SCHOOL OF BUILT ENVIRONMENT, ENGINEERING AND COMPUTING

LEEDS BECKETT UNIVERSITY

Comparative Analysis of Machine Learning Algorithms for Predicting Household Energy Consumption and Feature Extraction

By: Nikita Vrajesh Savaliya

(Student ID: 77319835)

Dr. Taimur Bakhshi

Submitted to Leeds Beckett University in partial fulfilment of the requirements for the degree of MSc Advanced Computer Science

September - 2023

Candidate's Declaration

I, Nikita Vrajesh Savaliya, confirm that this dissertation and the work presented in it are my own achievement.

Where I have consulted the published work of others this is always clearly attributed.

Where I have quoted from the work of others the source is always given with the exception of such quotations this dissertation is entirely my own work.

I have acknowledged all main sources of help.

I have read and understand the penalties associated with Academic Misconduct.

Signed:

Nikita Savaliya

Date:10/09/2023

Student ID No:77319835

Acknowledgements

I would like to convey my heartfelt thanks to the persons who played a key part in the successful completion of my study and my path toward acquiring a master's degree.

First and foremost, I want to offer my deepest thanks to my supervisor, Dr. Taimur Bakhshi. His unfailing support, essential direction, encouragement, and patience were vital over the life of this endeavours.

Additionally, I would like to convey my thanks to my family for their continual support throughout our research trip. I apologise to my son for any extra grumpiness you may have witnessed while I was focused in my thesis. And to my husband Vrajesh Savaliya, I genuinely appreciate your constant support, which has been my pillar of strength.

Abstract

The residential energy consumption has significant attention within the realm of urban energy management. This heightened interest is driven by data-driven machine learning algorithms that leverage consumption data to forecast overall energy usage trends. The research underpins its investigative endeavours with a meticulous comparison of different machine learning algorithms. Different ML algorithms are meticulously examined using the Kaggle dataset of Household Electric Power Consumption. The overarching objective of this study is to identify the algorithm that exhibits the highest level of effectiveness , predictive capabilities and timing complexity.

This research takes a comprehensive approach by analysing both regression and classification tasks using a shared dataset. Various regression models, including linear regression, decision trees, lasso regression, ridge regression, random forest, and XGBoost, are evaluated using a diverse set of metrics. Simultaneously, classification models such as logistic regression, decision trees, random forests, support vector machines, k-nearest neighbor, and XGBoost are rigorously assessed for accuracy, precision, recall, F1 score, and training time. The best classification model in terms of accuracy is XGBoost, whereas the best regression model is linear regression decided by the Root Mean Squared Error (RMSE) score. In computational analysis, K-Nearest Neighbors excels in classification, while Ridge Regression shines in regression tasks. Furthermore, the study goes beyond model selection, integrating Explainable Artificial Intelligence (XAI) techniques like SHAP values to unravel significant features driving linear regression and XGBoost models. This multifaceted approach underscores the holistic nature of model selection, encompassing performance benchmarks, computational intricacies, and interpretability facets, thus reinforcing the study's overall robustness.

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Networks
ANFIS	adaptive network-based fuzzy inference system
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
CART	Classification and Regression Trees
CC	coefficient of correlation
CO2	carbon dioxide
CV	Cross-validation
DT	Decision Tree
EDA	Exploratory data analysis
ERF	Extreme Random Forest
EPC	energy performance certificates
EU	European Union
HVAC	Heating, Ventilation, and Air Conditioning.
ID3	Iterative Dichotomiser 3
IEA	International Energy Agency
LIME	local interpretable model-agnostic explanations
LSTM	Long Short-Term Memory network

MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	machine learning
MLR	Multiple linear regression
MSE	mean squared error
RAE	relative absolute error
RF	Random Forest
RMSE	root mean square error
RRSE	root relative squared error
SCADA	supervisory control and data acquisition
SG	Smart Grid
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
tCO2e	tonnes of carbon dioxide equivalent
XAI	eXplainable Artificial Intelligence
XGboost	Extreme Gradient Boosting
ZigBee	Zonal Intercommunication Global-standard

Chapter 1: Introduction

1.1 Overview

1.1.1 Importance of building energy analysis:

In contemporary society, energy stands as an indispensable requisite, exerting a notable influence upon facets of societal stability, quality of life, social well-being, as well as the robust progression and advancement of a nation's economic prowess and growth (Nti et al., 2020; Sharifzadeh et al., 2017). A subset of analysts contends that a direct causative connection between increased energy usage and economic growth exists, thereby directing attention toward the domain of energy consumption. (Asghar, 2008) Moreover, the statistical information depicted in the graph below elucidates a notable upswing in worldwide energy consumption, displaying an estimated threefold escalation from around 7000 to 25500 terawatt-hours during the interval from 1980 to 2021 (EIA, 2022).

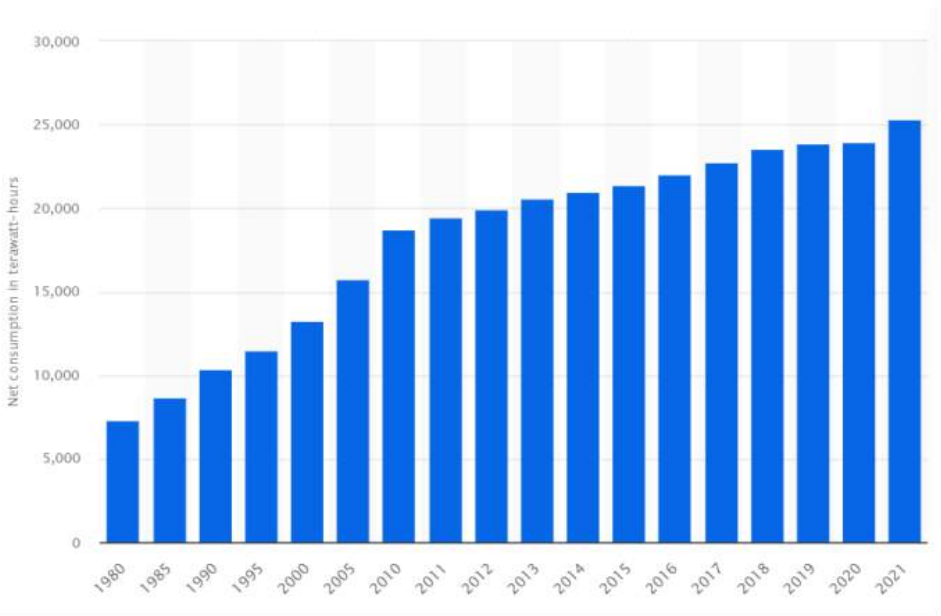

Fig 1.1: Electricity consumption worldwide (EIA,2022)

Significant energy is utilized across the three primary economic sectors: industry, transportation, and buildings.(Chou et al., 2014) Furthermore, the residential sector utilises more energy than the commercial sector. (Fayaz et al., 2019) The significant 27% of the global energy used that comes from the residential sector (Nejat et al., 2015)

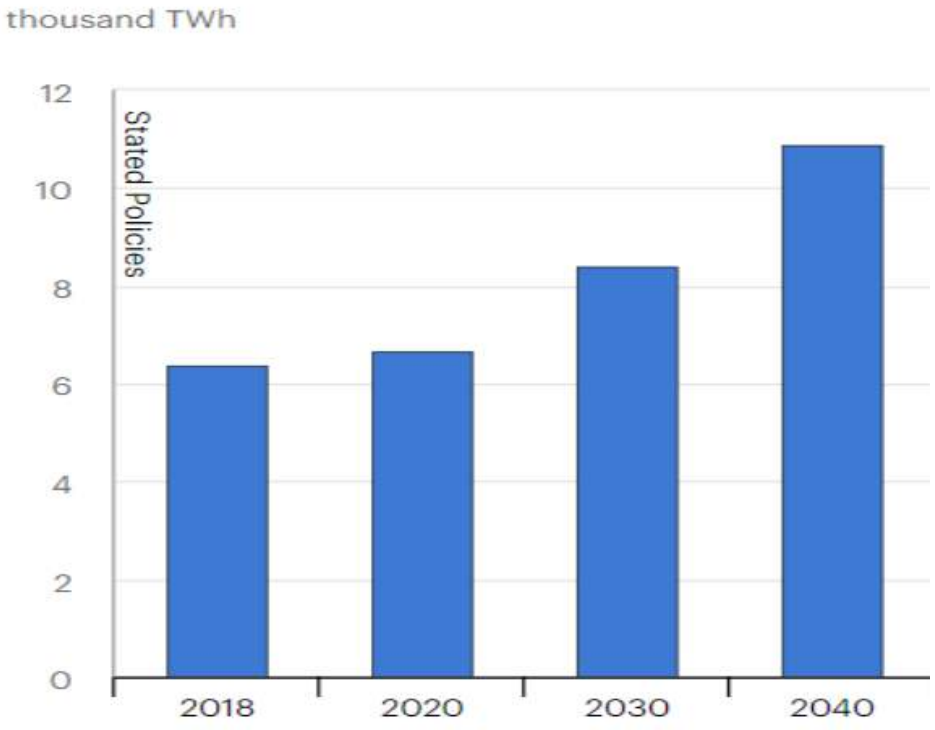

Fig 1.2 : Electricity demanded by Residential sector, 2018-2040(IEA, 2019)

According to the Stated Policies Scenario in the 2019 World Energy Outlook published by the International Energy Agency (IEA,2019), it is anticipated to propel the projected rise in global household power consumption at roughly 37% per year through 2040 shows in figure-1.2. As a result, precise power demand forecasting is crucial for assuring a consistent supply of electricity.

Furthermore, within the context of contemporary urban environments such as Hong Kong, it becomes evident that a substantial 60% of carbon emissions stem from the domain of electricity generation, with buildings contributing significantly by constituting 89% of the aggregate electricity consumption (Leung et al., 2012). The study by Gordic et al. focused on 31 European nations, including 27 EU members and 4 non-members. The average annual carbon emissions from households in these countries were 0.09 to 6.44 tCO2e (tonnes of carbon dioxide equivalent) to 1.36 tCO2e, according to this study. (Gordic et al., 2023). Consequently, the enhancement of energy efficiency and the overall energy performance of buildings assumes paramount

importance, not only to address the mounting requisites for supplementary energy provision but also to effectively alleviate CO2 emissions. (Chou et al., 2014)

1.1.2 Machine Learning in Analysis and Prediction of energy consumption:

An extensive branch of computer science known as artificial intelligence (AI) is focused on constructing intelligent machines and systems that can carry out tasks that usually need human brains. A subset of artificial intelligence (AI) known as machine learning (ML) allows a system to learn by itself without human participation. (Vijayan, 2022; Ardabili et al, 2019)

These ML models are being used in a number of industries since they are advantageous and operate like a function that best converts input data to output. Notably, Machine learning algorithms may be used to estimate energy usage with a high degree of accuracy. (Mosavi et al., 2019).

In recent years, different ML strategies have been presented in residential and commercial construction for estimate of heating and cooling loads, energy consumption, and performance under varied scenarios. (Seyedzadeh et al., 2018). It is remarkable that Regression analysis has traditionally been the most widely used modelling tool for projecting energy demand. (Tso et al., 2007) Moreover, linear regression is employed as a standard measurement and verification approach.(Ding et al., 2021)

Researchers predict that the availability of a building energy system with accurate forecasting would result in a 10%–30% decrease in building energy use. (Olu-Ajayi et al., 2022). Moreover, a better knowledge of the variables impacting residential energy usage may be acquired using the application of machine learning approaches to stakeholders and policymakers.(Burnett et al., 2022; Tso et al., 2007)

Over the past decade, there has been an enormous increase in the development of machine learning algorithms for estimating energy use in buildings, with various

approaches arising (Pham et al., 2020). These methods generally cover the use of Artificial Neural Networks (ANN), Support Vector Machine (SVM), and Decision Tree (DT) approaches for forecasting building energy demand. (Olu-Ajayi et al., 2022; Maarif et al., 2023) These data-driven ML algorithms have the benefit of not requiring precise physical knowledge about the buildings, energy infrastructure, or surroundings. This approach improves the accuracy and efficiency of forecasting energy consumption patterns in dwellings. (Ardabili et al., 2019)

To solve these forecasting issues efficiently, load forecasting systems are frequently divided into three separate time ranges: short term, medium-term, and long-term forecasting.

1. Long-term forecasting, which spans the period from one to twenty years out, is essential for strategic planning, the construction of extra generation, and the calculation of transmission capacity. It aids in decision-making about future investments and the expansion of the electricity system.
2. Medium-term forecasting, which takes a month to a year into account, is used for maintenance strategy and sharing power agreement schedules. It facilitates both the optimization of maintenance duties and the management of power transfers between diverse entities.
3. The goal of short-term load forecasting is to predict the demand for power over the hours to weeks ahead. The use of immediate energy storage, real-time control, unit scheduling, fuel purchasing plans, and quick maintenance planning all depend on it. Accurate short-term load estimates enable the greatest resource utilization, planning of resources, and power production control.

Each type of load forecasting has a distinct function and is essential to the effective management of the power system, including planning and decision-making. (Kyriakides et al., 2007; Sikiric et al., 2013)

1.1.3 eXplainable AI:

The area of eXplainable Artificial Intelligence (XAI) is devoted to the study and development of strategies and approaches with the objective of improving the comprehension of AI system outputs for human users. (Adadi et al., 2018) The field of XAI has expanded significantly during the last ten years. (Maarif et al., 2023)

In 2004, the phrase was originally established by Van Lent et al. to characterise their system's capacity to explain the decisions and actions done by AI-controlled entities in simulation games. (Van et al., 2004) In recent years, the concept of eXplainable Artificial Intelligence (XAI) has attracted fresh attention from both academic researchers and industry practitioners. (Adadi et al., 2018)

Many machine learning models operate as complex systems, making it impossible for humans to understand the predictions they make. Predictive models' lack of accountability and transparency has had tremendous effects and is still having an impact. The urgent need for improved interpretability and accountability in predictive modelling has been highlighted by cases of people being incorrect bail decisions resulting in the release of high-risk offenders, unfairly denied parole, deceptive ML-based pollution assessments deeming heavily contaminated air as safe for inhalation, and suboptimal allocation of precious resources across domains such as criminal justice, healthcare, energy reliability, finance, and others. (Rudin, 2019)

While many research in residential energy consumption prediction emphasises on the accuracy, neglecting model interpretability, knowing a model's judgements is as crucial as its accuracy. Relying simply on prediction model accuracy for validation is insufficient. (Kim et al., 2020) To instil faith in the consequences, energy planners need to understand how a model arrives at its conclusions. Thus, in the properly apply AI in energy forecasting, a balance between accuracy and interpretability is essential, ensuring that choices are based on credible predictions and intelligible explanations. (Maarif et al., 2023)

In view of the vital relevance of model interpretability in residence energy consumption prediction, it's worth mentioning that the Shapash Python library has emerged as a notable eXplainable Artificial Intelligence (XAI) technique. Designed to increase the comprehensibility of machine learning models, Shapash provides a helpful toolkit. This library provides the establishment of a visualization dashboard for the efficient development and display of machine learning model outcomes (Sparsh, nd) (Kuzlu et al., 2020). By applying such XAI approaches as Shapash, the search for the delicate

balance between accuracy and interpretability in energy forecasting becomes more reachable, supporting both credible projections and intelligible explanations.

1.2 Aim and Objectives

Aim:

The aim of this research is to compare various machine learning algorithms for predicting household electricity consumption. Additionally, the study aims to identify and extract significant features or factors that significantly influence residential electricity usage.

Objectives:

1. To perform an extensive assessment of the literature on the application of machine learning for energy prediction and to look at present approaches that employ machine learning for predictive analysis.
2. Provide insights into the peak consumption periods, seasonal variations, or specific events that affect energy usage.
3. Determine which machine learning algorithm performs the best in order to determine the most accurate and appropriate model for predicting residential energy use.
4. Compare accuracy as well as computational complexities among these algorithms.
5. Identifying essential factors to anticipate house energy efficiency by employing the most efficient Machine Learning techniques.

1.3 Research Questions:

1. Which machine learning algorithm is the most effective for a specific energy forecasting application?
2. What performance metrics can be used to assess the effectiveness of machine learning algorithms in energy forecasting?
3. How does the quality of the dataset and preprocessing choices impact the performance of machine learning models?

4. What are the strengths and limitations of algorithms within their specific contexts?
5. What are the key performance indicators for assessing the accuracy of ML algorithms?

1.4 Limitation of this work:

The work's weakness is that it is reliant on the quality and availability of data on home energy usage. The data utilised impacts how accurate and trustworthy a machine learning algorithm comparison will be. This research takes use of a dataset from Kaggle, a website that runs machine learning contests. Although Kaggle datasets may be interesting and a great starting point for research, they may not accurately mirror real-world circumstances. The use of synthetic or simulated data from Kaggle may not properly portray the complexity and nuances of actual household power consumption patterns. As a consequence, it is necessary to carefully examine the study's results since they may not be instantly relevant to real settings.

1.3 Outline

There are several chapters in this thesis, ranging from 1 to 5.

Chapter 1-Introduction

The introductory section provides background and rationale for the research, emphasizing the need for precise power consumption forecasting for energy management and conservation. The research objectives and questions are presented, as well as the constraints of the study. A summary of the thesis structure is also provided.

Chapter 2-Literature Review

The literature review section provides an in-depth review of existing work on machine learning techniques for predicting power usage. Previous research on predicting home power use is presented, as well as a comparison of other machine learning techniques applied to comparable issues. The purpose of this section is to identify research gaps that the current study proposes to fill.

Chapter 3- Methodology

The methodology section describes the data gathering and analysis procedures utilised in the study. This section should include describing the various machine learning methods that were compared in terms of precision, recall, F1-Measure scores, computational time and many more.

Chapter 4- Research Design and Implementation

The Research design and implementation section discuss the machine learning algorithms for examining the patterns of residential home energy use. The technical components and configurations of the algorithms, model comparison, model selection, and etc., are thoroughly covered in this chapter.

Chapter 5-Discussion

The findings of this thesis contribute to the understanding of the suitability of different machine learning algorithms for analysing residential home energy usage patterns. This discussion involves a comparative analysis of our research findings with the existing literature, highlighting similarities and differences.

Chapter 6- Project Management

The project management section discusses the overall planning, organization, and coordination of the study. It provides insights into how the research was structured, resources were allocated, timeliness were established, and milestones were set.

Chapter 7- Conclusion and Future Work

In this section should summarize the main findings of the study and highlight any limitations or future directions for research.

Chapter 2: Literature Review

The aim of this section is to provide a comprehensive review of the research done in the area of Forecasting residential energy usage and extract the important features. Its objective to explore various topics related to energy prediction algorithms and present the results. This chapter will compare methods and explore earlier work in detail while concentrating on ML models to estimate energy use.

2.1 Comparison of ML models:

Tso et al, has done the empirical research to explore three modelling techniques (traditional regression analysis, Decision Tree, and Neural Networks) for predicting electricity energy consumption. Data collected from two-phase survey and model selection is based on the square root of average squared error. The empirical application shows that the Decision Tree and Neural Network models are promising alternatives to stepwise regression for understanding energy consumption patterns and predicting consumption levels. During the summer phase, the decision tree model proves more accurate and identifies significant factors such as flat size, household members, and ownership of an air-conditioner. In the winter phase, housing type and ownership of electric water heater and range hood become significant factors. The study highlights the emergence of data mining as an analysis tool and concludes that the three modelling techniques are generally comparable in predicting energy consumption with slight variations in accuracy. (Tso et al., 2007)

Another study by Ahmad et al., two energy prediction model feed-forward back-propagation artificial neural network and random forest were evaluated for their ability to forecast the hourly HVAC energy consumption of a hotel in Madrid, Spain. Both models' prediction accuracy slightly increased when social aspects, such as the number of visitors, were taken into account. Overall, with a root-mean-square error of 4.97 and 6.10, respectively, ANN performed somewhat better than Random Forest. Random Forest(RF), on the other hand, does internal cross-validation with a minimal set of tuning parameters and has an advantage in processing multi-dimensional complicated

data that is usual in buildings. Both models have equivalent predicative abilities and might be used for applications involving building energy. (Ahmad et.al, 2017)

The project sought to forecast electricity usage every 10 minutes and/or every hour in order to enhance the operation of power distribution networks in Tetouan, Morocco. Different machine learning models were examined, including the feedforward neural network, Random Forest, Decision Tree, and Support Vector Machine for regression with radial basis function kernel. For 2017, historical data was extracted from SCADA every 10 minutes. The Random Forest model outperformed the other models by having the lowest prediction errors. The introduction of calendar and weather predictive factors, with hour and temperature being the most predictive, helped to improve the models' accuracy. The dataset utilised was unique and had never been used for this purpose previously. The authors intend to broaden the analysis to include additional Moroccan power producers and distribution firms, including those working with renewable energy, as well as undertake a financial study to assess the economic implications of the projections. (Salam et al., 2018)

Using extensive data from various types of buildings, the study evaluates four data-driven approaches for live energy projections. Processes for pre-treating data take care of problems with data dependability. The effectiveness of mathematical algorithms is examined and contrasted using monitoring data and results that are projected. The results indicate that, despite its complexity, the Artificial Neural Network technique does not always produce the best accuracy. The quickest approach, Gaussian Process Regression, has a lesser degree of accuracy. In the settings under study, Support Vector Machine and Multivariate Linear Regression approaches often outperform each other. All methods achieve the ASHRAE-recommended levels of accuracy, with calculation durations per prediction ranging from less than one second to 22 seconds. With its high accuracy and quick calculation times, MLR and SVM algorithms are particularly suitable for buildings with complicated and erratic occupancy schedules and energy usage patterns. (Zeng et al., 2019)

In a study by Xiang et al., five distinct models are analysed in the research study by Xiang et al. to forecast the energy consumption of household appliances. Support

Vector Machine (SVM), K Nearest Neighbour (KNN), Random Forest (RF), Extreme Random Forest (ERF), and Long Short-Term Memory network (LSTM) are just a few of the models that are examined. As a consequence, the deep learning approach, LSTM, has the lowest root mean square error (RMSE), 21.36, and the greatest R2 score, 0.97. Cross-validation techniques are employed in the training and evaluation of the models. The analysis found that deep learning methods like LSTM are more accurate in estimating household appliance energy usage. (Xiang et al., 2020)

Different machine learning methods such as ANN, ANFIS, SVM, MLR were examined for estimating Cyprus's electricity load in research by Solyali et al. The goal was to determine the best reliable technique for predicting both short- and long-term power usage. A dataset used in this research is real historical data from 2016-2017 after the data had undergone preprocessing. The four Machine Learning models are used on it to estimate electricity generation in Cyprus for long-term and short-term analysis. They discovered that SVM performed most effectively for long-term predictions, while the ANN model was superior for short-term analysis. The findings emphasize the close relationship between energy demand, the economy, and the environment, emphasizing the significance of precise power load forecasting for long-term energy planning. (Solyali ,2020)

In Bashir et al.'s study, a number of machine learning techniques, including Support Vector Machines, K-Nearest Neighbour, Logistic Regression, Naive Bayes, Neural Networks, and Decision Tree classifier, were examined for predicting SG stability. The UCI machine learning repository provided the SG dataset that was utilized. The Decision Tree method beat other algorithms, according to experimental data, reaching 100% precision, 99.9% recall, 100% F1 score, and 99.96% accuracy. This emphasizes how crucial it is to choose the right machine-learning methods for forecasting SG stability. Despite the small size of the dataset employed in this study, real-time SGs produce enormous volumes of data. (Bashir et al.,2021)

Using data from 2370 buildings in Chongqing, this research investigates the use of machine learning algorithms for forecasting and analysing energy use in public buildings. There were six machine learning methods used, with XGBoost proving to be

the most precise and effective predictor. The study emphasises the value of feature selection by showing that the top 10 characteristics have a substantial impact on model performance and that adding more features results in declining benefits. In terms of efficiency and goodness of fit, ensemble learning algorithms, particularly XGBoost, fared better than the competition. The dataset's quality is emphasised, putting the focus on the need of sufficient variation across features for successful supervised learning. Looking forward, this research provides recommendations for enhancing energy management in urban buildings and urges the prudent use of machine learning algorithms together with domain knowledge. (Ding et al., 2021)

The Rambabu et al.'s study focuses on predicting household energy consumption using supervised-based ML algorithms and weather-related features. The dataset comprises home temperature and humidity readings taken at 10-minute intervals over 4.5 months using a ZigBee Wireless sensor network. Various machine learning algorithms, such as Linear Regression, Lasso Regression, Random Forest, Extra Tree Regressor, and XG Boost, were performed and evaluated using R-square as the metric. Tree-based models, particularly Extra Trees Regressor, performed best, showing little linear correlation between temperature/humidity features and energy consumption. The time of day was found to be crucial in energy consumption patterns. The tuned model achieved an R-square score of 0.7449, explaining 74.5% of the variance on the test set. Overall, tree-based models are recommended for datasets with features having no linear correlation with the objective variable. (Rambabu et al, 2022)

Another study by Reddy et al. suggests a machine learning-based strategy for hourly electricity forecasting. Using historical electricity data from a utility company is used, with preprocessing to handle missing data and outliers. Number of machine learning models are examined such as Linear Regression, K Nearest Neighbours(KNN), XGBOOST, Random Forest(RF), and Artificial Neural Networks(ANN). Different Performance measures used, including Mean Absolute Error, Root Mean Squared Error, and R2. The results demonstrate that the proposed approach successfully forecasts electricity consumption, with the KNN model performing the highest an accuracy rate of 90.92%. (Reddy et al, 2023)

2.2 eXplainable AI:

Tsoka et al.'s research proposes a novel method for classifying building energy performance certificates (EPC) that departs from conventional direct measuring techniques and uses artificial neural network (ANN) models. With enough input data, the research builds EPC classification with a stunning 99% accuracy by using previous EPC data and experiences from industrialized nations. Less important input elements are found using explainable artificial intelligence (XAI) methods like LIME and SHAP, and they may be removed without substantially affecting the accuracy of the ANN model. The case studies, which use historical data from Lombardy, Italy, show how accurate ANN models can be. In case study 1, key influencing factors like CO2 emissions and net surface area achieve 93% and 89% accuracy in EPC classification, respectively. In case study 2, winter AC non-renewable energy performance does the same. (Tsoka et al., 2022)

The study by Maarif et al. addresses the urgent need for precise energy consumption predictions, especially for businesses looking to acquire and use energy efficiently. In order to address this, the study presents a model for estimating energy consumption that makes use of long short-term memory (LSTM) and explains artificial intelligence (XAI) for parameter analysis. The models are thoroughly tested using a public energy use dataset from a steel business, exceeding prior findings with much lower root mean squared error (RMSE) scores. In order to shed light on the variables influencing energy consumption projections, the interpretability analysis using XAI reveals significant characteristics, such as leading current reactive power and time of day. (Maarif et al.,2023)

The relevance of energy consumption prediction in the construction industry, which accounts for a significant amount of global energy usage and greenhouse gas emissions, is discussed in the work by Zang et al. An explainable deep learning model is created to fill the research gap regarding building attributes, geometry, and urban morphology. The Shapley Additive Explanation method and the Light Gradient Boosting Machine are used in this model to provide insights into energy performance prediction. The study, which uses data from Seattle as a case study, demonstrates the important

importance of urban morphology and building geometry in energy consumption prediction, resulting to a 33.46% increase in accuracy over using building attributes alone. (Zhang et al., 2023)

s

Ref	Model	Length of datasets	Region	Performance evaluation	Year
Tso et. Al, 2007	Regression Analysis, Decision Trees, Neural Networks	365 days (1999-2000)	Hong Kong	RASE	2007
Ahmad et al,2017	Artificial Neural Network, Random Forest	472 days (14/01/2015 TO 30/04/2016)	Spain.	RMSE, MAPE, MAD and R squar	2017
Salam et al., 2018	Feedforward Neural Network with Backpropagation Algorithm, Random Forest, Decision Tree, Support Vector Machine	365days (01-01-2017 to 31-12-2017)	Tetouan city	RSME AND MAE	2018
Zeng et,al, 2019	Artificial Neural Network, Gaussian Process Regression, Support Vector Machine, Multiple Linear Regression	30 days	Shanghai, China,	RMSE, NMBE, *R*-square, Computation time	2019
Solyali et. Al,2020	Artificial Neural Network, Multiple Linear Regression, Adaptive Neuro-Fuzzy Inference System, Support Vector Machine	730 days (2016, 2107)	Cyprus	RMSE	2020
Xiang et. Al, 2020	Support Vector Machine, K Nearest Neighbor,	137 day	Belgium	*R*-square, RMSE	2020

	Random Forest, Extreme Random Forest, Long Short-Term Memory Network				
Bashir, et,al,2021	Support Vector Machine, K-Nearest Neighbor, Logistic Regression, Naive Bayes, Neural Networks, Decision Tree classifier	1000 instances	-	Precision, recall, f1 score, accuracy	2021
Rambabu et al, 2022	Linear Regression, Lasso Regression, Random Forest, Extra Tree Regressor, XG Boost,	136 days (4.5 months)	-	*R*-square	2022
Reddy et al,2023	Linear Regression, K Nearest Neighbours, XGBOOST, Random Forest, Artificial Neural Networks	40,000 instances	-	MAE, RMSE, *R*-squares	2023

Table1: Comparative Analysis of Various Research Papers Investigating ML Models

According to the research findings described in Table 1, only a few energy consumption prediction algorithms between two and five in each study, have been compared. Although this narrow emphasis enables more in-depth study, it also highlights possible holes that may be filled by investigating other methods for improved prediction accuracy or taking into account more sophisticated models. In addition, Zeng et al.'s paper (2019) makes a contribution by comparing the computation times for four machine learning approaches to estimate energy consumption. However, the study's use of a limited dataset raises questions about the application of the findings. Another noteworthy observation is the constrained scope of performance metrics utilized in this study. The next sections that follows tries to fill in that research gap.

Chapter 3: Methodology

Methodology is the methodical application of rules and processes to research utilising a range of theoretical stances. It also goes through the implementation techniques we employed to create a fully working application with minimal usage restrictions.

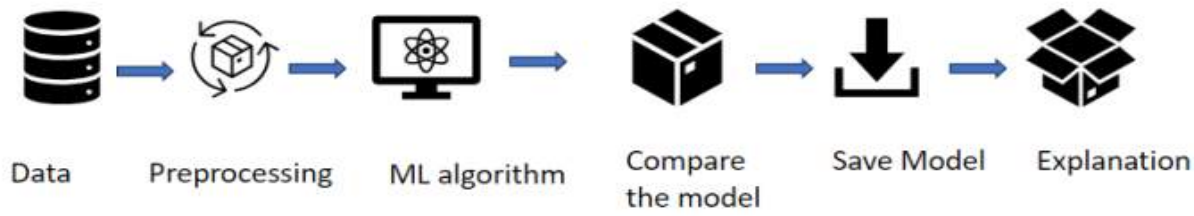

Fig 3.1: Flowchart of methodology

The defined schematic of the recommended methodology, which acts as the framework directing this investigation, is shown in Fig 3.1. The methodology is clearly divided into five essential parts, each of which caters to a different aspect of the research. These portions include Data description, Data Preprocessing and Feature Engineering, ML models, Comparison of performance matrices and save model, and the last part is Explanation A wide variety of regression and classification techniques are used within the scope of this research. Linear regression, Lasso regression, Ridge regression, Decision Tree, Random Forest, and Support Vector Machine (SVM), among others.
The research methodology employs a dual analytical strategy by using both the classification and regression paradigms. The goal of this work is to accurately estimate "Active Energy Consumption(G_ActivePower)," a crucial statistic obtained using following formula:

$$G\_ActivePower = \sum_{k=0}^{n} P\_active_k * (1000/60) - Sub1_k - Sub2_k - Sub3_k \quad (3.1)$$

Where, G_ActivePower= Active Energy Consumption, P_active=Global Active power, Sub1=Submetering 1 value, Sub2=Submetering 2 value, Sub3=Submetering 3 value

The following is a description of the procedural framework:

- Regression Analysis: The structuring of the "Active Energy Consumption" estimation as a regression problem is crucial to the technique. The computed Active Energy Consumption is the targeted variable in this situation. Using powerful machine learning methods, a rigorous forecast is made for this continuous variable. In order to measure the accuracy of the predictions, the performance of the regression models is quantitatively assessed using well-established assessment metrics such the Decision Coefficient (R2 Score), Root Square error (RSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Time complexity, Correlation Coefficient (CC), Relative Absolute Error (RAE), Root Relative Square Error (RRSE).
- Classification Analysis: A parallel thread inside the technique simultaneously focuses on getting the "Active Energy Consumption" feature ready for classification projects. In this preliminary stage, label encoding, a revolutionary method that gives continuous values related to energy consumption categorical characteristics, is strategically applied. This categorization is determined by the energy consumption's relation to the total energy, where values below 3.8 (representing first quartile) are classified as "Low." Meanwhile, values up to 10.3667 (encompassing third quartile) fall into the "Medium" category, while the remaining data is allocated to the "High" category.

There is a synergistic interaction between classification and regression approaches for predicting residential energy use. Energy patterns are classified into categories like "Low," "Medium," or "High," reflecting various usage patterns. By considering variables like time, date, and measuring their influence on use, regression then makes predictions about numerical energy consumption figures. This combination strategy optimizes energy forecasts for effective resource management and well-informed decision-making by validating insights, improving accuracy, and providing a thorough grasp of both category patterns and numerical projections.

The next step in this methodology is the selection of the best models. This necessitates careful selection in order to find the most efficient technique for both classification and regression tasks. After that save the classification and regression models as 'cls_model.pkl' and 'reg_model.pkl', respectively. The "Explainable AI" (XAI) technique is then put into practise. The ability to extract important factors that

significantly affect how effectively the models forecast the future is made feasible by XAI techniques by providing a thorough grasp of the internal workings of the selected models. This exploratory work not only aids in elucidating the variables affecting the classification and regression results, but it also enhances interpretability by offering meaningful data that aids in guiding future decision-making and improvement processes.

3.1 Data description:

The dataset under investigation in this research is sourced from Kaggle and pertains to Household Electric Power Consumption. (Kaggle, n.d.) This substantial dataset, enclosed within a compressed zip package, encompasses a voluminous compilation of 2,075,259 measurements meticulously gathered across the temporal expanse from December 2006 to November 2010, effectively spanning 47 months. For the purpose of this analysis, a meticulously curated subset comprising a total of one million measurements has been meticulously selected for comprehensive examination. The illustrative representation in the subsequent figure expounds upon the sample distribution and underlying structure of the dataset. Characterized by its multifaceted composition, the dataset encapsulates diverse electrical parameters alongside submetering readings, in addition to temporal attributes such as date and time. Embedded within this dataset are distinctive electrical attributes, including global voltage (V), global intensity (I), global active power (P_active), and global reactive power (P_reactive). Furthermore, the dataset incorporates submetering measurements such as Sub metering 1 (Sub1), Sub metering 2 (Sub2), and Sub metering 3 (Sub3), each contributing to the holistic portrayal of energy consumption patterns.

	Date	Time	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
0	16/12/2006	17:24:00	4.216	0.418	234.84	18.4	0.0	1.0	17.0
1	16/12/2006	17:25:00	5.360	0.436	233.63	23.0	0.0	1.0	16.0
2	16/12/2006	17:26:00	5.374	0.498	233.29	23.0	0.0	2.0	17.0
3	16/12/2006	17:27:00	5.388	0.502	233.74	23.0	0.0	1.0	17.0
4	16/12/2006	17:28:00	3.666	0.528	235.68	15.8	0.0	1.0	17.0

Fig 3.2: Features of Household consumption Data

3.1.1 Exploratory data analysis:

Any series of observations made at certain periods and often displayed as a time series plot are referred to as a time series. The observations in this graph are represented as a time function. A discrete set may exist for the set of periods T at which recorded power consumption data may be found. In this scenario, results are presented at random intervals or continuously, just as when data is continually gathered. The basic goal of time series analysis is to forecast the future values of the series by identifying a hidden signal, figuring out how the data is produced, simulating the series independent realisations, and predicting how the series will behave in the future.

There are two primary categories for time series forecasting. The first is a multivariate analysis, which looks at two or more variables over time, while the second is a univariate study, which looks at a single variable over time. Based on historical values, time series forecasting is used to make predictions about the future.

3.2 Data preprocessing

In the next stage the dataset goes through the following processes for Preprocessing:

1. Missing value imputation to handle null values in the dataset.
2. Feature engineering to make the data easier to understand.
3. Data Splitting
4. Data Scaling to improve the performance, stability, and interpretability of machine learning models.

3.2.1 Missing Value Imputation

The next crucial stage is the expert handling of null values inside the dataset after the process of data importing. A graphical depiction fig 3.3 that clearly illustrates the existence of a significant percentage of missing values are approximately 1.25%. Notably, even though the dataset fully captures a chronology of historical timestamps, it is sometimes clear that certain timestamps do not correlate to the corresponding measured values. The presence of a "?" character, which essentially acts as a stand-in placeholder inside the timestamp column, visually illustrates this shortcoming. The presence of missing data points is further emphasised by the fact that the identical placeholder appears in the dataset as "NaN".

In order to fulfil the analytical aims, a deliberate method is utilized to decrease the possible effect of these missing variables on the prediction model. This method

comprises deleting null values to establish an environment appropriate for accurate prediction modelling.

```
df.isnull().sum()
```
✓ 0.8s

```
Date                       0
Time                       0
Global_active_power     3990
Global_reactive_power   3990
Voltage                 3990
Global_intensity        3990
Sub_metering_1          3990
Sub_metering_2          3990
Sub_metering_3          3990
dtype: int64
```

Fig 3.3: Sum of Null Value before removing null values.

3.2.2 Feature Engineering:

For an improved grasp of the data and enhanced outcomes, it is beneficial to make certain updates to the dataset, such as adding or removing columns from the table. (Wilfling, 2023). In the current scenario, the dataset solely comprises electricity consumption values for each minute for 4 years. In order to attain more robust results, a data frame is constructed that encompasses the total consumption and average consumption of electricity alongside corresponding dates. This supplementary data frame is subsequently appended to the original dataset.

Time-based feature engineering is essential in consumption of electricity because it allows us to capture temporal patterns and relationships within data. The consumption of energy demonstrates daily, weekly, and seasonal changes, which are impacted by things like the time of day, day of the week, month, weekend , day and many more. Some more features are also included for enhancing the accuracy and comprehensiveness of energy consumption prediction:

1. Active energy consumption calculation: An accurate depiction of the overall energy spent inside the home is made possible by the implementation of the

"Active_energy_consumption" (G_ActivePower) feature, which aggregates the active power consumption across sub-meters. This feature encapsulates energy utilization more effectively than using individual power readings from different sub-meters. The formula is shown below:

$$G\_ActivePower = \sum_{k=0}^{n} P\_active_k * (1000/60) - Sub1_k - Sub2_k - Sub3_k \quad (3.2)$$

2. Rolling_mean_power : This feature helps to more easily discover trends over short time periods by calculating the rolling mean of "Global_active_power" over a 6-hour time period. This feature assists by eliminating noise and revealing underlying consumption patterns that may not be seen in raw data.

$$Rolling\_mean\_power_i = \frac{1}{w} \sum_{k=i-w+1} P\_active_k \quad (3.3)$$

Here 'w' is the window size=6

An efficient method for pattern recognition, anomaly detection, and energy use optimization in energy management is the application of rolling averages to data on energy consumption.

3. Derived ratios and differences: The creation of derived features like "Active_power_ratio," "Power_factor," and "Power_diff" encapsulates intricate connections between active power, voltage, and other variables. The efficiency, power quality, and power relationships within the household's energy usage are shown by these ratios and discrepancies. The equations of these are followed:

$$Active\_power\_ratio_i = \frac{P\_active_i}{V_i} \quad (3.4)$$

$$Power\_factor_i = \frac{P\_active_i}{(V_i * I_i)} \quad (3.5)$$

$$Power\_diff_i = P\_active_i - P\_Reactive_i \quad (3.6)$$

4. Voltage Characteristics: Calculating numerous voltage-related characteristics, including "Voltage_mean," "Voltage_std," "Voltage_range," "Voltage_min," and "Voltage_max," offers a thorough knowledge of voltage behaviour. These

traits represent the stability, volatility, and range of the voltage, which may have an effect on total energy use and power management.

$$\text{Voltage\_mean}(\mu) = \frac{V_i + Sub1_i + Sub2_i + Sub3_i}{4} \quad (3.7)$$

$$\text{Voltage\_std} = \sqrt{\frac{1}{4}(V_i - \mu)^2 + (Sub1_i - \mu)^2 + (Sub2_i - \mu)^2 + (Sub3_i - \mu)^2} \quad (3.8)$$

Where μ is Voltage_mean

$$\text{Voltage\_range}_i = \text{Voltage\_mean}_i - \text{Voltage\_std}_i \quad (3.9)$$

$$\text{Voltage\_min}_i = \min(V_i, Sub1_i, Sub2_i, Sub3_i) \quad (3.10)$$

$$\text{Voltage\_max}_i = \max(V_i, Sub1_i, Sub2_i, Sub3_i) \quad (3.11)$$

When analysing electricity consumption, feature engineering is used to uncover insights via complex relationships and temporal patterns to improve prediction accuracy for improved energy management and decision-making.

```
# Feature Engineering
# Time Based Features Engineering

# Add a new column for the hour of the day
df["Hours"] = df["datetime"].dt.hour

# Create a new column for the day of the week
df["Day"] = df["datetime"].dt.weekday

# Create a new column for the month of the year
df["Month"] = df["datetime"].dt.month

# Create a new column for the day of the month
df["Day_of_month"] = df["datetime"].dt.day

# Create day of the week feature (0=Monday, 6=Sunday)
df['Day_of_week'] = df['datetime'].dt.dayofweek
# Calculate the active energy consumption
df["Active_energy_consumption"] = df["Global_active_power"] * 1000 / 60 - df["Sub_metering_1"] - df["Sub_metering_2"] - df["Sub_metering_3"]

# Create binary feature for weekend
df['is_weekend'] = (df['Day_of_week'] >= 5).astype(int)

# Calculate rolling mean of global active power over 3-hour window
df['rolling_mean_power'] = df['Global_active_power'].rolling(window=6).mean()

# Additional Feature Engineering
df["Active_power_ratio"] = df["Global_active_power"] / df["Voltage"]
df["Power_factor"] = df["Global_active_power"] / (df["Voltage"] * df["Global_intensity"])
df["Power_diff"] = df["Global_active_power"] - df["Global_reactive_power"]
df["Voltage_mean"] = df[["Voltage", "Sub_metering_1", "Sub_metering_2", "Sub_metering_3"]].mean(axis=1)
df["Voltage_std"] = df[["Voltage", "Sub_metering_1", "Sub_metering_2", "Sub_metering_3"]].std(axis=1)
df["Voltage_range"] = df["Voltage_std"] - df["Voltage_mean"]
df["Voltage_min"] = df[["Voltage", "Sub_metering_1", "Sub_metering_2", "Sub_metering_3"]].min(axis=1)
df["Voltage_max"] = df[["Voltage", "Sub_metering_1", "Sub_metering_2", "Sub_metering_3"]].max(axis=1)
```

Fig 3.4: Additional features

The exclusion of columns such as "Date," "Time," and "Datetime" from this dataset appears to be a deliberate action aimed at eliminating redundant temporal information.

3.2.3 Data Splitting

After data preprocessing the dataset was randomly divided into separate training and testing subsets for this experiment. To be more precise, 80% of the dataset is allotted to the training subset in order to assist model training, while the remaining 20% is held back for the testing subset to enable model assessment. The size of the dataset after splitting that shown below in fig 3.5:

```
X_train shape: (796804, 22)
X_test shape: (199201, 22)
y_train shape: (796804,)
y_test shape: (199201,)
```

Fig 3.5: splitting dataset

3.3.4 Data Scaling

Most machine learning algorithms initially examine a set of characteristics that are presented to them before generating predictions. These algorithms often compute the spacing between data points to gain more accurate conclusions from the data. The approach is more likely to execute effective and rapid training when feature values demonstrate closeness to one another, as opposed to datasets where feature values or data points show substantial dissimilarities. In these instances, the algorithm takes extra time to analyse the data, which often leads to less accuracy in the final model's predictions. (Lee et al., 2019)

The dataset includes anomalous information associated to very high-power usage. To determine the probability density function for each feature in the dataset and show the results in Figure 3.6 to highlight the breadth of the features in the dataset. Notably, a Gaussian function with a centre of 240 volts may be used to simulate the voltage values (V) in a normal environment, which generally vary between 230 and 250 volts.

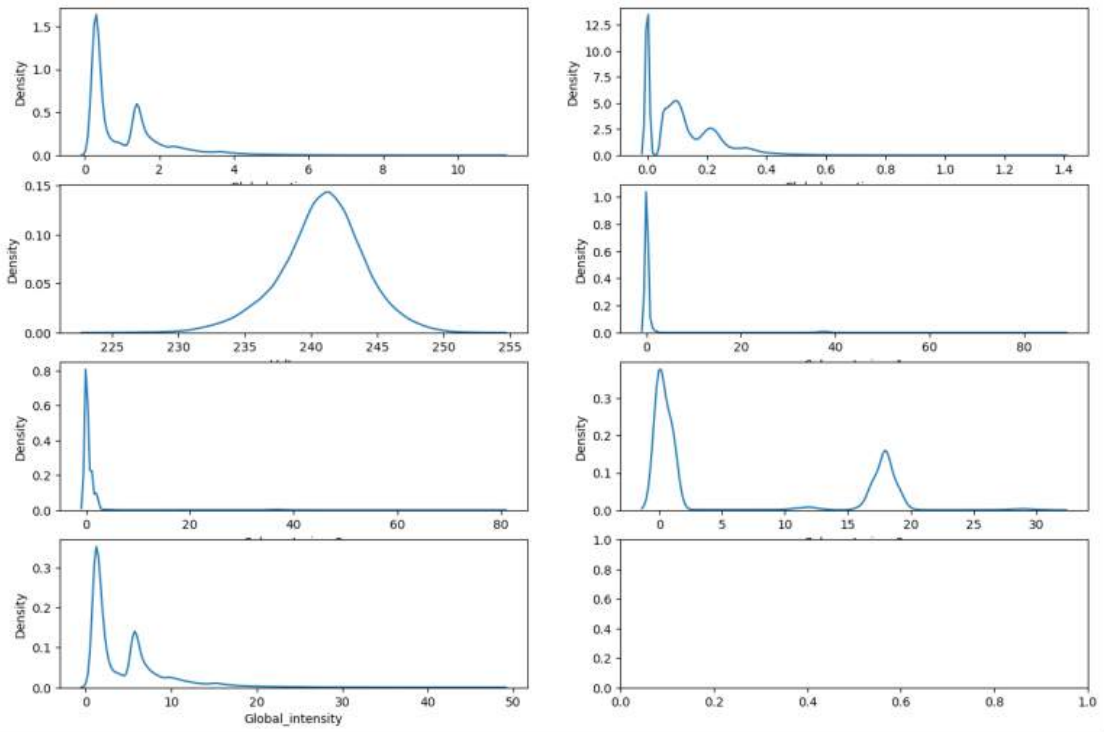

Fig 3.6: Probability density function of every feature in dataset.

As illustrated in Figure 3.6, large differences in features were observed in the dataset, including a variety of values and units, with the voltage feature standing out significantly. Standardisation, one of the scaling procedures, was utilised for this research to resolve the dataset's significant differences in magnitudes, values, or units.

Standardization gives a cure to the dataset's clear inequalities in magnitudes, values, and units. This strategy concentrates on changing the data in a way that aligns their distributions by subtracting the mean(μ_p) and dividing by the standard deviation (σ_p)(Ellen et al.,2021). Consequently, each feature is re-scaled to contain a mean of zero and a standard deviation of one. This method promotes comparability and interpretability of the data by minimising the impact of differing scales. The formula is also presented here:

$$z = \frac{x - \mu_p}{\sigma_p} \tag{3.12}$$

Following the standardization process, the dataset is rendered more amenable to subsequent analyses, promoting a comprehensive exploration of relationships among the features, thus contributing to robust modelling and insightful conclusions.

3.3 Machine Learning Model:

A variety of supervised machine learning models were trained for the goal of conducting a robustness evaluation once the preprocessing and data splitting processes were finished. This included both the use of regression and classification models.

- **Multiple Linear regression:**

Linear regression (LR) stands as one of the most renowned and familiar models within both the realms of statistics and machine learning (ML). In statistics, linear regression has been meticulously developed and examined to elucidate the interplay between numerical input variables and output variables, a concept that has been embraced by the field of ML.(Solyali. 2020) multiple input variables (x) are involved, the endeavour takes on the name of multiple linear regression (MLR). Consequently, the MLR model endeavours to delineate how a sole response variable (y) is contingent upon multiple predictor variables (x) through linear dependence. In our particular investigation, the presence of multiple input variables (x) designates our approach as the MLR model. (Rambabu et al., 2022) The equation is below:

$$y=\theta_0+\theta_1 * x_1+\theta_2 * x_2+\ldots+\theta_n * x_n \quad (3.13)$$

Where y = response variable, x = features, n = number of features, θ_n= nth feature weight.

In multiple regression modelling, the least-squares technique is typically employed as an estimate method. (Tso et al., 2007). This regression technique is frequently chosen since they are straightforward to apply and offer a clear understanding of the model's parameter meanings. (Amir et al, 2020)

Advantages	Disadvantages
Renowned for parameter interpretability and user-friendliness. (Amir et al, 2020)	to establish correlations but not causal mechanism. (Tso et al., 2007)

Relatively precise forecasts when the foundational assumptions are met.(Solyali ,2020)	Multicollinearity prevents separating individual variable effects. (Amir et al, 2020)

Table 3.1: MLR Advantages and Disadvantages

- **Ridge Regression**

Ridge Regression is a regularised form of Linear Regression in which the cost function is provided in equation 3.14. It can also be referred to as Tikhonovr egularization. (Rambabu et al, 2022)

$$\alpha \sum_{k=0}^{n} \theta_k^2 \quad (3.14)$$

The Learning algorithms have a duty to adjust to the data and keep weights as low as possible as necessary. The cost function should only be subjected to regularisation during the training phase, it is vital to highlight. The unregularized performance metric may be used to evaluate the model's performance once it has been trained in accordance with equation (3.15). (Rambabu et al, 2022)

$$f(\theta) = MSE(\theta) + \alpha/2 \sum_{k=0}^{n} \theta_k^2 \quad (3.15)$$

Advantages	**Disadvantages**
By using regularisation, it combats overfitting. (Rambabu et al, 2022)	it doesn't produce zero coefficients might be a difficulty when removing pointless variables. (Morgul et al, 2021)
Effectively addresses multicollinearity by lowering coefficient sizes. (Morgul et al, 2021)	It performs poorly when the true relationship is highly nonlinear. (Rambabu et al, 2022)

Table 3.2 Ridge regression Advantages and Disadvantages

- **Lasso Regression**

Another type of regularised linear regression is the Least Absolute Shrinkage and Selection Operator Regression (Lasso Regression), often known as L1 regularization.

Similar to Ridge Regression, it incorporates a regularisation parameter into the cost function. However, Lasso Regression acts on the weight vector and uses the l1 norm rather than the square of the l2 norm. When working with high-dimensional datasets, LASSO regression is essential because it can efficiently reduce overfitting and provide models that are easy to understand. The equation is below (3.16):

$$f(\theta) = MSE(\theta) + \alpha \sum_{k=0}^{n} \theta_k^2 \quad (3.16)$$

- **Decision Tree**

A supervised learning tree-structure-based support tool known as a decision tree is utilized in both classification and regression models. (Tso et al., 2007) By stratifying areas of predictor space that don't overlap, one may theoretically estimate the value of the output variable using a reversed tree-like structure having leaves at the bottom. The inverted structure is characteristic of decision trees, and internal nodes are those that divide a space into sub-spaces. Branches are the pieces that link the nodes (Amir et al., 2020)

There are a number of decision tree algorithms, including C4.5, C5.0, and CART (Classification and Regression Trees), ID3 (Iterative Dichotomiser 3), and others. (Salam et al.,2018) Because CART closely resembles C4.5 and can support numerical target variables for regression tasks without needing the construction of rule sets. This particular scenario, the choice of the CART algorithm is driven by its suitability. By choosing the feature and threshold that maximise information gain at each node, the CART algorithm builds binary trees

Advantages	**Disadvantages**
It uses in both regression and classification model (Tso et al., 2007; Amir et al., 2020)	Can be computationally expensive (Mohamed et al., 2017)
handle outliers and missing data without difficulty. (Mohamed et al., 2017)	Prone to overfitting on the training data. (Amir et al., 2020)
The implementation process is simple. (Mohamed et al., 2017)	extended training period

Table 3.3 DT Advantages and Disadvantages

- **Random Forest**

The random forest ensemble method makes use of the combined power of several decision tree algorithms to forecast variable values. This technique may be used for both classification and regression problems to increase the accuracy of decision trees. (Salam et al.,2018) The maximum samples for Random Forests, which are effectively an ensemble of Decision Trees, are determined by the quantity of the training data. The Random Forest technique, in contrast to the search of the optimum feature, adds more unpredictability by choosing the best feature from a pool of features that were created arbitrarily. This results in the growth of additional trees, which raises bias while lowering variance. (Rambabu et al, 2022)

The average is produced by random forest after the construction of k regression trees. As k trees are developed, the following are considered regression predictors:

$$f(x) = 1/k \sum_{n=1}^{k} T(x) \quad (3.17)$$

Where x is p-dimensional vector and T(x) is decision tree

A subset of trees from the forest is randomly picked to create a new training set. The out-of-bag samples are made up of a group of trees that were not chosen. At each split node of a decision tree, random features are chosen rather than all features being taken into account. To create a random forest, this process is repeatedly repeated . The forecast of the random forest is the sum of the predictions from each individual tree, and it performs better than individual tree predictions (Salam et al.,2018).

Advantages	**Disadvantages**
The implementation process is simple. (Salam et al., 2018)	Size of the produced models is large. (Amir et al., 2020)
Efficiently operates on extensive databases. (Rambabu et al, 2022)	In correlated feature situations, it randomly picks one characteristic as a predictor,

	lowering the value of the others. (Amir et al., 2020)

Table 3.4 RF Advantages and Disadvantages

- **Extreme Gradient Boosting (XGBOOST) regression**

Gradient boosting machines are approached in a novel and creative way with Chen and Guestrin's technique known as XGBOOST, especially for building regression and classification trees.(Chen et al., 2016). Recently, XGBoost has won in Kaggle events centred on structured or tabular data. Gradient-boosted decision trees may be quickly and effectively implemented using XGBoost, which excels in both speed and performance. (Rambabu et al., 2022). In order to develop a strong learner through repetitive training methods, XGBOOST works on the notion of "boosting," which combines the predictions of weak learners. This method improves computing efficiency while simultaneously reducing overfitting. (Fan et al., 2018.). This is accomplished while maintaining the fastest feasible processing performance by simplifying the goal functions that allow the prediction and regularisation terms to be combined. The XGBOOST training procedure smoothly incorporates automatic parallel estimate. The approach uses additive learning techniques, starting with a learner that was first trained using the whole input dataset. The second model is then trained using the residuals to correct the weak learner's flaws. Up until the stopping requirement is met, this iterative process is repeated several times. By adding the predictions from each individual learner, the model's final prediction is determined. (Chen et al, 2016) The following equations (3.18) define the total prediction function at this phase:

$$f_i^n = \sum_{j=1}^{n} f_k(x_i) = f_i^{n-1} + f_n(x_i) \tag{3.18}$$

Here, f_i^n = prediction of phase n, x_i = input variable and $f_n(x_i)$= learner phase of i

The XGBoost model formulates the following analytical expression to evaluate the model's quality based on the original function in order to solve overfitting without compromising computational speed:

$$obj(\theta) = \sum_i^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (3.20)$$

Here l = lost function and Ω= regularization parameter

Advantages	Disadvantages
It represents optimal computational speed. (Rambabu et al., 2022) (Fan et al., 2018)	Complex model (Fan et al., 2018)
high-performance (Rambabu et al., 2022)	

Table 3.5 XGBOOST Advantages and Disadvantages

- **Logistic regression**

One of the most often used machine learning methods is logistic regression. (Bashir et al., 2021). Logistic regression is designed to handle classification jobs where the output variable has categorical data. This approach entails carefully shaping a logistic function or curve to fit the properties of the dataset, then forecasting the output.

The input data "X" which contains characteristics like x0, x1, x2,..., xn and corresponding weights "W" which includes w0, w1, w2,..., wn are added together to calculate the value of "Z" in the classification technique of logistic regression. Equation 3.21, which offers the formula for determining "Z," serves as an illustration of this procedure. (Zou et al., 2019)

$$Z = w_0x_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n \quad (3.21)$$

Advantages	Disadvantages
It's computationally efficient and works well with huge datasets. (Zou et al., 2019)	It might have problems finding complicated, nonlinear relationships in the data.(Zou et al., 2019)

Table 3.6 Logistic Regression Advantages and Disadvantages

- **Support Vector Machine**

Support Vector Machines (SVM) are often used in research to address classification and regression issues. Researchers chose SVM because of its potential to obtain tremendous levels of accuracy while consuming little in the way of computer resources. (Bashir et al., 2021) It was initially presented in the late 1960s, but it wasn't given much attention until lately. (Salam et al.,2018)

Its primary method is finding the best line or hyperplane that closely follows the data and minimises the difference between expected and actual values. As shown in figure 3.7, these margin lines may be changed such that one extends to the nearest positive position and the other to the nearest negative position (Bashir et al., 2021). SVM makes use of a kernel function's capacity to transform the input data into a higher-dimensional space, making it easier to capture complex correlations. (Zeng et al., 2019) Notably, the method includes regularisation settings to prevent inclinations towards overfitting. The SVM process involves carefully choosing the best kernel function and regularisation parameters, training the model, and then producing predictions for new data. SVM is a strong algorithm that can successfully handle a variety of regression jobs.

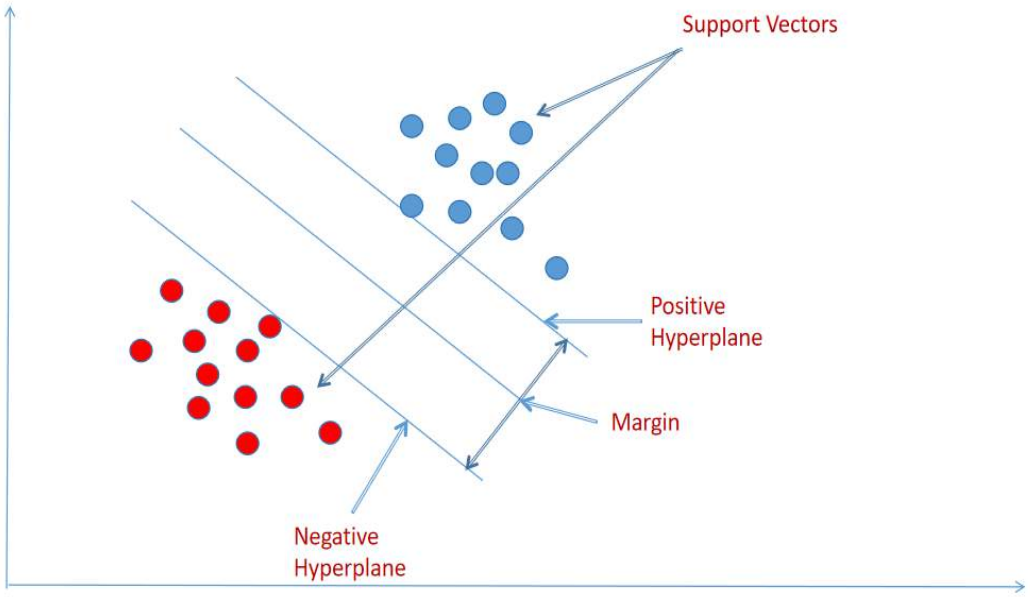

Fig: 3.7 Diagram of Support Vector Machin

Support Vector Machine is being utilised more frequently in research as a result of its efficiency in solving non-linear issues with diverse data sizes. (Zhao et al, 2012)

Advantages	Disadvantages
Require a minimal computational power (Bashir et al., 2021)	It lies in their interpretability.(Mohamed, 2017)
It works efficiently with high-dimensional data. (Mohamed, 2017)	It incur significant computational costs.(Solyali, 2020; Fan et al., 2018)
handle both continuous and categorical data (Bashir et al., 2021)	Its non-parametric nature results in limited transparency in its outcomes. (Mohamed, 2017)
Nonlinear relationships within the data are captured. (Salam et al.,2018)	

Table 3.7 SVM Advantages and Disadvantages

- **K-nearest Neighbour**

The machine learning technique known as KNN, or K-nearest Neighbours, may be used for both classification and regression problems. KNN relies on similarity metrics like Euclidean, Manhattan, and Minkowski distances. (Bashir et al., 2021) It compares each new data point with existing ones using these distance measures, and the closest sample is employed to classify the new data point. The choice of the K value has fundamental relevance due to its large impact on algorithm performance. Selecting a low K value might result in overfitting since the model can become overly sensitive to data noise. On the other hand, an extremely high K value can cause the model to be oversimplified and miss the inherent patterns in the data. (Mohamed, 2017) The following ((3.22), (3.23), (3.24)) are the equations for k-NN:

$$Eculidean\ equation = \sqrt{\sum_{k=1}^{n} (a_k - b_k)^2} \quad (3.22)$$

$$Manhattan\ equation = \sum_{k=0}^{n} |a_k - b_k| \quad (3.23)$$

$$Minkowski = \left(\sum_{k=0}^{n} (|a_k - b_k|)^r \right)^{1/r} \quad (3.24)$$

In the proposed approach, k was set to 5 with a default leaf-size of 30.

Advantages	Disadvantages
Straightforward and readily implementable	Requires a substantial amount of memory to accommodate all the training examples.
Training proceeds at a rapid pace with zero associated cost.	The computational expense associated with it is quite significant.
It has the capability to handle noisy data effectively.	

Table 3.8 KNN Advantages and Disadvantages(Mohamed, 2017)

3.4 Comparison of Performance Metrics:

At this stage, evaluate the classification models' performance using measures like as accuracy, precision, recall, and F1 score. Analyse the regression models simultaneously using measures such as R-squared (R2), mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), coefficient of correlation(CC), relative absolute error(RAE), and root relative squared error(RRSE).

3.4.1 Performance Metrics for Regression Model:

In the context of the regression task requiring the prediction of continuous values within a predefined range for energy consumption, numerous metrics, including MAE, MAPE, MSE, and RMSE, are often applied to analyse prediction results. Many papers deliberately highlight measures that show off their method's advantages while ignoring

others in order to overcome possible reporting biases in algorithm performance. However, a more thorough methodology has been used in this research. To assess and contrast various models and to provide a more comprehensive view of model performance, a range of well-established measures have been used. The analysis has seven assessment measures, which will be covered in the paragraphs that follow.

In this context, "n" represents the number of data points,y_k stands for the actual value, ȳ stands for the average value and y'_k denotes the predicted value.

- **Mean Absolute Error (MAE)**

The mean absolute error (MAE) is a key regression error measure to understand. The residual is calculated for each data point by taking into account just the absolute value. This strategy stops positive and negative residuals from cancelling each other out. The average of each of these residuals is then determined. In essence, MAE illustrates the usual residual size. The following is how MAE is expressed mathematically:

$$MAE = \frac{1}{n}\sum_{k=0}^{n} |y_k - y'_k| \quad (3.25)$$

- **Mean Square Error (MSE)**

The average square of the difference between the actual and anticipated data is used to compute mean square error. The mathematical illustration is displayed as follows.

$$MSE = \frac{1}{n}\sum_{k=0}^{n} (y_k - y'_k)^2 \quad (3.26)$$

- **Mean Absolute Percentage Error (MAPE)**

The mean absolute percentage error (MAPE) is the same as the mean absolute error (MAE), expressed as a percentage. It is a statistical evaluation of a system's precision. The following is the mathematical representation:

$$MAPE = \frac{1}{n}\sum_{k=0}^{n} \left|\frac{y_k - y'_k}{y_k}\right| * 100 \quad 3.27$$

- **Root Mean Squared Error (RMSE)**

The mean square error's square root is referred to as the root mean square error. the average standard deviation of the mistakes made when making a forecast based on a dataset.(Xiang et al., 2020) The following is a mathematical representation:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{3.28}$$

- **Decision Coefficient (R2 score)**

R-squared (R2) is a measure of how much of the variance in the dependent variable can be ascribed to how the independent variable is interpreted. R2 typically tends to be higher in the context of a regression model when the evaluation metric RMSE is less. This correlation means that a higher R2 indicates a more effective model. The following is the mathematical representation:

$$R^2 = 1 - \frac{\sum_{k=0}^{n} (y_k - y'_k)^2}{\sum_{k=0}^{n} (y_k - \bar{y}_k)^2} \tag{3.29}$$

- **Correlation Coefficient**

The following equation (3.30) is the Pearson correlation coefficient (CC) is calculated for any model:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{3.30}$$

Where y_i and x_i represents the actual observed value in the data, ŷ and x’ signifies the forecasted value generated by the model, and n denotes the sample number.

In essence, the value of r_{xy} comprises from -1 to 1, with a value of 1 signifying that a linear equation can correctly reflect the link between X and Y. In this scenario, every data point lines up perfectly along a line. The regression slope defines the correlation's

sign (- or +), where a + implies that Y increases as X rises and a - signifies the reverse. When r_{xy} is equal to 0, it signifies that there is no link between X and Y. Depending on the context and aims of the experiment, numbers closer to 1 or -1 imply a more robust model. Intermediate values (i.e., 0 <r_{xy}< 1 and -1 <r_{xy}< 0) reveal partial correlations. (Cebekhulu et al., 2022)

- **Relative Absolute Error (RAE)**

The ratio of the cumulative absolute error obtained from a model to the cumulative absolute error of a simple predictor where the basic predictor is simply the mean of the target values that is used to estimate the relative absolute error (RAE). The RAE is determined in the following method mathematically:

$$\mathrm{RAE} = \frac{\sum_{k=1}^{n} |y'_k - y_k|}{\sum_{k=1}^{n} |\bar{y} - y_k|} * 100\% \tag{3.31}$$

A benefit of the RAE measure, in contrast to the root mean square error (RMSE), resides in its impartial handling of each mistake. It does this by solely examining the absolute value of mistakes, without squaring them. As a consequence, the RAE becomes especially beneficial for examining systems that are resilient and unaffected by outliers, making it a superior option over RMSE in such circumstances. (Cebekhulu et al., 2022)

- **Root Relative Square Error**

The root relative square error (RRSE) is defined as the square root of the total squared errors divided by the total squared errors of a simple predictor. Once again, the average of the target values serves as the simple predictor. The provided RRSE is.

$$\mathrm{RRSE} = \sqrt{\mathrm{RAE} = \frac{\sum_{k=1}^{n} (y'_k - y_k)^2}{\sum_{k=1}^{n} (\bar{y} - y_k)^2} * 100\%} \tag{3.32}$$

The RRSE lowers the error to a similar magnitude range as the RAE by calculating the square root of the relative squared error. But unlike the RAE, the RRSE penalises

outliers with large error values, making it straightforward to discover models with plausible outliers. (Cebekhulu et al., 2022)

3.4.2 Performance Metrics for Classification Model:

To evaluate the classification model, accuracy, precision(P), recall(R), and F1 score will be utilised as the metrics. Equations (3.33) through Equations (3.36) outline the formulas for Accuracy, Precision, Recall, F1 score, and Precision.

$$Accuracy = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \quad (3.33)$$

$$Precision(P) = \frac{Tp}{Tp + Fp} \quad (3.34)$$

$$Recall(R) = \frac{Tp}{Tp + Fn} \quad (3.35)$$

$$F1\ Score = \frac{2 * P * R}{P + R} \quad (3.36)$$

Where Tp= True Positive, Tn = True Negative, Fp= False positive, Fn=False Negative The proportion of accurately anticipated positives to all correctly predicted positives is known as precision. The proportion of accurately anticipated positive instances to all positive examples is known as recall, which is often referred to as sensitivity. According to Poh et al.'s, the F1 score is a balanced measure indication based on Precision and Recall.(Poh et al., 2019) Precision, Recall, F1 score, and Accuracy are computed for each model.

3.4.3 Save Model

After comparing all models using various performance metrics, they are saved as

"reg_model.pkl" and "cls_model.pkl" in the "model" folder for regression and classification tasks, respectively. In this research, the best model is selected based on the Root Mean Squared Error (RMSE) value for regression tasks and accuracy for classification tasks.

3.5 Model explanation

During the exploratory modelling phase, Explainability techniques that have received a lot of attention are utilised to establish model dependability. The final regression and classification models were easier to understand because to the use of the tree-based SHAP method in this instance. Techniques like summary charts, dependent plots, contributions plots, and others were utilised to measure the contribution of each input variable to the predictions.

3.5.1 SHAP:

Shapely values are used by SHapley Additive Explanation (SHAP) to describe the contribution of each feature to the prediction. (Lundberg et al, 2017). To determine how well each group of agents performs, SHAP uses ideas from coalitional game theory. This is how the SHAP method is described:

$$f(k') = \emptyset_0 + \sum_{n=1}^{N} (\emptyset_n k'_n) \tag{3.37}$$

Where $\emptyset_n$= attribute of nth feature, k' = coalition vector, N = Total no of features. In order to determine Shapley values, SHAP makes the assumption that certain feature values are assumed to be present while others are assumed to be missing. To compute SHAP value using following equation:

$$\phi_j = \sum_{S \subseteq Z \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{j\}) - f_x(S)] \tag{3.38}$$

Where S represents a feature subset and Z signifies the set of all input features, and (z' = 1) along with E [$f(x) \mid x_s$] represents the expected value of the function conditioned on the subset S of input features, the attribution of ϕj values to each feature using game theory can be construed as a value function of players within S.

The Python version of SHAP provides a tool for visualising each characteristic and its significance. Additionally, it functions with Python's Scikit-learn package's tree-based models. (Kuzlu et al., 2020)

3.6 Ethical Considerations

This study solely utilised publicly available datasets and previously published research outcomes. As part of this research effort, there were no surveys of participants from the outside. No private repositories or packages were included in the study's scope; all included packages and repositories were of a public nature. Although this study did not contain any human subjects or sensitive data, it is necessary to designate it as belonging to category 1 of research ethics. The formal request for ethical approval for this dissertation was made on 8th June 2023, and it was granted the same day. The research practise module of the programme's ethical requirements was rigorously observed throughout the period of this inquiry.

Chapter 4: Research Design and Implementation

4.1 Research Design:

This empirical research makes use of secondary data from Kaggle. The research proceeds from a rigorous phase of exploratory data analysis (EDA) to a later data processing phase before entering the domain of machine learning (ML) modelling. It's interesting that most research initiatives generally utilise regression or classification modelling approaches individually on their own datasets. This effort, however, is the first to concurrently apply classification and regression models to the same dataset. Regression is a predictive modelling tool that makes it simpler to estimate energy usage correctly based on historical data. For policymakers seeking exact estimates to properly distribute energy resources and apply regulatory measures, this method is particularly significant. The classification approach, however, allows the system to classify energy use into multiple levels, including high, medium, and low categories. For policymakers, this category facilitates the process by allowing them to estimate energy demands without having to go into actual numerical amounts.

The analysis of current literature indicates a considerable body of research devoted to the use of preset machine learning algorithms, frequently near to five algorithms, while applying a wide set of performance criteria for evaluation. The accuracy of predictions made using regression models is often evaluated using generally accepted metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2). However, this research differentiates itself by evaluating six machine learning algorithms and widens its evaluation beyond the traditional measures. It integrates additional variables, including processing time, the Coefficient of Correlation (CC), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Relative Root Mean Squared Error (RRSE), among others, to give a thorough assessment of model performance. This multidimensional approach provides a more thorough grasp of the efficacy and applicability of these models across varied situations and domains.

Following the deployment of machine learning models to the dataset, the selection of the optimum model in both regression and classification tasks is carried out.

Subsequently, Explainable Artificial Intelligence (XAI) approaches are utilised to extract key characteristics. This procedure assists to show the model's efficiency by emphasising the critical contribution of these recognised attributes.

4.2 Implementation of Exploratory Data Analysis

The data is a multivariate time series. There are different line plots for each of the eight variables in this study. This provides us with a very high level of four years of one-minute observations. According to uni-variant graph, some observations are following:

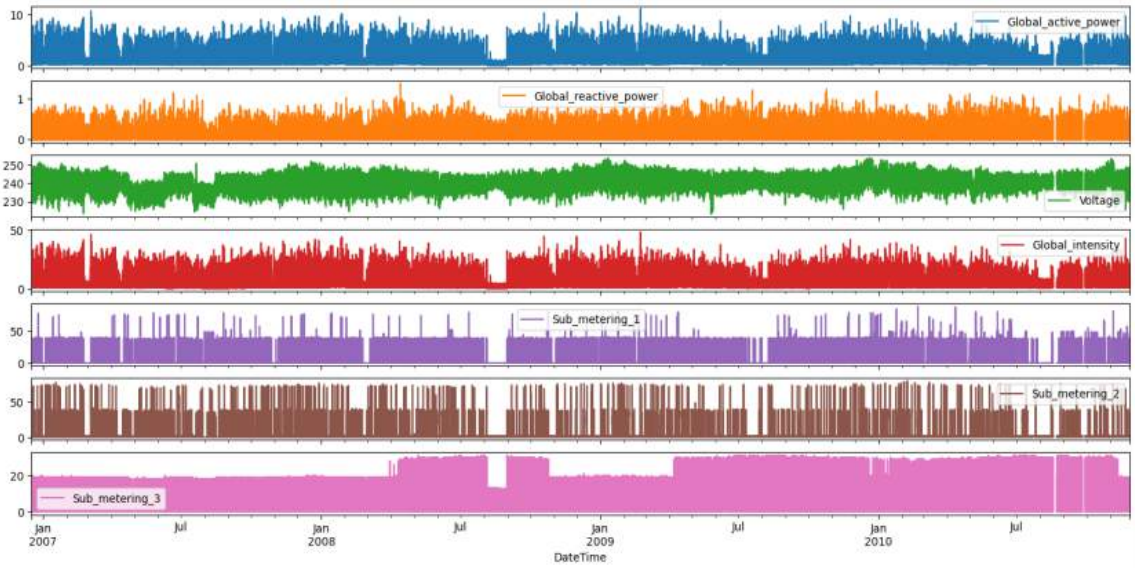

Fig 4.1 : Line plot for Global_active_power, Global_reactive_power, Voltage, Sub_metering1, Sub_metering2, Sub_metering3.

In the yearly trend of global active power, certain repeating trends are noticeable, such as significant drops in consumption around March and August months. In Global Reactive Power exhibits an annual seasonal trend but exhibits less volatility than global active power. Voltage has a repeating seasonal pattern across all years, with variation between 250 and 255 volts at the start of each year, dropping around July, and following an increase towards the end of the year. The Global Intensity does peak in the start of each year, then rises again towards the end. There are three sub-meters namely Sub_metering1, Sub_metering2 and Sub_metering3 are reflecting a different part of the house such as kitchen area, laundry area and electric water-heater & an air-conditioner respectively. A constant trend is observed in the Sub_metering_1 a feature over all

years, occasionally punctuated by spikes indicating increased kitchen usage. Like Sub_metering_1, data in Sub_metering_2 follow periodic patterns throughout the year, with spikes signifying intense laundry area usage during specific periods. In the case of Sub_metering_3, readings in 2007 show a consistent pattern, showing balanced use of the electric water heater and air conditioner over their respective seasons. In 2008, rising temperatures prompted an increase in utilisation of both devices around the middle of April. Temperature changes are significant and sustained from May 2009 to November 2010, resulting in substantial usage of both appliances during both the summer and winter seasons.

The time-wise pattern of total power usages describing the following Bar-graphs.

4.2.1 Yearly Analysis:

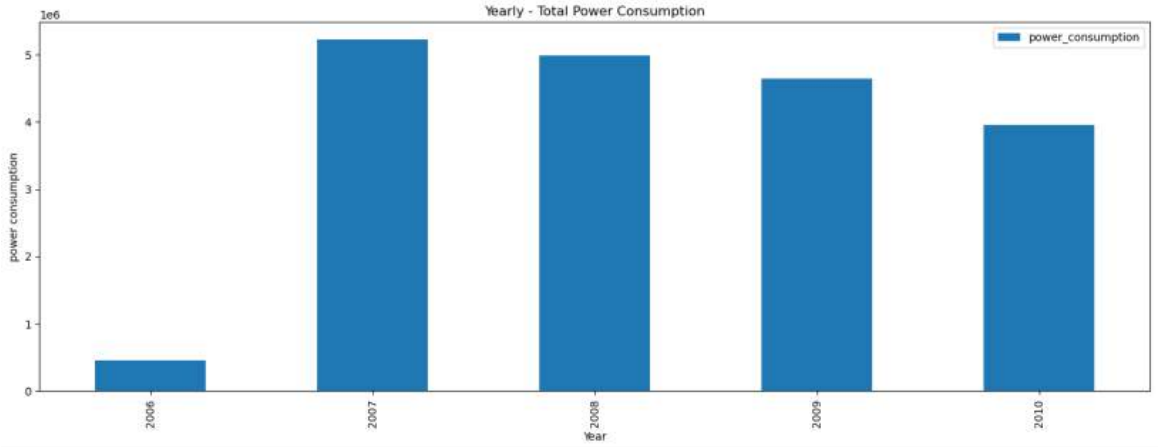

Fig 4.2: Year-wise Total Power Consumption

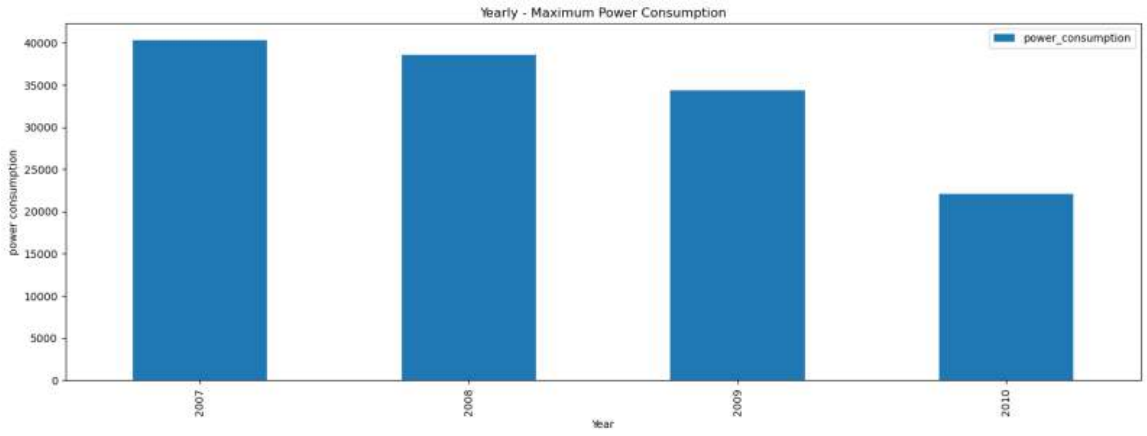

Fig 4.3: Year-wise Maximum Power Consumption

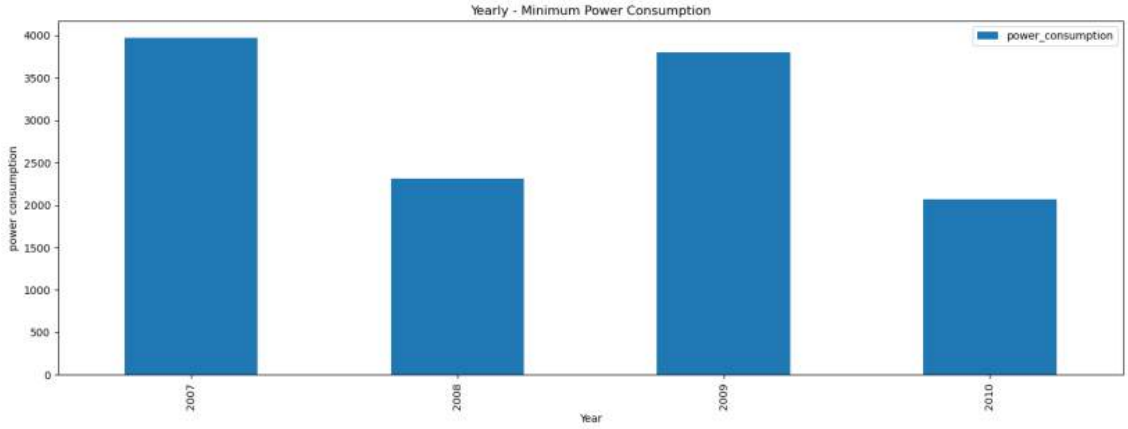

Fig 4.4: Year-wise Minimum Power Consumption

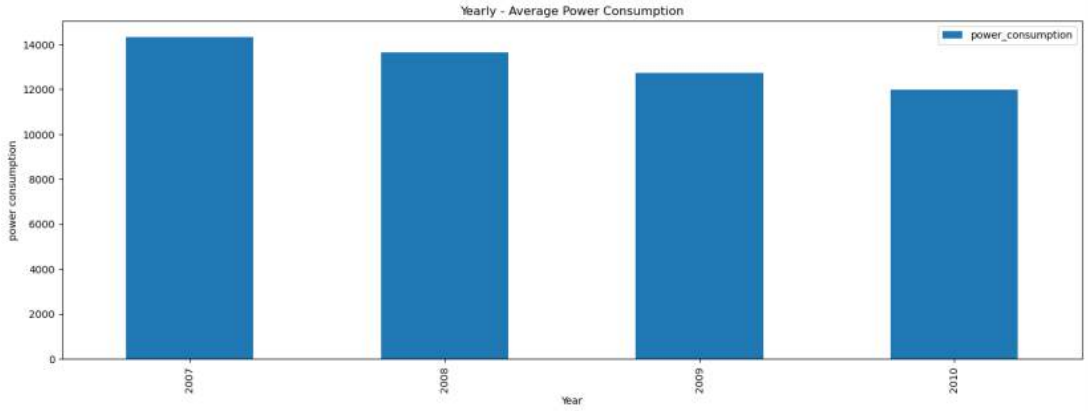

Fig 4.5: Year-wise Average Power Consumption

Due to the substantial difference in data availability, with only sixteen records in 2006 compared to over 330 records in previous years, the results for 2006 are not considered owing to their inconsistency with the data from the other years. The overall year-wise power usage ranges from 4,000 to approximately 6000 kilowatt-hours. The highest electricity usage was reported in the year 2007, while the lowest was recorded in 2010. Surprisingly, the average power usage remains quite steady overall years, ranging from around 12,000 to 15,000 watt-hours.

According to Liu et al.'s study, the UK's electricity usage has been increasing from 1993, reaching a high in 2005, and then consistently declining until 2019.(Liu et al.,2022). This pattern is also evident in this context, where the year 2007 registers the highest consumption and gradually declines until 2010.

4.2.2 Monthly Analysis:

Further zooming in on the total power consumption for each of the twelve months across the four-year period can be shown here.

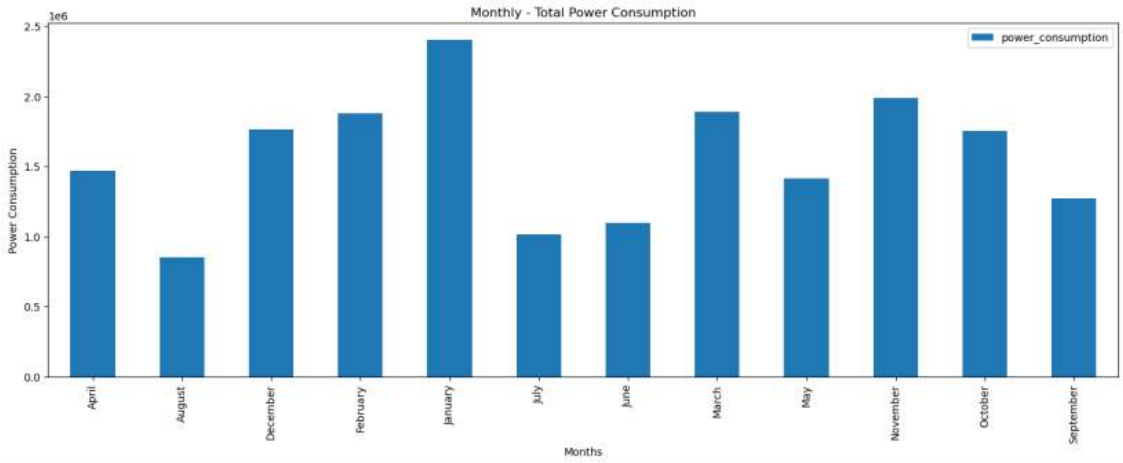

Fig 4.6: Month-wise Total Power Consumption

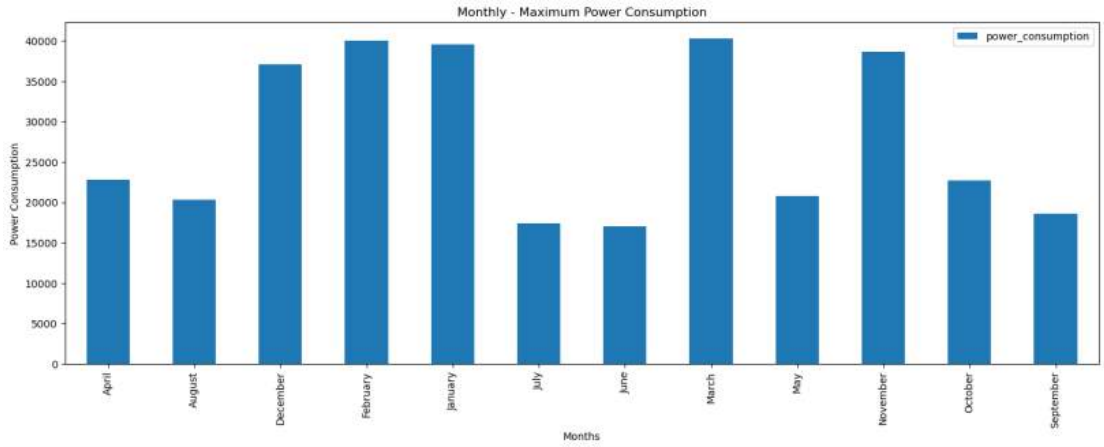

Fig 4.7: Month-wise Maximum Power Consumption

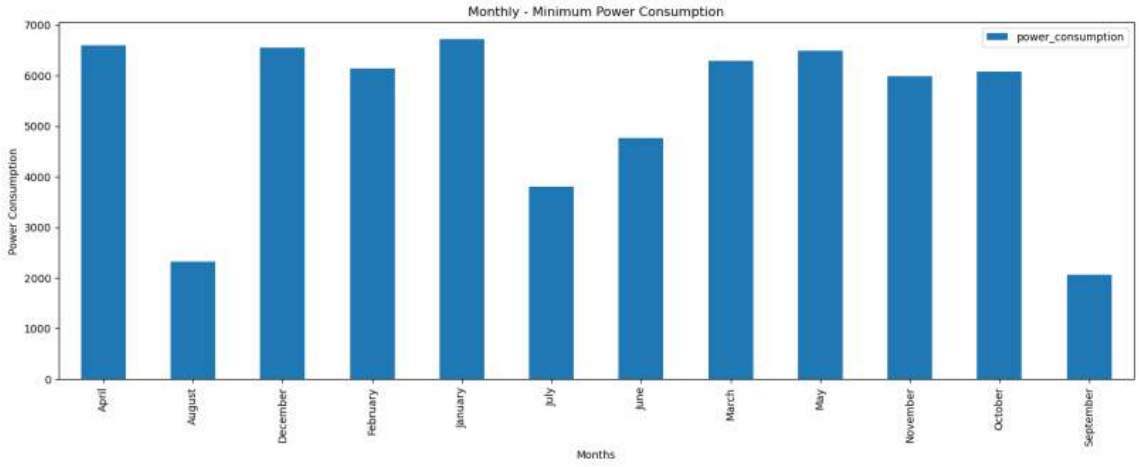

Fig 4.8: Month-wise Minimum Power Consumption

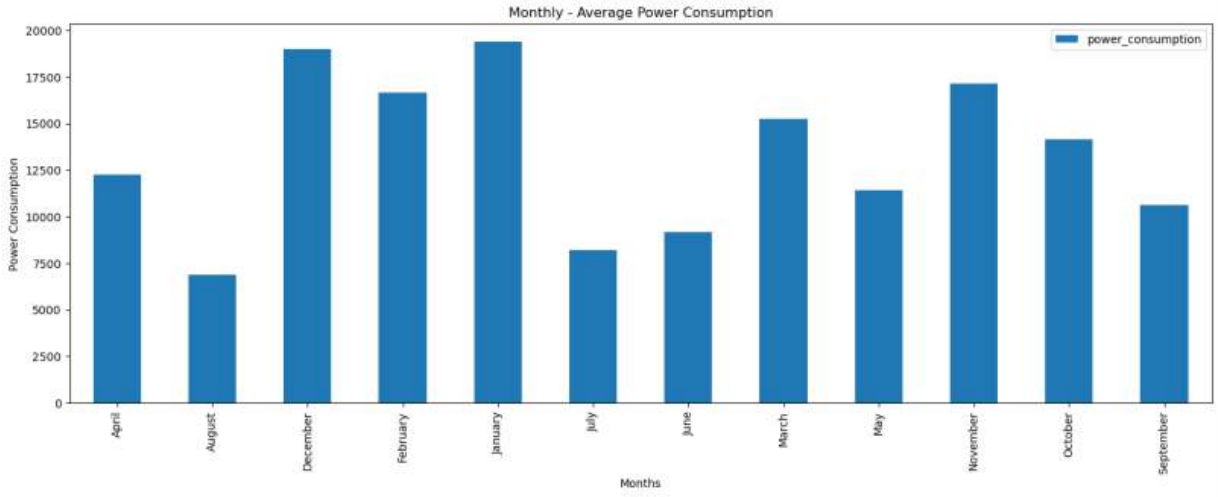

Fig 4.9: Month-wise Average Power Consumption

The overall power usage ranges between around 900 and 2400 kilowatt-hours. Across all years, September has the lowest power usage, averaging roughly 2000 watt-hours. The biggest power use, on the other side, is noted in January-February, with an average of roughly 61,000 watt-hours.

Over the course of a year, the average power usage ranges between 7000 and 20,000 watt-hours. When the Monthly Average Graph is examined, it is clear that power usage begins to fall about March, reaches its lowest point in August or September, and then progressively increases after that. Notably, there is a definite trend of reduced consumption during the months of June, July, and August. These times may correlate to vacations or times when the house was not occupied, resulting in much lower power use.

On the other side, increased consumption during the months of December, January, and February. This might be linked to the winter season. Because of the fewer daylight hours during the winter months, families frequently need more energy for heating and lighting. Furthermore, in certain cultures, the holiday season falls at this time, resulting in increased energy use for cooking, decorations, and hosting parties. All of these variables may have contributed to the reported rise in electricity use over these months. This contrasts with Zhang et al.'s study, which demonstrates the opposite pattern, indicating the highest electricity demand from May to October if the winter season is present.(Zhang et al.s, 2018)

4.2.3 Weekly Analysis:

Zooming in further, a closer examination of power consumption on a weekly basis within each year can be described here:

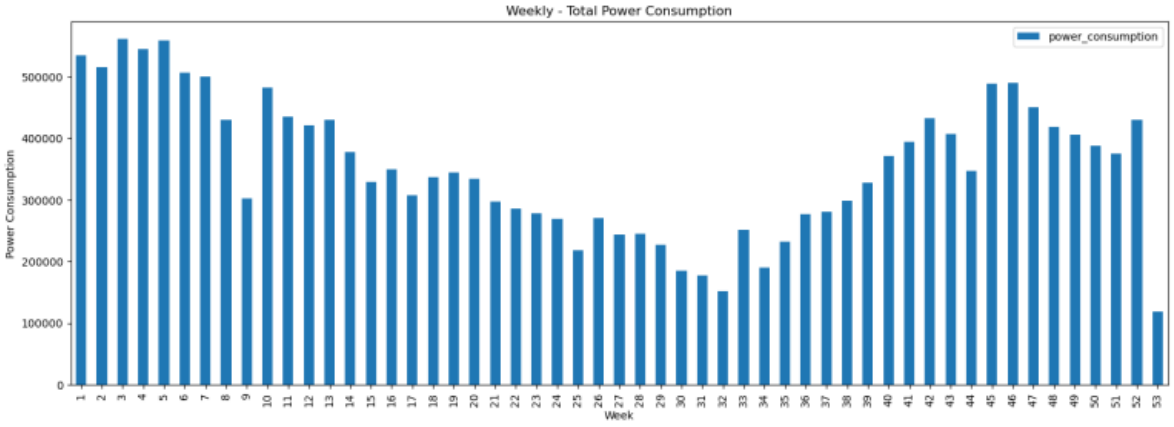

Fig 4.10: Week-wise Total Power Consumption

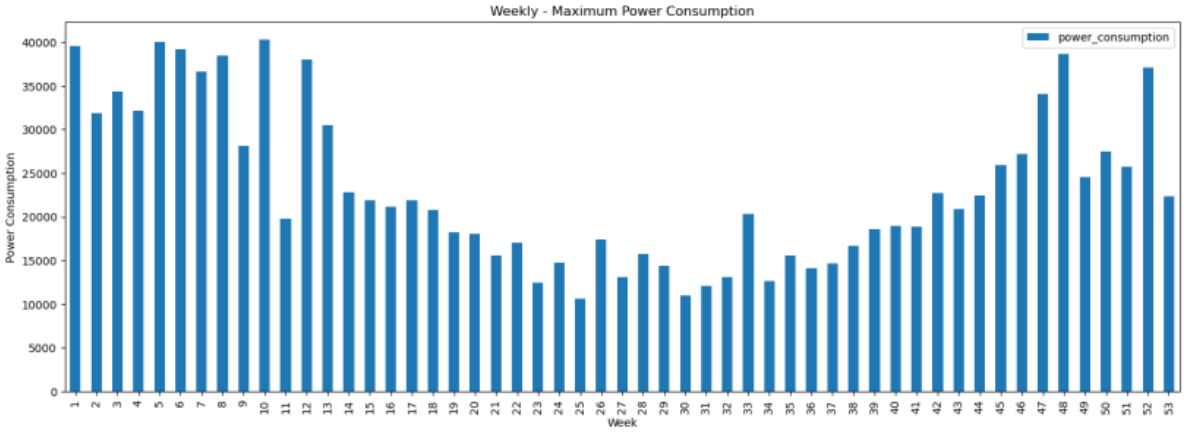

Fig 4.11: Week-wise Maximum Power Consumption

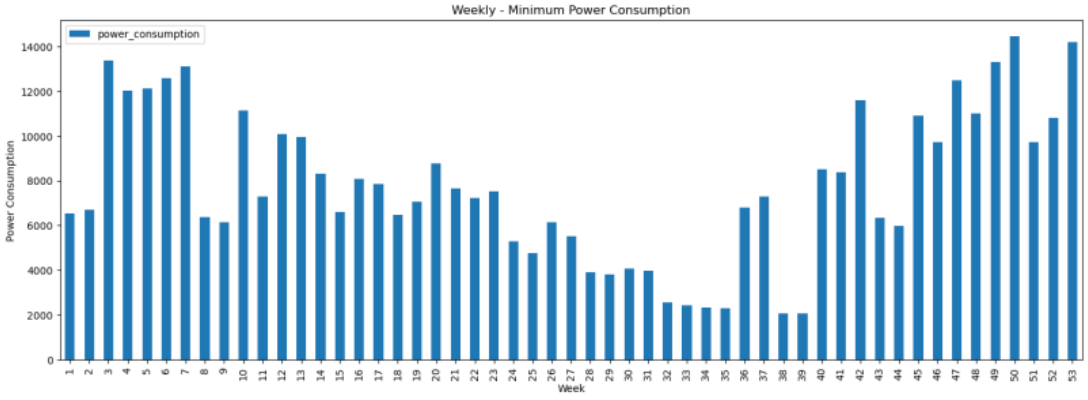

Fig 4.12: Week-wise Minimum Power Consumption

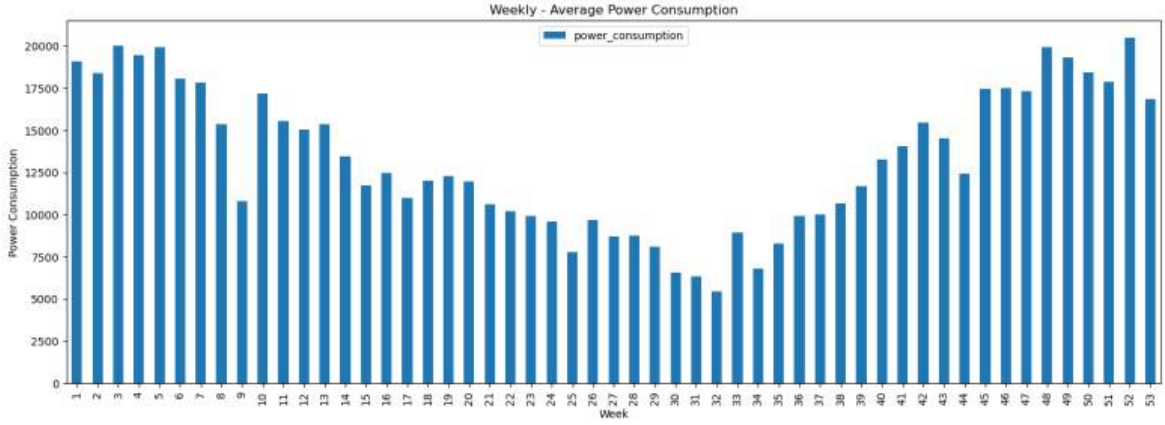

Fig 4.13: Week-wise Average Power Consumption

This graph demonstrates total Power consumption, a notable increase during the initial and final few weeks, exceeding 500 kilowatt-hours, followed by a decline in the intervening weeks. The highest usage occurs in weeks 1st , 5th , 6th , 47th , 48th , and 52th , all falling within the months of January, February, and December, with consumption levels exceeding 35,000 units. On the other hand, Energy consumption is notably lower during weeks 32th, 33th, 34th,35th, and 38th, which correspond to the months of June, July and august which is below the 15 Kilowatt-hour.

4.2.4 Week-days Analysis:

To delve deeper, a more detailed examination of the total power consumption for each day of the week throughout the four-year duration can be presented.

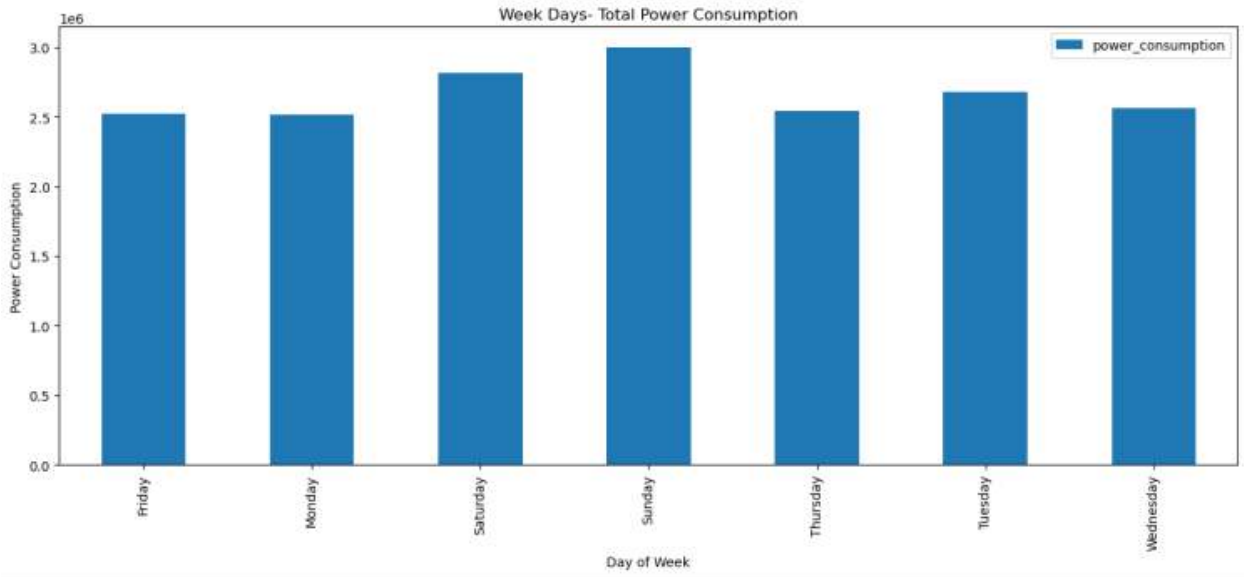

Fig 4.14: Weekday-wise Total Power Consumption

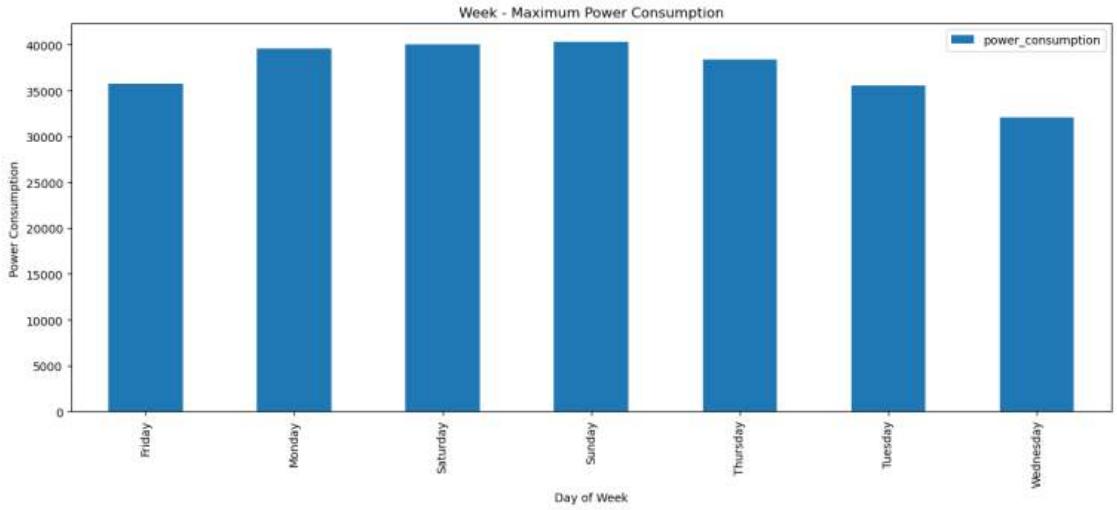

Fig 4.15: Weekday-wise Maximum Power Consumption

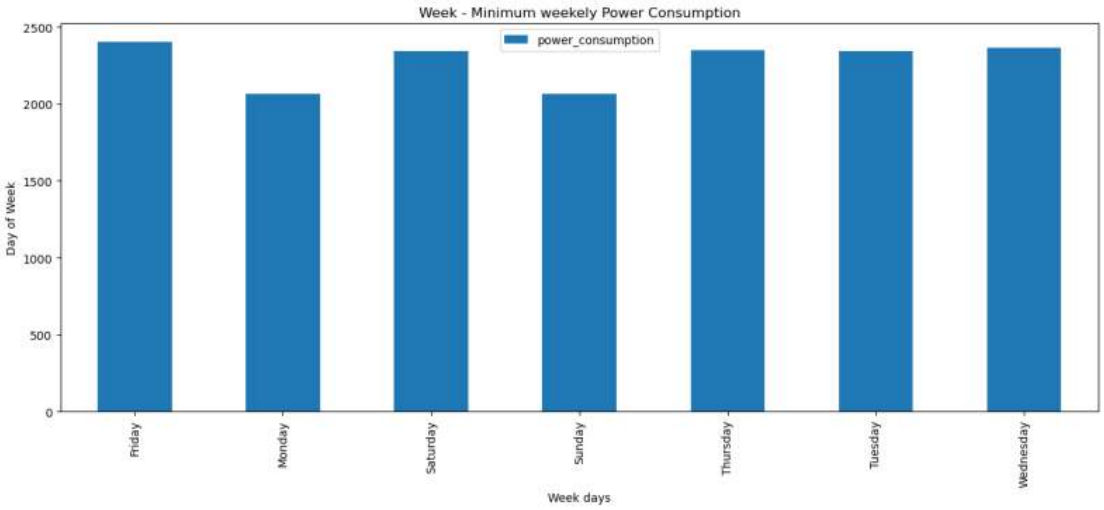

Fig 4.16: Weekday-wise Minimum Power Consumption

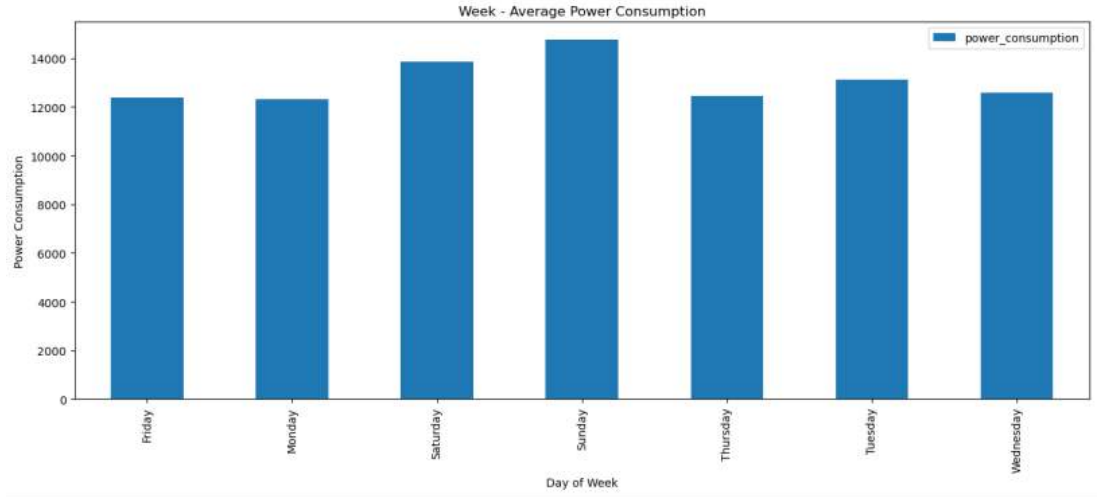

Fig 4.17: Weekday-wise Average Power Consumption

Electricity usage tends to be higher on Weekends (Saturdays and Sundays) compared to the other days of the week (Salam et al., 2018), this trend also observed in the above bar graph. Amber et al., who conducted research on Electricity consumption forecasting models for administration buildings, suggest a potential linear correlation between the number of working days and power use in commercial dwellings. (Amber et al., 2015). This correlation becomes evident as people spend more time at home during weekends, participating in activities that demand power, such as cooking, using electronic devices, watching TV, and doing housework. Furthermore, families may be at home together, resulting in an increase in overall energy use. This combination of greater house

occupancy and other activities may have contributed to the reported higher weekend power demand.

4.2.5 Hourly Analysis:

Lastly, to further refine this analysis by zooming in to closely examine power consumption on an hourly basis.

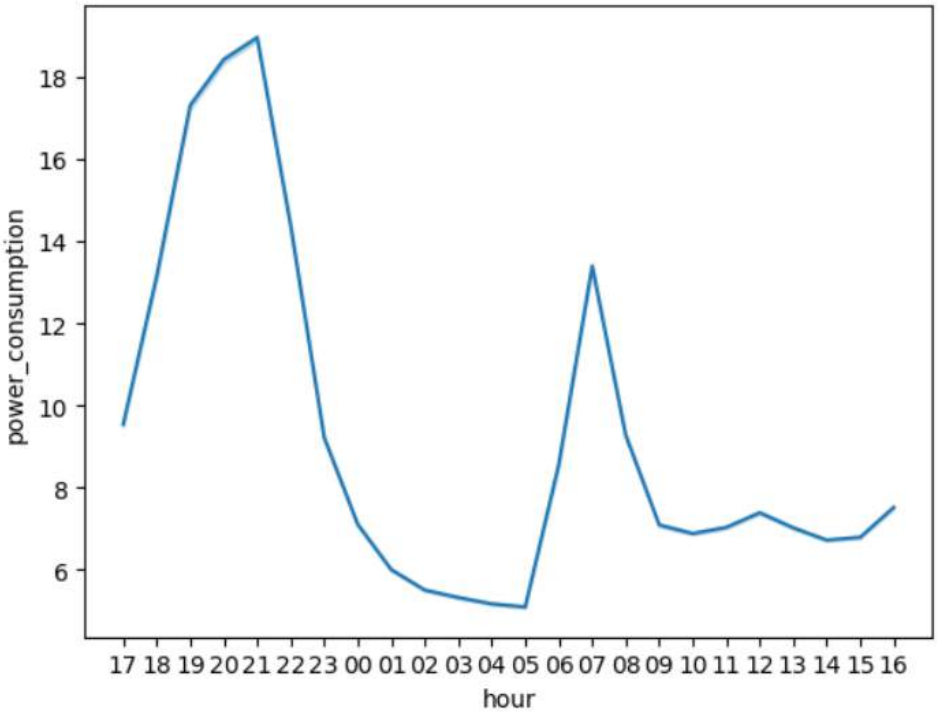

Fig 4.18: Hourly Power Consumption

The graph displays dual peaks, indicating two distinct periods of heightened activity for power consumption. A notable insight from the graph is the primary household consumption occurring predominantly between 19:00 - 21:00(evening peak) and 6:00 - 9:00(morning peak). This result also corresponds with the findings of Rambabu et al.'s study (Rambabu et al., 2020). The morning peak corresponds to the morning rush when people get up, get ready for the day, and go out to work or school. Activities like using appliances, taking a shower, and making breakfast all contribute to higher energy consumption during this time. Similar to the morning peak, the evening peak coincides with when people commonly go home from work or school and start using appliances, watching TV, and taking part in other types of entertainment, all of which use more energy.

4.3 Comparison of ML Models:

4.3.1 Regression Model:

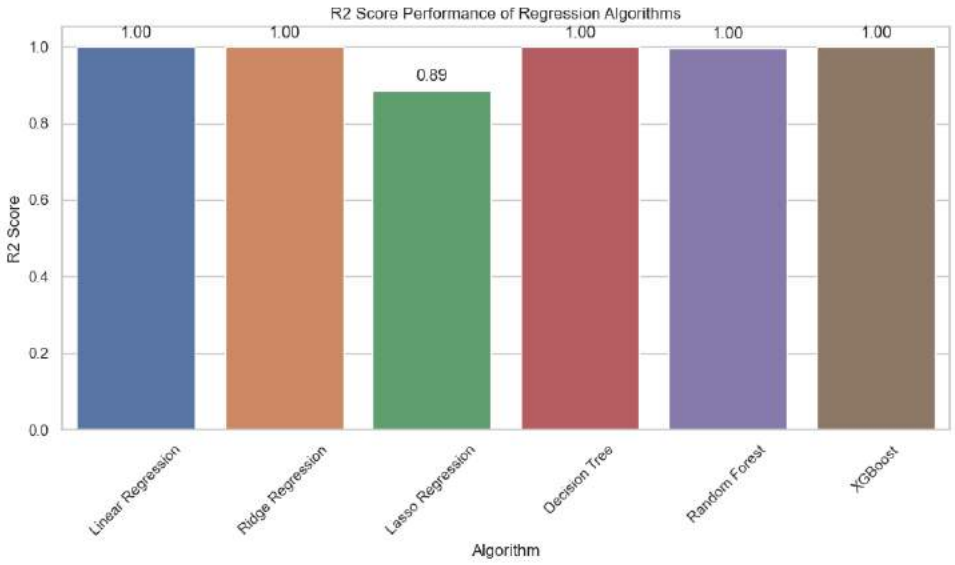

Fig 4.19: Comparison of R2 score in Regression Model

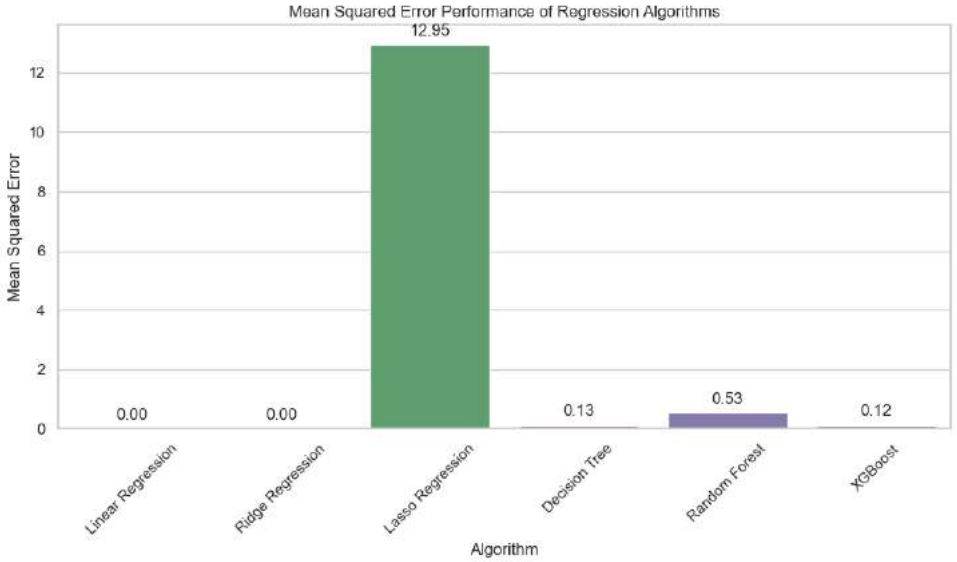

Fig 4.20: Comparison of Mean Squared Error in Regression Model

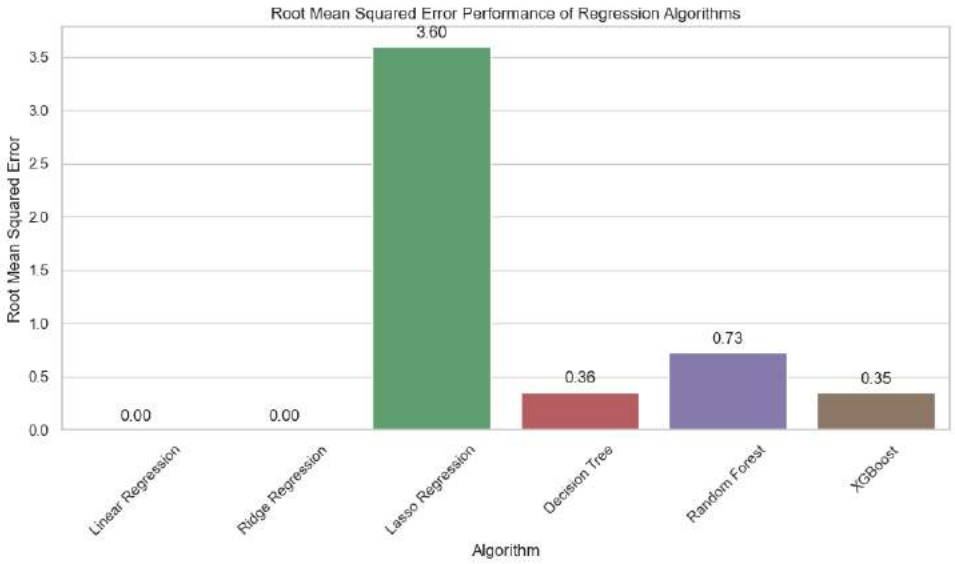

Fig 4.21: Comparison of Root Mean Squared Error in Regression Algorithms

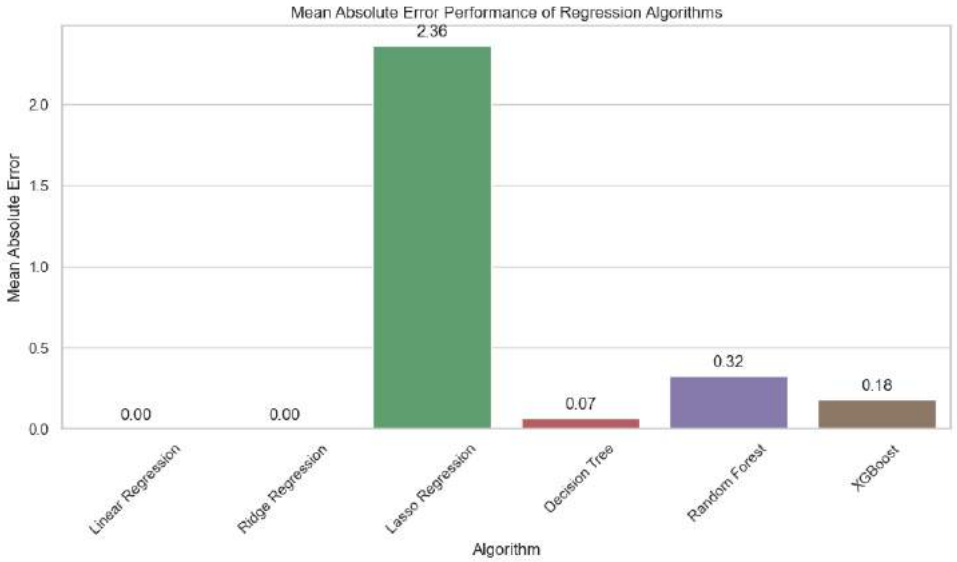

Fig 4.22: Comparison of Mean Absolute Error in Regression algorithm

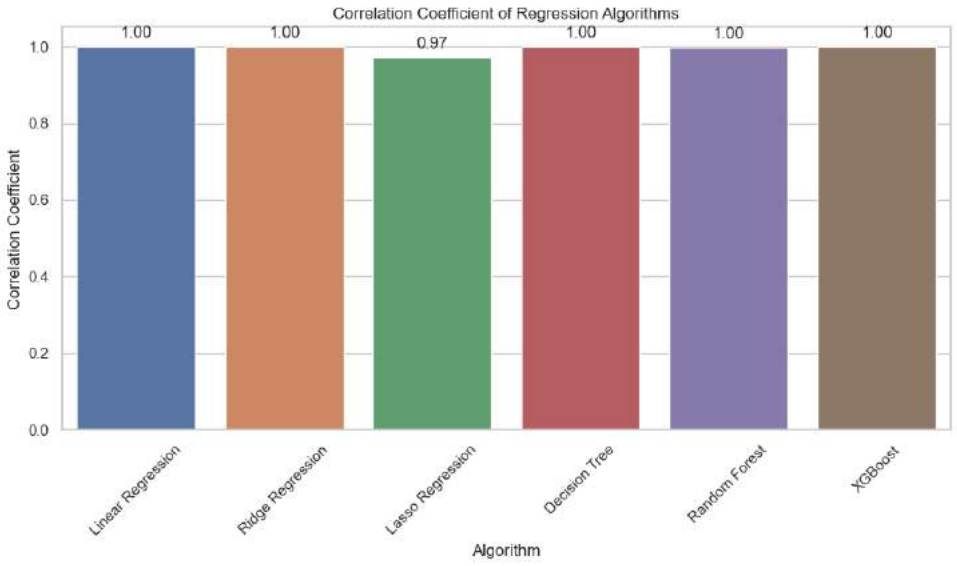

Fig 4.23: Comparison of Correlation Coefficient of Regression Algorithm.

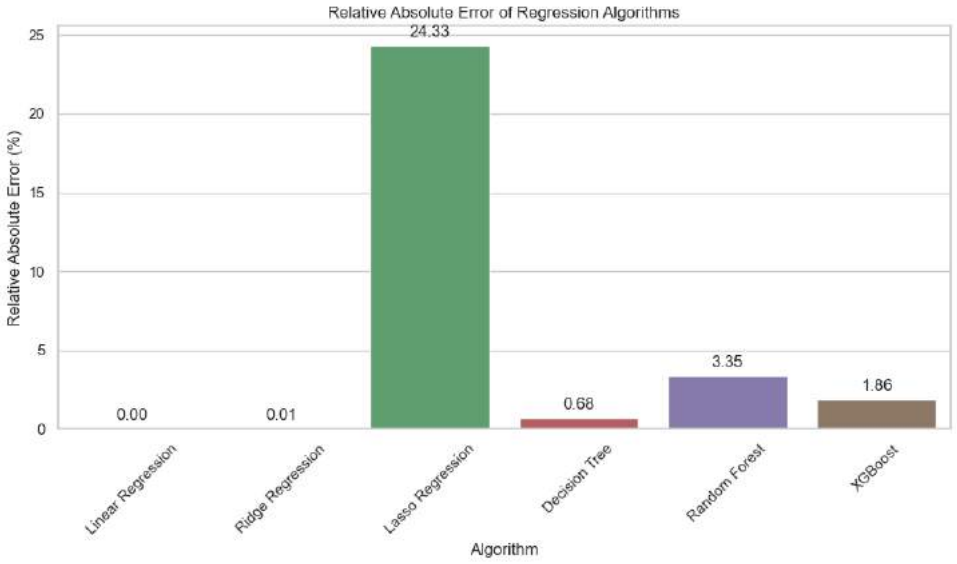

Fig 4.24: Comparison of Relative Absolute Error of Regression algorithm

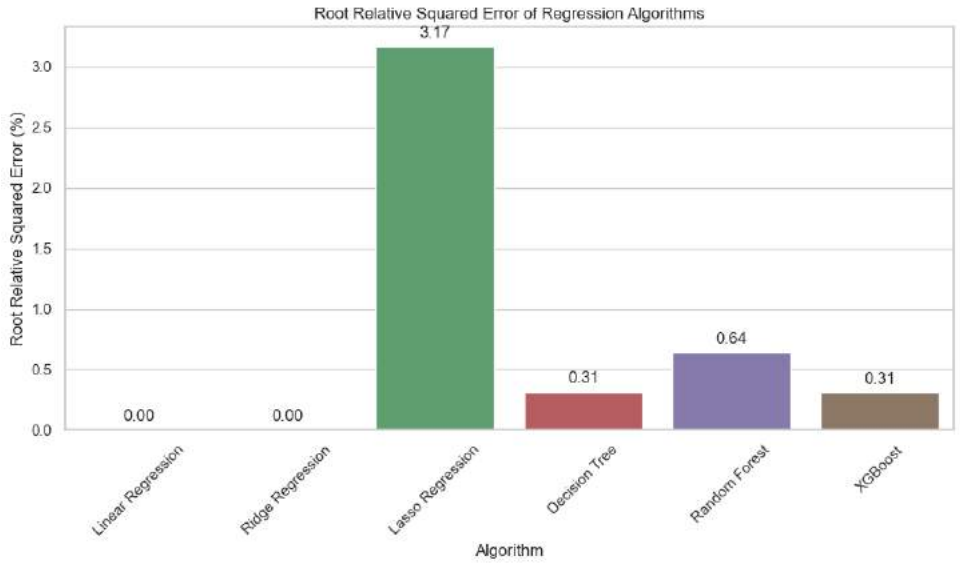

Fig 4.25: Comparison of Root Relative Squared Error in Regression algorithms

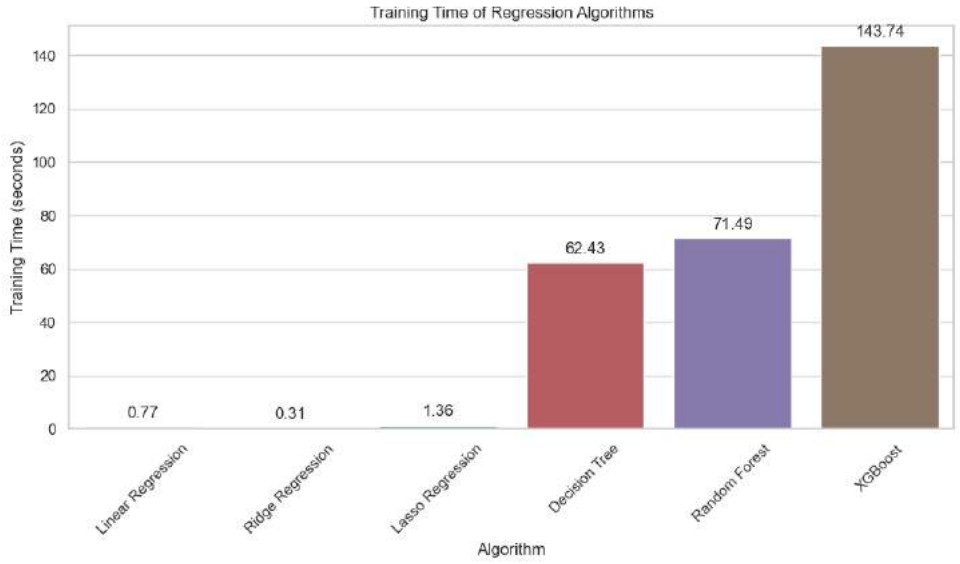

Fig 4.26: Comparison of Time Complexity in Regression model

	Algorithm	R2 Score	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	Correlation Coefficient	Relative Absolute Error(%)	Root Relative Squared Error(%)	Training Time (seconds)
0	Linear Regression	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	1.6510
1	Ridge Regression	1.0000	0.0000	0.0008	0.0005	1.0000	0.0051	0.0007	0.6113
2	Lasso Regression	0.8860	12.9518	3.5989	2.3622	0.9711	24.3349	3.1685	2.5355
3	Decision Tree	0.9989	0.1268	0.3561	0.0662	0.9994	0.6819	0.3136	114.1452
4	Random Forest	0.9953	0.5314	0.7290	0.3249	0.9977	3.3469	0.6418	103.3473
5	XGBoost	0.9989	0.1222	0.3496	0.1806	0.9995	1.8602	0.3078	193.3285

Fig 4.27: Summary of Regression Model Performance Using Key Metrics

The Fig 4.27 provides a thorough analysis of the performance indicators for several algorithms used in the fields of regression and machine learning. In particular, for both linear regression and ridge regression, the R2 Score, a statistic used to measure the amount of variability explained by a particular model, reaches its maximum value of 1.0. This accomplishment allegedly points to a very precise model fitting. The fact that these models have zero Mean Squared Error (MSE) values, however, calls for a nuanced interpretation. This situation raises questions about possible overfitting, a problem made worse by the perfect correlation value of 1.0. The latter denotes a continuous linear link between the expected and actual results, raising doubts about the generalizability of the models beyond the sample data. This fascinating convergence of high R2 Scores, zero MSE values, and perfect correlation highlights the complicated interaction between model complexity, variance explanation, and the need of being on the lookout for overfitting tendencies.

According to this work, Ridge Regression offers a balanced approach by successfully reducing possible overfitting via a small Mean Squared Error (MSE) and regulated error metrics. Although Lasso Regression has a good R2 score of 0.9020, it also has a somewhat high MSE of 12.2229, indicating both predictive potential and error trade-offs. Decision Tree, Random Forest, and XGBoost all achieved impressively high R2 scores that were very close to 1.0, underscoring their prowess in identifying data patterns. Random Forest and XGBoost showed robust performance despite slightly larger errors, while Decision Tree stands out with noticeably lower error metrics all around. This sophisticated research offers important insights into algorithmic behaviour, assisting in the formulation of well-informed decisions in scenarios involving predictive modelling.

A substantial correlation coefficient nearing 1.0 within the context of regression models signifies a robust association between the predictor variables and the target variable, thereby implying the model's capacity to elucidate a noteworthy fraction of the variability present within the target variable. However, when correlation coefficients approach their maximum value (1.0), special attention should be given to the danger of overfitting. This underscores the necessity to systematically examine the model's performance when presented with fresh, unobserved data to establish its efficacy in generalising. Additionally, variations in correlation coefficients between these different

modelling methodologies may be utilised as a signal for the relative intensity of regularisation and feature selection.

Higher R2 scores don't always imply low mistakes, thus it's vital to take into account how they relate to error measures. Additionally, overfitting problems should be evaluated in light of the data's context and the modelling strategies used. Linear Regression and Ridge Regression are noticeably quicker than more complicated algorithms like XGBoost. The relative training durations also provide light on computing efficiency. it conclude that, tree-based algorithms tend to perform poorly, whereas linear algorithms demonstrate good performance

A comprehensive analysis of all relevant factors is crucial in the search for the ideal algorithm. Linear regression, which is renowned for its quick computing efficiency and ideal predictive accuracy, stands out as an enticing option. However, the absence of error measures demands a careful assessment to consider possible overfitting. Instead, Decision Tree offers a possible option because of its balanced performance and much lower error metrics. This decision depends on a thorough analysis of the application's contextual requirements, computing efficiency, prediction accuracy, and error metrics. To ensure that the algorithms' claimed prediction skills are accurate and to ascertain the best possible solution, rigorous validation against an independent dataset is essential.

```
Best algorithm based on R2 Score: Linear Regression
Best algorithm based on Mean Squared Error: Linear Regression
Best algorithm based on Root Mean Squared Error: Linear Regression
Best algorithm based on Mean Absolute Error: Linear Regression
Best algorithm based on Time: Ridge Regression
Best algorithm based on Coefficient of Determination: Linear Regression
Best algorithm based on Relative Absolute Error: Linear Regression
Best algorithm based on Relative Relative Error: Linear Regression
```

Fig:4.28 Comparison output of Regression model

In pursuit of identifying the most suitable regression method, an intricate comparison of diverse performance indicators is undertaken. This involves an iterative examination of a comprehensive Data Frame encompassing metrics from various algorithms. The optimal method is determined by evaluating criteria such as the highest R2 Score, lowest Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), shortest training time, highest correlation coefficient, lowest Relative

Absolute Error (%), and lowest Root Relative Squared Error (%). The method manifesting the optimal confluence of these metrics is selected. Subsequently, the algorithm yielding the lowest RMSE is chosen for training on the complete training set. This chosen model is then leveraged to generate predictions on an independent test dataset, with the resultant predictions stored as "y_pred_best." This meticulous approach contributes to a systematic evaluation of algorithmic efficacy and facilitates judicious selection for the generation of reliable prognostications.

Opting for the Linear Regression model, despite the allure of the Ridge Regression's superior training time, constitutes a valid and judicious decision. While expedited training times hold inherent advantages, the preference for the Linear Regression model stems from its prowess in diverse performance metrics, as indicated by its consistently superior scores across multiple evaluation criteria. This selection underscores a comprehensive perspective that not only acknowledges efficiency but equally prioritizes the model's overarching predictive accuracy and congruence with the data. In essence, this decision exemplifies a balanced amalgamation of computational efficiency and model efficacy within the context of the selection process.

4.3.2 Classification Model:

From a classification perspective, it explores the distribution of the "Active_energy_consumption" characteristic over three different ranges: "Low," "Medium," and "High."

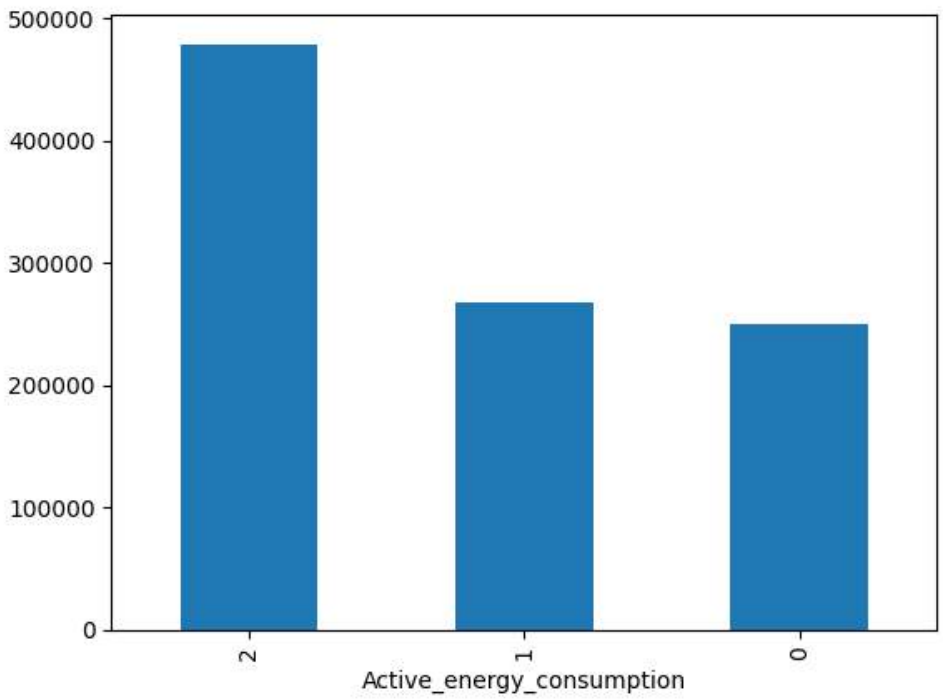

Fig 4.29: Count plot for Active energy consumption

A bar graph (Fig: 4.29), which provides a graphical depiction of the frequency of each category, is used to graphically describe this distribution. According to the statistics, there are 125,001 cases that fall into the "Low" category, 133,696 instances that fall into the "Medium" category, and 237,370 instances that fall into the "High" category. The categorical distribution within the dataset is clearly visualised in this graphical style, effectively illuminating the frequency of occurrences within each defined category. The distributional properties of the "Active_energy_consumption" feature across the defined categories are more clearly understood thanks to the analytical approach's illumination of the dataset's innate distributional patterns.

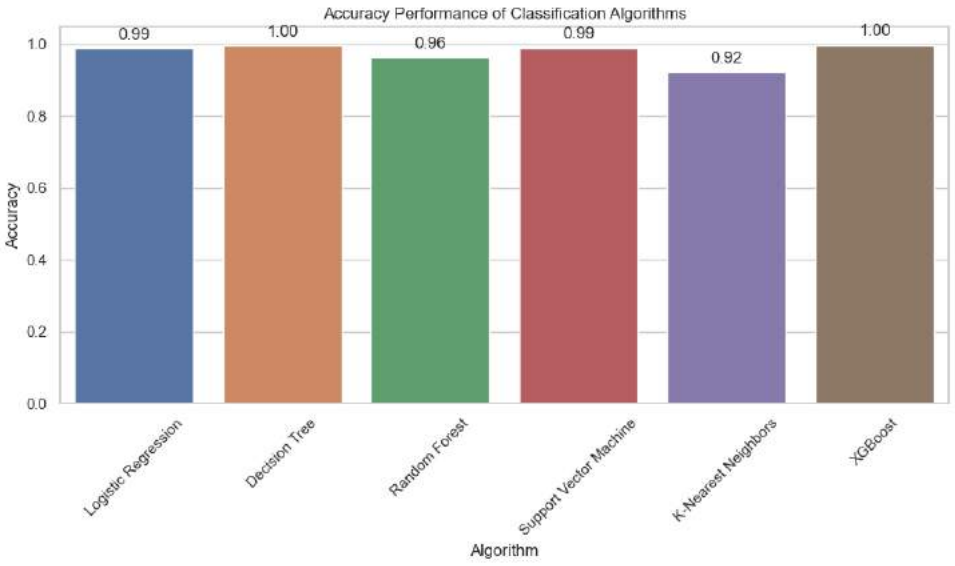

Fig 4.30: Comparison of Accuracy in Classification algorithms

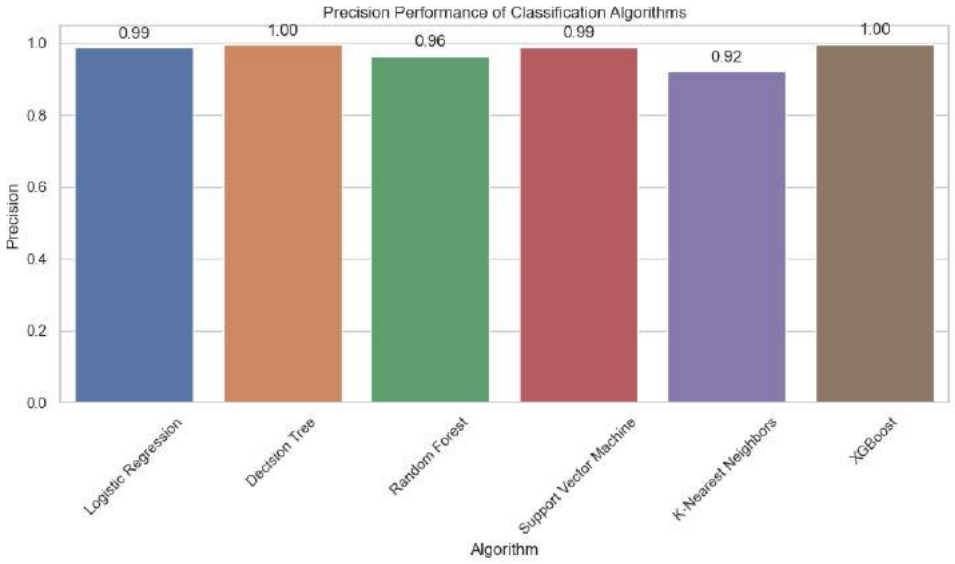

Fig 4.31: Comparison of Precision in Classification algorithms

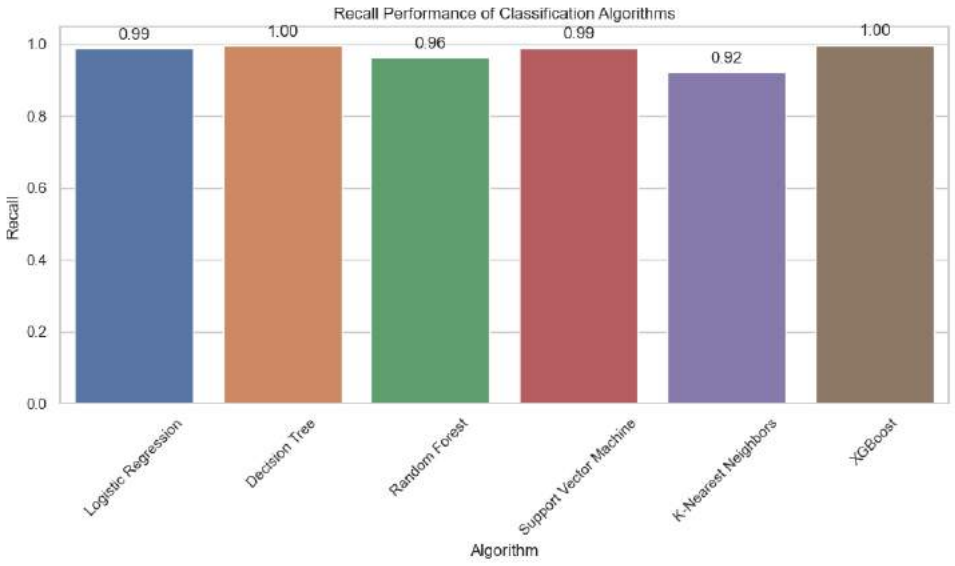

Fig 4.32: Comparison of Recall in Classification algorithms

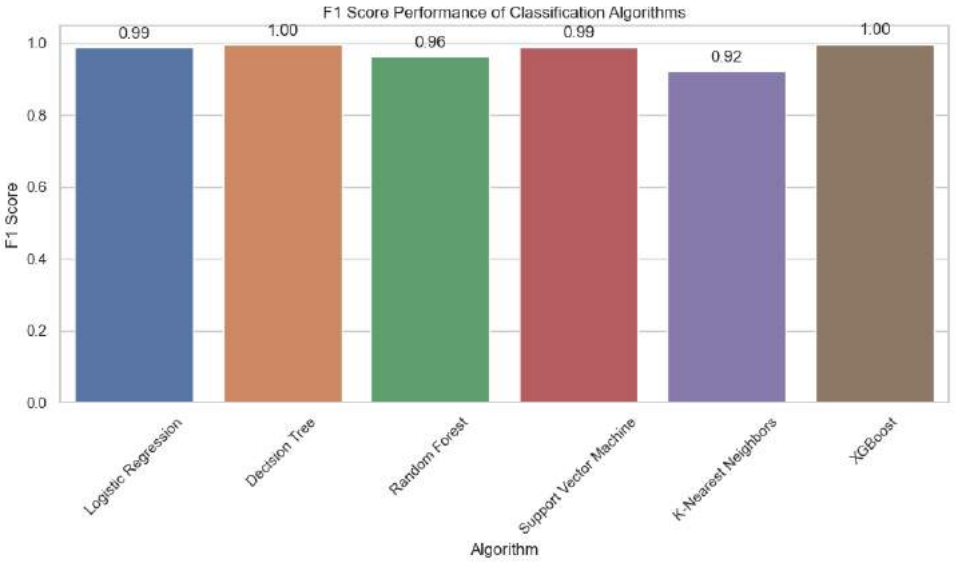

Fig 4.33: Comparison of F1 Score in Classification algorithms

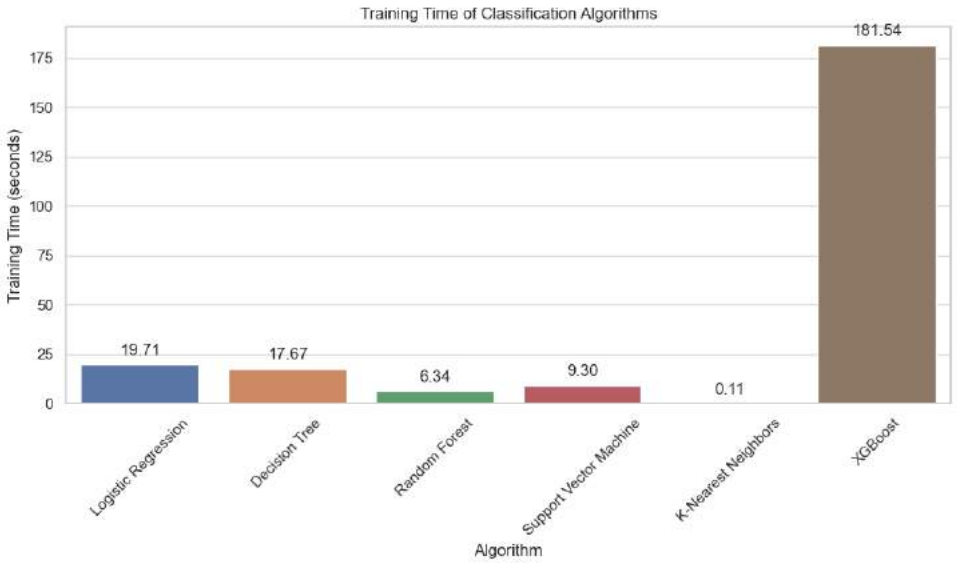

Fig 4.34: Comparison of Training time in Classification algorithms

	Algorithm	Accuracy	Precision	Recall	F1 Score	Training Time (seconds)
0	Logistic Regression	0.988258	0.988336	0.988258	0.988272	19.705976
1	Decision Tree	0.995191	0.995191	0.995191	0.995191	17.669070
2	Random Forest	0.963399	0.964306	0.963399	0.963238	6.344111
3	Support Vector Machine	0.988855	0.988859	0.988855	0.988854	9.297903
4	K-Nearest Neighbors	0.922059	0.922263	0.922059	0.922146	0.108816
5	XGBoost	0.997058	0.997064	0.997058	0.997058	181.536316

Fig 4.35: summary of Classification Algorithms based on Accuracy, Precision, Recall, F1 Score, and Training Time

The presented tabular data and graphs provide a comprehensive analytical assessment of various classification algorithms, each evaluated across multiple performance metrics including Accuracy, Precision, Recall, F1 Score, and Training Time. Logistic Regression demonstrates an Accuracy of 0.988258, Precision of 0.98836, Recall of 0.988258, and F1 Score of 0.988272, with a corresponding Training Time of 19.705976 seconds. Decision Tree surpasses with an Accuracy of 0.995191, Precision of 0.995191, Recall of 0.995191, and F1 Score of 0.995191, achieved within a Training Time of 17.669070 seconds. Random Forest offers an Accuracy of 0.963399, Precision of 0.964306, Recall of 0.963399, and F1 Score of 0.963238, accompanied by a Training Time of 6.344111 seconds. Support Vector Machine exhibits an Accuracy of 0.988855, Precision of 0.988859, Recall of 0.988855, and F1 Score of 0.988854, with a Training

Time of 9.297903 seconds. K-Nearest Neighbors achieves an Accuracy of 0.922059, Precision of 0.922059, Recall of 0.927903, and F1 Score of 0.922146, notably requiring a shorter Training Time of 0.108816 seconds. XGBoost emerges as the leader with an Accuracy of 0.997058, Precision of 0.997064, Recall of 0.997058, and F1 Score of 0.997058, although its Training Time is comparatively longer at 181.54 seconds. This comprehensive analytical evaluation discerns the algorithms' performance nuances, reflecting on their accuracy, precision, recall, and F1 Score trade-offs in conjunction with their corresponding training times. The insights gleaned from this meticulous assessment offer valuable guidance in selecting the most appropriate algorithm for classification tasks, weighing the computational time against the achieved classification performance.

```
Best algorithm based on Accuracy: XGBoost
Best algorithm based on Precision: XGBoost
Best algorithm based on Recall: XGBoost
Best algorithm based on F1 Score: XGBoost
Best algorithm based on Time: K-Nearest Neighbors
```

Fig: 4.36 Final Output for classification Model

It is vital in the evaluation and choosing of machine learning algorithms. The optimum algorithm is initially selected for specific performance parameters including Accuracy, Precision, Recall, F1 Score, and Training Time. The algorithm that has each metric's largest value is determined in this identification technique by looking at a specified DataFrame called metrics_df, which is expected to comprise a range of performance metrics connected to different algorithms. The method then prints out the names or identifiers of these top algorithms after getting them from the DataFrame, giving helpful information about which algorithm performs best given a specified set of evaluation criteria.

The best algorithm is then picked by the code and stored into the variable best_algorithm based on Accuracy. By making this tactical option, it is assured that the algorithm with the greatest Accuracy will be utilised to generate predictions. The best algorithm is then picked, and it is trained using the complete training dataset so that it may profit from the given training data. In order to generate predictions on a new test

dataset, it ultimately uses the trained best technique, storing the results in the variable y_pred_best. Essentially, it automates the crucial processes of algorithm selection, training, and prediction, functioning as a major step in the evaluation and optimisation of models within machine learning workflows.

In all important performance criteria, such as Accuracy, Precision, Recall, and F1 Score, it is clear from the presented figure 4.37 that the XGBoost algorithm constantly beats competitors. This shows that XGBoost is the best-performing algorithm for the given job in terms of predicted accuracy and overall effectiveness.

However, the K-Nearest Neighbours (KNN) algorithm emerges as the most effective option when time complexity is taken into account. In situations when computational efficiency is important, KNN is a better choice than XGBoost because to its reduced temporal complexity.

4.3 Model explanation:

4.3.1 Regression model:

The eXplainable Artificial Intelligence (XAI) approaches are used in this work to clarify the underlying processes of both the ideal regression and classification models, which have been conserved within the "load model" variable accordingly.

SHAP calculates global feature relevance by averaging SHAP values throughout the dataset. Each SHAP value in this research gives information on the positive or negative impact of specific factors to forecasting home power use. The SHAP values for the Linear regression model are shown in Figure 7 to show the relevance of each feature. A larger SHAP number suggests a greater influence on model output, whether good or negative.

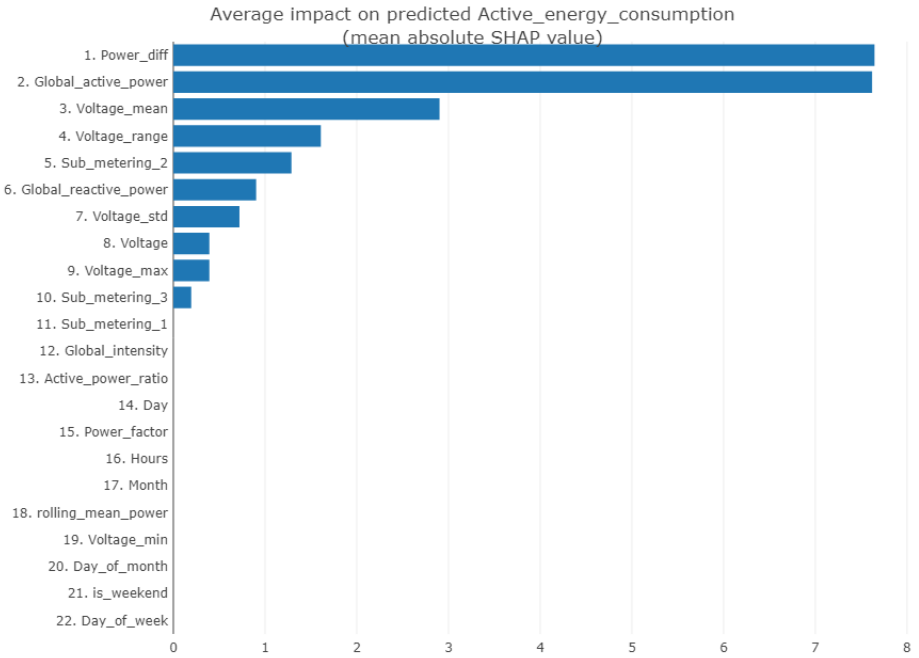

Fig 4.37: SHAP values with impact on Regression model.

The feature "power_diff" demonstrates the highest absolute SHAP value, standing at 15. This significant number emphasizes its crucial function in the model's predictions. The output is significantly variable when "power_diff" is changed, indicating that it contains important information that greatly influences the model's conclusions. The feature "voltage_mean" has a similar absolute SHAP score of 8, indicating that it has a significant influence on the model's predictions. Voltage_mean variations have a significant impact on the model's output, although having a less noticeable impact than "power_diff" and "voltage_max" does. In the hierarchy of feature significance, the feature "voltage_range" and "voltage_max" demands an absolute SHAP value of 3.5 and 2.5 respectively, reflecting its middle place.

On the other hand, features like "global_intensity," "power_factor," "months," and "hours," which have absolute SHAP values of 0, have no influence on model rojections. The output of the model is not considerably affected by changes to these characteristics.

Another advantage of utilising the SHAP XAI tool is its ability to create partial dependency charts, which assist in understanding the impact of one or two characteristics on the anticipated result of the model. To explore the target feature's connection with other features by graphing its SHAP value, indicating if the

relationship is linear, monotonic, or more complex. Figure 4.38 shows, for example, that vertical dispersion at a certain Active_energy_consumption value indicates interaction effects with Power_diff. The figure shows a positive, virtually linear relationship between the target variable (Active_energy_consumption) and power_diff, suggesting frequent interaction.

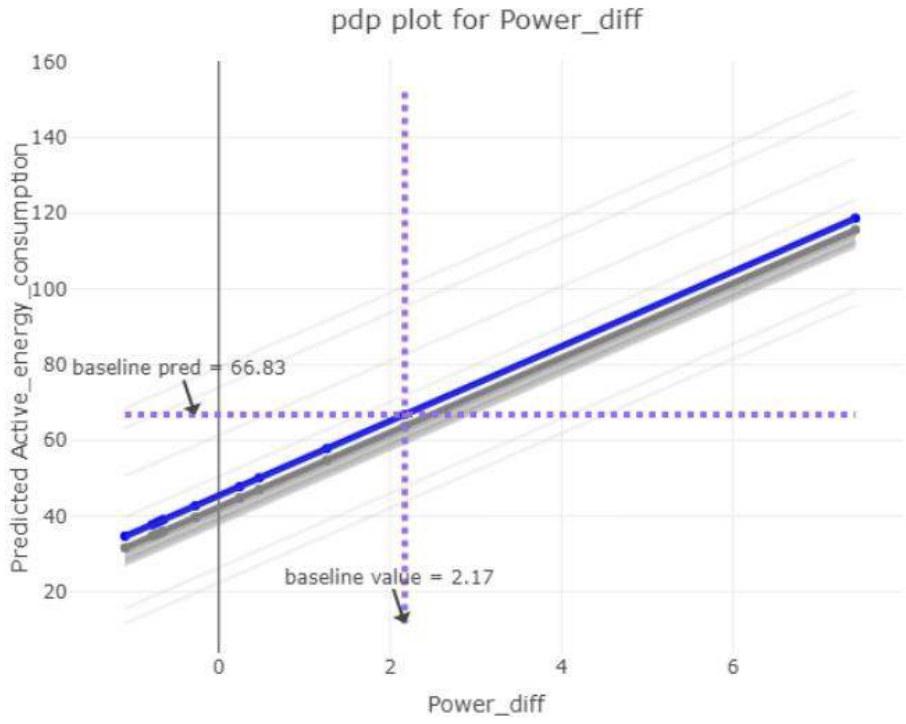

Fig 4.38: Partial Dependence plot of Active_energy_consumption.

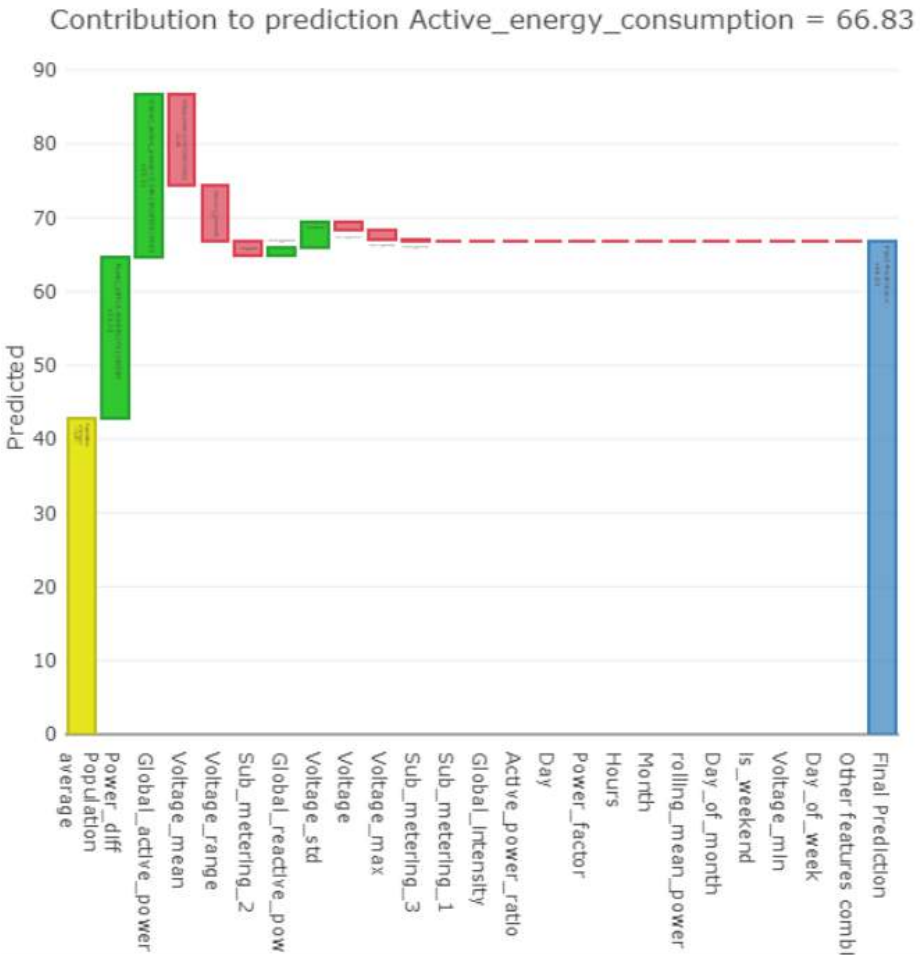

Fig 4.39: Contribution plot

The influence of numerous variables on the prediction process is highlighted for the particular index 14831, as shown in the Contributions plot (4.39). Beginning with the "Average of population" at a considerable value of 42.87, the adventure continues with the "Power_diff" characteristic, which offers a large positive effect of +21.73, significantly amplifying the forecast. Similarly, "Global_active_power" plays an important impact, providing a positive effect of +22.11, raising the projection even higher. However, the complexities emerge with "Voltage_mean," where its positive influence of +2.930723627532035 results in a -12.28 drop in forecasts, demonstrating the complexities of its participation. This story is distinguished by a symphony of positive and negative impacts, as several elements interact to produce the ultimate forecast of 66.83 for the index 14821.

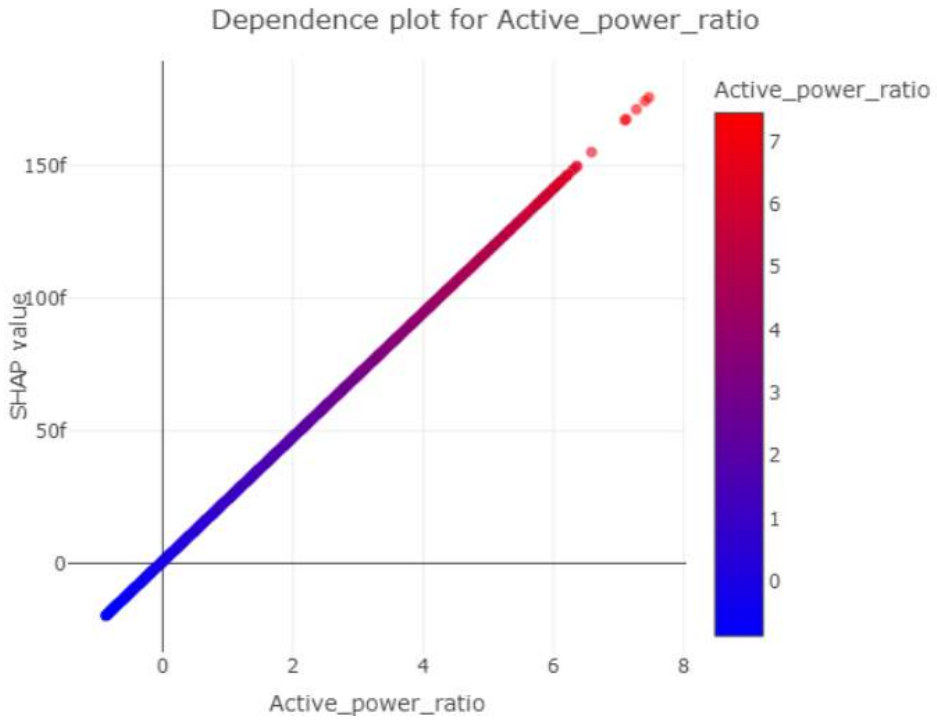

Fig 4.40: relationship between target value and SHAP value

The above fig 4.40 shows a relation between target value and SHAP value. A linear regression line passing through the origin point (0,0) delineates a direct and linear relationship between the property "active_power_ratio" and its associated SHAP values. This arrangement denotes a straightforward and proportionate correlation between the magnitudes of the attribute's values and the related SHAP values, as well as the resulting model predictions. The regression line's passage through the origin emphasises that when the "active_power_ratio" attribute is instantiated with a value of 0, both the SHAP value and the predictive output of the model assume a value of 0, emphasising the dependable and predictable influence of this specific attribute on the model's resultant predictions.

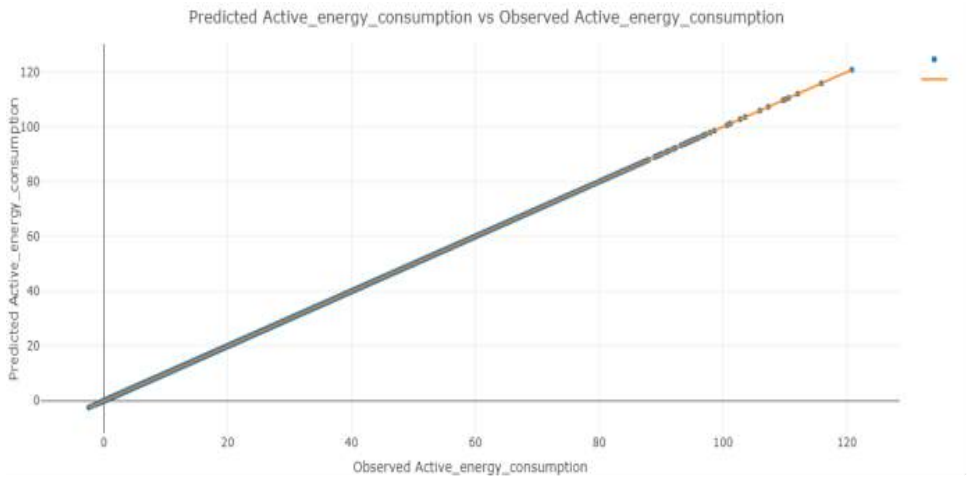

Fig 4.41: Predicted vs Actual

The fig: shown compares the actual and projected values of the parameter "Active_energy_consumption." A remarkable alignment along the diagonal line is seen, indicating an ideal situation in which the model's predictions completely correspond with the actual values. This diagonal alignment represents a perfect predictive model in which every point on the plot is exactly on this line, indicating a perfect match between the model's estimates and the actual observed outcomes.

4.3.2 Classification Model:

The presented graph delves into the realm of eXplainable Artificial Intelligence (XAI) in the context of a classification model.

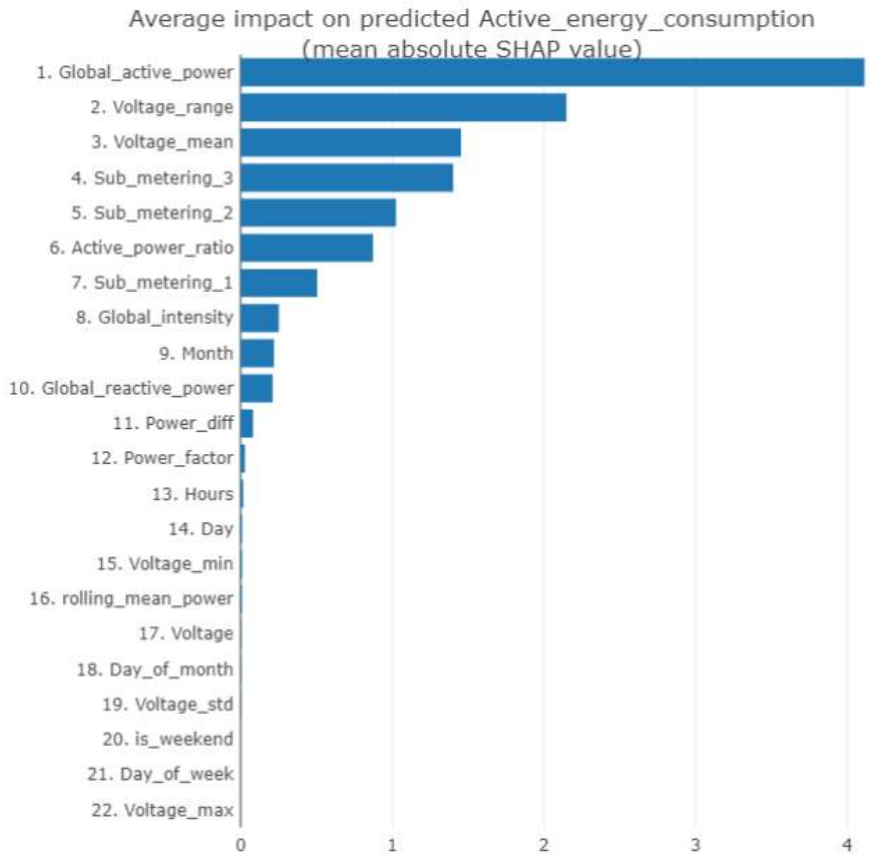

Fig 4.42: Importance of features in classification model

The representation of mean absolute SHAP values in the graph provides insight into the collective influence on the predicted Active Energy Consumption. Notably, characteristics such as "Global_active_power" and "voltage_range" have significant mean absolute SHAP values of 4.1 and 2.1, respectively, indicating their importance in determining forecasts. "Voltage_mean" and "sub_metering_3" come next, each having a tiny influence of 1.5. Other elements, such as "sub_metering_2," "active_power_ratio," and "global_intensity," have varied affects between 0 and 1. Similarly, "Voltage_max" and "Day_of_week" have significant influence. "voltage_std" on the other hand, remains marginalised with a mean absolute SHAP value of 0, highlighting its limited effect on expected outcomes.

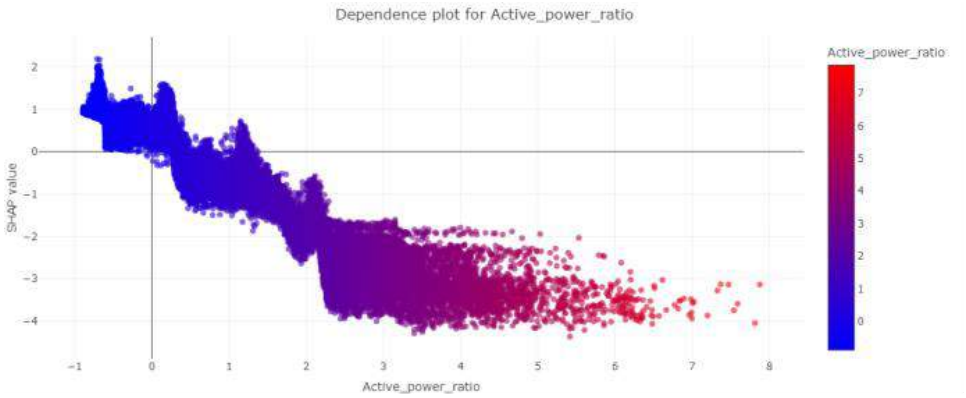

Fig 4.43: SHAP dependence graph for target feature

In more specific, SHAP dependence graph for "Active_power_ratio" depicts an intriguing contrast to the conventional regression model. Specifically, it reveals a counterintuitive relationship where an increase in the "Active_power_ratio" value leads to a decrease in the corresponding SHAP value. This divergence from the anticipated linear trend observed in typical regression models highlights a unique and potentially nonlinear interaction between the "Active_power_ratio" feature and the model's output. This phenomenon underscores the importance of considering complex interactions when interpreting the impact of features on the model's predictions through SHAP values.

label	probability	logodds
0*	99.4 %	5.189
1	0.0 %	-9.55
2	0.5 %	-5.202

* indicates observed label

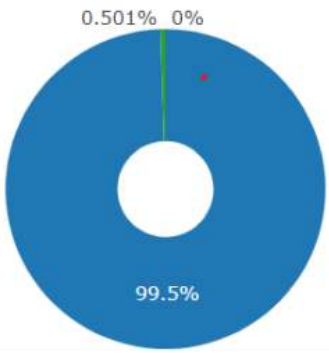

Fig 4.44: prediction chart at index- 18431

In the specific context of classification, with a particular focus on index 18431, the dataset provided encompasses discrete labels, probabilities, and logodds values. Each label corresponds to a distinct category denoted as "low," "medium," or "high." It's pertinent to highlight that the "low" category notably presents a substantial probability of 99.4% alongside a positive logodds value of 5.189. This conveys a high level of

certainty from the model in assigning instances to this category. In contrast, the "medium" and "low" categories exhibit probabilities of 0.0% and 0.5%, respectively, accompanied by corresponding negative logodds values of -9.55 and -5.202. This observed disparity in both probabilities and logodds values underscores a relatively diminished level of model confidence when ascribing instances to the "high" and "medium" categories.

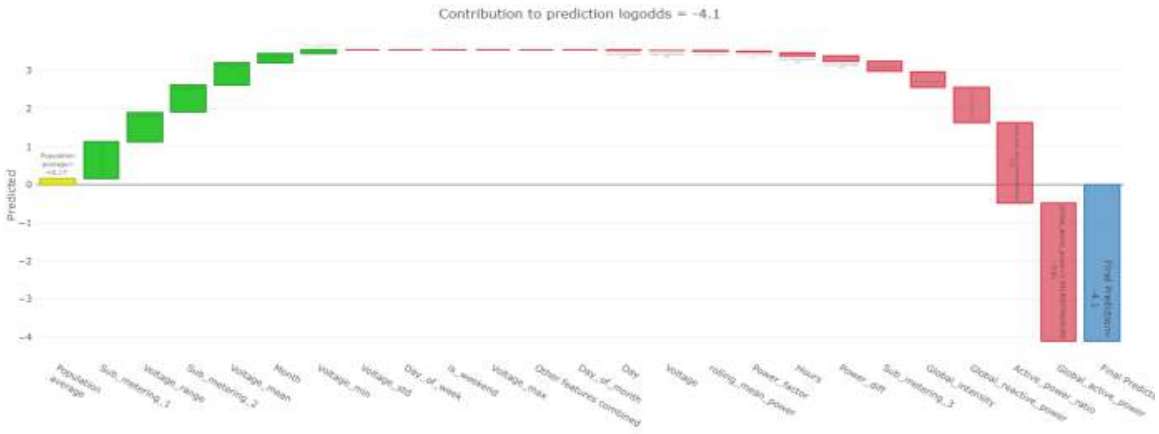

Fig 4.45: Contribution of features plot

The contribution plot at index 18431 gives a detailed look at the impact of numerous variables on the final forecast. With a value of +0.17, the characteristic "Average of population" contributes favourably, meaning that a rise in this feature leads to a higher anticipated result. In contrast, "Global_active_power" has a substantial negative contribution of -3.61, meaning that larger values of this characteristic lead to poorer forecasts. With a value of +0.77, the attribute "Voltage_range" contributes favourably, indicating that cases with a broader voltage range are likely to result in higher forecasts. Similarly, "Voltage_mean" has a positive coefficient of +0.6, indicating that greater average voltage levels are connected with higher forecasts. "sub_metering_3", on the other hand, has a negative influence with a coefficient of -0.27, indicating that larger levels of this characteristic are associated with worse projected outcomes. With a coefficient of +0.7, the characteristic "sub_metering_2" contributes favourably, suggesting that higher values correlate to greater forecasts. "Active_power_ratio" has a -2.12 coefficient of influence on prediction, meaning that greater values of this attribute are related to worse forecasts. In contrast, "sub_metering_1" contributes favourably with a coefficient of +0.96, suggesting that greater values improve forecasts.

"Global_intensity" has a negative coefficient of -0.4, indicating that greater intensity levels lead to poorer anticipated results. With a coefficient of +0.24, the characteristic "month" contributes positively, indicating that specific months are connected with higher projected values. Furthermore, with a coefficient of -0.93, "global_reactive_power" has a negative influence, suggesting that larger values of this attribute are associated with worse forecasts. With a coefficient of -0.15, "Power_diff" contributes negatively, implying that greater values of this characteristic lead to poorer predictions. "Power_factor," "Hours," "voltage_min," and "rolling_mean_power" have less of an influence. The sum of these factors results in a final prediction of -4.1, which represents the model's projection for the individual occurrence under discussion. This contribution plot elucidates the numerous linkages inside the predictive model by providing an in-depth view of how each attribute jointly effects the final forecast.

```
Accuracy: 0.9970582476995598
Precision: 0.9970638387009088
Recall: 0.9970582476995598
F1 Score: 0.9970581363518078
Classification Report:
              precision    recall  f1-score   support

         Low       1.00      1.00      1.00     50153
      Medium       1.00      0.99      1.00     53335
        High       1.00      1.00      1.00     95713

    accuracy                           1.00    199201
   macro avg       1.00      1.00      1.00    199201
weighted avg       1.00      1.00      1.00    199201
```

Fig 4.46: Confusion Matrix

Subsequent to the SHAP contribution graph, a confusion matrix is also presented, conveying essential information about the model's classification performance. This metrics and classification report offer a comprehensive assessment of the model's performance. The accuracy of 0.997 indicates that the model correctly predicts the class for approximately 99.7% of instances. Precision, which stands at 0.997, signifies the proportion of correctly predicted positive instances among all instances predicted as positive. Similarly, recall, also at 0.997, denotes the fraction of actual positive instances that were correctly identified by the model. The F1 score of 0.997 reflects a harmonic mean between precision and recall, providing a balanced measure of the model's accuracy across classes. The classification report further elaborates on the performance for each class. For the "Low" class, precision, recall, and F1 score are all 1.00, indicating high accuracy. In the "Medium" class, precision remains high at 1.00, while recall and F1 score are slightly lower but still impressive at 0.99. The "High" class displays similar precision, recall, and F1 score values of 1.00. Overall, the model's accuracy and performance across all classes are notably high, as indicated by the macro and weighted average scores of 1.00 in the classification report. This collective

evaluation demonstrates the model's robust ability to classify instances accurately and consistently across the different categories.

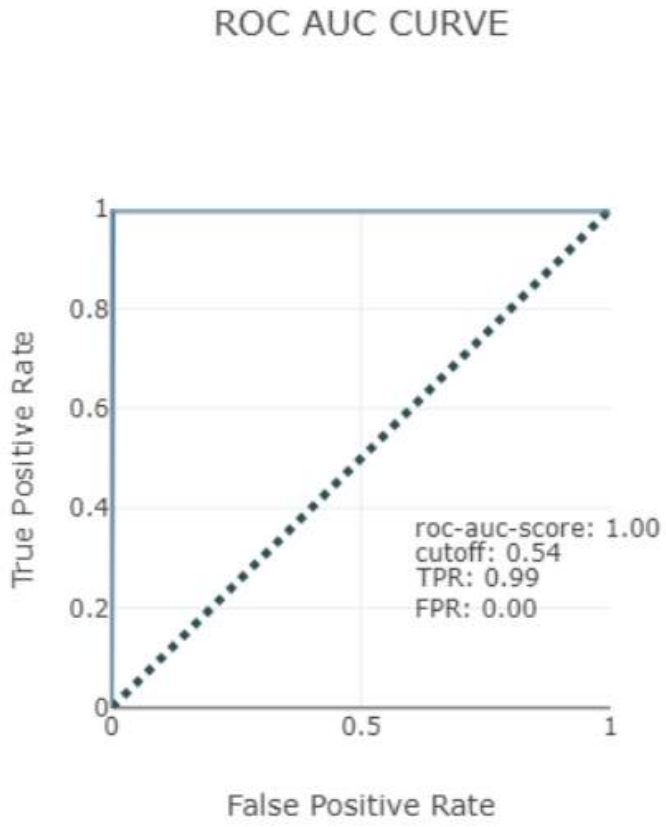

Fig 4.47: ROC AUC Plot

The ROC curve with an AUC value of 1.00 shows that the classification model performed very well. The threshold of 0.54 selected specifies the point at which occurrences are classed as positive or negative. The TPR of 0.99 indicates the model's great capacity to properly identify positive cases, whilst the FPR of 0.00 indicates a small incidence of labelling negative examples as positive. These measures illustrate the model's exceptional sensitivity, specificity, and accuracy in discriminating between positive and negative cases, bolstering its substantial predictive potential.

Chapter 5: Research Outcomes and Discussion

At the foundation of this research lies the crucial difficulty faced by government policymakers: the accurate projection of future energy demand. The research employs a mix of regression and classification techniques to handle this vital problem, with the ultimate purpose of offering policymakers essential tools for intelligent decision-making in energy management and budget allocation.

The Kaggle website provided the dataset that was used in this investigation. Despite the dataset's lack of specific country labels, it is strikingly similar to the UK household data when compared to the findings from the literature review, closely resembling the dataset used in Liu et al.'s thorough investigation of yearly Exploratory Data Analysis (EDA) on actual UK household data. Notably, this study supports Salam et al. and Amber et al.'s results, which emphasise how much more power is used on weekends. Additionally, a closer look, notably an hourly-based EDA, uncovers startling parallels to Rambabu's findings, which explains the recurrent morning and evening peaks in power usage patterns. It is crucial to remember, nevertheless, that Zhang's data runs counter to this established monthly trend, necessitating more research and study.(Liu et al., 2022; Salam et al., 2018 ; Amber et al.,2015)

After conducting Exploratory Data Analysis (EDA), the research proceeds to compare the performance of various regression models, Linear Regression and Ridge Regression both performed exceptionally well, achieving a perfect R-squared (R2) score of 1 and zero values for Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Coefficient of Correlation (CC), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE). This result illustrates their remarkable total energy consumption forecast accuracy. The findings of this study, however, vary from those of Rambabu et al. and Ding et al. in an exciting manner as research demonstrates that tree-based algorithms do not defeat linear models. The best approach, according to Rambabu's analysis, is Extra Tree Regressor, followed by Decision Tree and Linear Regression, which performs less well. Rambabu et al.'s

research also considers other aspects like wind speed, temperature, and humidity, yet in this analysis showed that Multiple Linear Regression is the best alternative. The study highlights the consistency of these results across both research endeavors by agreeing with Rambabu's observation that training time is shorter for Linear Regression and longer for XGBoost.(Rambabu et al., 2022; Ding et al., 2021)

When comparing classification models, it is evident that XGBoost beats all others in terms of a range of performance measures, with the exception of training time, where it lags behind. The performance of Decision Tree is equivalent to that of XGBoost. This conclusion is congruent with those produced by Bashir et al., who discovered that the Decision Tree classification method performed better than a variety of other approaches, such as Support Vector Machines (SVM), k-Nearest Neighbours (KNN), Nave Bayes, Logistic Regression, and Neural Networks. This closeness in the research results illustrates the robustness of the Decision Tree approach, which is congruent with the findings of the study.(Bashir et al., 2021)

Subsequent to determining the best-performing models in both regression and classification tasks, the research applies eXplainable Artificial Intelligence (XAI) approaches to extract essential characteristics. Notably, the identical index, 18431, is applied in both models, demonstrating that "power_diff" and "Voltage_mean" emerge with the greatest Shap values, suggesting their vital relevance. In contrast, "Global_active_power" and "voltage_range" dominate the classification model with the biggest Shap values, whilst "power_diff" has a considerably smaller influence than these features. These findings jointly throw light on the many attributes that are significant for each given model, showing their individual importance and contribution to model performance.

However, it's vital to realise that this study effort does not incorporate environmental factors, which are crucial for generating exact forecasts. Furthermore, the usage of a dataset from Kaggle may not give findings that truly represent real-world settings or scenarios.

Chapter 6 : Project Management

This chapter discusses the complete planning, organization, and alignment of the research endeavour.

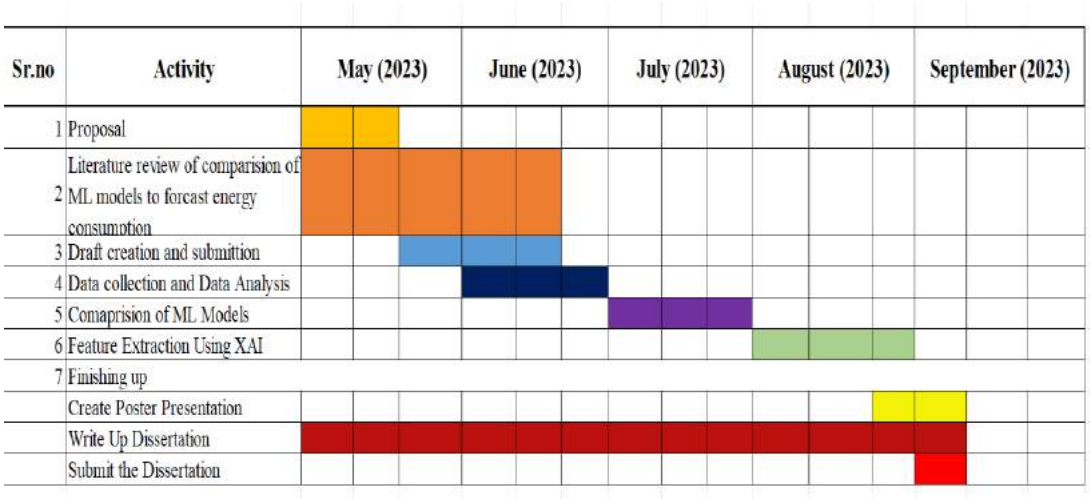

Fig 6.1: Project Planning

First Phase(Proposal): This first phase concentration is on the preliminary work and planning necessary for the proposal creation process.

Second Phase(Literature review): During this phase, the major duty entails obtaining information regarding current activity and efforts.

Third Phase(Draft Creation and submission): Writing an introduction and delivering a brief summary of the literature on the comparison of machine learning models for energy consumption prediction is part of the scope of this phase. It also requires submitting the work for evaluation and review.

Fourth Phase (Data collection and Analysis): During this specific phase, the emphasis switches towards data collection and the use of relevant procedures for analysis.

Fifth Phase(Comparison of ML models): The major purpose of this phase is to apply machine learning models and then analyse the results using multiple performance measures.

Sixth Phase(Feature extraction using XAI): In this step, the purpose is to identify the most impactful or best characteristics for the machine learning model.

Seventh Phase(Finishing Up): This step comprises evaluating and editing the report and preparing for the final submission, which may include developing a poster presentation.

Chapter 7 : Conclusion and Future Work

7.1 Conclusion:

This work presents a comprehensive approach to energy cost management for households and policymakers by merging regression and classification techniques. It employs past energy consumption data to create accurate regression predictions and simplify categorizations via classification.

The study examines multiple machine learning models for energy consumption prediction and categorization as low, medium, or high. While all models showed good performance, Ridge Regression and Linear Regression consistently outperformed the others in terms of various performance metrics in regression model. When it comes to recognising data patterns, Decision Tree, Random Forest, and XGBoost demonstrated significant capabilities. Decision Tree fared wonderfully in properly identifying energy consumption levels, but XGBoost exceeded it in terms of Precision, Recall, and F1 Score. Furthermore, within the realm of computational analysis, K-Nearest Neighbors excels as the optimal choice for classification tasks, while Ridge Regression demonstrates its prowess as the superior selection for regression purposes. Additionally, the analysis of Exploratory Data Analysis (EDA) provides further insights into energy consumption patterns, enriching the overall understanding of this critical domain.

The study also employs explainable AI (XAI) tools to grasp the factors impacting model predictions, including contribution plots and SHAP values. It underlines the relevance of specific qualities in both classification and regression settings.

In conclusion, this study delivers important information on patterns of energy utilisation and the performance of machine learning algorithms. Effective energy management strategies and decision-making may be affected by these results, which may aid both families and policymakers.

7.2 Future work:

Future studies may increase the reliability of the results by including building and weather data in the study. The influence of weather on energy consumption is important, and incorporating this data could offer us a fuller view of how external factors affect energy use. Similar to this, taking into consideration building factors like insulation, occupancy, and architectural design may produce a more accurate prediction of energy usage unique to a single structure. For homes and governments trying to effectively regulate energy expenses, this expanded information is likely to deliver more accurate and relevant insights.

References

1. Amiri, S. S., Mostafavi, N., Lee, E. R., & Hoque, S. (2020). Machine learning approaches for predicting household transportation energy use. *City and Environment Interactions*, *7*, 100044.
2. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.
3. Amber, K. P., Aslam, M. W., & Hussain, S. K. (2015). Electricity consumption forecasting models for administration buildings of the UK higher education sector. Energy and Buildings, 90, 127-136.
4. Ahmad, Muhammad Waseem, Monjur Mourshed, and Yacine Rezgui. "Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption." Energy and buildings 147 (2017): 77-89.
5. Asghar, Z. (2008). Energy-GDP relationship: a causal analysis for the five countries of South Asia. Applied Econometrics and International Development, 8(1).
6. Ardabili, S., Mosavi, A., & Várkonyi-Kóczy, A. R. (2019, September). Systematic review of deep learning and machine learning models in biofuels research. In International Conference on Global Research and Education (pp. 19-32). Cham: Springer International Publishing.
7. Bashir, A. K., Khan, S., Prabadevi, B., Deepa, N., Alnumay, W. S., Gadekallu, T. R., & Maddikunta, P. K. R. (2021). Comparative analysis of machine learning algorithms for prediction of smart grid stability. International Transactions on Electrical Energy Systems, 31(9), e12706.
8. Burnett, J. W., & Kiesling, L. L. (2022). How do machines predict energy use? Comparing machine learning approaches for modeling household energy demand in the United States. Energy Research & Social Science, 91, 102715.
9. Cebekhulu, E., Onumanyi, A. J., & Isaac, S. J. (2022). Performance analysis of machine learning algorithms for energy demand–supply prediction in smart grids. *Sustainability*, *14*(5), 2546. s
10. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
11. Chou, J. S., & Bui, D. K. (2014). Modeling heating and cooling loads by artificial intelligence for energy-efficient building design. Energy and Buildings, 82, 437-446.

12. Ding, Y., Fan, L., & Liu, X. (2021). Analysis of feature matrix in machine learning algorithms to predict energy consumption of public buildings. Energy and Buildings, 249, 111208.
13. EIA. (December 2, 2022). Net electricity consumption worldwide in select years from 1980 to 2021 (in terawatt-hours) [Graph]. Available from: https://www.statista.com/statistics/280704/world-power-consumption/ [Accessed: 10 June 2023].
14. Elen, A., & Avuçlu, E. (2021). Standardized Variable Distances: A distance-based machine learning method. Applied Soft Computing, 98, 106855.
15. Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., ... & Xiang, Y. (2018). Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. *Energy conversion and management*, *164*, 102-111.
16. Fayaz, M., Shah, H., Aseere, A. M., Mashwani, W. K., & Shah, A. S. (2019). A framework for prediction of household energy consumption using feed forward back propagation neural network. Technologies, 7(2), 30.
17. Gordic, D., Nikolic, J., Vukasinovic, V., Josijevic, M., & Aleksic, A. D. (2023). Offsetting carbon emissions from household electricity consumption in Europe. Renewable and Sustainable Energy Reviews, 175, 113154.
18. IEA (2019). *Electricity – World Energy Outlook 2019 – Analysis*. [online] IEA. Available from: https://www.iea.org/reports/world-energy-outlook-2019/electricity. [Accessed: 10 June 2023].
19. kaggle (n.d.). *Household Electric Power Consumption*. [online] Available at: https://www.kaggle.com/datasets/uciml/electric-power-consumption-data-set. [Accessed: 20 May 2023] s
20. Kim, M., Jun, J. A., Song, Y., & Pyo, C. S. (2020, October). Explanation for building energy prediction. In 2020 International Conference on Information and Communication Technology Convergence (ICTC) (pp. 1168-1170). IEEE.
21. Kuzlu, M., Cali, U., Sharma, V., & Güler, Ö. (2020). Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. *IEEE Access*, *8*, 187814-187823.
22. Kyriakides, E., & Polycarpou, M. (2007). Short term electric load forecasting: A tutorial. Trends in neural computation, 391-418.

23. Learning, U.M. (2016) Household Electric Power Consumption, Kaggle. Available at: https://www.kaggle.com/datasets/uciml/electric-power-consumption-data-set (Accessed: 05 June 2023).
24. Lee, S., Jung, S., & Lee, J. (2019). Prediction model based on an artificial neural network for user-based building energy consumption in South Korea. Energies, 12(4), 608.
25. Leung, M. C., Norman, C. F., Lai, L. L., & Chow, T. T. (2012). The use of occupancy space electrical power demand in building cooling load prediction. Energy and Buildings, 55, 151-163.
26. Liu, Y., & Li, J. (2022). Annual electricity and energy consumption forecasting for the UK based on back propagation neural network, multiple linear regression, and least square support vector machine. Processes, 11(1), 44.
27. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
28. Maarif, M. R., Saleh, A. R., Habibi, M., Fitriyani, N. L., & Syafrudin, M. (2023). Energy Usage Forecasting Model Based on Long Short-Term Memory (LSTM) and eXplainable Artificial Intelligence (XAI). Information, 14(5), 265.
29. Mohamed, A. E. (2017). Comparative study of four supervised machine learning techniques for classification. *International Journal of Applied*, *7*(2), 1-15.
30. Morgul Tumbaz, M. N., & İpek, M. (2021). Energy demand forecasting: avoiding multi-collinearity. *Arabian Journal for Science and Engineering*, *46*(2), 1663-1675.
31. Mosavi, A., & Bahmani, A. (2019). Energy consumption prediction using machine learning; a review.
32. Nejat, P., Jomehzadeh, F., Taheri, M. M., Gohari, M., & Majid, M. Z. A. (2015). A global review of energy consumption, CO2 emissions and policy in the residential sector (with an overview of the top ten CO2 emitting countries). *Renewable and sustainable energy reviews*, *43*, 843-862. s
33. Nti, I. K., Teimeh, M., Nyarko-Boateng, O., & Adekoya, A. F. (2020). Electricity load forecasting: a systematic review. Journal of Electrical Systems and Information Technology, 7(1), 1-19.
34. 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.
35. Pham, A. D., Ngo, N. T., Truong, T. T. H., Huynh, N. T., & Truong, N. S. (2020). Predicting energy consumption in multiple buildings using machine learning for

improving energy efficiency and sustainability. Journal of Cleaner Production, 260, 121082.

36. Poh, S. C., Tan, Y. F., Cheong, S. N., Ooi, C. P., & Tan, W. H. (2019). Anomaly detection on in-home activities data based on time interval. Indonesian Journal of Electrical Engineering and Computer Science, 15(2), 778-785.
37. Rambabu, M., Ramakrishna, N. S. S., & Polamarasetty, P. K. (2022). Prediction and Analysis of Household Energy Consumption by Machine Learning Algorithms in Energy Management. In E3S Web of Conferences (Vol. 350, p. 02002). EDP Sciences.
38. Reddy, G. V., Aitha, L. J., Poojitha, C., Shreya, A. N., Reddy, D. K., & Meghana, G. S. (2023). Electricity Consumption Prediction Using Machine Learning. In E3S Web of Conferences (Vol. 391, p. 01048). EDP Sciences
39. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5), 206-215.
40. Salam, A., & El Hibaoui, A. (2018, December). Comparison of machine learning algorithms for the power consumption prediction:-case study of tetouan city–. In 2018 6th International Renewable and Sustainable Energy Conference (IRSEC) (pp. 1-5). IEEE.
41. Seyedzadeh, S., Rahimian, F. P., Glesk, I., & Roper, M. (2018). Machine learning for estimation of building energy consumption and performance: a review. Visualization in Engineering, 6, 1-20.
42. Sikiric, G., Avdakovic, S., & Subasi, A. (2013). Comparison of machine learning methods for electricity demand forecasting in bosnia and herzegovina. Southeast Europe Journal of Soft Computing, 2(2).
43. shapash.readthedocs.io. (n.d.). Welcome to Shapash's documentation ! — Shapash 2.3.5 documentation. [online] Available at: https://shapash.readthedocs.io/en/latest/index.html [Accessed 10 Aug. 2023].
44. Sharifzadeh, M., Lubiano-Walochik, H., & Shah, N. (2017). Integrated renewable electricity generation considering uncertainties: The UK roadmap to 50% power generation from wind and solar energies. Renewable and Sustainable Energy Reviews, 72, 385-398.
45. Solyali, D. (2020). A comparative analysis of machine learning approaches for short-/long-term electricity load forecasting in Cyprus. Sustainability, 12(9), 3612.
46. Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy, 32(9), 1761-1768.

47. Tsoka, T., Ye, X., Chen, Y., Gong, D., & Xia, X. (2022). Explainable artificial intelligence for building energy performance certificate labelling classification. Journal of Cleaner Production, 355, 131626.
48. Van Lent, M., Fisher, W., & Mancuso, M. (2004, July). An explainable artificial intelligence system for small-unit tactical behavior. In Proceedings of the national conference on artificial intelligence (pp. 900-907). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.
49. Vijayan, P. (2022, July). Energy consumption prediction in low energy buildings using machine learning and artificial intelligence for energy efficiency. In 2022 8th international youth conference on energy (IYCE) (pp. 1-6). IEEE.
50. Wilfling, S. (2023). Augmenting data-driven models for energy systems through feature engineering: A Python framework for feature engineering. arXiv preprint arXiv:2301.01720.
51. Xiang, L., Xie, T., & Xie, W. (2020). Prediction model of household appliance energy consumption based on machine learning. In Journal of Physics: Conference Series (Vol. 1453, No. 1, p. 012064). IOP Publishing.
52. Zeng, A., Liu, S., & Yu, Y. (2019). Comparative study of data driven methods in building electricity use prediction. Energy and Buildings, 194, 289-300.
53. Zhang, X., Grolinger, K., & Capretz, M. A. (2018, December). Forecasting Residential Energy Consumption Using Support Vector Regressions. In Proceedings of the IEEE International Conference on Machine Learning and Applications, Orlando, FL, USA (pp. 17-18).
54. Zhang, Y., Teoh, B. K., Wu, M., Chen, J., & Zhang, L. (2023). Data-driven estimation of building energy consumption and GHG emissions using explainable artificial intelligence. Energy, 262, 125468.
55. Zou, X., Hu, Y., Tian, Z., & Shen, K. (2019, October). Logistic regression model optimization and case analysis. In 2019 IEEE 7th international conference on computer science and network technology (ICCSNT) (pp. 135-139). IEEE.
56. EIA. (December 2, 2022). Net electricity consumption worldwide in select years from 1980 to 2021 (in terawatt-hours) [Graph]. Available from: https://www.statista.com/statistics/280704/world-power-consumption/ [Accessed: 10 June 2023].