following the preprocessing of the dataset, it was partitioned into training and testing sets, and cross-validation methods were applied in order to validate the models and prevent overfitting.

The decision of which algorithms to use was an essential part of the process, and some of the algorithms that were taken into consideration for the job were linear regression, decision trees, random forest, Bagging regressor. These algorithms were selected because of their applicability to regression issues and their capacity to comprehend intricate connections within the data.

We performed an in-depth analysis and comparison of the outcomes that were achieved by applying the chosen algorithms to the dataset including data from the years 2020 to 2023. In order to evaluate the precision of the predictions, the $R^{2,}$ RMSE and MSE metrics were applied. The data showed that Decision Tree continually exhibited competitive performance across the years, showing its consistency and usefulness for this application. This was proved by the fact that it had the best overall performance.

In conclusion, the goals of this research project were effectively accomplished through the development of a model based on machine learning for predicting the amount of electricity consumed in a structure that has many residential units. The research highlighted the need of precise data pretreatment, careful consideration of method choice, and careful attention paid to result analysis. The Decision Tree Algorithm has been shown to be an accurate method for estimating energy consumption, showing its potential for real-world use in improving energy management and consumption predictions in comparable settings.

It is possible that as the area of machine learning continues to advance, more tweaks and improvements to the model might be investigated. If successful, this could result in even more accurate predictions and insights. This research lays the foundation for future studies in energy consumption estimation and opens avenues for the integration of advanced techniques to address the challenges of sustainable energy usage in residential complexes. Additionally, this research lays the foundation for future studies in energy consumption estimation.

REFERENCES

[1] Atanassov, A., Todorov, V., Iliev, A., & Doukovska, L. (2019). Energy Consumption Prediction in Residential Buildings Using Machine Learning. Proceedings of the International Conference on Embedded Systems and Applications, 183-190.

[2] Hussain, S., Javaid, N., Alrajeh, N., Alamri, A., & Guizani, N. (2020). Energy Consumption Prediction in Commercial Buildings Using Machine Learning. Sensors, 20(3), 830.

[3] International Energy Agency. (2019). Iraq Energy Outlook 2019. Retrieved from https://www.iea.org/reports/iraq-energy-outlook-2019

[4] Ministry of Electricity, Iraq. (2018). Annual Statistical Report for 2018. Baghdad: Government Printing Office.

[5] Atanassov, K., Mehandjiev, N., & Ivanova, A. (2019). Machine Learning for Residential Energy Consumption Prediction. In Proceedings of the International Conference on Big Data and Machine Learning (pp. 143-152).

[6] Khan, S. U., & Wu, J. (2018). A Survey on the Technologies for Smart Metering and Smart Grid. Renewable and Sustainable Energy Reviews, 82, 3029-3038.

[7] Sutharshan, S., & Jirutitijaroen, P. (2018). A Comparative Study of Supervised Learning Algorithms for Residential Energy Consumption Prediction. In Proceedings of the 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 1172-1176).

[8] Khader, M., & Al-Naymat, G. (2020). Density-based algorithms for big data clustering using mapreduce framework. *ACM Computing Surveys*, *53*(5), 1–38. https://doi.org/10.1145/3403951

[9] Patel, D., & Srinivasan, D. (2019). Energy Consumption Prediction in Residential Buildings Using Machine Learning: A Case Study in a Smart City. Procedia Computer Science, 152, 214-221.

[10] Cantín, M. (2018). World Medical Association Declaration of helsinki: Ethical principles for medical research involving human subjects. reviewing the latest version. *International Journal of Medical and Surgical Sciences*, *1*(4), 339–346. https://doi.org/10.32457/ijmss.2014.042

[11] European Commission. (2018). General Data Protection Regulation (GDPR). Retrieved from https://gdpr.eu/.

[12] National Institutes of Health. (2018). Protecting Human Research Participants. Retrieved from https://phrp.nihtraining.com/users/login.php.

[13] World Health Organization. (n.d.). *Standards and operational guidance for Ethics Review of health-related research with human participants*. World Health Organization. https://www.who.int/publications-detail-redirect/9789241502948

[14] Abera, F. Z., & Khedkar, V. (2020). Machine Learning Approach Electric appliance consumption and peak demand forecasting of residential customers using Smart Meter Data. *Wireless Personal Communications*, *111*(1), 65–82. https://doi.org/10.1007/s11277-019-06845-6

[15] Bandyopadhyay, A., & Bhattacharya, A. (2021). Residential appliance usage patterns from overall energy consumption data: A statistical machine learning approach. *Volume 8A: Energy*.https://doi.org/10.1115/imece2021-70122

[16] Gao, Y., Schay, A., & Hou, D. (2018). Occupancy detection in smart housing using both aggregated and appliance-specific power consumption data. *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. https://doi.org/10.1109/icmla.2018.00210

[17] Constantin, D., & Clipici, E. (2017). A New Model for Estimating the Risk of Bankruptcy of the Insurance Companies Based on the Artificial Neural Networks. 17th International Multidisciplinary Scientific GeoConference SGEM 2017, 21, 85-94. doi: 10.5593/ sgem2017/21/S07.012.

[18] Ignatiadis, D., Henri, G., & Rajagopal, R. (2019). Forecasting residential monthly electricity consumption using smart meter data. *2019 North American Power Symposium (NAPS)*. https://doi.org/10.1109/naps46351.2019.9000285

[19] Kaligambe, A., Fujita, G., & Keisuke, T. (2022). Estimation of unmeasured room temperature, relative humidity, and CO2 concentrations for a smart building using machine learning and Exploratory Data Analysis. *Energies*, *15*(12), 4213. https://doi.org/10.3390/en15124213

[20] Kim, K., Ohsugi, S., & Koshizuka, N. (2021). Machine learning model for frailty Detectxion using electric power consumption data from smart meter. *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*. https://doi.org/10.1109/dsaa53316.2021.9564127

[21] Matijasevic, T., Antic, T., & Capuder, T. (2022). Machine learning-based forecast of secondary distribution network losses calculated from the Smart Meters Data. *2022 7th International Conference on Smart and Sustainable Technologies (SpliTech)*. https://doi.org/10.23919/splitech55088.2022.9854276

[22] Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F., & Ajayi, S. (2022). Building energy consumption prediction for residential buildings using Deep Learning and other machine learning techniques. *Journal of Building Engineering*, *45*, 103406. https://doi.org/10.1016/j.jobe.2021.103406

[23] Oprea, S.-V., & Bara, A. (2019). Machine learning algorithms for short-term load forecast in residential buildings using smart meters, sensors and Big Data Solutions. *IEEE Access*, *7*, 177874–177889. https://doi.org/10.1109/access.2019.2958383

[24] Ravinder, M., & Kulkarni, V. (2023). Intrusion detection in smart meters data using Machine Learning Algorithms: A research report. *Frontiers in Energy Research*, *11*. https://doi.org/10.3389/fenrg.2023.1147431

[25] Smart Meter App - Web Service. (n.d.). http://185.181.55.92:8081/

[26] Sobrino, E. M., Santiago, A. V., & Gonzalez, A. M. (2019). Forecasting the electricity hourly consumption of residential consumers with smart meters using machine learning algorithms. *2019 IEEE Milan PowerTech*. https://doi.org/10.1109/ptc.2019.8810902

[27] Tang, W., Wang, H., Lee, X.-L., & Yang, H.-T. (2022). Machine Learning Approach to uncovering residential energy consumption patterns based on socioeconomic and smart meter data. *Energy*, *240*, 122500. https://doi.org/10.1016/j.energy.2021.122500

[28] Home. LUNA. (2022, February 8). https://lunatr.com/en/

[29] Zhou, H., Hou, Z., Etingov, P., & Liu, Y. (2019). Machine-learning-based investigation of the associations between residential power consumption and weather conditions. 2019 3rd International Conference on Smart Grid and Smart Cities (ICSGSC). https://doi.org/10.1109/icsgsc.2019.00-13

[30] Asnil, & Elfizon. (2018). *APLIKASI Fuzzy Logic Untuk PENGENDALIAN Motor Compressor Pada Air Conditioner Berbasis ATMEGA 8535*. https://doi.org/10.31227/osf.io/9edpx

[31] Alfonsi, A. (2020). *Adets User Manual*. https://doi.org/10.2172/1770868

[32] Blank, J., & Goldblatt, M. I. (2018). Masters program hernia pathway: Laparoscopic Inguinal Hernia. *The SAGES Manual of Hernia Surgery*, 23–33. https://doi.org/10.1007/978-3-319-78411-3_3

[33] Li, W., Zhao, L., Bo, Y., Wang, W., Wang, M., Liu, S., Liu, R., & Wang, X. (2021). Robust transmission expansion planning model considering multiple uncertainties and active load. *Global Energy Interconnection*, *4*(5), 476–484. https://doi.org/10.1016/j.gloei.2021.11.009

[34] Ebtehaj, I., Bonakdari, H., Gharabaghi, B., & Khelifi, M. (2023). Short-term precipitation forecasting based on the Improved Extreme Learning Machine Technique. *ECWS-7 2023*. https://doi.org/10.3390/ecws-7-14237

[35] Nägeli, C., Jakob, M., Catenazzi, G., & Ostermeyer, Y. (2020). Policies to decarbonize the Swiss Residential Building Stock: An agent-based building stock modeling assessment. *Energy Policy*, *146*, 111814. https://doi.org/10.1016/j.enpol.2020.111814

[36] Rasouli, M., Sabzehgar, R., & Reza Teymour, H. (2018). An efficient approach for measurement-based composite load modeling. *2018 IEEE Energy Conversion Congress and Exposition (ECCE)*. https://doi.org/10.1109/ecce.2018.8558273

[37] Shang, Zhigang; Li, Mo; An, Yanyan; Yu, Han; Wang, Yi; Qin, Jixing. In: 2022 IEEE International Conference on Unmanned Systems (ICUS) Unmanned Systems (ICUS), 2022 IEEE International Conference on. :1518-1522 Oct, 2022.

[38] Dehghanian, P., Zhang, B., Dokic, T., & Kezunovic, M. (2019). Predictive risk analytics for weather-resilient operation of Electric Power Systems. *IEEE Transactions on Sustainable Energy*, *10*(1), 3–15. https://doi.org/10.1109/tste.2018.2825780

[39] Yan, Y., & Zhang, Z. (2021). Cooling, heating and electrical load forecasting method for integrated energy system based on SVR model. *2021 6th Asia Conference on Power and Electrical Engineering (ACPEE)*. https://doi.org/10.1109/acpee51499.2021.9436990

[40] Thommessen, C., Soltysik, S., & Roes, J. (2022). Heat load forecasting for district heating systems using Neural Networks. *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. https://doi.org/10.1109/iceccme55909.2022.9988718

[41] Şahin, U. (2021). Future of renewable energy consumption in France, Germany, Italy, Spain, Turkey and UK by 2030 using optimized fractional nonlinear grey bernoulli model. *Sustainable Production and Consumption*, *25*, 1–14. https://doi.org/10.1016/j.spc.2020.07.009

[42] Papaioannou, N., Tsimpiris, A., Talagozis, C., Fragidis, L., Angeioplastis, A., Tsakiridis, S., & Varsamis, D. (2023). Parallel feature subset selection wrappers using K-means classifier. *WSEAS TRANSACTIONS ON INFORMATION SCIENCE AND APPLICATIONS*, *20*, 76–86. https://doi.org/10.37394/23209.2023.20.10

[43] Xiao, L., Wang, C., Dong, Y., & Wang, J. (2019). A novel sub-models selection algorithm based on Max-relevance and min-redundancy neighborhood mutual information. *Information Sciences*, *486*, 310–339. https://doi.org/10.1016/j.ins.2019.01.075

[44] Chabouni, N., Belarbi, Y., & Benhassine, W. (2020). Electricity load dynamics, temperature and seasonality nexus in Algeria. *Energy*, *200*, 117513. https://doi.org/10.1016/j.energy.2020.117513

[45] Zhao, H. (2021). *Ann-Based Day-Ahead Short-Term Load Forecasting*. https://doi.org/10.32920/ryerson.14656170.v1

[46] Zare-Noghabi, A., Shabanzadeh, M., & Sangrody, H. (2019). Medium-term load forecasting using support vector regression, feature selection, and symbiotic organism search optimization. *2019 IEEE Power & Energy Society General Meeting (PESGM)*. https://doi.org/10.1109/pesgm40551.2019.8973726

[47] Loggia, R., Flamini, A., Massaccesi, A., Moscatiello, C., Galasso, A., & Martirano, L. (2023). Electrical load profiles for residential buildings: Enhanced bottom-up model (EBM). *2023 International Conference on Clean Electrical Power (ICCEP)*. https://doi.org/10.1109/iccep57914.2023.10247473

[48] Nichiforov, C., Stamatescu, G., Stamatescu, I., & Făgărăşan, I. (2019). Evaluation of sequence-learning models for large-commercial-building load forecasting. *Information*, *10*(6), 189. https://doi.org/10.3390/info10060189

[49] Mahendra, Moch. F., & Azizah, N. L. (2023). *Implementation of Machine Learning to Predict the Weather Using a Support Vector Machine*. https://doi.org/10.21070/ups.2889

[50] Benson, S. A., & Ogunjuyigbe, J. K. (2018). Impact of weather variables on electricity power demand forecast using fuzzy logic technique. *Nigerian Journal of Technology*, *37*(2), 450. https://doi.org/10.4314/njt.v37i2.21

[51] [51] Zhao, J., Duan, Y., & Liu, X. (2018). Uncertainty analysis of weather forecast data for cooling load forecasting based on the Monte Carlo method. *Energies*, *11*(7), 1900. https://doi.org/10.3390/en11071900

[52] Bedi, J., & Toshniwal, D. (2018). Empirical mode decomposition based deep learning for electricity demand forecasting. *IEEE Access*, *6*, 49144–49156. https://doi.org/10.1109/access.2018.2867681

[53] Cleophas, T. J., & Zwinderman, A. H. (2021). Regression trees. *Regression Analysis in Medical Research*, 383–391. https://doi.org/10.1007/978-3-030-61394-5_22

[54] T. Anjali, K. Chandini, K. Anoop and V. L. Lajish, "Temperature Prediction using Machine Learning Approaches," 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kannur, India, 2019, pp. 1264-1268, doi: 10.1109/ICICICT46008.2019.8993316.

[55] Griffin, S. (2020). *Spatial Downscaling Disease Risk Using Random Forests Machine Learning*. https://doi.org/10.21079/11681/35618

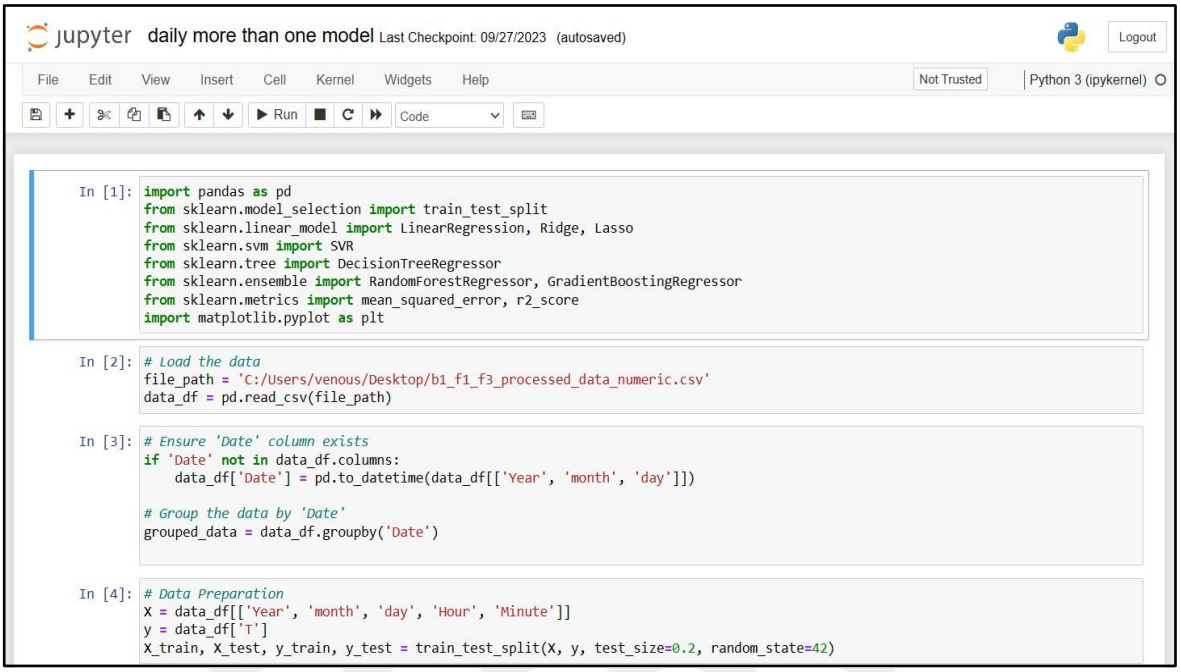
[56] Grace, R. Kingsy; Priyadharshini, M. Indira. In: 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT) Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2023 Second International Conference on. :1-4 Apr, 2023.

[57] Moon, J., Kim, Y., Son, M., & Hwang, E. (2018). Hybrid short-term load forecasting scheme using random forest and Multilayer Perceptron. *Energies*, *11*(12), 3283. https://doi.org/10.3390/en11123283

[58] Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Elements for building supervised Statistical Machine Learning Models. *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, 71–108. https://doi.org/10.1007/978-3-030-89010-0_3

APPENDIX A

APPENDIX A.1 SCREENSHOT FOR PYTHON CODE OF DAILY PREDICT



jupyter daily more than one model Last Checkpoint: 09/27/2023 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

In [11]:
```
# Overlay of predictions from different models with true values (for a subset of data)
subset_length = 40  # Adjust as needed

plt.figure(figsize=(14,8))
plt.plot(y_test.reset_index(drop=True)[:subset_length], label='True Values', marker='o', linewidth=2.5, color='black')

for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    plt.plot(y_pred[:subset_length], label=f'Predicted ({name})', marker='x', linestyle='--')

plt.xlabel('Index')
plt.ylabel('Power Consumption')
plt.title('Overlay of Predictions from Different Models with True Values')
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()
```

jupyter daily more than one model Last Checkpoint: 09/27/2023 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

In [12]:
```
# Prediction for specific date and hour
selected_date = pd.to_datetime(grouped_data['Date'].unique()[1])
selected_date = selected_date[0]
selected_hour = 14
selected_date = pd.to_datetime("2022-05-15")

year = selected_date.year
month = selected_date.month
day = selected_date.day

data_for_prediction = pd.DataFrame({
    'Year': [year],
    'Month': [month],
    'Day': [day],
    'Hour': [selected_hour],
    'Minute': [0]
})

# Now we can run the prediction loop
for name, model in models.items():
    prediction = model.predict(data_for_prediction)
    print(f"The prediction from {name} for {selected_date.date()} at {selected_hour}:00 is: {prediction[0]}")
```

```
The prediction from Linear Regression for 2022-05-15 at 14:00 is: 41434.95001716107
The prediction from Ridge for 2022-05-15 at 14:00 is: 41435.090226621134
The prediction from Lasso for 2022-05-15 at 14:00 is: 41435.0359568485
The prediction from SVR for 2022-05-15 at 14:00 is: 45911.25949568448
The prediction from Decision Tree for 2022-05-15 at 14:00 is: 40596.261
The prediction from Random Forest for 2022-05-15 at 14:00 is: 40588.32312999998
The prediction from Gradient Boosting for 2022-05-15 at 14:00 is: 40566.98930726328
```

In [ ]:

APPENDIX A.2 SCREENSHOT FOR PYTHON CODE OF WEEKLY PREDICT

jupyter weekly_pred Last Checkpoint: 09/23/2023 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

In [1]:
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, TimeSeriesSplit, cross_val_score
from sklearn.linear_model import LinearRegression, Lasso, Ridge, ElasticNet
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.neural_network import MLPRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import matplotlib.pyplot as plt
```

In [2]:
```
# Step 1: Load the data
data = pd.read_csv('C:/Users/venous/Desktop/weekly.csv')
```

In [3]:
```
# Step 2: Add lagged features
data['Lag_1'] = data['Power Consumption'].shift(1)
data['Lag_2'] = data['Power Consumption'].shift(2)
data['Lag_3'] = data['Power Consumption'].shift(3)
data.dropna(inplace=True)
```

In [4]:
```
# Step 3: Prepare the data for training
X = data[['Subscriber No', 'Week Number', 'Lag_1', 'Lag_2', 'Lag_3']]
y = data['Power Consumption']
X.loc[:, 'Subscriber No'] = X['Subscriber No'].astype('category').cat.codes
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, shuffle=False)
tscv = TimeSeriesSplit(n_splits=5)
```

```
C:\Users\venous\anaconda3\lib\site-packages\pandas\core\indexing.py:1745: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  isetter(ilocs[0], value)
```

jupyter weekly_pred Last Checkpoint: 09/23/2023 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

In [5]:
```
# Step 4: Initialize the models
models = {
    "Linear Regression": LinearRegression(),
    "Decision Tree": DecisionTreeRegressor(),
    "Random Forest": RandomForestRegressor(),
    "Gradient Boosting": GradientBoostingRegressor(),
    "Support Vector": SVR(kernel='linear'),
    "Lasso Regression": Lasso(),
    "Ridge Regression": Ridge(),
    "Elastic Net": ElasticNet(),
    "K-Nearest Neighbors": KNeighborsRegressor(),
    "Neural Networks": MLPRegressor(max_iter=1000, random_state=42)
}
```

In [6]:
```
# Step 5: Train and evaluate the models
model_performance = {}
for name, model in models.items():
    cv_scores = cross_val_score(model, X_train, y_train, cv=tscv, scoring='r2')
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    mae = mean_absolute_error(y_test, y_pred)
    rmse = np.sqrt(mean_squared_error(y_test, y_pred))
    r2 = r2_score(y_test, y_pred)
    model_performance[name] = {"MAE": mae, "RMSE": rmse, "R2": r2, "Cross-validated R2": np.mean(cv_scores)}
```

jupyter weekly_pred Last Checkpoint: 09/23/2023 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

In [7]:

```
# Step 6: Visualize and compare the results

# Individual model predictions vs true values
for name, model in models.items():
    y_pred = model.predict(X_test)
    plt.figure(figsize=(10,6))
    plt.plot(y_test.reset_index(drop=True), label='True Values', marker='o')
    plt.plot(y_pred, label=f'Predicted Values ({name})', marker='x')
    plt.xlabel('Index')
    plt.ylabel('Power Consumption')
    plt.title(f'True vs Predicted Power Consumption ({name})')
    plt.legend()
    plt.grid(True)
    plt.show()
```

jupyter weekly_pred Last Checkpoint: 09/23/2023 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

In [8]:

```
# Overlay of predictions from different models with true values (for a subset of data)
subset_length = 40  # Adjust as needed
plt.figure(figsize=(14,8))
plt.plot(y_test.reset_index(drop=True)[:subset_length], label='True Values', marker='o')
for name, model in models.items():
    y_pred = model.predict(X_test)
    plt.plot(y_pred[:subset_length], label=f'Predicted Values ({name})', marker='x')
plt.xlabel('Index')
plt.ylabel('Power Consumption')
plt.title('Overlay of Predictions from Different Models with True Values')
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()
```

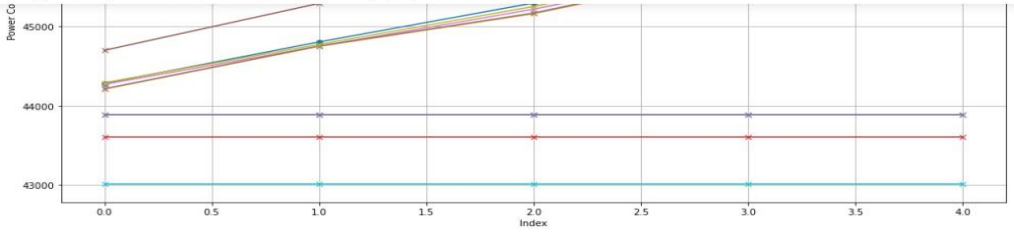
jupyter weekly_pred Last Checkpoint: 09/23/2023 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code



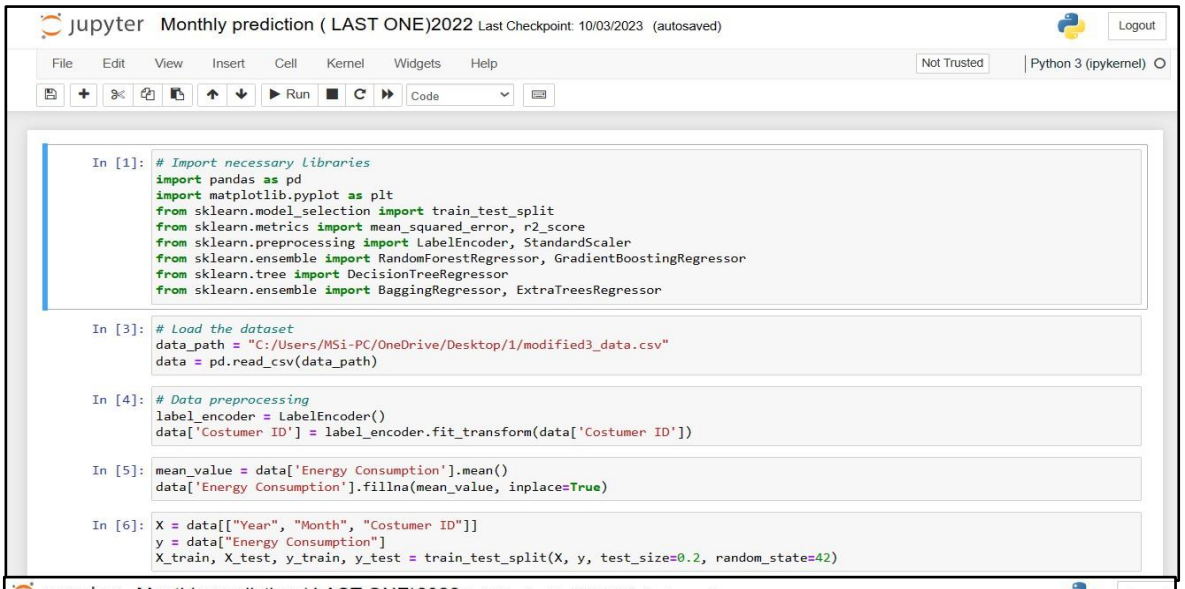
In [9]:

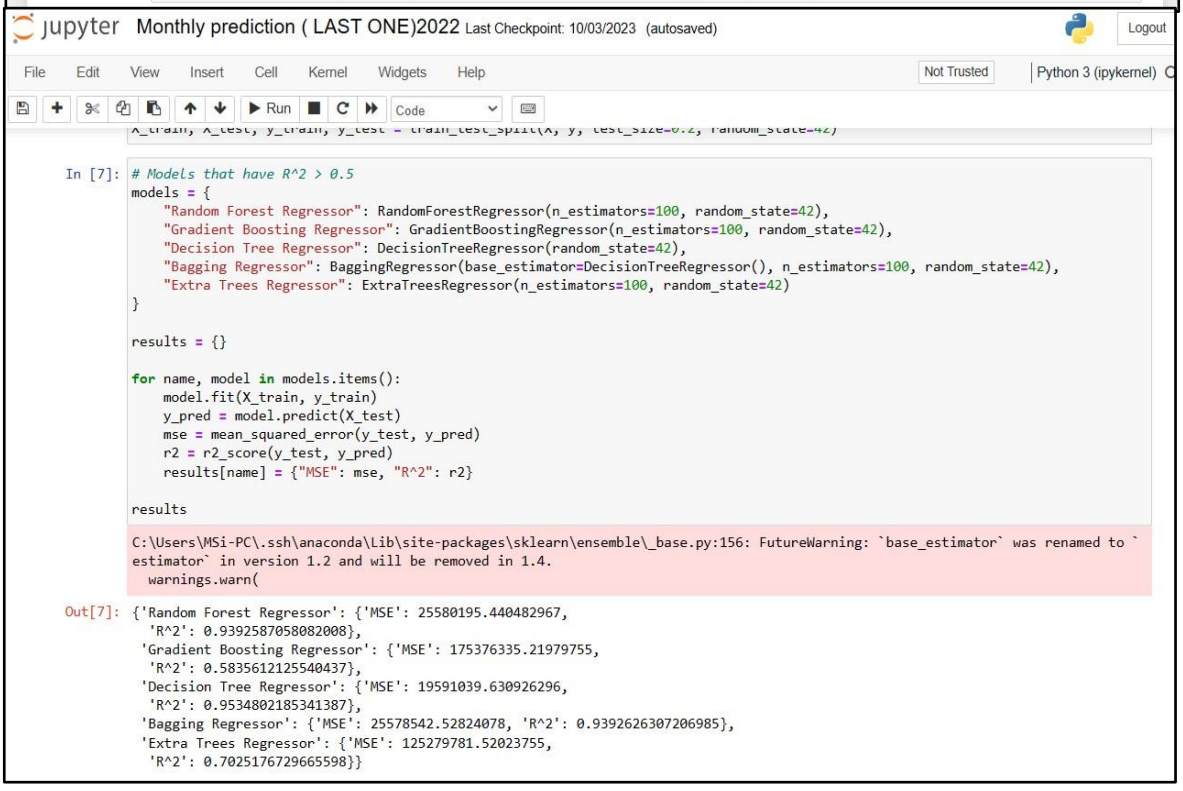
```
# Print the performance metrics
for name, performance in model_performance.items():
    print(f"{name} - MAE: {performance['MAE']:.2f}, RMSE: {performance['RMSE']:.2f}, R2: {performance['R2']:.2f}, Cross-validated
```

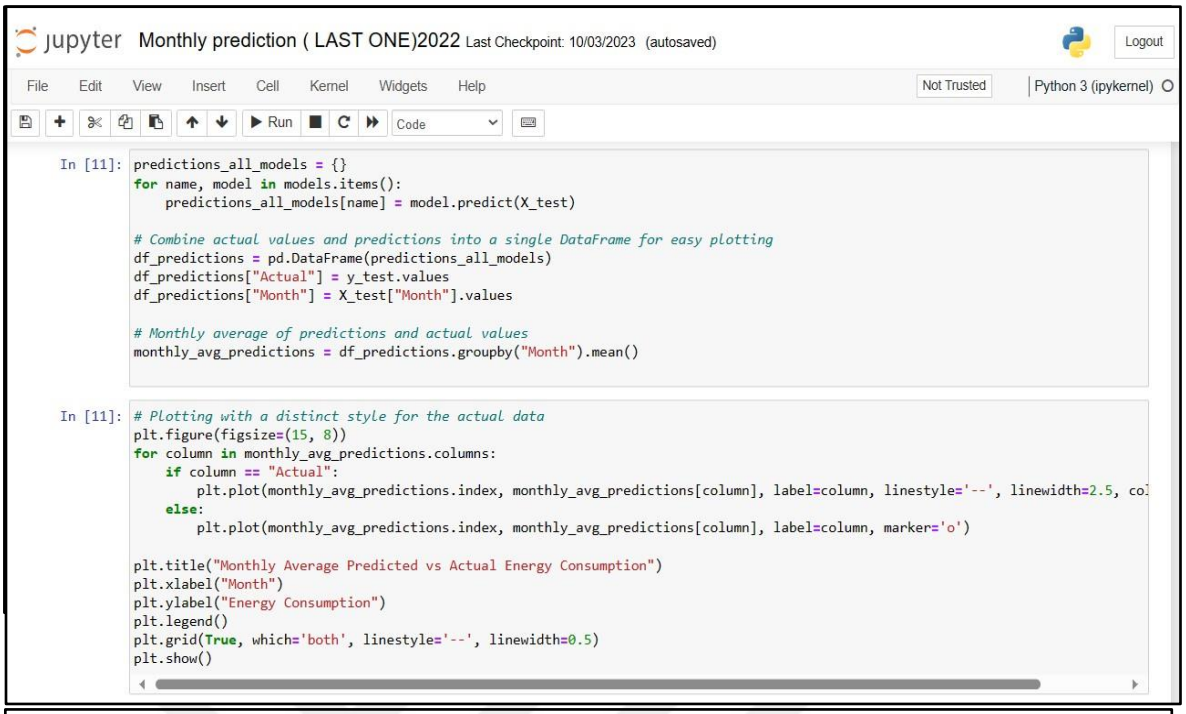
```
Linear Regression - MAE: 56.77, RMSE: 73.10, R2: 0.99, Cross-validated R2: -1.57
Decision Tree - MAE: 1360.50, RMSE: 1509.70, R2: -4.32, Cross-validated R2: -5.89
Random Forest - MAE: 1642.93, RMSE: 1768.46, R2: -6.30, Cross-validated R2: -12.24
Gradient Boosting - MAE: 1360.63, RMSE: 1509.82, R2: -4.32, Cross-validated R2: -5.89
Support Vector - MAE: 557.06, RMSE: 570.90, R2: 0.24, Cross-validated R2: -1.53
Lasso Regression - MAE: 39.52, RMSE: 46.57, R2: 0.99, Cross-validated R2: -0.19
Ridge Regression - MAE: 52.75, RMSE: 68.04, R2: 0.99, Cross-validated R2: 0.55
Elastic Net - MAE: 45.35, RMSE: 51.53, R2: 0.99, Cross-validated R2: 0.58
K-Nearest Neighbors - MAE: 2233.78, RMSE: 2327.66, R2: -11.65, Cross-validated R2: -38.85
Neural Networks - MAE: 1424.70, RMSE: 1426.38, R2: -3.75, Cross-validated R2: -104.97
```

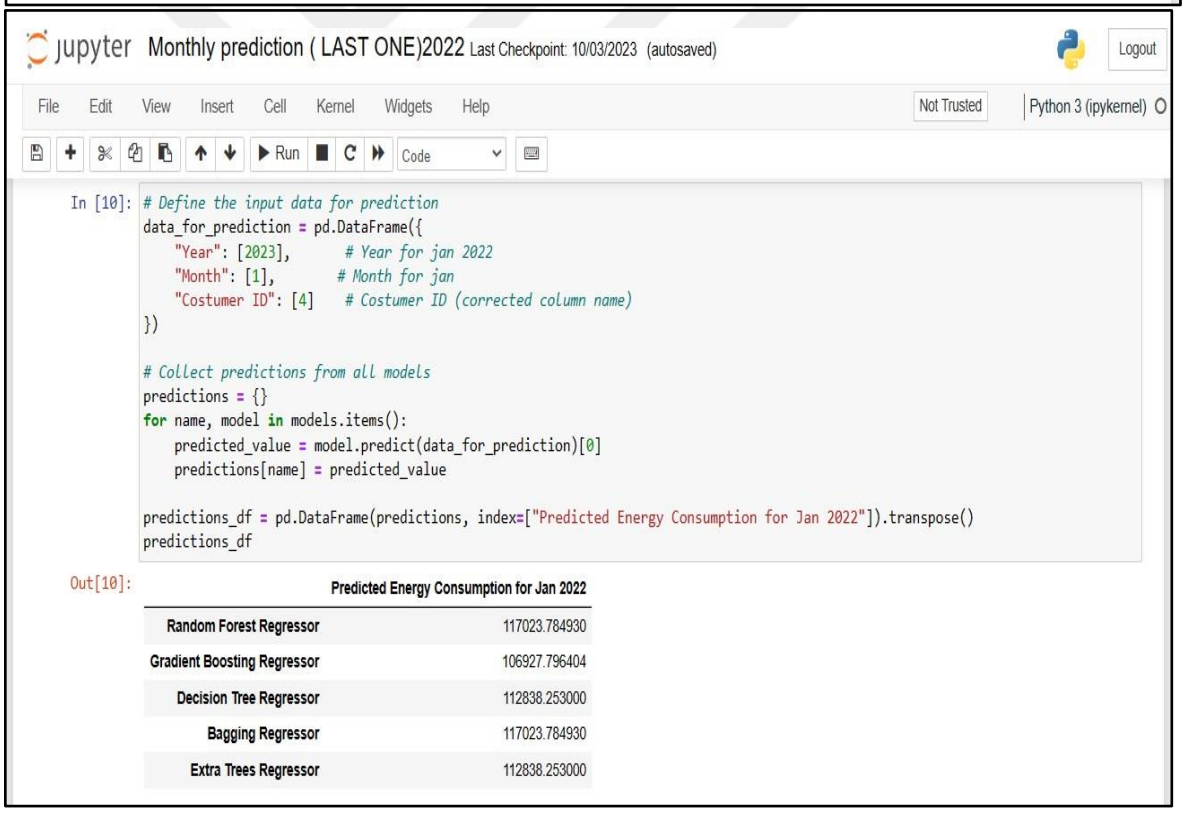
APPENDIX A.3 SCREENSHOT FOR PYTHON CODE OF MONTHLY PREDICT

Here are example for 2022 and smiliar to 2020, 2021 and 2023 due to same code and only year change :



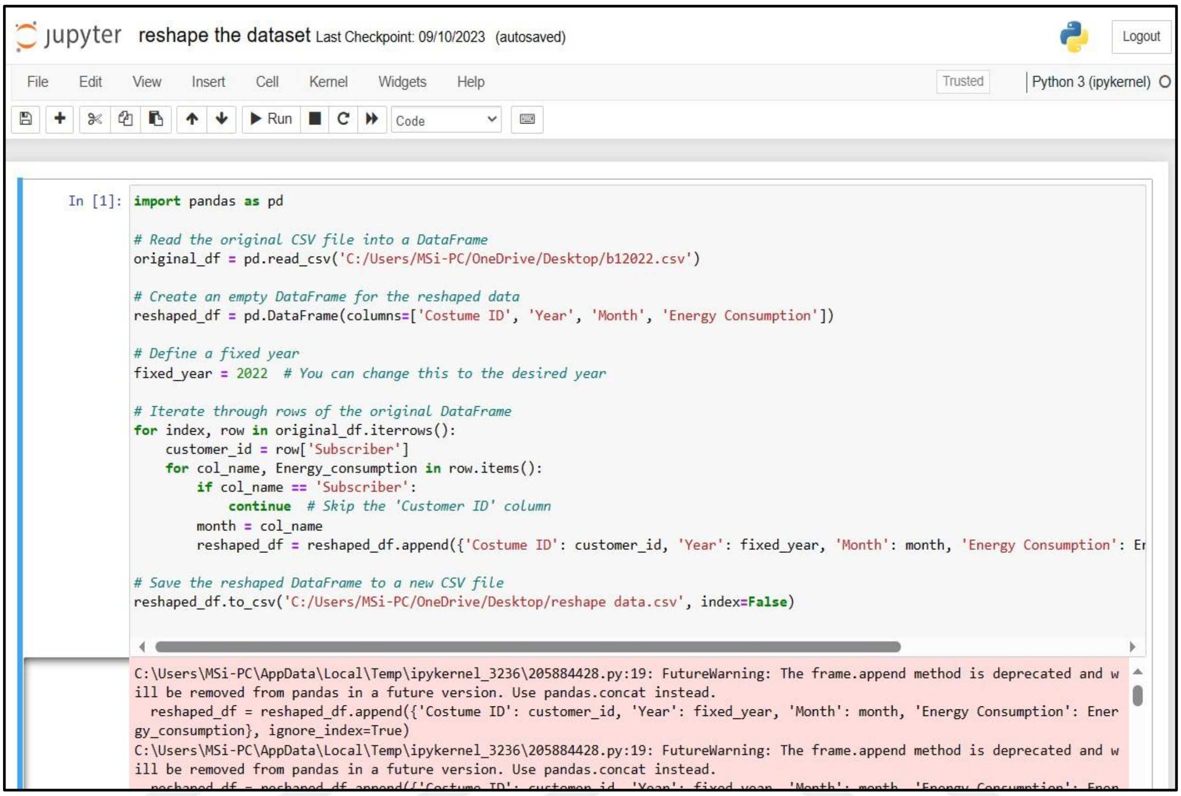






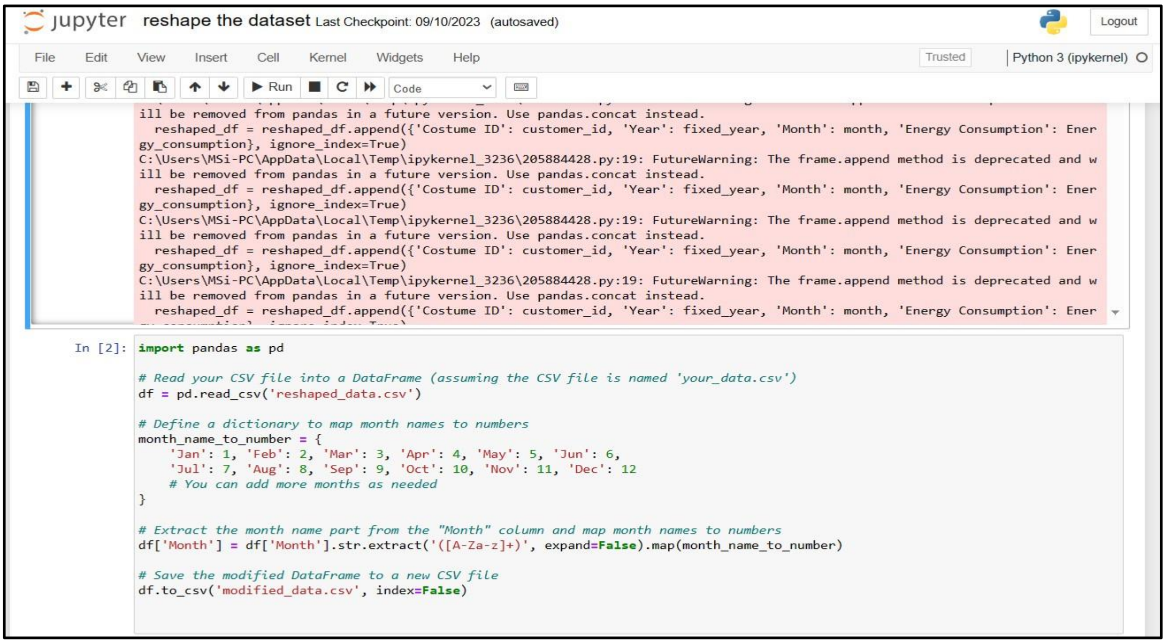
	Predicted Energy Consumption for Jan 2022
Random Forest Regressor	117023.784930
Gradient Boosting Regressor	106927.796404
Decision Tree Regressor	112838.253000
Bagging Regressor	117023.784930
Extra Trees Regressor	112838.253000

APPENDIX A.4 SCREENSHOT FOR PYTHON CODE OF RESHAPE AND MODIFY DATASET




```
import pandas as pd

# Read the original CSV file into a DataFrame
original_df = pd.read_csv('C:/Users/MSi-PC/OneDrive/Desktop/b12022.csv')

# Create an empty DataFrame for the reshaped data
reshaped_df = pd.DataFrame(columns=['Costume ID', 'Year', 'Month', 'Energy Consumption'])

# Define a fixed year
fixed_year = 2022  # You can change this to the desired year

# Iterate through rows of the original DataFrame
for index, row in original_df.iterrows():
    customer_id = row['Subscriber']
    for col_name, Energy_consumption in row.items():
        if col_name == 'Subscriber':
            continue  # Skip the 'Customer ID' column
        month = col_name
        reshaped_df = reshaped_df.append({'Costume ID': customer_id, 'Year': fixed_year, 'Month': month, 'Energy Consumption': Er

# Save the reshaped DataFrame to a new CSV file
reshaped_df.to_csv('C:/Users/MSi-PC/OneDrive/Desktop/reshape data.csv', index=False)
```

```
C:\Users\MSi-PC\AppData\Local\Temp\ipykernel_3236\205884428.py:19: FutureWarning: The frame.append method is deprecated and w
ill be removed from pandas in a future version. Use pandas.concat instead.
  reshaped_df = reshaped_df.append({'Costume ID': customer_id, 'Year': fixed_year, 'Month': month, 'Energy Consumption': Ener
gy_consumption}, ignore_index=True)
C:\Users\MSi-PC\AppData\Local\Temp\ipykernel_3236\205884428.py:19: FutureWarning: The frame.append method is deprecated and w
ill be removed from pandas in a future version. Use pandas.concat instead.
```




```
ill be removed from pandas in a future version. Use pandas.concat instead.
  reshaped_df = reshaped_df.append({'Costume ID': customer_id, 'Year': fixed_year, 'Month': month, 'Energy Consumption': Ener
gy_consumption}, ignore_index=True)
C:\Users\MSi-PC\AppData\Local\Temp\ipykernel_3236\205884428.py:19: FutureWarning: The frame.append method is deprecated and w
ill be removed from pandas in a future version. Use pandas.concat instead.
  reshaped_df = reshaped_df.append({'Costume ID': customer_id, 'Year': fixed_year, 'Month': month, 'Energy Consumption': Ener
gy_consumption}, ignore_index=True)
C:\Users\MSi-PC\AppData\Local\Temp\ipykernel_3236\205884428.py:19: FutureWarning: The frame.append method is deprecated and w
ill be removed from pandas in a future version. Use pandas.concat instead.
  reshaped_df = reshaped_df.append({'Costume ID': customer_id, 'Year': fixed_year, 'Month': month, 'Energy Consumption': Ener
gy_consumption}, ignore_index=True)
C:\Users\MSi-PC\AppData\Local\Temp\ipykernel_3236\205884428.py:19: FutureWarning: The frame.append method is deprecated and w
ill be removed from pandas in a future version. Use pandas.concat instead.
  reshaped_df = reshaped_df.append({'Costume ID': customer_id, 'Year': fixed_year, 'Month': month, 'Energy Consumption': Ener
```

In [2]:

```
import pandas as pd

# Read your CSV file into a DataFrame (assuming the CSV file is named 'your_data.csv')
df = pd.read_csv('reshaped_data.csv')

# Define a dictionary to map month names to numbers
month_name_to_number = {
    'Jan': 1, 'Feb': 2, 'Mar': 3, 'Apr': 4, 'May': 5, 'Jun': 6,
    'Jul': 7, 'Aug': 8, 'Sep': 9, 'Oct': 10, 'Nov': 11, 'Dec': 12
    # You can add more months as needed
}

# Extract the month name part from the "Month" column and map month names to numbers
df['Month'] = df['Month'].str.extract('([A-Za-z]+)', expand=False).map(month_name_to_number)

# Save the modified DataFrame to a new CSV file
df.to_csv('modified_data.csv', index=False)
```

APPENDIX A.5 SCREENSHOT OF OFFICIAL APPROVAL LETTER FOR OPTAINED DATASET

Ethical considerations are fundamental to the successful and responsible conduct of the energy consumption estimation project. By upholding the principles of informed consent, data privacy, confidentiality, minimizing harm, transparency, and accountability, the research team was ensure the protection of participants' rights and welfare. Adhering to ethical guidelines was contribute to the credibility and validity of the study's findings and foster trust among participants and stakeholders. Below is the official approval letter:

Republic of Iraq
Ministry of Electricity
Training & Energy researches Office

جمهورية العراق
وزارة الكـــــهرباء
دائرة التدريـــب وبحوث الطـــاقة

وزارة الكهرباء
MINISTRY OF ELECTRICITY

العدد : ٦٤٥٠/٢٢
التاريخ : ٢٠/٦/٢٠٢٣

الى/ شركة العراق المعاصر المحترمون

م/ بيانات

تحية طيبة ...

يرجى تفضلكم ببيان إمكانية تزويد طالبة الماجستير السيدة (نور مالك صفاء/ معاون رئيس مهندسين) المنسوبة الى الشركة العامة لتوزيع كهرباء بغداد/ فرع توزيع كهرباء الصدر بالبيانات الخاصة بمنظومة (Smart Meter) لمجمع بوابة العراق الخاص بشركتكم وذلك لأهمية هذا الموضوع وحاجتها في تنفيذ بحث الماجستير الخاص بإدخال المقاييس الذكية في قطاع التوزيع .

شاكرين تعاونكم معنا ... مع التقدير

الخبير الدكتور
نصير كريم قاسم
ع. / المدير العام
٢٠٢٣/٦/٢٠

نسخة منه الى:
- مكتب السيد المدير العام المحترم/ للتفضل بالاطلاع ... مع التقدير.
- شعبة المتابعة والبريد الالكتروني/ لما يلزم ... مع التقدير .
- اضبارة الكتب الصادرة .
- قسم بحوث الطاقة ... مع الأوليات .

البحث العلمي الطريق الانسب لحل المشاكل الفنية
-Email :70-planning.dept.mgr@moelc.gov.iq

بـوابـــة
BAGHDAD

العـــدد : IG1-512
التاريخ :8/5/ 2023

الى / وزارة الكهرباء /دائرة التدريب وبحوث الطاقة

م/ تزويد بيانات

تحية طيبة..

نحن شركة بوابة بغداد للاستثمارات العقارية الحاصلة على الاجازة الاستثمارية المرقمة 28/لسنة 2010 .. لا مانع لدينا من تزويد طالبة الماجستير (نور مالك صفاء) التي تشغل صفة معاون رئيس مهندسين في الشركة العامة لتوزيع كهرباء بغداد فرع الصدر بكافة البيانات الخاصة بمنظومة السمارت ميتر الخاصة بمجمعنا (مجمع بوابة العراق السكني) وحسب الطلب المرفق طيا.

المرفقات
كتاب وزارة الكهرباء دائرة التدريب والبحوث
- بيانات العدادات الذكية / مجمع بوابة العراق

مدير الشركة
م. حيدر علي كنان

شركة بوابة بغداد
للاستثمارات العقارية محدودة المسؤولية
Baghdad Gate Comp.
For Real Estate Investments Ltd.

نسخة منه
- الادارة للحفظ