A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF ÇANKIRI KARATEKİN UNIVERSITY

ESTIMATION OF COMMERCIAL BUILDING ENERGY CONSUMPTION WITH MACHINE LEARNING

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRONICS AND COMPUTER ENGINEERING

BY

YOUSIF MURSHID MUHEALDDIN MUHEALDDIN

ÇANKIRI

2023

ESTIMATION OF COMMERCIAL BUILDING ENERGY CONSUMPTION WITH MACHINE LEARNING

By Yousif Murshid Muhealddin MUHEALDDIN

July 2023

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science

Advisor : Asst. Prof. Dr. Selim BUYRUKOĞLU
Co-Advisor : Asst. Prof. Dr. Mohammed Rashad Baker BAKER

Examining Committee Members:

Chairman : Asst. Prof. Dr. Selim BUYRUKOĞLU
Electronics and Computer Engineering
Çankırı Karatekin University

Member : Asst. Prof. Dr. Fuat TÜRK
Electronics and Computer Engineering
Kırıkkale University

Member : Asst. Prof. Dr. Mustafa KARHAN
Electronics and Computer Engineering
Çankırı Karatekin University

Approved for the Graduate School of Natural and Applied Sciences

Prof. Dr. Hamit ALYAR
Director of Graduate School

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Yousif Murshid Muhealddin MUHEALDDIN

ABSTRACT

ESTIMATION OF COMMERCIAL BUILDING ENERGY CONSUMPTION WITH MACHINE LEARNING

Yousif Murshid Muhealddin MUHEALDDIN

Master of Science in Electronics and Computer Engineering

Advisor: Asst. Prof. Dr. Selim BUYRUKOĞLU

Co-Advisor: Asst. Prof. Dr. Mohammed Rashad Baker BAKER

July 2023

The estimation of energy consumption in commercial buildings holds immense significance in the pursuit of sustainable energy management and the efficient allocation of resources. As energy demands continue to rise, optimizing energy consumption becomes crucial for reducing carbon footprints and enhancing cost-effectiveness. In response to these challenges, this research endeavors to leverage the power of machine learning (ML) techniques to accurately predict energy consumption in commercial buildings. By employing ML algorithms, this study seeks to improve energy efficiency, lower operational costs, and facilitate informed decision-making in building energy management. To achieve these objectives, the research employs eight ML algorithms, namely Support Vector Machines (SVM), Random Forest (RF), Extra Trees, Linear Regression, Lasso, Gradient Boosting Regressor (GBR), Multilayer Perceptron (MLP), and a novel stacking model. Each of these algorithms is well-known for its predictive capabilities and ability to handle various types of data. The research methodology encompasses the development and assessment of predictive models using a robust and extensive dataset, carefully collected from commercial buildings. Before applying the ML algorithms, the dataset undergoes rigorous preprocessing to ensure data quality and enhance model performance. Normalization, data cleaning, and feature selection techniques, such as ANOVA and Relief F, are employed to eliminate noise and irrelevant information, making the dataset suitable for training and evaluation. The research splits the data into training and testing sets in a balanced ratio of 70:30, ensuring that the models are trained on a substantial portion of the data while being evaluated on unseen samples.

This balanced data splitting allows for unbiased evaluation and comparison of the performance of the ML algorithms. Remarkably, the results reveal that the stacking model, a novel approach that synergizes multiple models' strengths, outperforms all other ML algorithms and state-of-the-art approaches. The stacking model showcases an impressive Mean Absolute Error (MAE) of 0.01286, Mean Squared Error (MSE) of 0.00054, and Root Mean Squared Error (RMSE) of 0.02328. These exceptional performance metrics demonstrate the stacking model's potential as a highly effective tool for accurately estimating energy consumption in commercial buildings. The research findings are further validated by comparing the results with related works in the field. The stacking model consistently outperforms existing approaches, reaffirming its superiority in accurately predicting energy consumption. This model's success can be attributed to its ability to combine diverse models, leveraging their complementary strengths to capture complex relationships within the energy consumption data. The adoption of the stacking model as the best-performing ML algorithm in energy consumption estimation carries significant implications for sustainable energy practices in the built environment. Accurate energy consumption prediction empowers building managers to make informed decisions on energy allocation, implement energy-saving measures, and optimize energy usage. By enabling data-driven energy management, the stacking model supports the advancement of sustainable energy practices, contributing to the reduction of carbon emissions and the promotion of environmentally friendly building operations. In conclusion, this research showcases the immense potential of machine learning techniques in accurately estimating energy consumption in commercial buildings. The stacking model's exceptional performance, when compared with other ML algorithms and state-of-the-art approaches, highlights its role as a transformative tool in achieving energy efficiency and sustainable energy management. As the world continues to prioritize sustainability, the implementation of the stacking model can significantly contribute to building a greener and more energy-efficient future.

2023, 76 pages

Keywords: Commercial building, Energy consumption, Machine learning

ÖZET

MAKİNE ÖĞRENİMİ İLE TİCARİ BİNA ENERJİ TÜKETİMİ TAHMİNİ

Yousif Murshid Muhealddin MUHEALDDIN
Elektronik ve Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Dr. Öğr. Üyesi Selim BUYRUKOĞLU
Eş Danışman: Dr. Öğr. Üyesi Mohammed Rashad Baker BAKER
Temmuz 2023

Ticari binalarda enerji tüketiminin tahmini, sürdürülebilir enerji yönetimi ve kaynakların etkin dağıtımı açısından büyük önem taşır. Enerji talepleri artmaya devam ederken, enerji tüketimini optimize etmek, karbon ayak izini azaltmak ve maliyet etkinliğini artırmak için hayati önem taşır. Bu zorluklara yanıt olarak, bu araştırma, ticari binalardaki enerji tüketimini doğru bir şekilde tahmin etmek için makine öğrenmesi (ML) tekniklerinin gücünden yararlanmayı hedeflemektedir. ML algoritmalarını kullanarak, bu çalışma enerji verimliliğini artırmayı, işletme maliyetlerini düşürmeyi ve bina enerji yönetiminde bilinçli karar verme yeteneğini kolaylaştırmayı amaçlamaktadır. Bu hedeflere ulaşmak için, araştırma, Support Vector Machines (SVM), Random Forest (RF), Extra Trees, Linear Regression, Lasso, Gradient Boosting Regressor (GBR), Multilayer Perceptron (MLP) ve yeni bir stacking modeli olmak üzere sekiz ML algoritması kullanmaktadır. Bu algoritmaların her biri, çeşitli veri türlerini işleme yeteneği ve tahmini kabiliyetleriyle tanınmıştır. Araştırma metodolojisi, dikkatlice toplanan ticari binalardan geniş ve sağlam bir veri seti kullanarak tahminsel modellerin geliştirilmesi ve değerlendirilmesini kapsar. ML algoritmaları uygulanmadan önce, veri seti veri kalitesini garanti etmek ve model performansını artırmak için titiz bir ön işlemeye tabi tutulur. Normalizasyon, veri temizleme ve özellik seçme teknikleri olan ANOVA ve Relief F kullanılır, gürültüyü ve gereksiz bilgileri ortadan kaldırarak veri setini eğitim ve değerlendirme için uygun hale getirir. Araştırma, modellerin verilerin önemli bir kısmı üzerinde eğitilirken görünmeyen örneklerde değerlendirilmesini sağlamak için veriyi eşit bir oranda 70:30 olarak eğitim ve test setlerine böler. Bu dengeli veri bölme, ML algoritmalarının performansının

tarafsız değerlendirilmesini ve karşılaştırılmasını sağlar. Şaşırtıcı bir şekilde, sonuçlar, birden fazla modelin güçlerini birleştiren yeni bir yaklaşım olan stacking modelinin, tüm diğer ML algoritmalarını ve en son teknikleri geride bıraktığını ortaya koymaktadır. Stacking modeli, etkileyici bir Ortalama Mutlak Hata (MAE) değeri olan 0.01286, Ortalama Kare Hata (MSE) değeri olan 0.00054 ve Karekök Ortalama Kare Hata (RMSE) değeri olan 0.02328 sergilemektedir. Bu olağanüstü performans metrikleri, stacking modelinin ticari binalardaki enerji tüketimini doğru bir şekilde tahmin etmek için son derece etkili bir araç olarak potansiyelini göstermektedir. Araştırma bulguları, sonuçların alandaki ilgili çalışmalarla karşılaştırılarak daha da doğrulanmaktadır. Stacking modeli, enerji tüketimini doğru bir şekilde tahmin etme konusunda sürekli olarak mevcut yaklaşımları geride bırakarak üstünlüğünü teyit etmektedir. Bu modelin başarısı, çeşitli modelleri birleştirme yeteneğine ve enerji tüketimi verileri içindeki karmaşık ilişkileri yakalama kabiliyetine dayandırılabilir. Enerji tüketimi tahmininde en iyi performans gösteren ML algoritması olarak stacking modelinin benimsenmesi, inşa edilmiş çevrede sürdürülebilir enerji uygulamaları için önemli sonuçlar doğurur. Doğru enerji tüketimi tahmini, bina yöneticilerinin enerji tahsisine ilişkin bilinçli kararlar almasını, enerji tasarrufu önlemleri uygulamasını ve enerji kullanımını optimize etmesini sağlar. Veriye dayalı enerji yönetimini sağlayarak, stacking modeli sürdürülebilir enerji uygulamalarının ilerlemesine destek olur, karbon emisyonlarının azaltılmasına ve çevre dostu bina operasyonlarının teşvikine katkıda bulunur. Sonuç olarak, bu araştırma, ticari binalarda enerji tüketiminin doğru bir şekilde tahmin edilmesinde makine öğrenmesi tekniklerinin muazzam potansiyelini sergilemektedir. Stacking modelinin diğer ML algoritmaları ve en son yaklaşımlarla karşılaştırıldığında olağanüstü performansı, enerji verimliliği ve sürdürülebilir enerji yönetiminin sağlanmasında dönüştürücü bir araç olarak rolünü vurgulamaktadır. Dünya sürdürülebilirliği önceliklendirirken, stacking modelinin uygulanması, daha yeşil ve enerji verimli bir geleceğin inşasına önemli ölçüde katkı sağlayabilir.

2023, 76 sayfa

Anahtar Kelimeler: Ticari bina, Enerji tüketimi, Makine öğrenimi

PREFACE AND ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Asst. Prof. Dr. Selim BUYRUKOĞLU and Asst. Prof. Dr. Mohammed Rashad Baker BAKER for their patience, guidance and understanding.

Yousif Murshid Muhealddin MUHEALDDIN
Çankırı-2023

CONTENTS

LIST OF ABBREVIATIONS

ANN	Artificial neural network
ARIMA	Auto regressive integrated moving average
BMR	Basal metabolic rate
BTS	Base transceiver stations
CBECS	Commercial buildings energy consumption survey
DNN	Deep neural network
ET	Effective temperature
FA-HELF	Fast and accurate hybrid electrical energy forecasting
FDI	Fault discrimination information
FPGA's	Field Programmable gate array
FPR	False positive rate
HVAC	Heating ventilating air-conditioning
MAPE	Mean absolute perecentage error
MLP	Multilayer perceptron
MSE	Mean squared error
RMSLE	Normalized mean bias error
NP	Non-deterministic polynomial
PMV	Predicted mean vote
RF	Random forest
RMSE	Root meen square error
SVM	Support vector machine
TS	Thermal sensation

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Energy consumption in commercial buildings are the most popular research topics in the world. Sustainable development is a broad and complex concept that has grown as one of the important issues in the construction industry. The ideas of sustainability is to raise the quality of life. People can live in a healthy environment by improving social, economic and environmental conditions (Aguilar *et al.* 2022). The sustainability project is designed to reuse resources in an efficient manner with the environment and protect resources. The goals of sustainable development emphasize the need to be careful in the consumption of resources in the country. Therefore preparing and compiling standards and principles of green economy, productivity and green management. The government and green organizations are essential in order to wisely save limited resources and preserve natural resources for future generations (Yahman and Setyagama 2022).

Data mining is a technique that was developed specifically to address a specific issue. This approach operates on a substantial quantity of data and carries out the necessary analysis; lastly, a number of repetitive patterns that could be the winning card are extracted from the data. Following that, it will be necessary to identify links between the various patterns, and eventually, a number of significant solutions will be presented for this task. The process of data mining involves gleaning useful information from vast stores of previously mined data. One approach to resolving issues is known as "data mining," which involves sifting through copious volumes of information in order to recognize recurrent patterns (Gray *et al.* 2014). By establishing links between these patterns, it subsequently offers answers to the problems. The data mining is one of the most important in the energy filed. Predictive energy analysis, which can reduce commercial buildings' energy use, has drawn a lot of interest from public bodies, utilities, and building owners (Cao *et al.* 2016, Pérez-Lombard *et al.* 2008). To explore and assess the energy performance of buildings, a variety of methods have been utilized, most notably physics-based models like Energy Plus (Crawley *et al.* 2001,Yezioro *et al.* 2008).

Researchers have created a number of statistical approaches for building energy modeling in order to supplement physics-based models and forecast either overall energy usage or

energy use broken down by end-use (Zhao and Magoulès 2012). Regression-based methods have been frequently employed for estimating energy use from building and occupant data and comparing to simulated outcomes (Deng *et al.* 2018). In order to anticipate heating demand, (Catalina *et al.* 2008) utilised polynomial regression. In a range of cases, it was discovered that the generated model was accurate with a maximum divergence of 5% from the simulated data. Modern machine learning methods have been utilised to resolve complex interactions and nonlinear relationships, such as Deep Neural Network (DNN) and Support Vector Machine (SVM) (Ahmad *et al.* 2014).

According to (Tsanas and Xifara 2012), random forest, a tree-based techniqueestimates the heating and cooling loads of Ecotect's building components more accurately than regression Autodesk, San Rafael and CA. To estimate hourly building cooling demand, (Li *et al.* 2009) built models based on Artificial Neural Network (ANN) and SVM which they then used to a Chinese office building. They observed that the prediction errors from the SVM-based model were almost 50% smaller than those from the DNN model. Energy Plus was compared to an DNN model that predicts energy usage by Neto and Fiorelli (Neto and Fiorelli 2008), and it was concluded that both techniques are appropriate candidates with comparable performance. Collectively, these studies show that using empirical methods to solve nonlinear relationships can produce results that are just as accurate as or even more accurate than those obtained using regression-based models to analyze simulated data. The evaluation of how effectively these sophisticated statistical models predict actual building energy performance as opposed to simulated results (De Wilde 2014).

Numerous case studies have employed sophisticated algorithms using past measurements to forecast the energy performance of buildings (Gonzalez and Zamarreno 2005). employed a feedback ANN for hourly electricity used. Based on five years' worth of utility bills and meteorological information, the SVM model with radial-basis function kernel (Dong *et al.* 2005) was used to forecast the monthly energy usage of four commercial buildings in Singapore. In an empirical investigation that forecasted the use of electricity in residential buildings, (Tso and Yau 2007) examined regression, decision tree, and DNN models and discovered that both performed better than regression

techniques. In general, recent case studies have proved that machine learning algorithms provide trustworthy outcomes. Additionally, because the majority of machine learning algorithms are nonparametric, it is not always necessary to make assumptions about probability distributions because they may model nonlinear interactions (Hollander *et al.* 2013). However, the majority of earlier studies have looked into instances that are particular to one structure or a small group of buildings in one place. The significance of various architectural and occupant factors on building energy performance utilizing these methods is yet unknown because generic prediction models based on machine learning techniques are still lacking (Nooruldeen *et al.* 2022).

1.1 Motivation

- Cost reduction: Accurate energy estimation helps identify areas for improvement, leading to cost reductions and improved energy efficiency for commercial buildings.
- Environmental sustainability: Estimating energy usage enables strategies to reduce consumption, lower carbon footprints, and contribute to sustainability goals.
- Resource management: Accurate estimation optimizes resource allocation, identifies inefficiencies, and enables targeted improvements in lighting, and appliances.
- Predictive maintenance: Energy estimation helps detect anomalies and system issues, preventing breakdowns, reducing downtime, and improving operational efficiency.
- Economic incentives: Accurate energy estimation quantifies potential savings, making it easier to access economic incentives for energy-efficient practices.

1.2 Aim of Study

Utilising machine learning methods to calculate energy usage is the goal of this thesis. Several methods, such as support vector machines, random forests, linear regression, extra trees, lasso regression, gradient boosting, multilayer perceptron and stacking regression will be employed to identify the most accurate and high-performing model compared to other approaches. In summary, the primary goal of this thesis is to achieve the following:

- Feature selection: the best and efective features select to use in the machine learning.
- Suitable ML methods: the best algorithm will be selected to the regression of energy consumbtion. These algorithms are single based, bagging and boosting based methods.

1.3 Objectives of Study

The main objectives of thesis can be noted as:

- Focusing on the most important features, feature selection enables machine learning to reduce the mistake rate when identifying infected data from electrical energy consumption data.
- Selecting machine learning methods to use and developing generalised prediction models for the actual energy consumption of subsystems in commercial office buildings.
- Comparing of the predicting abilities and statistical analysis of machine learning methods.

1.4 Research Questions

In this thesis different types of the questions have been solved. This research has some question such as:

RQ1: Does the feature selection techniques are valuable for this research

RQ2: Does ensemble learning model is better than single based algorithm?

RQ3: Does bagging ensemble algorithm better than boosting?

RQ4: Does stacking model is convenient for energy consumption and why?

1.5 Contributions

The main contribution of this research can be highlighted as the following points:

- Suitable feature determination: The study employs feature selection methods to select the most relevant features for accurate energy consumption estimation, leading to better predictions and understanding of energy patterns in commercial buildings.
- Through the implementation and testing of the proposed machines learning techniques, the findings provide valuable guidance for energy management, resource allocation, and the development using the stacking regression model to optimize energy usage and reduce costs.

1.6 Thesis Organisation

We have arranged this thesis in the manner that is shown below. Chapter 2 includes a study of several methodology for the thesis's optimization methods. The application of the chosen technique is shown in Chapter 3 along with examples of the methods and model analyses. Chapter 4 also presents and analyses the proposed algorithm for a system under consideration, and Chapter 5 examines the results and the direction of future study. Finally, chapter five will contain recommendations to guide future studies (Figure 1.1).

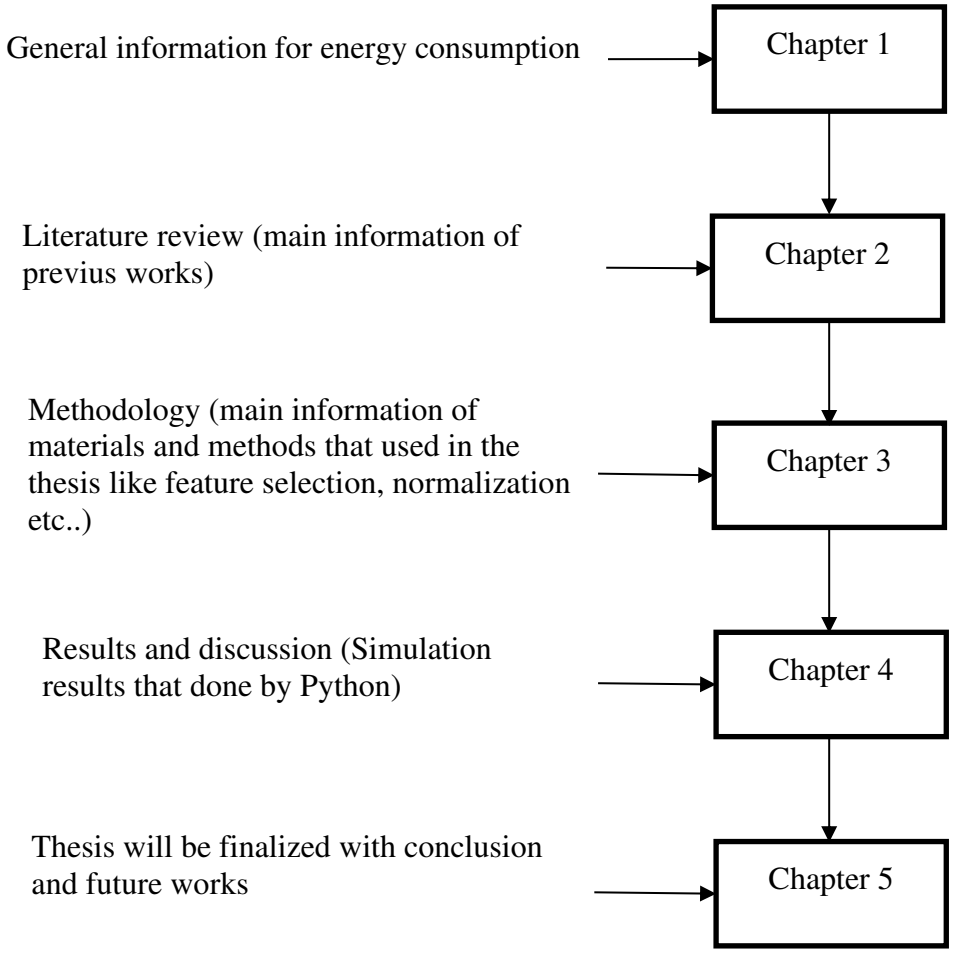

Figure 1.1 Thesis work flow chart

2. LITERATURE REVIEW

2.1 Introduction

All of the energy used to carry out an action, create something, or simply occupy a structure is referred to as energy consumption. Here are a few illustrations: total energy consumption in a factory can be calculated by looking at how much energy is used during a production process, such as when making car parts. In this chapter, the literature reiew of the articles have been investigated. The different types of the machine learning methods have been investigated. In next sections the energy consumption overwies has ben investigated. Also the machine learning methods are discused and how to use in energy consumption. Single based algorithms, bagging based algorithms, and boosting methods have been discussed.

2.2 Overview of Energy Consumption

There is strong evidence to imply that variations in energy use and climate-changing greenhouse gas emissions are directly related to diverse built environment designs (Nichols and Kockelman 2014, Guhathakurta and Williams 2015). According to 2009 National Academy-commissioned study, raising development density results in minor energy savings for transportation, and thus, a decrease in greenhouse gas emissions (Council 2009).

Energy consumption is the term used to describe the total amount of energy consumed to perform an action, generate something, or even occupy a space (Omer 2008). Here are a few examples: By examining how much energy is consumed during a production process, such as when making vehicle parts, the total energy consumption in a plant can be computed.

2.3 Efficiency of Feature Selection

Efficiency of feature selection techniques on energy consumption regression has been extensively studied in the literature. Various methods have been proposed and evaluated for their effectiveness in selecting relevant features that contribute to accurate energy consumption predictions

2.3.1 Relief f feature selection

The Relief algorithm (Urbanowicz *et al.* 2018) initially introduced by Kira and Rendell, has gained widespread adoption and has been extended in various forms, including Iterative Relief, RReliefF, and Multi-Relief. These methods have demonstrated superior performance compared to other feature selection techniques in terms of accuracy and efficiency. In energy consumption research, Relief-based algorithms have been successfully applied offering an invaluable approach to enhance the interpretability and performance of ML models. The feature selection process in the FA-HELF framework (Hafeez *et al.* 2020) , Relief-F and random forests are used to gauge feature importance. The feature selector evaluates each characteristic's significance before deciding whether to keep it or remove it depending on a threshold. The forecasting process is made more accurate and effective by using feature selection algorithms like Relief-F. These techniques aid in locating and choosing the dataset's most important features, which improves speed.

Solano *et al.* (2022) incorporates Relief feature selection as part of an ensemble feature selection methodology to effectively choose the most relevant variables for solar radiation forecasting. Relief is an algorithm that assesses the importance of each variable by considering its values in relation to the target variable. By combining Relief with other feature selection algorithms like Mutual Information and Random Forest, the ensemble feature selection method aims to identify crucial exogenous variables and their corresponding delay values, enabling accurate solar radiation forecasting.

2.3.2 ANOVA feature selection

ANOVA (Analysis of Variance) feature selection plays a crucial role in estimating energy power using ML algorithms. It enables the identification of significant features that greatly contribute to energy consumption prediction. By selecting the most informative features, ANOVA feature selection enhances the accuracy and interpretability of models, resulting in more effective energy consumption estimation.

Several studies have demonstrated the application of ANOVA feature selection in energy power estimation. For instance, in this study (Daud *et al.* 2015), ANOVA is utilized to identify the most pertinent features by comparing the means across different categories. By employing ANOVA, the researchers can pinpoint which features remain consistent and accurate despite the influence of diverse factors or disturbances. Moreover, ANOVA aids in distinguishing genuine and forged signatures. It is also applied in feature extraction for the classification of Parkinson's disease. Overall, ANOVA serves as a valuable statistical technique for analyzing and selecting features in this research endeavor.

By utilizing ANOVA, the researchers (Camposeco-Negrete 2013) analyzed the cutting power consumed and surface roughness Ra, partitioning the total variation into accountable sources and assessing the contribution of each factor (cutting parameters) to the observed variation. Through ANOVA analysis, the researchers determined the significance of each factor and its impact on the response evaluation. For cutting power consumed, the cutting velocity was found to be insignificant, indicating its negligible contribution to the response evaluation. However, for surface roughness Ra, all factors were deemed significant. ANOVA also provided the researchers with percentage values representing the influence of each factor, enabling the identification of more impactful factors among others. This information facilitated the determination of appropriate factor levels that offer economic benefits or operator convenience.

2.4 Energy Consumption with Machine Learning

The most effective method to identify energy usage in business buildings is helped by machine learning algorithms. (Robinson *et al.* 2017) present a novel method for estimating commercial building energy consumption from a constrained set of building parameters by training ML models using national data from the Commercial Buildings Energy Consumption Survey (CBECS)

The substantial link between climate change, indoor air quality, and building energy use has been shown in numerous research. The building sector, which accounts for 40% of a country's yearly energy consumption and 30% of its greenhouse gas emissions, is mostly responsible for this. Therefore, predictive analysis of building energy use can help to lower energy use in public buildings (Bilgen 2014).

The three box models are the white box, black box, and grey box (hybrid model) are three different ways to anticipate how much energy a building will need. Since the white box technique is founded on physical principles, the prediction model must incorporate characteristics and energy transfer mechanisms that actually have physical meaning. Building energy modeling tools, such as Energy Plus (Jia and Chong 2021), are currently being extensively utilized to study and assess the energy performance of buildings (Gourlis and Kovacic 2017). This approach frequently uses numerous calculations and comprehensive architectural information to calculate the energy usage of a handful of buildings. According to certain studies (Zhao and Magoulès 2012, Zheng *et al.* 2020), the simulation software's calculation results will be significantly impacted by the complexity of the building model and the specifics of the building information (Schlueter and Thesseling 2009).

With advancements in machine learning techniques and increased availability of sample datasets, it is now possible to explore the intricate relationship between actual energy usage and architectural features (Alahakoon and Yu 2016). It's common to undervalue the role that feature engineering has in model performance. We investigated the impact of attributes on model performance using machine learning approaches to predict the

energy consumption (Energy Usage Intensity, or EUI) of public buildings (Wang *et al.* 2018). The Recursive Feature Elimination (RFE) method was employed to eliminate redundant features and obtain the optimal feature subset. The value of the features was assessed using the learning methods, and only a select few features from the complicated feature matrix were shown to have a substantial impact on the machine learning model.

In this section, four methods has been explained. These methods are single based algorithms, bagging, boosting, and stacked ensemble methods.

2.4.1 Single based algorithms

Single-based algorithms only use one hand-based model. In this section, we examine various studies that apply single-based algorithms to address the issue of building energy usage.

To increase the effectiveness of energy management, a framework for quick and precise hybrid electrical energy forecasting (FA-HELF) has been developed. It has three modules: a forecaster based on SVM, an optimizer based on modified improved differential evolution (MEDE), and data pre-processing and feature engineering. Data from the ISO New England control area's hourly load is used to assess the framework's efficacy. Results show that the FA-HELF framework greatly increases forecast accuracy compared to spark distributed (SDPSO), F-RBF-CNN, and SSA-SVM-CS by 77.65%, 58.66%, and 57.39%, respectively.. Notably, the framework offers the advantage of simultaneously predicting convergence rate and accuracy. However, it is important to acknowledge that the framework's implementation entails high system complexity.

In (Amin and Hoque 2019) two popular techniques for predicting short-term load, SVM and Auto Regressive Integrated Moving Average (ARIMA), are compared in their ability to forecast day-ahead load. The dataset employed consists of projected electricity consumption patterns for 28 base transceiver stations (BTS) between March and May 2018. The SVM model is trained using electricity consumption data from the same period,

while the ARIMA model employs the identical data for forecasting. Although the SVM model exhibits superior accuracy compared to ARIMA, it encounters challenges in accurately predicting the magnitude of load peaks due to sudden shifts in consumer behavior. On the other hand, SVM excels at forecasting non-linear load components that are sensitive to change but falls short when it comes to predicting non-linear external factors like weather and days of the week. The calculated MSE values for SVM and ARIMA are 0.12644 and 4457.2, respectively.

Other researchers (Han *et al.* 2019) proposed a trust evaluation model for underwater acoustic sensor networks (UASNs) using machine learning algorithms. The dataset utilized in the study comprises three types of trust evidence: communication trust, packet trust, and energy trust. The methodology involves dividing the network into clusters and utilizing k-means and SVM algorithms to develop the trust evaluation model. Results indicate that the proposed model effectively addresses the challenge of limited evidence in sparse underwater deployment environments. A key advantage of this model is its capacity to enhance network security and prolong network lifetime. However, a limitation of the study is the insufficient exploration of the influence of underwater environmental factors on trust prediction accuracy. The success rate of the proposed model was found to be 97%.

Torabi *et al.* (2019) designed a short-term planning model electrical energy demand forecasting in Bandarabbas, Iran. The dataset used encompassed 18 months of energy consumption data from 2010 and 2011. The study followed the Cross Industry Standard Process (CRISP) for Data Mining and compared the performance of three models: SVM, ANN, and a novel hybrid model called CBA-ANN-SVM. The CBA-ANN-SVM model employed clustering to divide the data into subsets and applied SVM and ANN for forecasting within each subset. Results revealed that the novel model, with three clusters, achieved the lowest average percentage error (APE) of 1.297. The advantage of the CBA-ANN-SVM model lies in its combination of the strengths of both SVM and ANN, as well as the benefits of clustering. However, a drawback is that the user needs to specify the desired number of clusters as an input parameter for the algorithm. Future research could

explore the utilization of fuzzy methods and genetic algorithms or a hybrid approach to forecast each cluster.

Daneshfaraz *et al.* (2021) investigated how the aperture diameter of horizontal screens affects vertical hydraulic parameters using the SVM method. The study analyzed 164 experimental datasets, with the first 82 focusing on relative energy consumption and the remaining datasets on relative pool depth. Three different aperture diameter sizes were tested, and the SVM method was utilized to predict energy dissipation and relative pool depth. The findings indicate that the SVM method accurately predicts hydraulic parameters, and the best models were determined based on statistical criteria such as RMSE, R2, and KGE. In the first scenario, the SVM method achieved a training mode RMSE value of 0.00565 and a testing mode Root meen square error (RMSE) value of 0.00489 when predicting energy dissipation. In the second scenario, the SVM method yielded a training mode RMSE value of 0.095 and a testing mode RMSE value of 0.0899 for predicting relative pool depth. The advantage of the SVM method is its capability to handle high-dimensional datasets, although it is sensitive to the choice of kernel function.

Ahmad *et al.* (2017) compares the performance of Random Forest and ANN for predicting hourly electricity consumption in hotel buildings. It utilizes a dataset comprising historical energy consumption, weather, and hotel occupancy data for one year. The methodology involves developing and training the models using the dataset and evaluating their performance using metrics like RMSE. ANN showing higher accuracy on the testing dataset. Overall, ANN performed marginally better than RF with RMSE of 4.97 and 6.10 respectively. Proposed models offer advantages in handling complex, high-dimensional data, identifying trends, and analyzing what-if scenarios. However, they require substantial training data and may underperform if data doesn't represent the actual system. The study underscores the potential of machine learning algorithms for energy prediction in hotels, suggesting further research on ensemble-based algorithms and the use of big data technologies for training and deploying models.

Younes *et al.* (2021) presented a selection-based K-nearest neighbor (KNN) architecture for smart embedded hardware accelerators. The selection of the KNN algorithm was

based on its effective parallelization and ability to achieve accurate classification with reduced computational complexity. The architecture was evaluated using a dataset with two touch modalities (rolling and sliding) for touch modality classification. The KNN classifier, specifically using 3-nearest neighbors, achieved the highest classification accuracy of 89.8%. The proposed architecture demonstrated significant speed improvements compared to its software counterpart and consumed less energy than GPU implementations. The main advantage of the architecture is its optimal utilization of FPGA's parallelism capabilities, although careful selection of hardware resources is necessary to balance performance and resource utilization.

2.4.2 Bagging ensemble learning algorithms

This section provides information about the energy consumption studies developing with Bagging ensemble learning algorithms. Bagging is an ensemble technique that improves prediction accuracy and stability by developing several classifiers trained on portions of the training data spread at random. Each classifier's training set is produced by randomly selecting and replacing N examples from the original data. This process introduces diversity among the classifiers, mitigating the risk of overfitting and enhancing prediction robustness. Random Forest is an example of a bagging algorithm that creates multiple subsets and trains individual models on each subset. The predictions from these models are combined to make the final prediction. Bagging-based algorithms are effective for handling complex and noisy datasets and reducing the impact of outliers. While they increase computational complexity, bagging-based algorithms have demonstrated reliability and accuracy across various domains in machine learning tasks (Jafarzadeh *et al.* 2021).

The prediction model was an adjudication tree (Fadda *et al.* 2018, Liu *et al.* 2011). The subject's target value is based on the model's prediction from the subject's observations up to that decision. Branch and leaves are other names for the subject's observations and target values, respectively. Bagging is a method of minimizing the variance of an estimated prediction that works well with decision trees (Bühlmann *et al.* 2009). A huge group of independent trees are produced by the RF method, which averages their outputs

(Athey *et al.* 2019). Since every tree produced by bagging has an identical distribution, there are few ways to improve except reducing variation. In random forest method, By minimising correlation between trees while barely increasing variance, RF enhances bagging by creating trees using random input variable selection.

Chen *et al.* (2019) developed a ML classifier for predicting energy consumption levels using a real dataset obtained from a large hypermarket. The dataset comprised 12 months of energy consumption data and hourly temperature records. The data was preprocessed using the latest data class interval method and compared against conventional forecasting techniques. The proposed approach utilized a RF classifier, which yielded superior results in terms of classification accuracy and execution time compared to the conventional method. The model reach a classification accuracy of approximately 90% for three-class predictions. However, the accuracy drops to around 83% and 75% for five- and seven-class predictions, respectively. One benefit of the suggested strategy is its capacity to avoid overfitting and deliver more accurate forecasts. However, it should be noted that the proposed method requires greater computational resources compared to the conventional method. Overall, this study highlights the potential of ML techniques for accurate energy consumption prediction.

Wu *et al.* (2018) aimed to create a robust model using ML algorithms to predict thermal comfort. It was conducted in Changsha, China, involving 24 dormitory buildings and 813 participants. The dataset consisted of various parameters, including environmental conditions, microclimate data, and demographic information. The study specifically examined thermal behaviors like adjusting clothing, opening windows/doors, using air-conditioners, and employing fans. Three ML algorithms (Bagging, ANN, and SVM) were utilized to predict thermal sensation, and their performances were compared. The findings demonstrated that the Bagging approach outperformed the other two methods in terms of accuracy and linear relationship. The R2 values for the Bagging model were 0.4986, 0.9892, 0.9920, and 0.9900 for TS, prediction mean value (PMV), Standard effective temperature (SET) and effective temperature (ET). The Bagging approach offers advantages such as improved prediction performance and the ability to handle complex

and high-dimensional data. However, it should be noted that a significant amount of training data is required for the Bagging approach to be effective.

2.4.3 Boosting

This section provides information about the energy consumption studies developing with Boosting ensemble learning algorithms. Approaches emphasize the development of several classifiers. The training set used by each next member of the series is chosen based on how well the prior classifier(s) performed. Examples that were correctly predicted by the series' previous classifiers are chosen less frequently in boosting (Opitz and Maclin 1999). Boosting seeks to construct new classifiers to enhance the performance of the current ensemble for situations for which it is unable to anticipate. Notably, boosting resamples the training set while taking into consideration the success of the prior classifiers, in contrast to the bagging ensemble.

(Touzani *et al.* 2018) evaluates the accuracy of various ML models in predicting energy use in commercial buildings. The study utilizes a dataset containing hourly energy use data from 495 buildings over 12 months. Three machine learning models, namely gradient boosting machines (GBM), RF, and time-series ordinary least squares (TOWT), are employed, and their performance is assessed using metrics like R2, normalized mean bias error (NMBE), and coefficient of variation of the (CV(RMSE)). The findings reveal that GBM models outperform the others in terms of R2, particularly when utilizing k-fold block cross-validation. However, the GBM model's performance diminishes when the training period is shortened. The best R2 rate of the proposed model is 83.96. The advantage of this study lies in providing insights into the accuracy of different ML models for predicting energy use in commercial buildings. Nonetheless, a limitation is that the study only considers data from a single region and doesn't evaluate the models' performance on data from other regions.

The performance of three gradient boosting algorithms LightGBM, CatBoost, and XGBoost for predicting the short-term energy consumption of buildings is compared in this work (Touzani *et al.* 2018) using a synthetic dataset. The dataset also contains metre

readings for several forms of energy together with weather and building data. RMSLE, or Root Mean Squared Logarithmic Error, is the evaluation metric employed. The information is synthesised and may not accurately reflect real-world scenarios, which is a limitation of the study even though it offers a direct comparison of the three models. On this dataset, the XGBoost model fared the best, achieving an RMSLE score of 0.897.

The study (Zhang *et al.* 2023) showed that urban morphology and building geometry have a big impact on how much energy buildings use and how much greenhouse gas they emit. When these parameters were taken into account, prediction accuracy was 33.46% higher than when simply building attributes were taken into account. The largest influence on energy consumption was found to be the total gross floor area, whereas the biggest influence on greenhouse gas emissions was found to be natural gas. The Light Gradient Boosting Machine and the SHapley Additive exPlanation algorithm enabled the suggested model to attain great accuracy, with an average R2 value of 0.8435. Additionally, it fared better than competing strategies like Support Vector Regression, Random Forest, and eXtreme Gradient Boosting..

Ding *et al.* (2021) investigates how different datasets impact the performance of ML models. The study uses a dataset of 2370 public buildings in China, comprising 70 features. The methodology involves preprocessing the data, selecting a ML algorithm, splitting the dataset for training and testing, and evaluating model performance. The results highlight the significant role of the dataset in model performance, emphasizing the importance of feature selection and the distribution of feature and target values. The R2 for XGBoost was 0.82 and 0.625, respectively. This study provides valuable insights to enhance machine learning model performance in building energy consumption prediction. However, it should be noted that the dataset is limited to public buildings in China, and the results may not be generalized to other regions or building types.

2.4.4 Stacked

This section provides information about the energy consumption studies developing with Stacked ensemble learning algorithms. Stacking is a broad method in which a learner is

taught to mix individual learners. Individual learners are referred to as first-level learners in this context, whereas the combiner is referred to as a second-level learner, or meta-learner. Stacking, also known as Super Learning or Stacked Regression, is a type of technique that involves training a second-level "metalearner" to identify the best combination of base learners. Unlike bagging and boosting, the purpose of stacking is to group together strong, diverse groups of learners (Mabayoje *et al.* 2018).

Moon *et al.* (2020) introduces a novel stacking ensemble model, for day ahead electric load forecasting in a university building. The dataset used spans from January 2016 to December 2018, comprising 26,280 hourly electric energy consumption data points. The methodology involves employing deep neural network (DNN) models as level 0, and a Linear regressionmodel as a level 1 to combine their predictions. Results demonstrate that the their model outperforms other forecasting models in terms of MAPE and MAE with values 6.97, .84 respectivly. Their model's ability to effectively capture recent electric load patterns and alter the weights of each DNN model using the sliding window method is a significant benefit. However, a limitation is the challenge of accurately predicting electric energy consumption during weekends.

Li *et al.* (2021) proposed a fault-discrimination information (FDI)-based augmented stacking-based sensor failure detection approach for building energy systems. The research makes use of an original operational dataset from a building energy system that is segmented into training, validation, and test sets. The training set is used to train four base-learner models, and their parameters are optimised with the validation set to produce optimised base-learners. The suggested work has a false positive rate (FPR) of 0.83 and the Received Operating Characteristic (ROC) area under curve (AUC) larger than 0.8. The method offers the advantage of high detection performance and low false alarm rates. However, it may not consistently enhance overall fault detection performance. Nevertheless, the proposed method demonstrates promise for enhancing fault detection in building energy systems. Level 0 of the suggested method employed support vector machine, K-means, and autoencoder. Level 1 entails stacking the outputs of the base-learners and utilising the difference item that carries the FDI. The best base-learner is then used to build the meta-learner.

Bolandnazar *et al.* (2020) aimed to develop a machine learning model that predicts energy consumption in greenhouse cucumber production. The dataset used consists of various inputs, including electricity, natural gas, and labor, along with outputs like cucumber yield and energy consumption. Although the dataset size is unspecified, the methodology involves utilizing multiple Linear regression and ANN for level 0 and SVR model for level 1 for energy consumption prediction. To evaluate the models, they are trained and tested using k-fold cross-validation. The assessment of RMSE, MAPE, and R2 using k-fold cross-validation on different randomized data sets (50%, 60%, 70%, 80%, and 90%) revealed that the Radial basis function (RBF) model consistently excelled in predicting performance compared to other models. Notably, the RBF model achieved an impressive R2 value of 0.98 across various sizes of training data. The advantage of employing machine learning techniques is their capability to handle complex data and non-linear relationships. However, it should be noted that these techniques demand more computational resources and may overfit the data. Overall, this study highlights the potential of machine learning in accurately predicting energy consumption in greenhouse cucumber production.

Divina *et al.* (2018) presented a stacking ensemble method for time series forecasting, utilizing three base learning approaches (Evolutionary Algorithms-based regression trees, Artificial Neural Networks, and Random Forests) along with a top method (Generalized Boosted Regression Models). The dataset used comprises 9 years and 6 months of electricity consumption data from Spain, with measurements taken at 10-minute intervals, resulting in a dataset size of 497,832. The results demonstrate that the proposed ensemble approach outperforms the individual base methods, achieving RMSE values ranging from 0.012 to 0.015. The advantage of this method lies in its ability to leverage the strengths of different algorithms, while the potential drawback is the increased complexity of the model. The level 0 algorithms employed in this research include Evolutionary Algorithms-based regression trees, ANN, and RF, while the level 1 algorithm is Generalized Boosted Regression Models.

In this chapter the literature review of the methods has been presented. The different types of the ML methods has been investigated and the results of the literature are

presented for the energy consumption with Machine Learning methods such as single based algorithms, bagging, and boosting methods has been investigated. Also, the used types of feature selection techniques. Finally, with stacked ensemble learning for energy consumption this chapter has been finalized. The folowing section presensts detailed information about the proposed learning models for energy consumption.

3. MATERIALS AND METHODS

This chapter describes the steps of the methodology that we used. Figure 3.1 depicts the proposed block diagram. The section will outline the various steps in our pipeline, encompassing data collection, preprocessing like cleaning, normaliation, feature selection and data splitting, model training, and model evaluation.

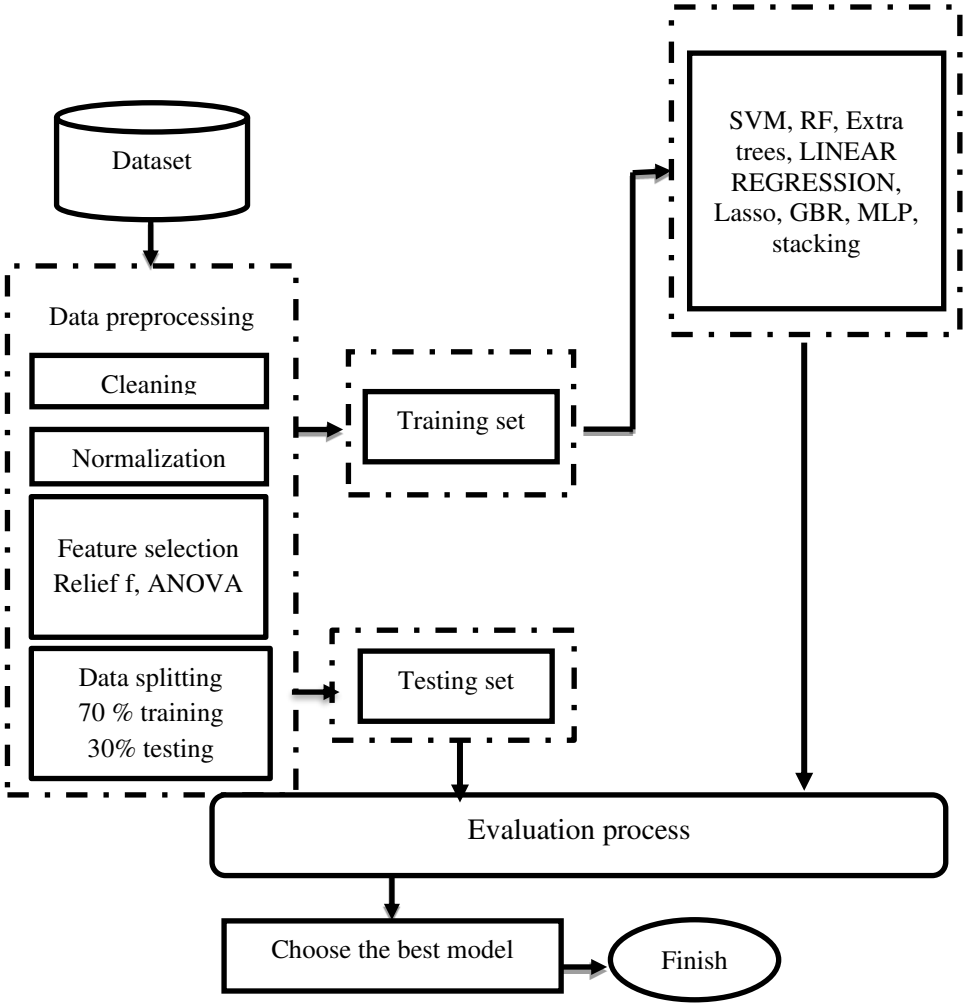

Figure 3.1 Proposed approach for the energy building consumption

3.1 Data Collection

In this study, we uses the energy consumption dataset from CBECS data (Deng *et al.* 2018). Over a 15-year period (from 2003 to 2018), the total number of buildings increased by 22%, while total floorspace increased by 35%, according to the CBECS. The overall

square footage increased by 11% between 2012 and 2018. There were fewer buildings from 2012 to 2018, although the increase was not statistically significant. From the initial CBECS in 1979 to the 2018 CBECS, the number of buildings climbed from 3.8 million to 5.9 million (56%), and the amount of commercial floorspace increased from 51 billion square feet to 96 billion square feet (89%). The total number of records is 1016, with 78 features and one output. Table 3.1 represent the first five columns from the dataset.

Table 3.1 Sample of the dataset used

CENDIV	SQFT	WLCNS	RFCNS	RFTILT	TOTALEUI
5	89000	3	4	0	30.397618
5	1500	3	4	0	8.525333
5	7900	2	2	0	24.213165
9	10000	1	2	0	32.7632
5	600000	3	1	0	153.781077

As observed in Table 3.1, it represnt only five raw samples from the dataset with only five features and the output but the the original dataset containes many instances which is equal to 1016 records and many features which is 78 features. The appendixes section at the end of this thesis contains a detailed description of the dataset's features as well as complete instances of the dataset.

3.2 Data Preprocessing

To gain a better understanding of the various facets of the issue we tackle, we employ numerous preparation stages and data analysis. We break down the preprocessing into components and provide detailed illustrations for each module. Then, we use the clean module, which is the first module and checks for outliers, redundant rows, null values, and redundant columns. Due to the simplicity of the data, the clean module is the simpler one. The next step is the normalization process, feature selection and finally the data splitting process. The preprocessing step is shown in Figure 3.2.

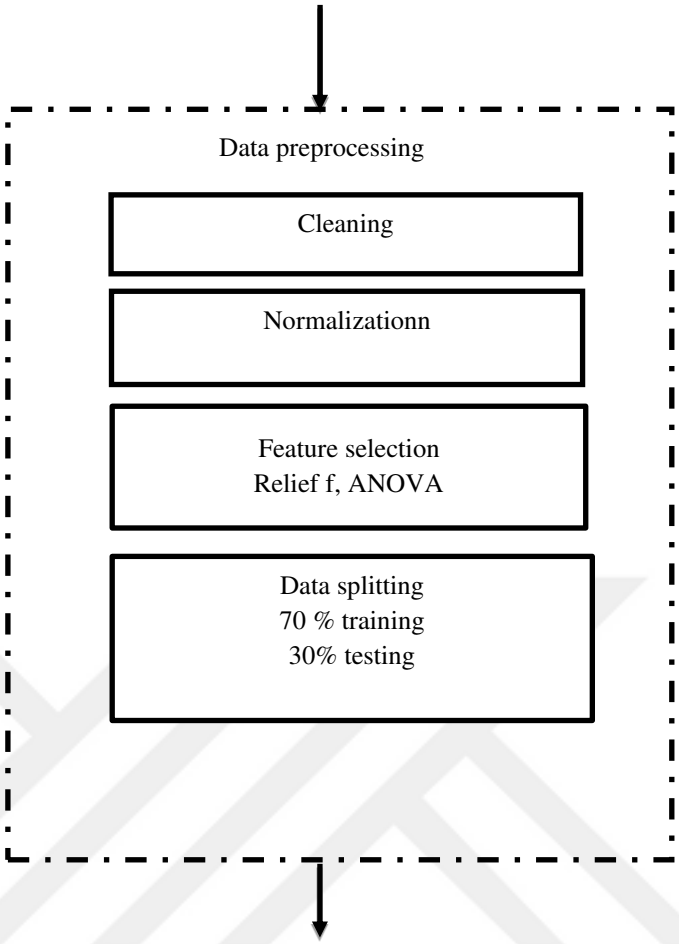

Figure 3.2 Preprocessing step

3.2.1 Data cleaning

Data cleaning ensures dataset quality by addressing null values and removing duplicate rows. For null values, techniques include identification, analysis, deletion (for minimal and random occurrences), and imputation (replacing missing values with estimates). To remove duplicate rows, identify them by comparing columns, delete duplicates, and consider key columns as unique identifiers. Caution and domain expertise are crucial to preserve data integrity during cleaning.

3.2.2 Normalization

Normalization is a technique used in data preprocessing to adjust numerical features to a common range. Min-Max normalization, also known as Min-Max scaling or feature scaling, is a common method of normalization (Patro and sahu 2015).

Here's how Min-Max normalization works:

1. Identify the range: Determine the minimum and maximum values for the feature to be normalized.
2. Define the new range: Determine the desired range for the normalized feature. Typically, this is chosen to be between 0 and 1.
3. Apply the formula: Subtract the minimum value from each value in the feature and divide it by the range (the gap between the maximum and minimum values). This transforms the values to a range between 0 and 1.

The formula for Min-Max normalization is as follows Equation (3.1)

$$Normalized = (value - min) / (max - min) \quad (3.1)$$

We found the dataset has no redundant columns or null values so, the size of the new dataset is the same like the original one.

3.2.3 Data splitting

In machine learning, data splitting is used to evaluate model performance and generalization. We used the common ratio is 70:30, dividing the dataset into a 70% for training set and a 30% for testing set. Steps involve shuffling the data, allocating 70% to training and 30% to testing, training the model on the training set, and evaluating its performance on the testing set. This procedure aids in estimating the model's accuracy on previously unseen data and identifying potential overfitting.

3.3 Feature Selection

In this thesis, the accurate features have been selected and tested. Also, the best features examined and found it in optimum features.

3.3.1 Relief f feature selection

The Relief algorithm works by estimating the quality of features by measuring the difference in feature values between nearest neighbors. It assigns weights to each feature based on how well they contribute to the prediction of the target variable. Features with higher weights are considered more important for the energy consumption regression task (Urbanowicz *et al.* 2018).

Relief is a popular feature selection algorithm that can be applied to energy consumption regression tasks. It is a filter-based method for determining the relevance and significance of characteristics by estimating their ability to discriminate between instances. Relief evaluates the weights of features based on the differences in feature values among nearest neighbors.

The Relief algorithm follows these steps:

Initialize feature weights: Assign an initial weight of zero to all features.

For each instance in the dataset:

1. Find the k nearest neighbors based on a distance metric.
2. Update the feature weights based on the differences in feature values between the instance and its neighbors. Features that have larger differences contribute more to the weight update.
3. Normalize the feature weights to ensure they are in the range [0, 1].

4. Select the top-ranked features based on their weights for energy consumption regression.

Relief feature selection has been applied to energy consumption regression in various research studies. By identifying the most influential features, Relief helps improve the accuracy of energy consumption predictions and enhances the understanding of factors affecting energy usage (Urbanowicz *et al.* 2018)

3.3.2 ANOVA feature selection

ANOVA (Analysis of Variance) feature selection is a commonly used technique for selecting relevant features in statistical investigation and ML tasks. ANOVA measures the statistical significance of the relationship between each feature and the target variable by analyzing the variance of the feature across different classes or groups.

The features are ranked in ANOVA feature selection based on their F-values, which represent the ratio of between-class variation to within-class variance. Higher F-values indicate a stronger relationship between the feature and the target variable (Brownlee 2019).

Here's an overview of the ANOVA feature selection process:

1. Group the data: Divide the dataset into different groups or classes based on the target variable.
2. Compute variance: Calculate the variance of each feature within each group.
3. Compute F-value: Calculate the F-value for each feature by contrasting between-class and within-class variance.
4. Rank the features: Rank the features based on their F-values. Higher F-values indicate stronger associations with the target variable.
5. Select top-ranked features: Select a predetermined number or a specific percentage of the top-ranked features for further analysis or model building.

ANOVA feature selection has been widely used in various domains, including biology, social sciences, and machine learning, to identify relevant features that contribute significantly to the target variable. By focusing on features with higher F-values, ANOVA helps in dimensionality reduction and improves the interpretability and performance of models.

3.4 Machine Learning Algorithms

In this section, eight machine learnng methods which are employed in this thesis are prsented. These methods are Support Vector Machine , Random Forest, Extra trees regression, Linear regression, Lasso regression, Gradient Boosting, Muli layer perceptron and stacking regression.

3.4.1 Support vector machine

These are one of the most widely employed classifiers. The core idea of SVM is using hyperplanes to separate different classes. Additionally, SVMs have the benefit of high precision levels when handling data that can be separated linearly, but they don't perform as well when dealing with data that can't be separated linearly. Additionally, the solution to this problem involves the use of kernel functions in order to move the data into a more expansive dimensional space. As a result, the data will be divided (in a linear manner). A key component of SVM is choosing an appropriate kernel function and modifying the kernel parameters.

A machine learning approach known as the SVM provides a sparse pattern of solutions as well as varied control over model complexity. The algorithm's primary driving force is to find the best generalization boundaries for regression, which is comparable to the hyperplane in the Prediction context (Drucker *et al.* 1996). For the implementation of SVR we used the default parametrs for SVR include the 'rbf' kernel, a degree of 3 for polynomial kernels, 'scale' for gamma, 0.0 for coef0, a tolerance of 0.001 (tol=0.001), C set to 1.0 for regularization, epsilon of 0.1 for error penalty, shrinking enabled

(shrinking=True), a cache size of 200 (cache_size=200), verbose set to False, and no maximum iteration limit (max_iter=-1).

3.4.2 Random forest

Random forest is an bagging ensemble machine learning algorithm used for regression tasks (Rodriguez-Galiano *et al.* 2015). It mixes several decision trees, each trained on a different set of data and attributes. The algorithm aggregates predictions through averaging, reducing overfitting and improving generalization. It assesses feature importance and supports hyperparameter tuning. Random Forest Regressor handles large and complex datasets, captures non-linear relationships, and provides interpretable insights. However, it may face challenges in extrapolation and computational complexity. Overall, it's a powerful and versatile algorithm for accurate and robust regression predictions. The default Random Forest parameters were n_estimators=100, criterion='mse,'min_samples_split=2,min_samples_leaf=1, min_impurity_decrease=0.0, and bootstrap=True.

3.4.3 Extra trees

Extra Trees Regression is a bagging ensemble learning algorithm and an extension of Random Forest. It is commonly used for regression tasks, predicting continuous numeric values (Mastelini *et al.* 2022).

Key characteristics and steps of Extra Trees Regression:

1. Random feature subsets: Each tree in the ensemble considers a random subset of features as potential splitting candidates at each node, promoting diversity.
2. Random splitting thresholds: Instead of optimizing thresholds based on impurity measures, Extra Trees Regression randomly selects splitting thresholds for candidate features, reducing bias.
3. Ensemble prediction: Predictions are obtained by averaging outputs from all trees in the ensemble.

Benefits of Extra Trees Regression include reduced variance, improved generalization, and resistance to overfitting. By introducing additional randomness, it explores the feature space effectively and captures complex relationships. Extra Trees Regression can handle diverse data types and is robust against outliers. However, interpreting feature importance can be challenging compared to single decision trees.

Overall, Extra Trees Regression is a powerful algorithm for accurate regression predictions, utilizing multiple randomized decision trees. For the implementation process, we utilized default parameters for Extra Trees. The parameters included max_depth=6 and random_state=0. Additionally, the parameters n_estimators=100 and min_samples_leaf=1.

3.4.4 Linear regression

Linear regression is a fundamental ML technique used for predictive modelling and determining the relationship between one or more independent variables and a dependent variable. It is based on the assumption of a linear relationship and seeks the best-fit line that minimises the discrepancies between observed and expected values (Yan 2011).

In linear regression, the algorithm learns the coefficients (intercept and slopes) that define the linear equation representing the relationship between the variables. The least squares method is used to minimise the sum of squared discrepancies between the observed and forecasted values.

Linear regressionis widely used due to its simplicity, interpretability, and efficiency. It serves as a baseline model and is particularly useful when the relationship between variables is expected to be linear. While linear regression assumes linearity, it can be extended to handle nonlinear relationships through techniques like polynomial regression. Additionally, variations such as ridge regression and lasso regression address issues like multicollinearity and overfitting.

Overall, linear regressionis a versatile and widely applicable machine learning technique that provides insights into the relationships between variables, facilitates prediction, and serves as a building block for more advanced regression methods. For the implemetation process, The default parameters for Linear regression are straightforward and do not require explicit values. Linear regression assumes the default value of 'ordinary least squares' as the optimization criterion and does not involve any regularization techniques.

3.4.5 Lasso regression

Lasso regression is a method of regularisation. It is used instead of regression techniques to create more accurate forecasts. This model employs the shrinkage technique. When data values are moved closer to a central point that acts as the mean, they are said to have been decreased.

As a logical extension to simple linear regression, the concept of "regularisation" is commonly utilised, particularly with models with many regressors. High variance in coefficient estimations is a given as the number of coefficients rises. The model would therefore detect more local curvature as a result of the high variation, increasing the prediction error (Deng *et al*. 2018).

This L1 norm penalised model employs the shrinkage and selection operator (also known as Lasso), a regularisation method, and an objective function.

When we have now added penalties to the coefficient estimations, a larger coefficient would be subject to a harsher penalty. The tuning parameter controls the size of the penalty. We used the default parameters for Lasso Regression, in addition to alpha=0.1, include fit_intercept=True, which allows the model to fit an intercept term; normalize=False, indicating that no feature scaling is performed by default; and precompute=False, meaning that no precomputation of the Gram matrix is done.

3.4.6 Gradient boosting

Gradient boosting is an iterative machine learning strategy that combines weak prediction models like decision trees to create a robust predictive model. It builds the model in stages, with each successive model attempting to correct the mistakes caused by earlier models. Gradient boosting optimises the cost function by iteratively picking weak hypotheses that align with the cost function's negative gradient. This strategy improves the model's prediction skills and can be used for regression and classification tasks. Its capacity to handle complex datasets and make accurate predictions has contributed to its broad use and appeal (Wu *et al.* 2018).

For the implementation process, default parameters were used for Gradient Boosting. These parameters include learning_rate=0.1 (controls the contribution of each tree), n_estimators=100 (number of boosting stages to perform), max_depth=3.

3.4.7 Muli layer perceptron (MLP)

MLP (Multilayer Perceptron) is a popular type of ANN used in ML. It is a feedforward neural network model that consists of multiple layers of interconnected nodes, or artificial neurons, organized in a sequential manner.

Here are the key aspects of MLP in machine learning (Olmedo *et al.* 2018)

1. Structure: The Multilayer Perceptron (MLP) is made up of an input layer, hidden layers, and an output layer. Neurons in each layer communicate with neurons in the next layer, forming a network of interconnected nodes.
2. Activation Function: MLP neurons use activation functions to introduce non-linearity. Sigmoid, hyperbolic tangent (tanh), and rectified linear unit (ReLU) are examples of common activation functions..

3. Forward Propagation: Forward propagation is the process through which input data is transferred through a neural network layer by layer. Each neuron's output is calculated using activation functions and weights.
4. Backpropagation: The backpropagation method is used by MLP to change network weights based on the difference between expected and actual outputs. It computes the gradient of the loss function in relation to the weights and then modifies the weights using an optimisation approach such as gradient descent..
5. Training: MLP is trained by exposing the network to training data, changing weights via backpropagation, and minimising a loss function. This method is repeated until the model converges or fulfils a stopping requirement.
6. Applications: MLP is adaptable and may be used for a variety of machine learning applications such as classification, regression, and pattern recognition. Because of its ability to capture complex non-linear interactions, it is well suited to solving issues with complex decision limits.

MLP is a powerful neural network model that enables the learning of complex patterns and relationships in data, making it a valuable tool in many machine learning applications. We used the default parameters for Multi-Layer Perceptron (MLP) algorithm, in addition to max_iter=500, include hidden_layer_sizes=(100,), activation='relu', solver='adam', which employs the Adam optimization algorithm for weight updates; and alpha=0.0001, representing the regularization parameter for L2 penalty.

3.4.8 Stacking regression

Stacking regression, also known as stacked generalisation, is a ML ensemble strategy for improving prediction performance by combining multiple regression models. It entails training numerous base regression models on the training data and then combining their predictions with a meta-regressor (Pavlyshenko 2018).

Here are the key aspects of stacking regression:

1. Base Regression Models (Level-0): Train multiple regression models using different algorithms or hyperparameters. Each model learns to make predictions based on input features.
2. Meta-Regressor (Level-1): Train a meta-regressor using predictions from the base regression models as input. The meta-regressor combines these predictions to generate a final prediction.
3. Training and Prediction: Train both the base regression models and the meta-regressor using the training data. Base models generate predictions for training data, which are used to train the meta-regressor. During prediction, base models generate predictions for new data, and these are fed into the meta-regressor for the final prediction.
4. Ensemble Combination: Combine predictions from base models using techniques like averaging, weighted averaging, or more advanced methods like stacking with cross-validation.
5. Stacking regression improves predictive performance by leveraging the strengths of different models and capturing complex relationships for more accurate predictions. Careful model selection, tuning, and validation are necessary to avoid overfitting and ensure robust performance.
6. Stacking can involve multiple layers of base models and meta-regressors for added complexity and flexibility, but this increases the training complexity and may require more computational resources.

So, stacking regression is a powerful ensemble technique that leverages the strengths of multiple regression models to enhance prediction accuracy and robustness.

The proposed stacking model consists of several machine learning algorithms at different levels. At level 0, we have Linear Regression, Lasso Regression, MLP, GBR, RF, and 'ExtraTreesRegressor'. These algorithms form the first layer and make individual predictions. The predictions from level 0 models are then combined, and their outputs serve as input to the second layer. At level 1, we have SVR which takes the combined predictions from the first layer and produces the final output. This stacking approach aims

to leverage the diverse predictions of different models to enhance the overall predictive performance.

3.5 Evaluation Metrics

In this section, the evaluation metrics has been investigated. For the evaluation metrics the MAE, MSE, RMSE, Actual value for one sample has been calculated.

3.5.1 Mean absolute error

MAE is a measure of the difference in error between two observations expressing the same phenomena. Comparisons of predicted against observed, subsequent time versus starting time, and one measuring technique versus another measurement technique are examples of Y versus X. MAE is derived by dividing the total number of absolute errors by the sample size (Estrada *et al.* 2020). Equation (3.2) shows the calcutation of MAE.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n} \tag{3.2}$$

It is thus an arithmetic average of the absolute errors $|e_i| = |y_i - x_i|$, where y_i is the prediction and x_i the true value. The MAE is the average absolute difference between forecasted and true values. It is measured on the same scale as the data and is a scale-dependent metric. MAE is commonly used in time series analysis as a measure of forecast error but can be mistaken for mean absolute deviation. This ambiguity exists in general.

3.5.2 Mean squared error

MSE is a well-known metric for calculating the average squared difference between predicted and actual values in regression situations. It measures how well a regression model matches the data quantitatively (Kamble and Deshmukh 2017).

The Equation (3.3) for calculating the MSE

$$\text{MSE} = (1/n) * \Sigma(y_i - \bar{y})^2 \quad (3.3)$$

Where MSE stands for Mean Squared Error, n stands for the total number of data points, y^i stands for the projected value for the i^{th} data point, and represents the actual (observed) value for the i^{th} data point.

The MSE is an important indicator for assessing the performance of regression models. A lower MSE suggests that the model's predictions are more accurate, implying a better fit to the data. However, it is important to interpret the MSE in conjunction with other performance metrics and consider the specific context and requirements of the regression problem.

3.5.3 Root mean squared error

RMSE is another widely used metric in regression analysis. It is the square root of the average of the squared differences between predicted and actual values. Equation (3.4) for calculating RMSE (Kamble and Deshmukh 2017).

$$\text{RMSE} = \sqrt{(\text{MSE})} \quad (3.4)$$

RMSE, short for Root Mean Squared Error, is a useful metric for assessing regression model accuracy. It measures the average magnitude of prediction errors in the same units as the target variable. A lower RMSE indicates better performance, as the model's predictions are closer to the actual values. However, RMSE is sensitive to outliers and can be affected by the scale of the target variable. It should be used in conjunction with other evaluation metrics and considered within the specific regression problem's context.

4. RESULTS AND DISCUSSION

In this thesis, eight different models are selected and tested to evaluate the proposed models. These models are linear regression model, Lasso regression, SVM, MLP, GBR, RF, Extra trees and stacking regressing model. In this thesis, the Python programming languages with hp Laptop based on the 2.4GHz CPU, RAM 16GHz, Core i5 is used.

4.1 Feature Selection Results

This section presents the names for the selected features for every type of feature selection for energy power. Table 4.1 represents the names for the selected features from the Relief f and ANOVA feature selection technique.

Table 4.1 The selected features from using feature selection technique

TECHNIQUE	#OF SELECTED FEATURES	SELECTED FEATURES
Relief f	40	'mainht', 'htvcav', 'htvvav', 'elcool', 'ngcool', 'coolp', 'maincl','clvcav', 'clvvav', 'emcs', 'rdhtnf', 'rdclnf', 'ecn', 'maint', 'enrgypln', 'pctermn', 'laptpn', 'prntrn', 'copiern', 'ltohrp', 'ltnhrp', 'fluor', 'cflr', 'bulb', 'halo', 'hid', 'led', 'sched', 'ocsn', 'wintyp', 'tint', 'refl', 'awn', 'skylt', 'dayltp', 'hdd65','cdd65', 'hvaceui', 'ltneui', 'pleui'
ANOVA		'enrgypln', 'hvaceui', 'rdclnf', 'rdhtnf', 'ownocc', 'pleui', 'coolp', 'ltneui', 'facil', 'open24', 'ownopr', 'nfloor', 'htvvav', 'ocsn', 'renrff', 'fkused', 'owntype', 'renins', 'clvvav', 'prntrn', 'rftilt', 'wlcns', 'flceilht', 'elht1', 'pctermn', 'wkhrs', 'ltohrp', 'ltnhrp', 'ngcool', 'fluor', 'nwker', 'cflr', 'htvcav', 'halo', 'sqft', 'nocc', 'renwll', 'refl', 'wintyp', 'prused'

As observed in Table 4.1, we choose the number of selected features as 40 after experimental results on variant numbers and the best results obtained using this number.

4.2 Results of Machine Learning Models

This section presents the outcomes of utilizing machine learning models for energy power estimation with and without employing feature selection techniques. Specifically, it

explores the utilization of Relief F and ANOVA as feature selection methods within these models.

4.2.1 Result of models without feature selection

This section represents the results of the proposed ML models without using the feature selection techniques and discussion about these results. Table 4.2 represents the results for the regression of energy power consumption using various algorithms without feature selection. The evaluation metrics include MAE, MSE and RMSE.

Table 4.2 Results of models without using feature selection

ALGORITHM NAME	MAE	MSE	RMSE
Linear regression	0.01369	0.00055	0.02358
Lasso	0.13326	0.02800	0.16734
SVM	0.01418	0.00056	0.02367
MLP	0.08357	0.00116	0.10811
GBR	0.01805	0.00078	0.02803
RF	0.02242	0.00126	0.03549
Extra trees Regressor	0.02224	0.00114	0.03386
Staking model	0.01286	0.00054	0.02328

As shown in Table 4.2, the Stacking model achieved the lowest MAE value of 0.01286, indicating the highest accuracy in predicting energy power consumption without feature selection. This highlights the effectiveness of the Stacking model in leveraging all available features and achieving accurate predictions.

Linear regression also performed well with a low MAE of 0.01369, indicating its ability to make accurate predictions without feature selection. Similarly, SVM achieved a relatively low MAE of 0.01418, showcasing its strong predictive capability in utilizing the full set of features.

For MSE and RMSE, the Stacking model again showed competitive performance with the lowest values of 0.00054 and 0.02328, respectively. This suggests that the Stacking

model, without feature selection, minimized the squared differences and provided accurate predictions for energy power consumption using all available features.

So, the Stacking model, linear regression, and SVM demonstrated the best performance in terms of MAE, MSE, and RMSE when no feature selection was applied. These algorithms effectively utilized the full set of features to achieve accurate predictions for energy power consumption. These results highlight the importance of leveraging all available information in the absence of feature selection techniques.

We found that the Stacking model without feature selection is considered the best algorithm for energy consumption regression due to its strong performance across various metrics, including low MSE and RMSE values, indicating accurate predictions with minimal deviation and this fulfil the answer for the RQ4.

4.2.2 Results of models using Relief f feature selection

This section represents the results of ML models using the Relief f feature selection technique according to the evaluation metrics. Table 4.3 introduces the results of the eight ML models using the Relief f feature selection technique for regression of energy power consumption using different algorithms. The evaluation metrics include MAE, MSE, and RMSE.

Table 4.3 Result of models using Relief f

ALGORITHM NAME	MAE	MSE	ROOT MSE
Linear regression	0.01754	0.00063	0.02521
Lasso	0.13257	0.02798	0.16729
SVM	0.01937	0.00069	0.02642
MLP	0.07008	0.00845	0.09197
GBR	0.01748	0.00063	0.02510
RF	0.02148	0.00093	0.03059
Extra trees Regressor	0.02155	0.00095	0.03083
Staking model	0.01438	0.00049	0.02232

As seen in Table 4.3, In terms of MAE, the Stacking model achieved the lowest value of 0.01438, indicating the highest accuracy in predicting energy power consumption. It was closely followed by linear regression and GBR, which had MAE values of 0.01754 and 0.01748, respectively.

For MSE and RMSE, the Stacking model again outperformed the other algorithms with the lowest values of 0.00049 and 0.02232, respectively. Linear regression and GBR also had relatively low MSE and RMSE values, indicating their effectiveness in minimizing the squared differences and providing accurate predictions.

On the other hand, Lasso regression and MLP had higher MAE, MSE, and RMSE values compared to other algorithms. This suggests that they had slightly lower predictive accuracy for energy power consumption.

Overall, the Stacking model outperformed all three evaluation parameters, suggesting its competence in accurately estimating energy power usage.

4.2.3 Results of models using ANOVA feature selection

This section represents the results of ML models using the ANOVA feature selection technique according to the evaluation metrics. Table 4.4 presents the results for the regression of energy power consumption using various algorithms with ANOVA feature selection. The evaluation metrics include MAE, MSE and RMSE.

Table 4.4 Results of models using ANOVA feature selection

ALGORITHM NAME	MAE	MSE	ROOT MSE
Linear regression	0.01326	0.00055	0.02353
Lasso	0.13326	0.02800	0.16734
SVM	0.01336	0.00055	0.02348
MLP	0.03974	0.00287	0.05359
GBR	0.01826	0.00078	0.02809
RF	0.02186	0.00120	0.03475

Table 4.4 Results of models using ANOVA feature selection (Continued)

ALGORITHM NAME	MAE	MSE	ROOT MSE
Extra trees Regressor	0.02306	0.00116	0.03411
Staking model	0.01272	0.00056	0.02384

As seen in Table 4.3, among the algorithms, the Stacking model achieved the lowest MAE value of 0.01272, indicating the highest accuracy in predicting energy power consumption when combined with ANOVA feature selection. This demonstrates the effectiveness of the Stacking model in selecting relevant features that contribute to accurate predictions.

Linear regression also performed well with a low MAE of 0.01326, showcasing its ability to make accurate predictions when combined with ANOVA feature selection. Similarly, SVM achieved a low MAE of 0.01336, indicating strong predictive capability when using the selected features.

For MSE and RMSE, the Stacking model again showed the lowest values of 0.00056 and 0.02384, respectively. This suggests that the Stacking model, in conjunction with ANOVA feature selection, minimized the squared differences and provided accurate predictions for energy power consumption.

Overall, the Stacking model, linear regression, and SVM demonstrated the best performance in terms of MAE, MSE, and RMSE when utilizing ANOVA feature selection. These algorithms effectively leveraged the selected features to achieve accurate predictions, indicating the importance of feature selection in improving regression performance for energy power consumption. We found that the feature selection is valuable only for stacking model with ANOVA for the MAE metric. Although the rest of results are better without using feature selection technique with the stacking model. And this fulfils the answer for the RQ1

After examining the results in Table 4.2, Table 4.3 and Table 4.4, the single based algorithm overcome the ensemble ML methods for the regression of the energy

consumption and this fulfil the answer for the RQ2. Also, the bagging ensemble algorithms like RF outperform the GBR algorithm which it's one of the bagging algorithms, and this fulfil the answer for RQ3.

4.3 Comparison between the proposed stacking models

Table 4.5 compares the results of regression models for energy power consumption without utilizing stacking models, with and without feature selection. MAE, MSE, and RMSE are the metrics used to evaluate the models.

Table 4.5 Comparison between the proposed stacking models

STACKING MODEL	MAE	MSE	RMSE
Without feature selection	0.01286	0.00054	0.02328
With Relief f feature selection	0.01438	0.00049	0.02232
With ANOVA feature selection	0.01272	0.00056	0.02384

As seen in Table 4.4, in the scenario without feature selection, the stacking model achieves an impressive performance with an MAE of 0.01286, MSE of 0.00054, and RMSE of 0.02328. This suggests that the stacking model, when used without feature selection, accurately predicts energy power consumption.

In contrast, when feature selection techniques are applied, the results vary slightly. The stacking model with Relief f feature selection achieves an MAE of 0.01438, MSE of 0.00049, and RMSE of 0.02232. On the other hand, the stacking model with ANOVA feature selection achieves an MAE of 0.01272 which is the best value for this metric overcoming the MAE for the stack model with Relief f and without feature selection, MSE of 0.00056, and RMSE of 0.02384. These results indicate that both feature selection techniques have a minor impact on the performance of the stacking model. We excluded the stacking model with relief f because its lesrning curve is not better

Among the scenarios of without fetures selection and with ANOVA, the stacking model without feature selection appears to be the best option. Its MSE and RMSE values are comparable or slightly better than the models with ANOVA feature selection. Therefore,

the stacking model without feature selection demonstrates superior predictive performance for energy power consumption regression in this analysis.

4.4 Learning Curves

Learning curves visualize the performance of machine learning models using training and testing datasets. By plotting a performance metric against the number of training examples, these curves provide insights into model behavior and performance. Learning curves consist of two lines representing the model's performance on the training and testing sets (Hutter 2021).

Key insights from learning curves include:

1. Model Performance: Learning curves show how a model's performance changes as more training examples are used. It helps identify overfitting and effective learning.
2. Bias and Variance: Learning curves provide insights into a model's bias and variance. High bias and variance can be detected by poor performance on the training and testing sets, respectively.
3. Dataset Size and Model Complexity: Learning curves help understand the impact of dataset size and model complexity on performance. More data may improve performance, while increased complexity may benefit the model if testing set performance improves.

Learning curves provide valuable insights into model performance, bias-variance trade-off, and generalization capabilities. This information assists in making informed decisions regarding model selection, dataset size, and model complexity.

4.4.1 Learning curves graphs without feature selection

This specific section presents the graphical representations of learning curves. These learning curves are constructed without employing any feature selection techniques. The

learning curve for Linear regressionis shown in Figure 4.1. The learning curve of Linear regression reveals valuable insights about the model's performance.

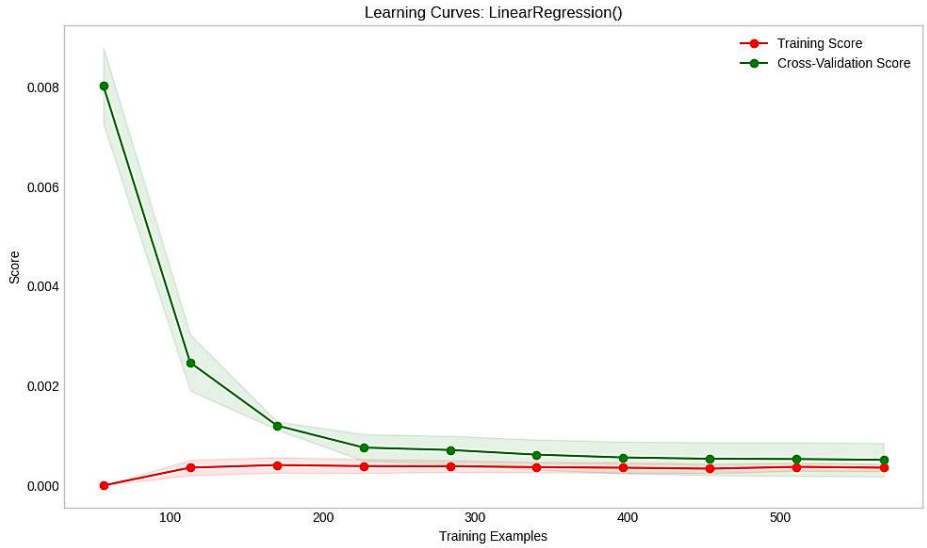

Figure 4.1 The learning curve for linear regression

As seen in Figure 4.1, the green-colored curve represents the cross-validation score, which initially starts at 0.008 and gradually converges to 0.001. This indicates that as more data points are included in the validation process, the model's predictive accuracy significantly improves. The decreasing trend of the cross-validation score demonstrates the model's ability to generalize well and make reliable predictions on unseen data.

On the other hand, the red-colored curve represents the training score, reflecting the model's performance on the training data. It is worth noting that the training score also converges to 0.001, which suggests that the model is effectively capturing the underlying patterns and relationships present in the training dataset. The convergence of the training score to a low value indicates that the model has learned the training data well and achieves a high degree of accuracy in reproducing the target values.

The learning curve provides evidence that the Linear regression model is performing effectively and exhibits both good generalization capabilities (as indicated by the cross-validation score) and strong training performance (as shown by the training score). The

convergence of both scores to 0.001 suggests that the model has achieved a desirable level of accuracy, indicating its robustness and reliability.

The learning curve for Lasso is shown in Figure 4.2.

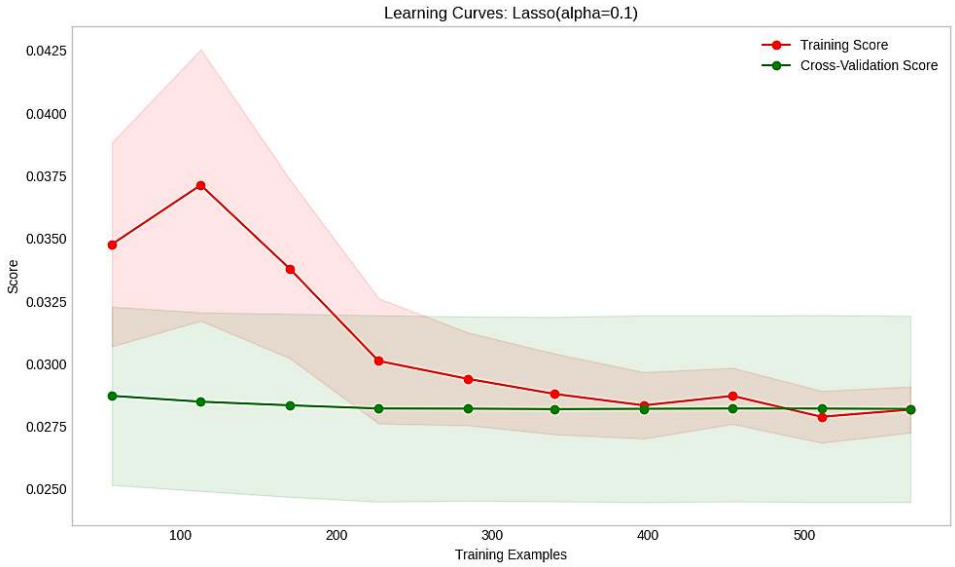

Figure 4.2 The learning curve for Lasso

The learning curve demonstrates that the lasso regression model performs well, as both the cross-validation and training scores converge to 0.0280 as seen in Figure 4.2. This indicates that the model achieves a desirable level of accuracy and exhibits both good generalization capabilities and strong training performance.

The learning curve for linear SVR is shown in Figure 4.3.

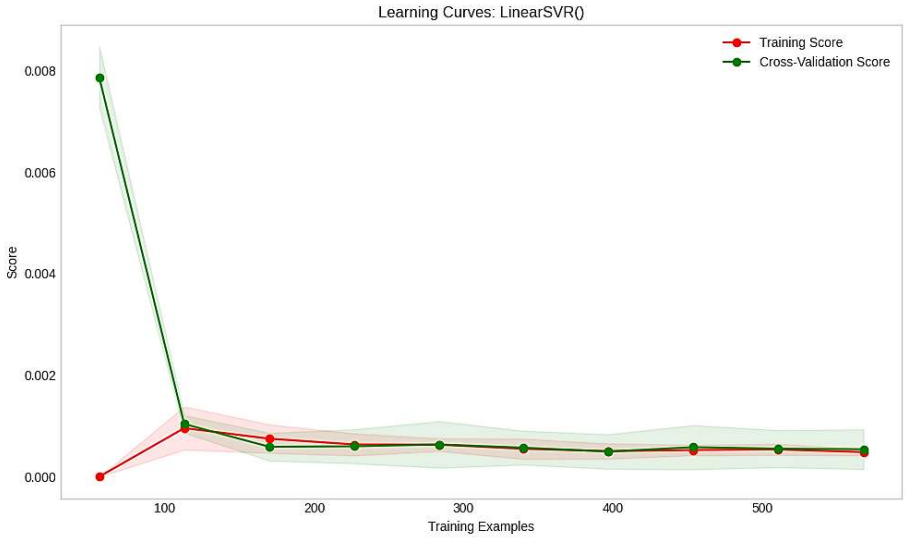

Figure 4.3 The learning curve for linear SVR

The green curve in Figure 4.3 represents the cross-validation score, which begins at 0.008 and gradually converges to 0.0013. Similarly, the red curve shows the training score, indicating that the method is well-trained and converges to the same value of 0.0013. The learning curve demonstrates that the SVR regression model performs exceptionally well. Both the cross-validation and training scores converge to 0.0008, indicating a high level of accuracy and strong performance. This suggests that the SVR model exhibits excellent generalization capabilities and reliably predicts the target variable. The convergence of both scores indicates the robustness and reliability of the SVR model in capturing complex relationships and making accurate predictions.

The learning curve for MLP regression is shown in Figure 4.4.

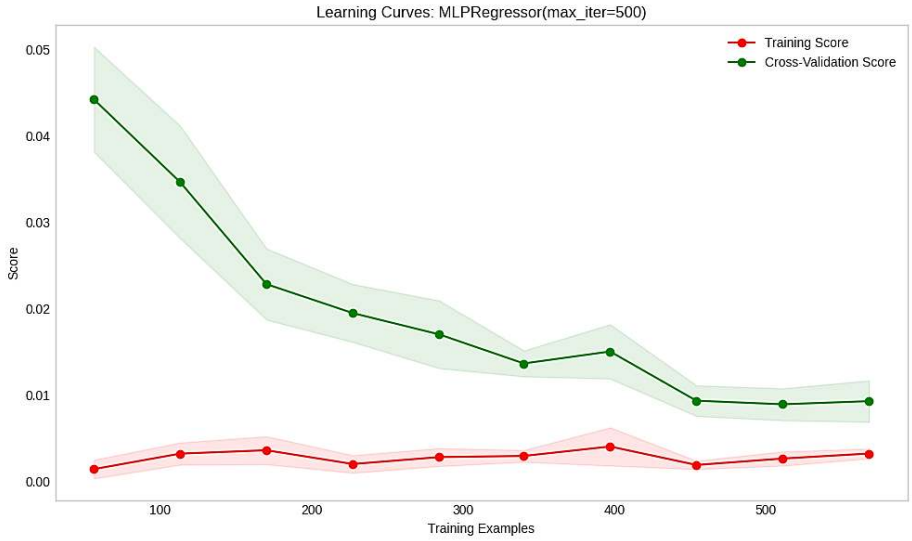

Figure 4.4 The learning curve for MLP regression

As observed in Figure 4.4, the green curve, representing the cross-validation score, shows a favorable trend, beginning at 0.005 and gradually converging to 0.0280. Likewise, the red curve depicting the training score demonstrates successful training and convergence at 0.0280. These outcomes indicate that the mLP regression method performs admirably, displaying commendable training and validation scores.

The learning curve for gradient boosting regression is shown in Figure 4.5.

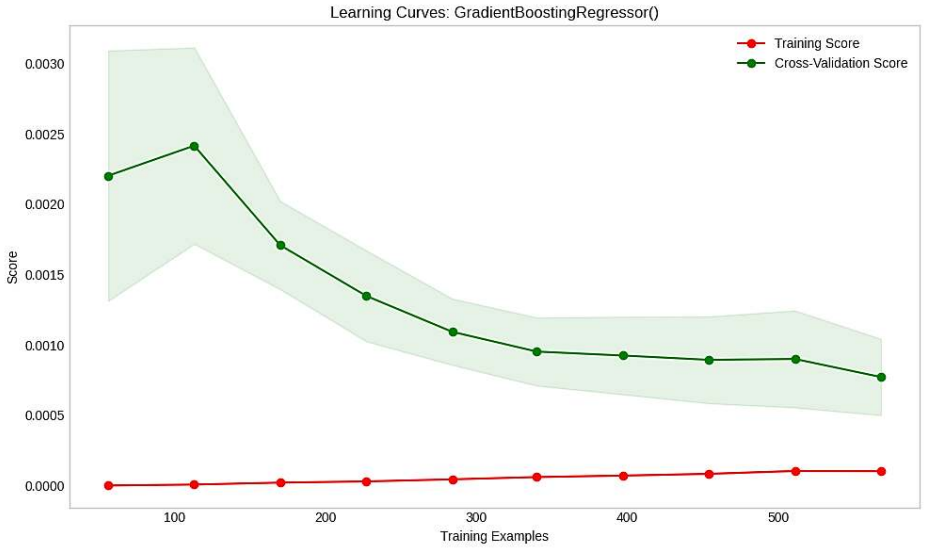

Figure 4.5 The learning curve for gradient boosting

The cross-validation score, represented by the ascending green curve, displays a favorable pattern, starting at 0.0022 and gradually converging to 0.0007. Similarly, the training score, indicated by the red curve, showcases successful training and convergence at 0.0001 as seen in Figure 4.5.

The learning curve for random forest regression is shown in Figure 4.6.

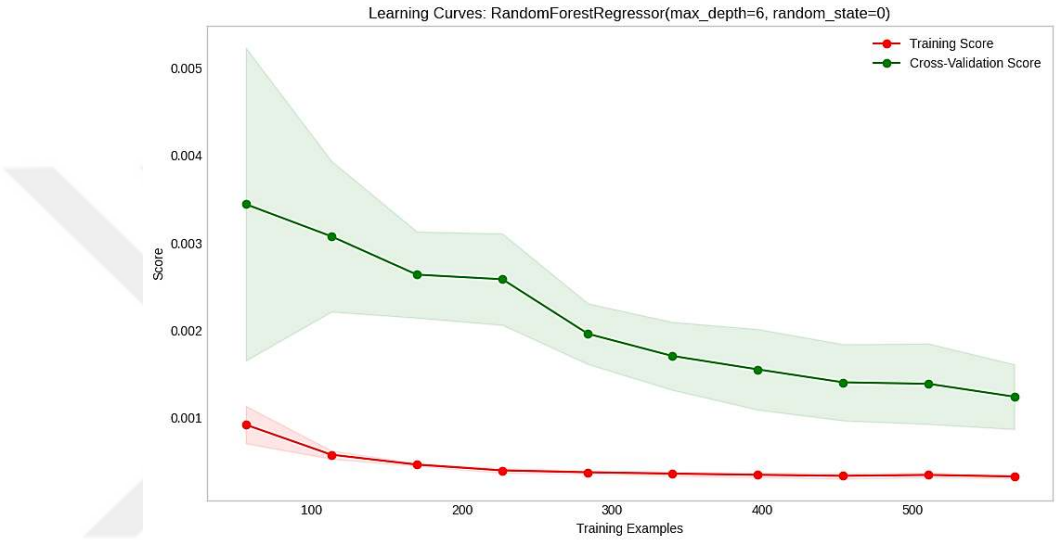

Figure 4.6 The learning curve for random forest regression

The green curve, representing the cross-validation score, exhibits a noticeable improvement, starting at 0.0035 and converging to 0.0012 as in Figure 4.6. Similarly, the red curve portraying the training score demonstrates a well-trained model that converges to an impressive value of 0.0001.

Figure 4.7 represents the learning curve for using Extra trees with max depth equal 6.

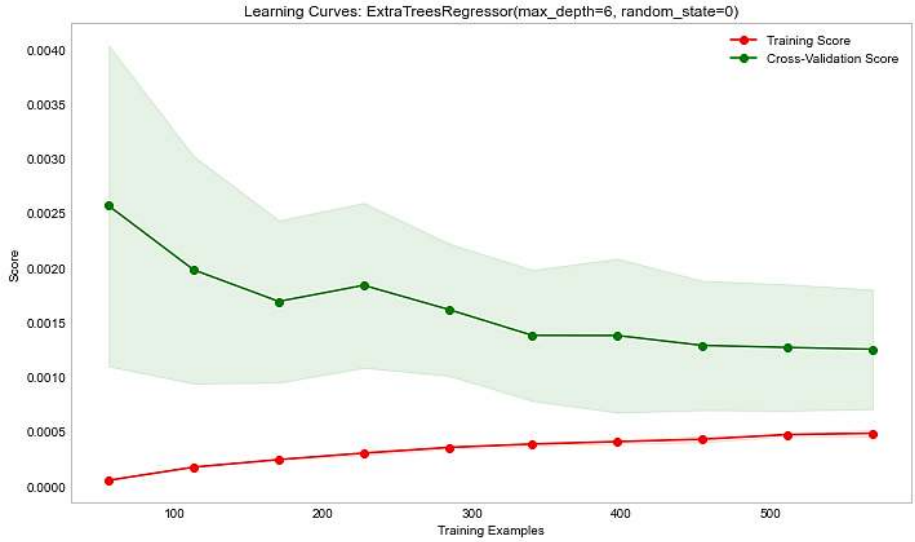

Figure 4.7 The learning curve for extra trees regression

The cross-validation score's green curve, which starts at 0.0026 and converges to 0.0013, shows a substantial improvement. The red training score curve shows a similarly well-trained model that converges to an amazing value of 0.0005 as in Figure 4.7.

Figure 4.8 represents the learning curve for using stacking regression model.

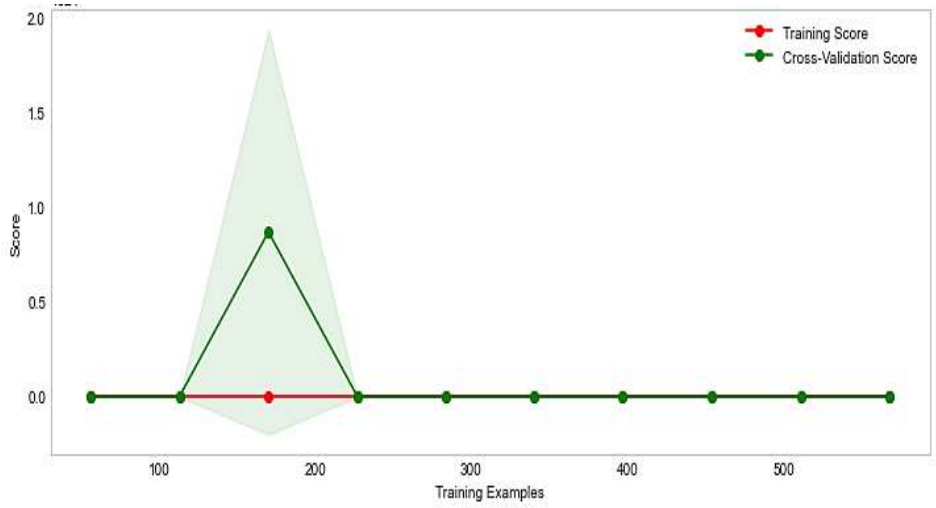

Figure 4.8 The learning curve for stacking regression

As depicted in figure 4.8, the two curves converges into 0.0, it typically indicates that the stacking model has achieved optimal performance.

4.4.2 Learning curves with Relief f feature selection

Figure 4.9 depicts the Linear regression model's learning curve. The Linear regression learning curve shows considerable gains in both the cross-validation and training scores.

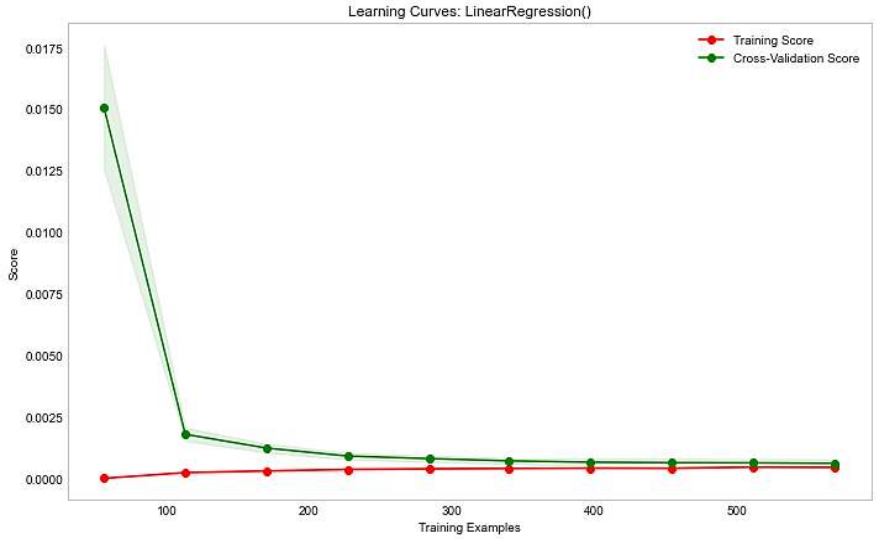

Figure 4.9 Learning curve for linear regression with Relief

The cross-validation score starts at 0.0150 and decreases, converging to an impressive value of 0.00002 as seen in Figure 4.9. The training score starts at 0.0150 and decreases to a remarkably low value of 0.00001, indicating effective learning and accurate predictions.

Figure 4.10 represents the learning curve for the lasso regression with Relief f feature slection. The learning curve for lasso regression demonstrates the progression of the cross-validation score and training score.

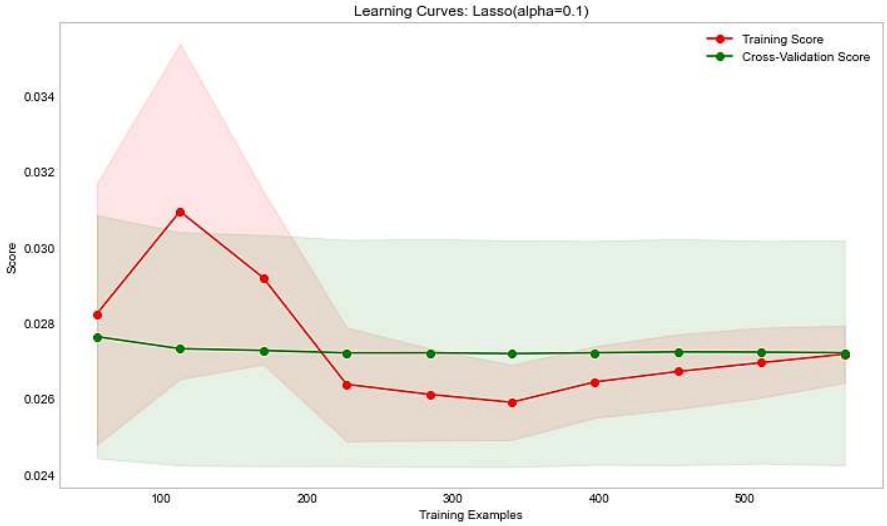

Figure 4.10 Learning curve for Lasso with Relief

As shown in Figure 4.10, The cross-validation score starts off at 0.0281 and rises steadily towards the outstanding result of 0.027. Similar to this, the training score starts out at 0.0265 and steadily drops until it reaches the incredibly low result of 0.0264.

Figure 4.11 represents the learning curve for SVR regression reveals the progression of the cross-validation score and training score.

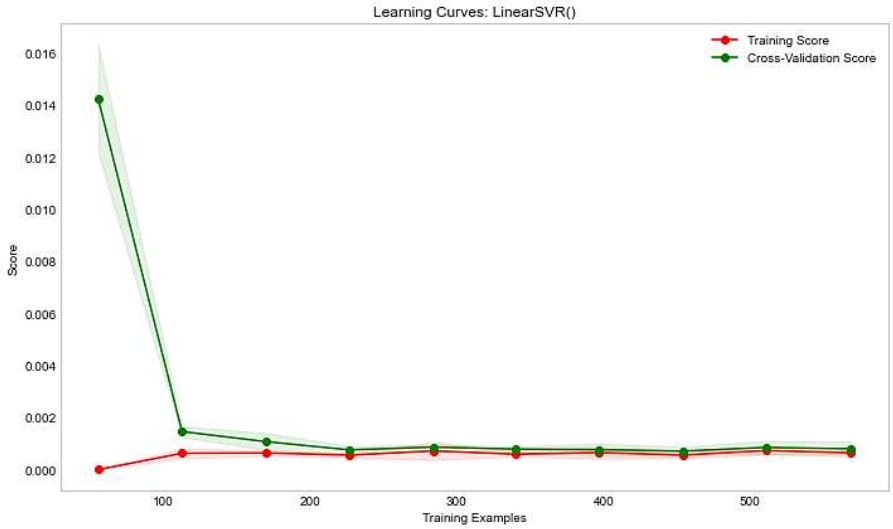

Figure 4.11 Learning curve for SVR with Relief

The cross-validation score in Figure 4.11 starts at 0.0145 and steadily converges to an impressive value of 0.001. On the other hand, the training score starts at 0.000 and gradually increases to a remarkably value of 0.001.

Figure 4.12 represenst the learning curve for the MLP regression with using Relief f feature slection technique.

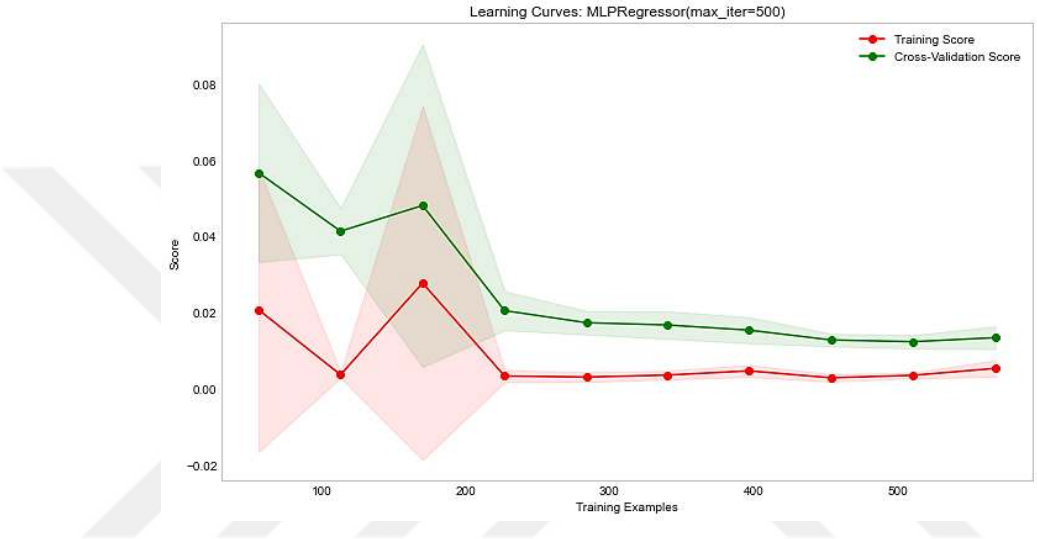

Figure 4.12 Learning curve for MLP with Relief

As observed in Figure 4.12, the cross-validation score begins at 0.055 and shows a remarkable convergence to an impressive value of 0.018. This indicates that the model's performance on unseen data improves significantly as more training examples are utilized, which is quite impressive. On the other hand, the training score initially starts at 0.022 and exhibits an increase, eventually reaching a remarkably low value of 0.01.

Figure 4.13 represents the learning curve for the GBR model using Relief f feature selection algorithm.

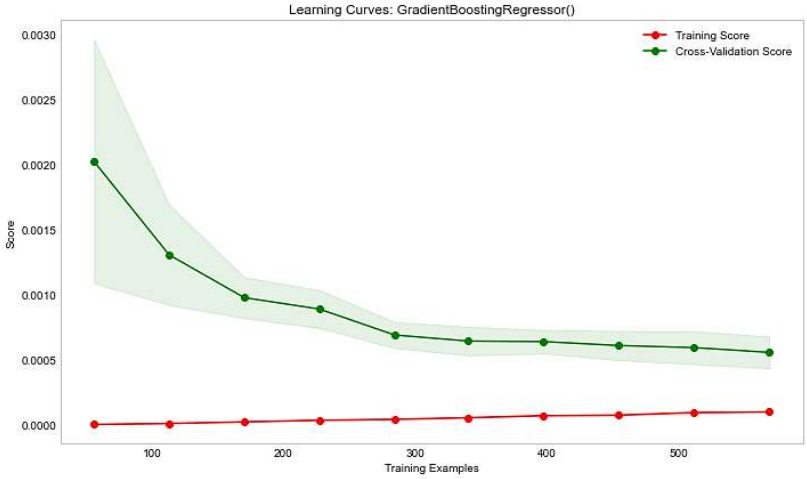

Figure 4.13 Learning curve for GBR with Relief

In the learning curve for GBR regression as in Figure 4.13, it is evident that the cross-validation score initiates at 0.0020 and exhibits a remarkable convergence to an impressive value of 0.00051. This indicates a substantial improvement in the model's performance on unseen data as more training examples are incorporated, which is truly impressive and indicative of effective learning and generalization. Additionally, the training score commences at 0.000 and undergoes an increase, eventually reaching a remarkably low value of 0.0002.

Figure 4.14 represnts the learning curve for RF regression with Relief f feature selection technique.

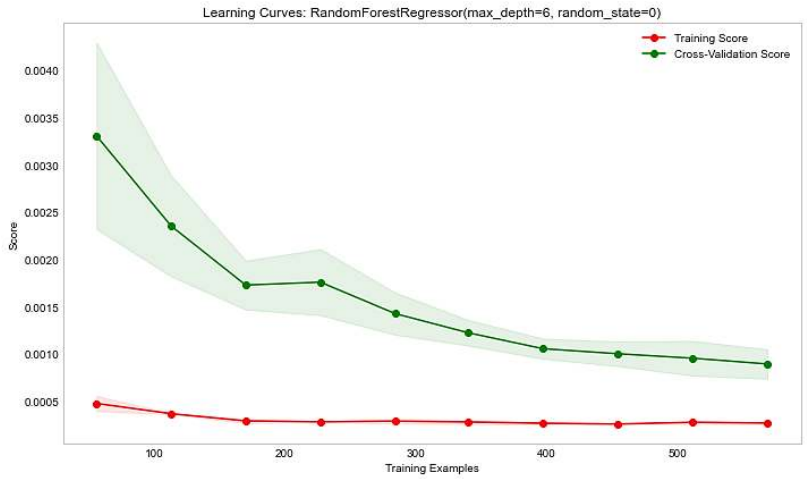

Figure 4.14 Learning curve for RF with Relief

The cross-validation score starts at 0.0033 and converges to 0.001 in Figure 4.14, indicating effective learning and generalization. Similarly, the training score starts at 0.0005 and gradually converges to 0.0003, reflecting the model's accurate predictions and high accuracy. This convergence to low values enhances the overall performance of the RF regression model, confirming its reliability and effectiveness.

Figure 4.15 represnt the learning curve for Extra trees using Relief f feature selection technique.

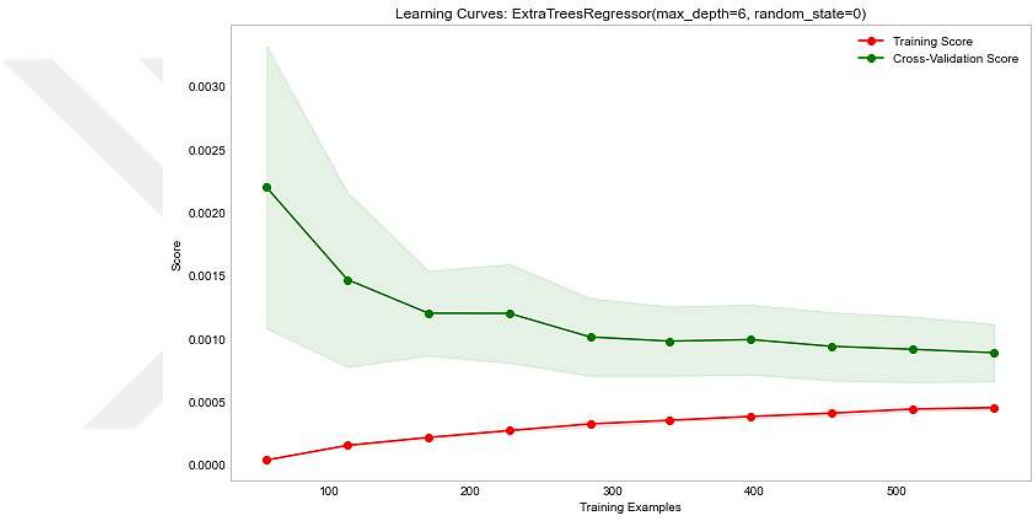

Figure 4.15 Learning curve for extra trees with Relief

The cross-validation score, which starts at 0.0022 and converges to 0.001, shows successful generalisation and learning. The model's high accuracy and precise predictions are also reflected in the training score, which starts at 0.00 and gradually converges to 0.0005 as in Figure 4.15.

Figure 4.16 represents the learning curve for stacking regression with Relief f feature selection technique.

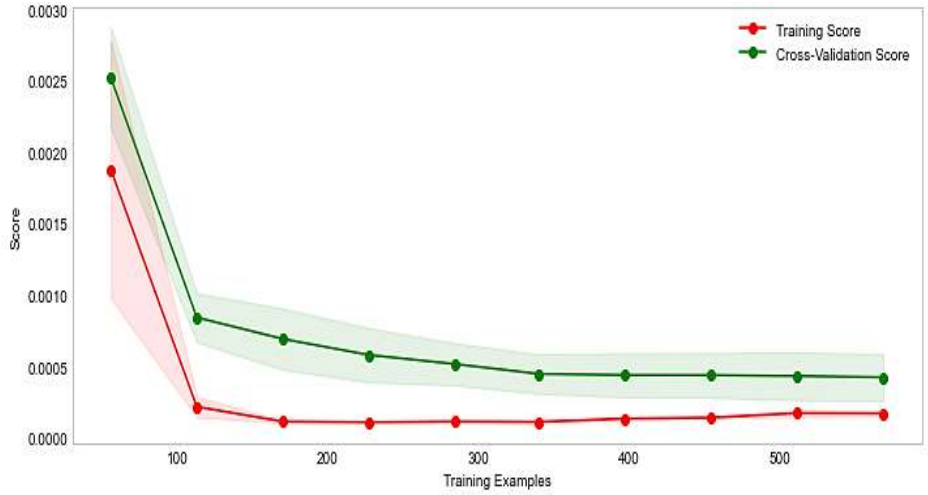

Figure 4.16 Learning curve for stacking model with Relief

As observed in Figure 4.16, The cross-validation score starts at 0.0025 and converges impressively to 0.0005, indicating effective learning and generalization. Also, the training score starts at 0.00157 and gradually converges to 0.0002. This indicates that the model effectively learns from the training data, achieving accurate predictions and a high level of accuracy.

4.4.3 Learning curves with ANOVA feature selection

Figure 4.17 represents the learning curve for Linear regression with ANOVA feature selection.

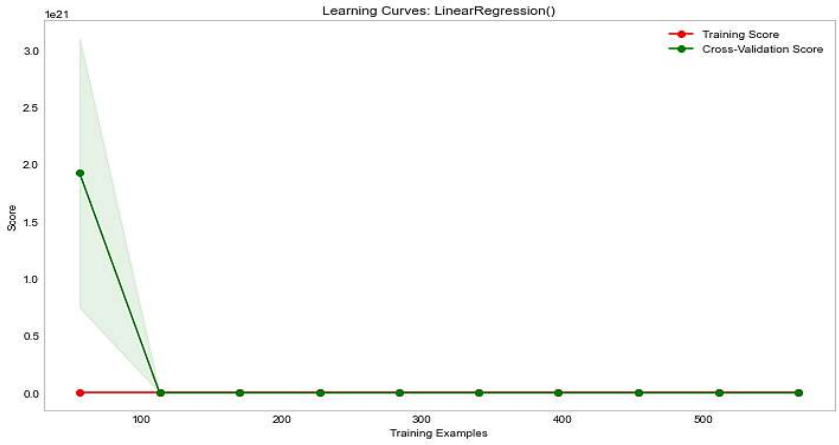

Figure 4.17 Learning curve for linear regression with ANOVA

As shown in Figure 4.17, the cross-validation score starts at 1.9 and converges to an impressive value of 0.0, indicating highly effective learning and generalization on unseen data. Furthermore, the training score initiates at 0.001 and gradually converges to 0.0. This indicates that the model effectively learns from the training data, achieving accurate predictions and a high level of accuracy.

Figure 4.18 represents the learning curve for lasso regression with ANOVA feature selection.

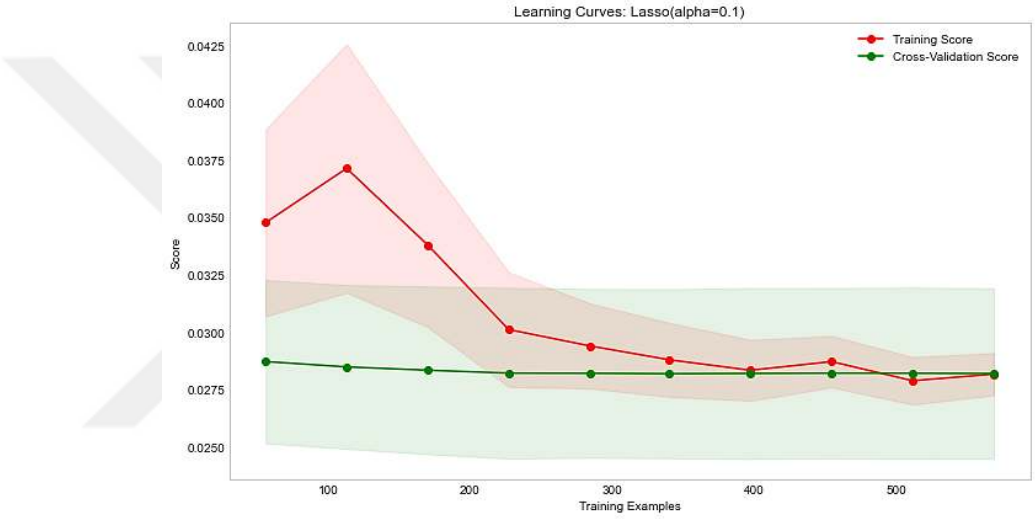

Figure 4.18 Learning curve for lasso regression with ANOVA

As in Figure 4.18, the cross-validation score begins at 0.034 and impressively converges to a value of 0.0285, indicating highly effective learning and generalization on unseen data. Moreover, the training score commences at 0.0285 and gradually converges to a value of 0.0283.

Figure 4.19 introduces the learning curve for SVR with ANOVA feature selection.

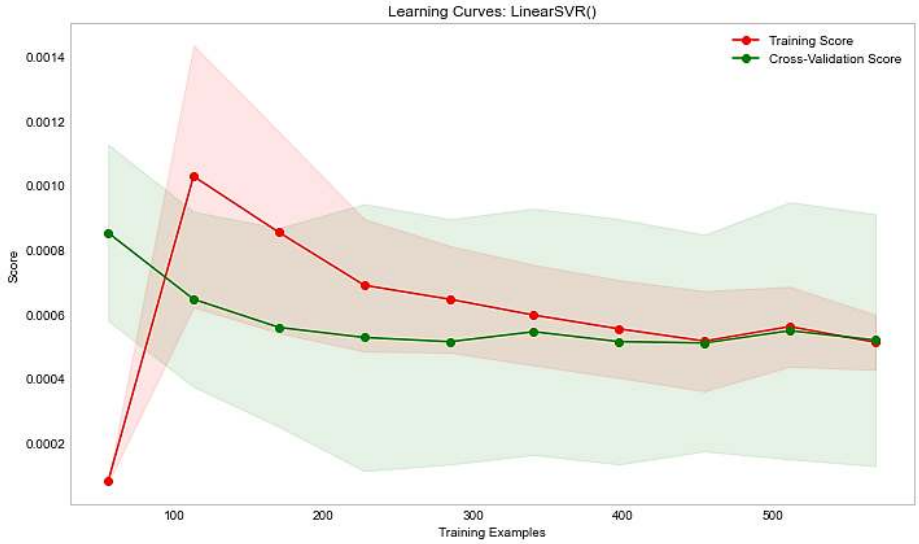

Figure 4.19 Learning curve for SVR with ANOVA

According to Figure 4.19, The cross-validation score, which starts out at 0.0088 and remarkably converges to a value of 0.0005, shows that generalisation and learning from unknowable data can be done quite successfully. The training score also begins at 0.00005 and gradually converges to 0.00005.

Figure 4.20 represnts the learning curve for MLP with ANOVA feature selection.

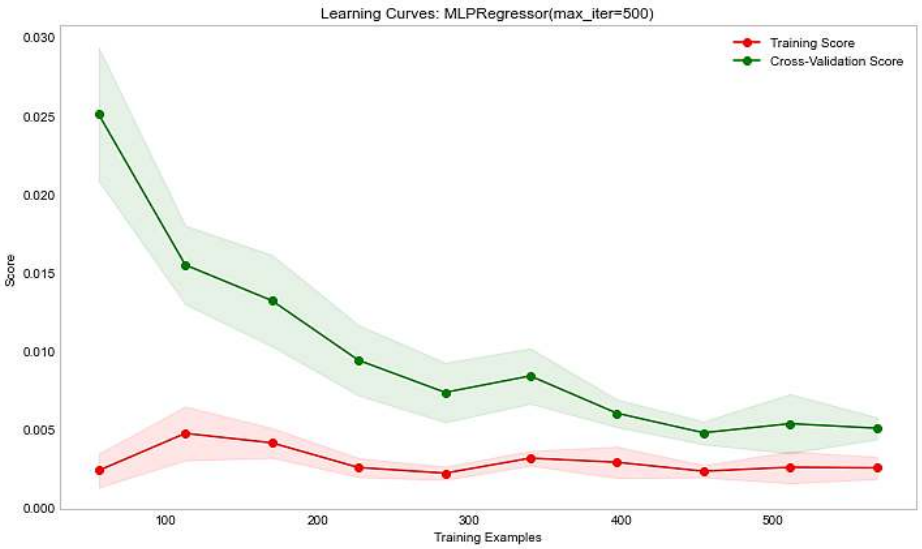

Figure 4.20 Learning curve for MLP with ANOVA

The cross-validation score starts at 0.025 and remarkably converges to a value of 0.005, as seen in Figure 4.20, suggesting extremely good learning and generalisation on unobserved data. In addition, the training score begins at 0.0025 and rapidly increases to 0.0029.

Figure 4.21 represents the learning curve for GBR with ANOVA feature selection

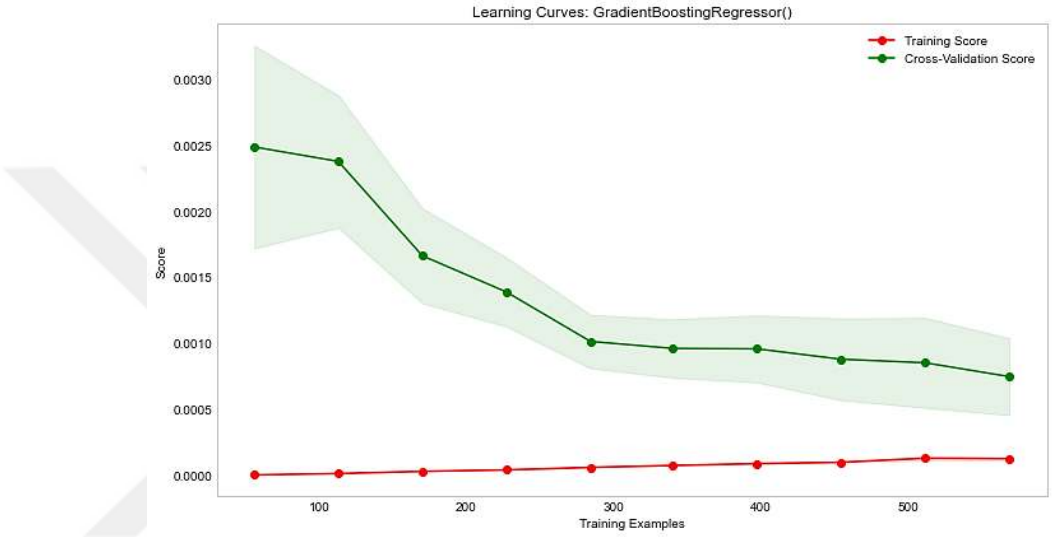

Figure 4.21 Learning curve for GBR with ANOVA

In Figure 4.21, it can be seen that the cross-validation score starts at 0.0025 and then displays exceptional convergence to arrive at a value of 0.0055. On the other hand, The training score also starts out at 0.0 and progressively rises to the remarkable number of 0.00004.

Figure 4.22 represents the learning curve for RF with ANOVA feature selection.

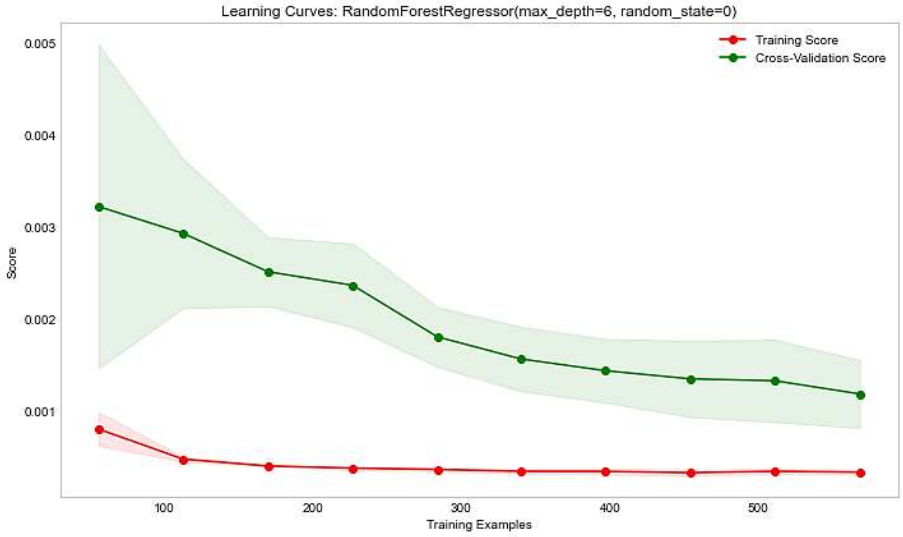

Figure 4.22 Learning curve for RF with ANOVA

The initial cross-validation score stands at 0.0033 and exhibits impressive convergence, reaching a value of 0.0012. Additionally, the training score commences at 0.0008 and gradually converges to a value of 0.0005 as in Figure 4.22.

Figure 4.23 represnts the learning curve for Extra trees with ANOVA feature selection.

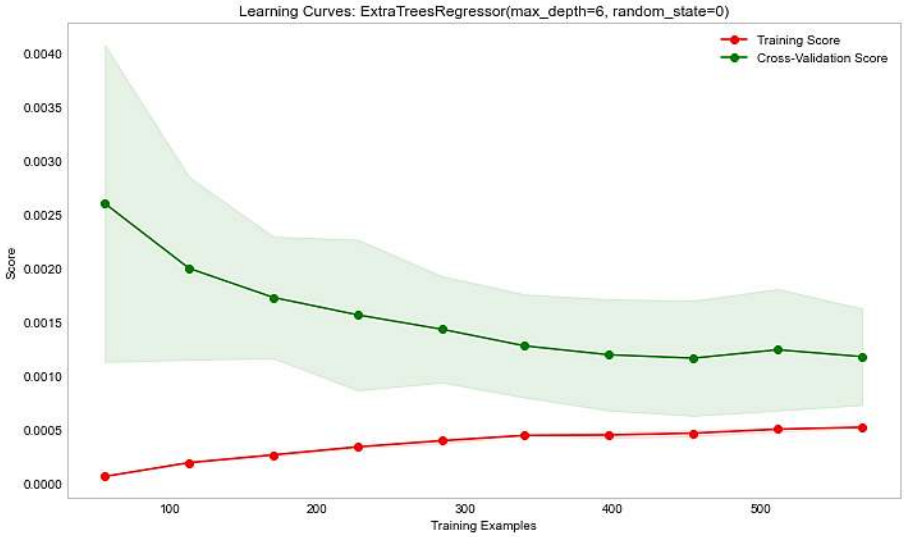

Figure 4.23 Learning curve for extra trees with ANOVA

The cross-validation score, which starts at 0.0027 and exhibits a convergence to a value of 0.0012 as shown in Figure 4.23, shows successful learning and generalisation on

unobserved data. Additionally, the training score starts at 0.0001 and converges gradually to a value of 0.00053.

Figure 4.24 represents the learning curve for Stacking regression with ANOVA feature selection

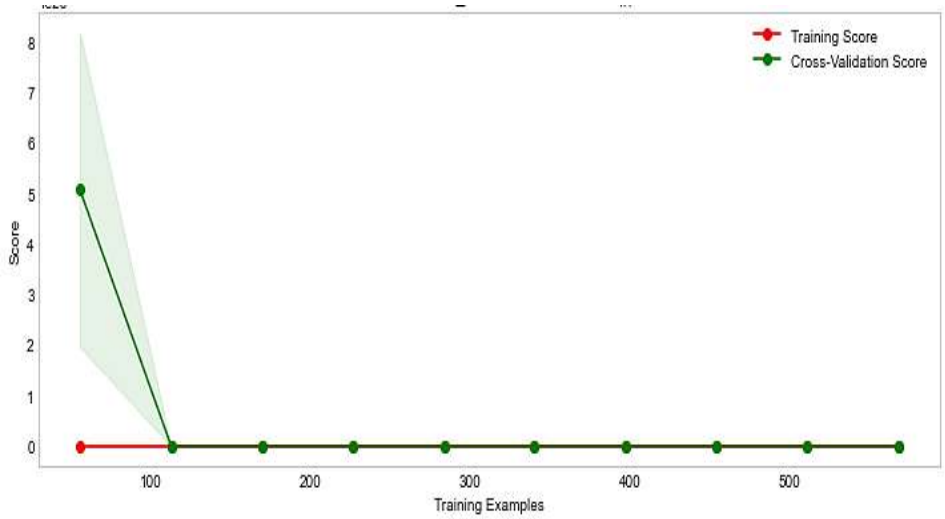

Figure 4.24 Learning curve for stacking regression with ANOVA

According to Figure 4.24, the cross-validation score initiates at 5 and impressively converges to a value of 0, indicating effective learning and generalization on unseen data. Furthermore, the training score begins at 0 and gradually converges to a value of 0. This signifies that the model effectively learns from the training data, resulting in accurate predictions and a high level of precision. The convergence of the training score to a value of 0 further enhances the overall performance of the Stacking regression model with ANOVA feature selection, highlighting its reliability and effectiveness in making precise predictions.

4.5 Comparison with the Related Works

Table 4.6 represents the results of different regression models applied to predict energy power consumption, with metrics such as MAE, MSE, and RMAE. The models used in the studies include SVM, Random Forest, Extra Trees Regression, Linear Regression,

Lasso Regression, Gradient Boosting, Stacking Regression, Bagging, MLP Regressor, XGBOOST, SVR, KNN Regressor AdaBoost, Adaboost, Gradient boosted trees, SVM-RBF, Decision Tree, SVM-Poly, ANN, Generalized linear model, Multivariate Adaptive regression splines, and others.

Table 4.6 Comparison of proposed models with others studies

REFERENCES	METHODS	MAE	MSE	RMAE
Proposed models without feature selection	Support Vector Machine	0.01369	0.00055	0.02358
	Random Forest	0.13326	0.02800	0.16734
	Extra Trees Regression	0.01418	0.00056	0.02367
	Linear Regression	0.08357	0.00116	0.10811
	Lasso Regression	0.01805	0.00078	0.02803
	Gradient Boosting	0.02242	0.00126	0.03549
	Multi-layer Perceptron	0.02224	0.00114	0.03386
	Stacking Regression	0.01286	0.00054	0.02328
Robinson *et al.* (2017)	Bagging	0.30 ± 0.01	-	-
	MLP Regressor	0.33 ± 0.01	-	-
	XGBOOST	0.30 ± 0.01	-	-
	Random Forest Regressor	0.33 ± 0.02	-	-
	Extra Trees Regressor	0.34 ± 0.02	-	-
	SVR	0.39 ± 0.01	-	-
	KNN Regressor AdaBoost	0.43 ± 0.01	-	-
	Adaboost	0.43 ± 0.03	-	-
Arjunan *et al.* (2020)	Gradient boosted trees	-	0.202	0.45
	Random Forest	-	0.6084	0.780
	SVM-RBF	-	0.0171	0.131
	Decision Tree	-	0.01822	0.135
	SVM-Poly	-	0.0234	0.153
	ANN	-	0.03422	0.185
Lokhandwala Nateghi (2018)	Generalized liner model	-	0.24	0.489
	Multivariate Adaptive regression splines	-	0.46	0.678
	SVM	-	0.51	0.714
	Random Forest	-	0.54	0.734
	ANN	-	0.60	0.774
Deng *et al.* (2018)	Linear regression	0.297	0.1513	0.389
	Lasso regression	0.271	0.1267	0.356
	SVM	0.257	0.1260	0.355
	ANN	0.271	0.1246	0.353
	Gradient boosting	0.267	0.1246	0.353
	Random forest	0.27	0.1253	0.354
Kumar Mohapatra *et al.* (2022)	Random Forest	0.22	0.29	0.538
Abdelkader *et al.* (2021)	Linear Regression	0.45	0.0232	1.523
	ANN	0.4323	0.0152	1.2312
Roth *et al.* (2020)	Lasso	-	0.95	0.974
	Random Forest	-	0.794	0.891

As observed in Table 4.6, the lowest MAE, MSE, and RMAE are observed in the proposed models without feature selection, specifically in the Stacking Regression model (Robinson *et al.* 2017). The Stacking Regression achieves an MAE of 0.01286, MSE of 0.00054, and RMAE of 0.02328. Stacking Regression is an ensemble method that combines multiple regression models to improve performance and increase prediction accuracy. It appears that this ensemble approach effectively leverages the complementary strengths of the constituent models, resulting in superior predictive capabilities for energy power consumption.

On the other hand, the highest MAE and MSE are found in the Generalized linear model (GLM) (Deng *et al.* 2018) with an MAE of 0.60 and MSE of 0.774. GLM is a traditional linear regression model that assumes a linear relationship between the features and the target variable. While it is interpretable and easy to implement, it may not be able to capture complex non-linear patterns in the data, leading to less accurate predictions compared to more advanced models like Random Forest or neural networks.

The comparison of the lowest and highest performing models highlights the importance of choosing appropriate regression models for energy power consumption prediction. Advanced techniques like ensemble methods, such as Stacking Regression and Random Forest, demonstrate improved accuracy compared to traditional linear models in this context. However, it's crucial to consider the specific dataset characteristics, feature selection, interpretability, and other constraints when selecting the most suitable model for a particular application.

Additionally, some studies report uncertainty measures (e.g., ± 0.01) for certain models, which indicates the variability in the model's performance. This variability could be due to factors such as different train-test splits, random initialization in neural networks, or hyperparameter tuning. Researchers should consider reporting such measures to provide a better understanding of the models' robustness and stability.

Finally, the Table 4.6 provides valuable insights into the performance of various regression models for energy power consumption prediction. The Stacking Regression

stands out as the most accurate model in the proposed models without feature selection, while the Generalized linear model performs less favorably due to its simplicity and limited capacity to capture complex relationships. Nevertheless, choosing the best model requires careful consideration of the specific requirements and characteristics of the prediction task.

5. CONCLUSIONS AND RECOMMENDATION

Buildings require energy for thermal comfort, air quality, lighting, hot water, and domestic appliances. Electricity is crucial in modern life, influencing technical, social, and economic factors. Demand planning is essential for effective capacity investments and energy plans. Policymakers prioritize accurate modeling and forecasting for quality and problem-free conditions. In the end of this thesis, the stacking model for energy power estimation has yielded promising results without using feature selection techniques, as obtained by the low MAE of 0.01286, the minimal MSE of 0.00054, and the relatively small RMSE of 0.02328. These measures show that the stacking model predicts energy power levels with a high degree of accuracy and precision. By combining multiple models in a stacked ensemble, the stacking approach has effectively leveraged the strengths of individual models, bagging and boosting algorithms, leading to improved performance and enhanced predictive capabilities. Overall, the success of the stacking model in energy power estimation highlights its potential as a valuable tool for optimizing energy management and facilitating more efficient and sustainable energy utilization in various domains.

For future work we can use the other optimization method such as metaheuristic methods, Genetic algorithm method to selet the best feature of electerical energy consumption data to recognize the electerical energy consumption data prediction.

Future work for energy consumption regression can focus on the following aspects:

- Model Fine-tuning: To discover the optimum model configurations, fine-tune hyperparameters using techniques such as grid search, random search, or Bayesian optimisation.
- Real-time Predictions: Implement the top-performing model in a real-time prediction system for proactive decision-making and efficient energy management, optimizing energy usage, reducing costs, and enhancing energy efficiency.

REFERENCES

Abdelkader, B. A. M., Sadek, H. and Hamdy, M. 2021. Buildings Energy Prediction Using Artificial Neural Networks. Engineering Research Journal, 171(0): 106–118.

Aguilar, J. E. R., Zang, L., and Fushimi, S. 2022. Towards sustainability in hospitality operations: how is quality of life and work–life balance related? Worldwide Hospitality and Tourism Themes, ahead-of-print.

Ahmad, A. S., Hassan, M. Y., Abdullah, M. P., Rahman, H. A., Hussin, F., Abdullah, H. and Saidur, R. 2014. A review on applications of ANN and SVM for building electrical energy consumption forecasting. Renewable and Sustainable Energy Reviews, 33: 102–109.

Ahmad, M. W., Mourshed, M. and Rezgui, Y. 2017. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy and Buildings, 147: 77–89.

Alahakoon, D. and Yu, X. 2016. Smart Electricity Meter Data Intelligence for Future Energy Systems: A Survey. IEEE Transactions on Industrial Informatics, 12(1): 425–436.

Amin, M. A. Al and Hoque, M. A. 2019. Comparison of ARIMA and SVM for short-term load forecasting. IEMECON 2019 - 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference, 205–210.

Arjunan, P., Poolla, K. and Miller, C. 2020. EnergyStar++: Towards more accurate and explanatory building energy benchmarking. Applied Energy, 276.

Athey, S., Tibshirani, J. and Wager, S. 2019. Generalized random forests. The Annals of Statistics, 47(2): 1148–1178.

Bilgen, S. 2014. Structure and environmental impact of global energy consumption. Renewable and Sustainable Energy Reviews, 38: 890–902.

Bolandnazar, E., Rohani, A. and Taki, M. 2020. Energy consumption forecasting in agriculture by artificial intelligence and mathematical models. Energy Sources, Part A: Recovery, Utilization and Environmental Effects, 42(13): 1618–1632.

BROWNLEE, 2019. Webstie. https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/. Date of access: 27-11-2019.

Bühlmann, P., Diggle, P., Fienberg, S., Gather, U., Olkin, I. and Zeger, S. 2009. Springer series in statistics,Book series, 221 page, Springer.

Camposeco-Negrete C. 2013. Optimization of cutting parameters for minimizing energy consumption in turning of AISI 6061 T6 using Taguchi methodology and ANOVA. Journal of Cleaner Production, 53: 195–203.

Cao, X., Dai, X. and Liu, J. 2016. Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade. Energy and Buildings, 128: 198–213.

Catalina, T., Virgone, J.and Blanco, E. 2008. Development and validation of regression models to predict monthly heating demand for residential buildings. Energy and Buildings, 40(10): 1825–1832.

Chen, Y.-T., Piedad Jr, E. and Kuo, C.-C. 2019. Energy consumption load forecasting using a level-based random forest classifier. Symmetry, 11(8): 956.

Council, N. R. 2009. Committee for the study on the relationships among development patterns vehicle miles traveled and energy consumption and transportation, Research Board.

Crawley, D. B., Lawrie, L. K., Winkelmann, F. C., Buhl, W. F., Huang, Y. J., Pedersen, C. O., Strand, R. K., Liesen, R. J., Fisher, D. E., Witte, M. J. and Glazer, J. 2001). EnergyPlus: Creating a new-generation building energy simulation program. Energy and Buildings, 33(4): 319–331.

Daneshfaraz, R., Aminvash, E., Ghaderi, A., Abraham, J. and Bagherzadeh, M. 2021. SVM performance for predicting the effect of horizontal screen diameters on the hydraulic parameters of a vertical drop. Applied Sciences (Switzerland), 11(9).

Daud, K., Abidin, A. F. and Nagindar Singh, H. S. 2015. ANOVA Based Feature Analysis and Selection in Power Quality Disturbances Identification. Applied Mechanics and Materials, 793: 510–515.

De Wilde, P. 2014. The gap between predicted and measured energy performance of buildings: A framework for investigation. Automation in Construction, 41: 40–49.

Deng, H., Fannon, D. and Eckelman, M. J. 2018. Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata. Energy and Buildings, 163: 34–43.

Deng, H., Fannon, D. and Eckelman, M. J. 2018. Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata. Energy and Buildings, 163: 34–43.

Ding, Y., Fan, L. and Liu, X. 2021. Analysis of feature matrix in machine learning algorithms to predict energy consumption of public buildings. Energy and Buildings, 249: 111208.

Divina, F., Gilson, A., Goméz-Vela, F., Torres, M. G. and Torres, J. F. 2018. Stacking ensemble learning for short-term electricity consumption forecasting. Energies, 11(4): 1–31.

Dong, B., Cao, C. and Lee, S. E. 2005. Applying support vector machines to predict building energy consumption in tropical region. Energy and Buildings, 37(5): 545–553.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. and Vapnik, V. 1996. Support vector regression machines. Advances in Neural Information Processing Systems, 9.

Estrada, M. del R. C., Camarillo, M. E. G., Parraguirre, M. E. S., Castillo, M. E. G., Juárez, E. M. and Gómez, M. J. C. 2020. Evaluation of several error measures applied to the sales forecast system of chemicals supply enterprises. International Journal of Business Administration, 11(4): 39.

Fadda, G., Brown, R. A., Longoni, G., Castro, D. A., O'Mahony, J., Verhey, L. H., Branson, H. M., Waters, P., Bar-Or, A. and Marrie, R. A. 2018. MRI and laboratory features and the performance of international criteria in the diagnosis of multiple sclerosis in children and adolescents: a prospective cohort study. The Lancet Child & Adolescent Health, 2(3): 191–204.

Gonzalez, P. A. and Zamarreno, J. M. 2005. Prediction of hourly energy consumption in buildings based on a feedback artificial neural network. Energy and Buildings, 37(6): 595–601.

Gourlis, G. and Kovacic, I. 2017. Building Information Modelling for analysis of energy efficient industrial buildings – A case study. Renewable and Sustainable Energy Reviews, 68: 953–963.

Gray, A., Ivezic, Z., Connolly, A. J. and VanderPlas, J. T. 2014. Statistics, Data Mining, and Machine Learning in Astronomy. Statistics, Data Mining, and Machine Learning in Astronomy.

Guhathakurta, S. and Williams, E. 2015. Impact of Urban Form on Energy Use in Central City and Suburban Neighborhoods: Lessons from the Phoenix Metropolitan Region. Energy Procedia, 75: 2928–2933.

Hafeez, G., Alimgeer, K. S., Qazi, A. B., Khan, I., Usman, M., Khan, F. A. and Wadud, Z. 2020. A hybrid approach for energy consumption forecasting with a new feature engineering and optimization framework in smart grid. IEEE Access, 8(i): 96210–96226.

Han, G., He, Y., Jiang, J., Wang, N., Guizani, M. and Ansere, J. A. 2019. A synergetic trust model based on SVM in underwater acoustic sensor networks. IEEE Transactions on Vehicular Technology, 68(11): 11239–11247.

Hollander, M., Wolfe, D. A. and Chicken, E. 2013. Nonparametric statistical methods. John Wiley and Sons.

Hutter, M. 2021. Learning curve theory, Machine Learning 26 page, NewYork.

Jafarzadeh, H., Mahdianpari, M., Gill, E., Mohammadimanesh, F. and Homayouni, S. 2021. Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral and polSAR data: A comparative evaluation. Remote Sensing, 13(21).

Jia, H. and Chong, A. 2021. eplusr: A framework for integrating building energy simulation and data-driven analytics. Energy and Buildings, 237: 110757.

Kamble, V. B. and Deshmukh, S. N. 2017. Comparision between accuracy and MSE,RMSE by using proposed method with imputation technique. Oriental Journal of Computer Science and Technology, 10(04): 773–779.

Kumar Mohapatra, S., Mishra, S., Tripathy, H. K. and Alkhayyat, A. 2022. A sustainable data-driven energy consumption assessment model for building infrastructures in resource constraint environment. Sustainable Energy Technologies and Assessments, 53.

Li, G., Zheng, Y., Liu, J., Zhou, Z., Xu, C., Fang, X. and Yao, Q. 2021. An improved stacking ensemble learning-based sensor fault detection method for building energy systems using fault-discrimination information. Journal of Building Engineering, 43: 102812.

Li, Q., Meng, Q., Cai, J., Yoshino, H. and Mochida, A. 2009. Applying support vector machine to predict hourly cooling load in the building. Applied Energy, 86(10): 2249–2256.

Liu, Y. Y., Yang, M., Ramsay, M., Li, X. S. and Coid, J. W. 2011. A Comparison of Logistic Regression, Classification and Regression Tree, and Neural Networks Models in Predicting Violent Re-Offending. Journal of Quantitative Criminology, 27(4): 547–573.

Lokhandwala, M. and Nateghi, R. 2018. Leveraging advanced predictive analytics to assess commercial cooling load in the U.S. Sustainable Production and Consumption, 14: 66–81.

Mabayoje, M. A., Balogun, A. O., Bajeh, A. O. and Musa, B. A. 2018. Software defect prediction: effect of feature selection and ensemble methods, FUW Trends in Science and Technology Journal 3(2): 518 – 522.

Mastelini, S. M., Nakano, F. K., Vens, C. and Carvalho, A. C. P. de L. F. de. 2022. Online extra trees regressor. IEEE Transactions on Neural Networks and Learning Systems.

Moon, J., Jung, S., Rew, J., Rho, S. and Hwang, E. 2020. Combination of short-term load forecasting models based on a stacking ensemble approach. Energy and Buildings, 216: 109921.

Neto, A. H. and Fiorelli, F. A. S. 2008. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. Energy and Buildings, 40(12): 2169–2176.

Nichols, B. G. and Kockelman, K. M. 2014. Life-cycle energy implications of different residential settings: Recognizing buildings, travel, and public infrastructure. Energy Policy, 68: 232–242.

Nooruldeen, O., Alturki, S., Baker, M. R. and Ghareeb, A. 2022. Time series forecasting for decision making on city-wide energy demand: A comparative study. 2022 International Conference on Decision Aid Sciences and Applications, DASA 2022, March, 1706–1710.

Olmedo, M. T. C., Paegelow, M., Mas, J. F. and Escobar, F. 2018. Geomatic approaches for modeling land change scenarios. an introduction. In Lecture Notes in Geoinformation and Cartography 2020.

Omer, A. M. 2008. Energy, environment and sustainable development. Renewable and Sustainable Energy Reviews, 12(9): 2265–2300.

Patro, S. G. K. and Sahu, K. K. 2015. Normalization: A preprocessing stage. Iarjset, 20–22.

Pavlyshenko, B. 2018. Using Stacking Approaches for Machine Learning Models. Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP 2018: 255–258.

Pérez-Lombard, L., Ortiz, J. and Pout, C. 2008. A review on buildings energy consumption information. Energy and Buildings, 40(3): 394–398.

Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M. A. and Pendyala, R. M. 2017. Machine learning approaches for estimating commercial building energy consumption. Applied Energy, 208: 889–904.

Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M. A. and Pendyala, R. M. 2017. Machine learning approaches for estimating commercial building energy consumption. Applied Energy, 208(May): 889–904.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. and Chica-Rivas, M. 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, 71: 804–818.

Roth, J., Lim, B., Jain, R. K. and Grueneich, D. 2020. Examining the feasibility of using open data to benchmark building energy usage in cities: A data science and policy perspective. Energy Policy, 139: 1–51.

Schlueter, A. and Thesseling, F. 2009. Building information model based energy/exergy performance assessment in early design stages. Automation in Construction, 18(2): 153–163.

Solano, E. S., Dehghanian, P and Affonso, C. M. 2022. Solar Radiation Forecasting Using Machine Learning and Ensemble Feature Selection. Energies, 15(19).

Torabi, M., Hashemi, S., Saybani, M. R., Shamshirband, S. and Mosavi, A. 2019. A hybrid clustering and classification technique for forecasting short-term energy consumption. Environmental Progress and Sustainable Energy, 38(1): 66–76.

Touzani, S., Granderson, J. and Fernandes, S. 2018. Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy and Buildings, 158(510): 1533–1543.

Tsanas, A. and Xifara, A. 2012. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy and Buildings, 49: 560–567.

Tso, G. K. F. and Yau, K. K. W. 2007. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy, 32(9): 1761–1768.

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S. and Moore, J. H. 2018. Relief-based feature selection: Introduction and review. Journal of Biomedical Informatics, 85: 189–203.

Wang, F., Li, K., Duić, N., Mi, Z., Hodge, B. M., Shafie-khah, M. and Catalão, J. P. S. 2018. Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns. Energy Conversion and Management, 171: 839–854.

Wu, Z., Li, N., Peng, J., Cui, H., Liu, P., Li, H. and Li, X. 2018. Using an ensemble machine learning methodology-Bagging to predict occupants' thermal comfort in buildings. In Energy and Buildings, Elsevier 173.

X Yan, C. T. X. S. 2011. Linear regression. munro's statistical methods for health care research: sixth edition, 339–370.

Yahman, Y. and Setyagama, A. 2022. Government policy in regulating the environment for development of sustainable environment in Indonesia. Environment, Development and Sustainability, 1–12.

Yezioro, A., Dong, B. and Leite, F. 2008. An applied artificial intelligence approach towards assessing building performance simulation tools. Energy and Buildings, 40(4): 612–620.

Younes, H., Ibrahim, A., Rizk, M. and Valle, M. 2021. An Efficient Selection-Based kNN Architecture for Smart Embedded Hardware Accelerators. IEEE Open Journal of Circuits and Systems, 2: 534–545.

Zhang, Y., Teoh, B. K., Wu, M., Chen, J. and Zhang, L. 2023. Data-driven estimation of building energy consumption and GHG emissions using explainable artificial intelligence. Energy, 262.

Zhao, H. and Magoulès, F. 2012. A review on the prediction of building energy consumption. Renewable and Sustainable Energy Reviews, 16(6): 3586–3592.

Zheng, Z., Wu, J., Lian, Z. and Zhang, H. 2020. A method to evaluate building energy consumption based on energy use index of different functional sectors. Sustainable Cities and Society, 53: 101893.

APPENDICES

APPENDIX 1. Samples from the dataset

APPENDIX 2. Dataset features description

APPENDIX 1. Samples from the dataset

Table 1. Samples from the dataset

CENDIV	SQFT	WLCNS	RFCNS	RFCOOL	RFTILT	BLDSHP	GLSSPC	EQGLSS	NFLOOR	FLCEILHT	ATTIC	YRCON	RENRFF	RENWLL	RENWIN
5	89000	3	4	0	1	2	5	1	5	8	0	1984	0	0	1
5	1500	3	4	0	2	1	1	0	1	8	1	1979	0	1	0
5	7900	2	2	0	3	2	3	0	1	12	1	2005	0	0	0
9	10000	1	2	0	1	3	3	0	1	12	0	1979	0	0	0
5	600000	3	1	0	1	2	5	1	30	15	0	1983	0	0	0
RENHVC	**RENLGT**	**RENELC**	**RENINS**	**ONEACT**	**FACIL**	**OWNTYPE**	**NOCC**	**OWNOCC**	**OWNOPR**	**OCCUPYP**	**OPEN24**	**OPNWE**	**WKHRS**	**NWKER**	**FKUSED**
1	0	1	0	1	0	2	27	3	1	80	0	1	68	170	0
0	1	1	0	1	1	3	1	1	1	100	0	1	60	2	0
0	0	0	0	1	0	3	2	1	1	90	0	1	70	27	0
0	0	0	0	0	1	2	1	2	1	100	0	0	40	7	0
1	0	0	0	1	0	2	30	2	1	78	0	1	65	925	1
PRUSED	**STUSED**	**ELHT1**	**NGHT1**	**HEATP**	**MAINHT**	**HTVCAV**	**HTVVAV**	**ELCOOL**	**NGCOOL**	**COOLP**	**MAINCL**	**CLVCAV**	**CLVVAV**	**EMCS**	**RDHTNF**
0	0	0	0	0	8	0	0	1	0	95	1	0	0	0	0
0	0	0	0	0	8	0	0	1	0	100	1	1	0	0	0
0	0	0	0	0	8	0	0	1	0	100	1	1	0	0	0
0	0	0	0	0	8	0	0	1	0	50	1	1	0	0	0
0	0	0	0	0	8	0	0	1	0	100	5	0	1	1	0
RDCLNF	**ECN**	**MAINT**	**ENRGYPLN**	**PCTERMN**	**LAPTPN**	**PRNTRN**	**COPIERN**	**LTOHRP**	**LTNHRP**	**FLUOR**	**CFLR**	**BULB**	**HALO**	**HID**	**LED**
1	0	1	0	75	10	50	5	100	100	1	1	1	1	1	1
1	0	0	0	2	0	4	0	75	0	1	0	1	0	0	0
1	0	1	0	12	1	9	1	100	0	1	0	0	0	0	0
1	0	0	0	7	3	1	1	70	0	1	0	0	0	0	0
1	1	1	1	830	400	60	45	90	10	1	1	1	1	1	1
SCHED	**OCSN**	**WINTYP**	**TINT**	**REFL**	**AWN**	**SKYLT**	**DAYLTP**	**HDD65**	**CDD65**	**HVACEUI**	**LNEUI**	**PLEUI**	**TOTALEUI**		
1	0	2	1	0	0	0	65	80	3776	15.08991	9.924899	2.406584	30.39762		
0	0	3	0	1	0	0	0	214	3506	2.761333	0.776	4.044	8.525333		
0	0	2	0	0	1	0	0	147	4675	11.42987	4.571646	4.393924	24.21317		
1	0	1	1	0	0	0	15	905	1351	10.8686	5.4348	4.625	32.7632		
1	1	1	1	1	0	0	0	80	4584	123.6783	10.19839	7.145487	153.7811		

APPENDIX 2. Dataset features description

Table 2. Dataset features description

ITEM	VARIABLE NAME	CATEGORİCAL VARİABLES	ITEM	VARIABLE NAME	CATEGORICAL VARİABLES
1	CENDIV	Census division	20	RENINS	Insulation upgrade
2	SQFT	Square footage	21	ONEACT	One activity in building
3	WLCNS	Wall construction material	22	FACIL	On a multi-building complex
4	RFCNS	Roof construction material	23	OWNTYPE	Building owner type
5	RFCOOL	Cool roof materials	24	NOCC	Number of businesses
6	RFTILT	Roof tilt	25	OWNOCC	Owner occupied or leased to tenants
7	BLDSHP	Building shape	26	OWNOPR	Owner responsible for operation
8	GLSSPC	Percent exterior glass	27	OCCUPYP	Occupancy percentage
9	EQGLSS	Equal glass on all sides	28	OPEN24	Open 24 hours a day
10	NFLOOR	Number of floors	29	OPNWE	Open on weekends
11	FLCEILHT	Floor to ceiling height	30	WKHRS	Total hours open per week
12	ATTIC	Attic	31	NWKER	Number of workers
13	YRCON	Year of construction	32	FKUSED	Fuel oil/diesel/kerosene used
14	RENRFF	Roof replacement	33	PRUSED	Bottled gas/LPG/propane used
15	RENWLL	Exterior wall replacement	34	STUSED	District steam used
16	RENWIN	Window replacement	35	ELHT1	Electricity used for main heating
17	RENHVC	HVAC equipment upgrade	36	NGHT1	Natural gas used for main heating
18	RENLGT	Lighting upgrade	37	HEATP	Heating percentage
19	RENELC	Electrical upgrade	38	MAINHT	Main heating equipment
39	HTVCAV	Heating ventilation: Central air	59	FLUOR	Fluorescent bulbs
40	HTVVAV	Heating ventilation: Central air	60	CFLR	Compact fluorescent bulbs
41	ELCOOL	Electricity used for cooling	61	BULB	Incandescent bulbs
42	NGCOOL	Natural gas used for cooling	62	HALO	Halogen bulbs
43	COOLP	Cooling percentage	63	HID	High Intensity discharge (HID) bulbs
44	MAINCL	Main cooling equipment	64	LED	Light-emitting diode (LED) bulbs
45	CLVCAV	Cooling ventilation: Central air	65	SCHED	Light scheduling
46	CLVVAV	Cooling ventilation: Central air	66	OCSN	Occupancy sensors
47	EMCS	Building automation system	67	WINTYP	Window glass type
48	RDHTNF	Heating equipment controlled by EMS	68	TINT	Tinted window glass
49	RDCLNF	Cooling equipment controlled by EMS	69	REFL	Reflective window glass
50	ECN	Economizer cycle	70	AWN	External overhangs or awnings
51	MAINT	Regular HVAC maintenance	71	SKYLT	Skylights or atriums
52	ENRGYPLN	Energy management plan	72	DAYLTP	Daylight percentage
53	PCTERMN	Percentage of energy cost paid by tenant	73	HDD65	Heating degree days
54	LAPTPN	Number of laptop computers	74	CDD65	Cooling degree days
55	PRNTRN	Number of printers	75	HVACEUI	HVAC energy use intensity

Table.2 Dataset features description (Continued)

ITEM	VARIABLE NAME	CATEGORİCAL VARİABLES	ITEM	VARIABLE NAME	CATEGORICAL VARİABLES
56	COPIERN	Number of copiers	76	LTNEUI	Lighting energy use intensity
57	LTOHRP	Percent lit when open	77	PLEUI	Plug load energy use intensity
58	LTNHRP	Percent lit off hours	78	TOTALEUI	Total energy use intensity