

School of Built Environment, Engineering and Computing

Leeds Beckett University

**Comparison of Machine Learning Algorithms in Building Energy Consumption Prediction**

By: Kobina Folson

Hadeel Jazaa

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

**August 2023**

# Candidate’s Declaration

I, Kobina Gumye Folson, confirm that this dissertation and the work presented in it are my own achievement.

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

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

I have acknowledged all main sources of help;

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

Signed: K.F

Date: 30/08/2023

Student ID No: 77296873

# Acknowledgements

I sincerely thank my family for their unflagging support and inspiration along this journey. Your unwavering faith in my talents and tolerance while I worked on this dissertation for endless hours mean the world to me.

I owe a great deal to my supervisor, Hadeel Jazaa, for their crucial advice, know-how, and commitment. Their astute comments and helpful criticism were essential in determining the focus and calibre of this research.

I also appreciate the teachers and lecturers who shared their wisdom during my academic career. I am sincerely grateful for the inspiration you have given me, as your lessons have served as the cornerstone on which I have constructed my understanding.

Last but not least, I want to express my gratitude to everyone who, directly or indirectly, gave their time and wisdom to this effort. This dissertation was completed successfully thanks to the combined efforts of many people.

# Abstract

Efficient energy utilization, reduction of carbon emissions, and optimal energy management are paramount objectives in the modern built environment. To realize these goals, the accurate prediction of building energy consumption emerges as a pivotal facet, necessitating the integration of advanced forecasting methodologies driven by machine learning (ML) algorithms. These algorithms have exhibited promising potential in enabling strategic energy planning and facilitating eco-conscious initiatives.

This dissertation undertakes a comprehensive exploration to equip stakeholders with the insights required for informed decision-making. The focal point lies in meticulously examining and assessing a spectrum of ML systems tailored to forecast building energy consumption. The research initiative embarks with a meticulous dissection of prevailing techniques within building energy forecasting, expounding upon the strengths and weaknesses inherent in diverse algorithmic paradigms. Subsequently, an intricate matrix of critical determinants—encompassing prediction accuracy, computational efficiency, interpretability, and scalability—is the foundation for identifying an elite cohort of ML algorithms primed for the task.

Intrinsic to this endeavour is the acquisition and curation of pertinent building energy consumption data, underscored by an assiduous approach to addressing data gaps, isolating outliers, and implementing judicious feature scaling methodologies. This meticulously prepared dataset forms the bedrock for orchestrating the chosen machine-learning algorithms. In this phase, refinements, including hyperparameter calibration and optimization techniques, are instrumental in elevating the predictive capabilities of the models.

The efficacy of these methodologies is evaluated through a multifaceted lens, encompassing the dual facets of model interpretability and computational efficiency. A comprehensive suite of metrics, ranging from the Mean Absolute Error (MAE) to the Root Mean Squared Error (RMSE), lends quantitative rigour to the assessment process. This holistic evaluation framework encapsulates quantitative accuracy and qualitative dimensions, empowering a nuanced understanding of model performance.

Each distinct ML approach for building energy prediction is subjected to exhaustive scrutiny, unravelling its virtues and limitations. By fostering an in-depth comprehension of the intricacies and implications of each algorithm, this research contributes to the practical integration of these methodologies within the complexities of real-world contexts. Consequently, this study advances the discourse on building energy prediction. It serves as a pragmatic guide for stakeholders navigating the landscape of sustainable energy management.

Contents

[Candidate’s Declaration ii](#_Toc145023987)

[Acknowledgements iii](#_Toc145023988)

[Abstract iv](#_Toc145023989)

[List of Figures ix](#_Toc145023990)

[List of Tables 10](#_Toc145023991)

[Chapter 1: Introduction 11](#_Toc145023992)

[1.1 Overview 11](#_Toc145023993)

[1.2 Problem Statement 11](#_Toc145023994)

[1.3 Aim and Objectives 11](#_Toc145023995)

[1.4 Outline 14](#_Toc145023996)

[This dissertation is organized as follows: 14](#_Toc145023997)

[Chapter 2: Literature Review 16](#_Toc145023998)

[2.1 Background 16](#_Toc145023999)

[2.2 Studies Performing Comparative Analysis of Machine Learning Algorithms 17](#_Toc145024000)

[2.2.1 Predictive Modelling for US Commercial Building Energy Use 18](#_Toc145024001)

[2.2.2 Tuning Machine Learning Models for Prediction of Building Energy Loads 19](#_Toc145024002)

[2.2.3 A Deep Learning Framework for Building Energy Consumption Forecast 20](#_Toc145024003)

[2.2.4 Building Energy Consumption Prediction: An Extreme Deep Learning Approach 20](#_Toc145024004)

[2.2.5 Machine learning applications in urban building energy performance forecasting: A systematic review 21](#_Toc145024005)

[2.3 Machine-Learning in Energy Prediction 22](#_Toc145024006)

[2.3.1 Neural Networks (NN) 22](#_Toc145024007)

[2.3.2 Support Vector Machines (SVM) 22](#_Toc145024008)

[2.3.3 Random Forest (RF) 23](#_Toc145024009)

[2.3.4 Decision Trees 24](#_Toc145024010)

[2.3.5 Linear Regression 24](#_Toc145024011)

[2.3.5 Gradient Boosting 25](#_Toc145024012)

[2.4 Scope and Significance of Study 29](#_Toc145024013)

[Chapter 3: Methodology 31](#_Toc145024014)

[3.1 CRISP Methodology Implementation 31](#_Toc145024015)

[3.2 Research Design 33](#_Toc145024016)

[3.3 Data Exploration 34](#_Toc145024017)

[3.1.2 Data Collection 40](#_Toc145024018)

[3.1.3 Data Pre-processing 45](#_Toc145024019)

[3.1.4 Algorithm Selection 45](#_Toc145024020)

[3.6 Performance Metrics 45](#_Toc145024021)

[3.7 Experimental Setup 45](#_Toc145024022)

[3.8 Ethical Considerations 46](#_Toc145024023)

[Chapter 4 : Product/Research Design and Implementation 48](#_Toc145024024)

[4.1 Software and Hardware Infrastructure 51](#_Toc145024025)

[4.1.1 Software Framework 51](#_Toc145024026)

[4.1.2 Hardware Configuration 52](#_Toc145024027)

[4.2 Challenges and Limitations 53](#_Toc145024028)

[4.1.1 Data Availability and Quality 53](#_Toc145024029)

[4.1.2 Model Complexity and Interpretability 53](#_Toc145024030)

[4.1.3 Algorithm Selection and Hyperparameter Tuning 54](#_Toc145024031)

[4.1.4 Generalisation and Transferability 54](#_Toc145024032)

[4.2 Code Implementation 54](#_Toc145024033)

[4.2.1 Data Cleaning: Handling Missing Values 54](#_Toc145024034)

[4.2.2 Data Sampling 58](#_Toc145024035)

[4.2.3 Feature Engineering: Data Scaling 60](#_Toc145024036)

[Chapter 5: Research Outcomes/Results/Discussion and Evaluation 63](#_Toc145024037)

[5.1 Evaluation Metrics 63](#_Toc145024038)

[Chapter 6: Project Management 67](#_Toc145024039)

[6.1 Project Overview 67](#_Toc145024040)

[6.2 Project Phases 67](#_Toc145024041)

[6.3 Project Resources 69](#_Toc145024042)

[6.4 Risk Management 70](#_Toc145024043)

[6.5 Communication and Collaboration 70](#_Toc145024044)

[6.6 Project Monitoring and Control 70](#_Toc145024045)

[Chapter 7: Conclusion and Future Work 71](#_Toc145024046)

[References 74](#_Toc145024047)

[Appendices 77](#_Toc145024048)

# List of Figures

[Figure 3‑1 Energy Efficiency 36](#_Toc145100730)

[Figure 3‑2 Correlation Matrix of Data 37](#_Toc145100731)

[Figure 3‑3 Distribution of Current and Potential Energy Ratings 38](#_Toc145100732)

[Figure 3‑4 Distribution of Tenure/Building Occupancy 39](#_Toc145100733)

[Figure 4‑1 Flowchart of ML Model Implementation 49](#_Toc145100734)

[Figure 4‑2 Load dataset into Dataframe using Pandas 55](#_Toc145100735)

[Figure 4‑3 Output of first 5 rows from dataset 55](#_Toc145100736)

[Figure 4‑4 Percentage of missing values for Columns 56](#_Toc145100737)

[Figure 4‑5 Drop columns with 50% or more missing 57](#_Toc145100738)

[Figure 4‑6 Fill in columns with missing values 57](#_Toc145100739)

[Figure 4‑7 Data Sampling 60](#_Toc145100740)

[Figure 4‑8 Data Encoding and Scaling 61](#_Toc145100741)

[Figure 4‑9 Feature Selection 62](#_Toc145100742)

[Figure 5‑1 Model MAE Comparison 66](#_Toc145100743)

[Figure 5‑2 Model RMSE Comparison 67](#_Toc145100744)

# List of Tables

[Table 2‑1 Strengths, Weaknesses and Performance of various ML Algorithms 27](#_Toc145102182)

[Table 3‑1 Sample Description of Data 34](#_Toc145102183)

[Table 3‑2 Description of Columns in Dataset 40](#_Toc145102184)

# Introduction

## Overview

Building energy efficiency is essential for sustainable development and minimising environmental impact. The need for energy increases along with the growth of the world's population, putting more strain on available energy sources and raising environmental concerns. In this setting, attaining energy efficiency, reducing carbon emissions, and improving energy management in the built environment depends on accurate building energy usage prediction.

Algorithms for machine learning (ML) have become effective tools for forecasting complicated patterns and making data-driven judgements. ML approaches have recently demonstrated potential in predicting building energy demand, allowing for better energy planning and cost-effective solutions. These algorithms generate prediction models that can help in optimising energy usage, decreasing waste, and improving overall energy sustainability by drawing on past data on energy consumption, weather patterns, occupancy rates, building attributes, and other pertinent parameters.

## Problem Statement

A comparative analysis of machine learning methods for predicting building energy consumption which is necessary to find the most precise and trustworthy models. This will also benefit researchers and practitioners in their own research to find the best strategy for predicting building energy by assessing and contrasting the performance of various algorithms.

## Aim and Objectives

This dissertation's main goal is to analyse and assess various ML systems for predicting building energy. We hope to accomplish the following specific goals with this study:

1. Conducting a comprehensive assessment of existing ML techniques through a thorough literature review to gauge their effectiveness in estimating building energy consumption.
2. Finding and preparing data from trusted sources and performing the needed data pre-processing, including handling missing values, outliers, and feature scaling, ensures dataset readiness for ML model training.
3. Selecting and Implementing appropriate ML algorithms and training them using the prepared dataset, optimising models, and fine-tuning hyperparameters to yield optimal results.
4. Evaluating the performance of the ML algorithms using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and considering model interpretability and computational efficiency.
5. Analysing the advantages and limitations of each ML approach for predicting building energy, facilitating informed decision-making by comprehending their strengths and weaknesses.

The outcome of this research will ultimately help other researchers in building more efficient models to predict building energy consumption by making use of the results to create a baseline for their own research. The findings can be used by policymakers and urban planners to create evidence-based regulations that support sustainable building methods and energy efficiency. Additionally, a deeper comprehension of the performance traits of various ML approaches in building energy consumption prediction will be helpful to researchers in the field.

## 1.4 Outline

This dissertation is organized as follows:

Chapter 2: Literature Review

This chapter reviews relevant literature on the prediction of building energy consumption, with a focus on studies that have employed machine learning techniques. The review will cover key findings, methodologies, and gaps in existing research, setting the stage for the present study.

Chapter 3: Methodology

This chapter describes the research design, data collection, and data analysis methods used in the study. It provides an in-depth explanation of the CRISP-DM methodology used to guide the research process, from business understanding to deployment.

Chapter 4: Data Analysis and Results

This chapter presents the results obtained from applying the chosen machine learning algorithms to the dataset. It covers the analysis of the data, including summary statistics and data visualizations, as well as the model training and evaluation process.

Chapter 5: Research Outcomes, Results, Discussion, and Evaluation

This chapter discusses the findings of the study, providing an interpretation of the results in the context of the research objectives and questions. It also evaluates the performance of the machine learning models and discusses their implications for predicting building energy consumption.

Chapter 6: Conclusion and Future Work

This chapter concludes the study, summarizing the key findings and discussing their implications. It also identifies limitations of the current research and proposes directions for future work in this field.

Appendices

The appendices include any supplementary material related to the study, such as detailed tables of results, code snippets, or additional data visualizations.

References

The references section lists all the sources cited in the dissertation, following the chosen citation style.

# Literature Review

This chapter thoroughly analyses the body of work on machine learning techniques for predicting building energy demand. It explores the approaches and algorithms used in earlier studies, highlighting their advantages and disadvantages. The chapter seeks to understand the most recent techniques in this area thoroughly.

## Background

Building energy forecasting is an essential component of contemporary infrastructure management that enables the optimisation of energy use, lowers costs, and supports sustainability initiatives. Effective energy usage forecasting can help develop comprehensive energy policies, energy-efficient designs, and Heating, ventilation, and air conditioning (HVAC) system control. A key challenge in enhancing the energy efficiency of buildings in dense urban areas is the lack of accurate energy performance prediction models that consider this urban context (Nutkiewicz et al., 2018).

Modelling plays a crucial role in leveraging artificial intelligence for intelligent building design, energy efficiency, and maintenance. As urbanization accelerates, efficient energy consumption in buildings becomes vital, and technologies like Building Information Modeling(BIM) and AI offer innovative solutions for comprehensive model information sharing and effective energy-efficient design (Ohueri et al., n.d.). These models frequently made explicit connections between the physical characteristics of buildings and their energy consumption. Such models, though they can give a thorough insight into energy consumption, are often labour and time-intensive to create. Additionally, they frequently need specific knowledge regarding the design and construction of the building, which is only sometimes readily available.

Building energy management has been significantly altered by the development of machine learning over the past few decades. Using computer models to "learn" from data, machine learning, a subset of artificial intelligence, enables systems to anticipate or decide without explicit programming.

Within Building Energy Management Systems (BEMSs), predicting building energy usage has many benefits. It makes Numerous applications possible, including Demand Side Management (DSM), intelligent control decisions, energy supply-demand equilibrium, behavioural analysis, and the best possible reactions to the current situation. Additionally, it improves energy flexibility by facilitating communication between buildings and smart grids, creating a stable energy environment. Recent studies have shown that this multidimensional strategy enables effective building operations and integration with more extensive energy networks (Khalil et al., 2022).

“Ultimately, accurate and cost effective wide-scale energy prediction is a vital step towards next-generation energy efficiency initiatives, which will require not only consideration of the methods, but the scales for which data can be distilled into meaningful information” (Jain et al., 2014). As a result, an expanding corpus of research is focused on investigating and contrasting the effectiveness of various machine-learning models in this application area.

In this study, we evaluate the literature on building energy prediction using machine learning models, compare their performance, and analyse their benefits and drawbacks. For researchers and professionals interested in using machine learning methods to anticipate building energy use, this will offer useful insights.

## Studies Performing Comparative Analysis of Machine Learning Algorithms

The body of research that has already been done to compare various machine learning algorithms for predictive modelling is examined in this section. This section looks into the methods used by researchers to assess the effectiveness, advantages, and disadvantages of various algorithms when used in various fields. This section provides helpful insights into the methods used to choose the best algorithms for particular prediction tasks by examining the methodology used, the datasets used, and the assessment criteria taken into account. This part strengthens the groundwork for the ongoing research project by synthesising information from previous studies to help construct a complete framework for assessing and choosing the best machine-learning methods for forecasting building energy use.

### Predictive Modelling for US Commercial Building Energy Use

In the study Predictive modelling for US commercial building energy use, A comparison of existing statistical and machine learning algorithms using CBECS microdata (Deng et al., 2018), they sought to investigate how machine learning algorithms might be used to forecast the energy performance of buildings, particularly Energy Use Intensity (EUI) in US commercial office buildings. The Commercial Building Energy Usage Survey (CBECS) 2012 microdata, which offered a plethora of information on building attributes, tenant behaviour, and energy usage, served as the dataset used for analysis.

The prediction accuracy of six regression or machine learning algorithms was examined. Regarding forecasting Total EUI, the Support Vector Machine and Random Forest algorithms showed accuracy and stability. However, although having 10-15% fewer prediction errors for Total EUI than linear regression, machine learning techniques were only marginally superior.

When calculating Plug Loads EUI, linear regression models somewhat outperformed machine learning techniques. This contradictory outcome recommended that caution be used when using sophisticated predictive algorithms in the CBECS dataset.

Using Random Forest, the study additionally examined the significance of specific variables. It was discovered that the top 10 predictors for Total EUI and sub-system EUIs (HVAC, plug loads, and lighting) varied, proving that various factors influence various energy end-uses.

The investigation showed that the CBECS variables lacked sufficient predictive power to represent actual energy use. The study recommended completing information gaps in areas including occupant behaviour, power management, building thermal efficiency, and their interactions to improve predictive modelling.

Overall, the study showed the promise of machine learning algorithms for predicting the energy performance of buildings. However, it also made clear the need for more advancements in data gathering and modelling methods to produce more precise estimates of actual energy use.

Outlining the study's key conclusions in brief:

* Support Vector Machine and Random Forest show potential in estimating Total EUI for commercial office buildings.
* When estimating Plug Loads EUI, linear regression models significantly outperform machine learning techniques.
* Various circumstances influence energy end-uses (HVAC, plug loads, and lights).
* It is challenging to anticipate actual energy usage using the CBECS factors precisely.
* Improvements to data gathering and information shortages may improve predictive modelling.

### Tuning Machine Learning Models for Prediction of Building Energy Loads

This study by (Seyedzadeh et al., 2019) they investigated how building heating and cooling loads can be predicted using machine learning (ML) approaches. It intended to develop a substitute for computationally expensive simulation tools by using past data already available to estimate future samples and make wise decisions.

The research used two simulated building energy datasets produced in EnergyPlus and Ecotect to examine the precision of well-known ML models in predicting heating and cooling demands. The study concentrated on nonlinear regression models, where the inputs cannot be combined linearly. These models included Gradient Boosted Regression Trees (GBRT), Random Forest (RF), Gaussian Process (GP) regression, Support Vector Machines (SVM), and Artificial Neural Networks (ANN).

The researchers tested several combinations of model parameters using a grid-search approach combined with cross-validation to ensure optimal performance. Sensitivity analysis methods were used to assess how essential input factors were to the effectiveness of ML models.

According to the study, model optimisation is key to raising forecast accuracy. When the accuracy of the adjusted models was compared to the original research findings, model optimisation's significance became apparent.

The sensitivity analysis revealed information on the input variables' relative weights, allowing removing unimportant ones to facilitate model fitting more quickly without sacrificing accuracy.

In order to provide a more effective design process with a more extensive search space during optimisation, the article highlighted the potential of surrogate models (data-driven models) to give quick and accurate alternatives to creating performance simulators.

The research thoroughly analysed various nonlinear ML models for estimating building energy loads. It offered helpful advice on model selection and tuning for increased precision and efficacy in designing and retrofitting high-performance buildings.

### A Deep Learning Framework for Building Energy Consumption Forecast

Using a real-time building energy consumption dataset from a four-story building at IIT-Bombay in India, the CNN-LSTM framework's efficiency and applicability were shown. Using well-known quality indicators, CNN-LSTM's performance was compared to other cutting-edge energy demand forecasting models. The outcomes emphasised CNN-LSTM's capability to grasp spatiotemporal dependencies in the energy consumption data, emphasising its capacity to give precise energy demand estimates (Somu et al., 2021).

### Building Energy Consumption Prediction: An Extreme Deep Learning Approach

An extreme deep learning method is suggested in the research report for precise building energy consumption prediction. The technique combines stacked autoencoders (SAEs) and the extreme learning machine (ELM) to take advantage of each component's advantages. From the data on energy usage, SAEs are utilised to extract pertinent aspects, and ELM serves as the predictor for precise forecasts. The model's input variables are chosen using partial autocorrelation analysis. When the suggested method is against well-known machine learning techniques like BPNN, SVR, GRBFNN, and MLR, it shows superior prediction performance in many circumstances. By providing accurate forecasts to building managers, our research helps improve energy usage (Li et al., 2017).

### Machine learning applications in urban building energy performance forecasting: A systematic review

The study emphasises the significance of predicting buildings' urban-scale energy efficiency and the requirement for effective energy planning. The study's projection of building energy performance is based on data from 2015 to 2018. It divides the literature into categories based on teaching techniques, structure types, energy types, input data, and time frame. It demonstrates a need for more research on forecasting at the metropolitan size instead of the level of a single structure. Additionally, it points out areas for future investigation, such as the neglect of building functionality and the consequences of climate change utilising machine learning and forecasts for the future. Furthermore, there is disagreement regarding the ideal criteria for precise machine learning-based forecasting (Fathi et al., 2020).

In the literature review, the full body of research on machine learning methods for forecasting building energy demand was examined. The advantages and uses of several methods, including Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest (RF), and K-Nearest Neighbours (KNN), were emphasised, and their use in predicting building energy use was carefully examined. The study also evaluated numerous algorithms' performance using analytical comparison assessments, highlighting their potential for producing precise estimations of energy demand. The study underscored how crucial machine learning is for increasing energy efficiency and reducing the environmental impact of building operations. A methodical approach and a synthesis of the data from the studies being taken into consideration might help to increase the impact of the review.

## Machine-Learning in Energy Prediction

Because machine-learning can handle complicated, nonlinear relationships and massive amounts of data, its significance in predicting building energy usage is growing. These tools enable the development of more reliable and accurate models than their conventional equivalents. Regression-based models, support vector machines, decision trees, and neural networks are fundamental machine-learning methods for predicting building energy consumption. Each model functions differently and has a unique set of benefits and restrictions.

In this section, a few algorithms which have been selected for this research on comparative analysis for building energy consumption prediction a looked at, laying out the reason, their advantages, and disadvantages and how they have been used in other research works.

### Neural Networks (NN)

Building energy prediction has emerged as a significant study area for improving energy efficiency in building management systems. Artificial neural networks (ANNs) have drawn attention because they can model complicated interactions without specialised knowledge. This study conducts a bibliometric analysis of ANNs in building energy prediction using 324 recent papers. It offers a thorough analysis of twelve ANN architectures and the uses for them in this field. The study addresses three significant issues: choosing the best ANN architecture, enhancing prediction performance, and dealing with data scarcity. Its objective is to provide academics with a comprehensive grasp of ANNs for predicting building energy use and to point out potential future research directions (Lu, C. et al., 2022).

### Support Vector Machines (SVM)

SVMs have shown promising results in building energy prediction challenges and have excellent generalisation capabilities. The main topic of the paper “Applying Support Vector Machines to Predict Building Energy Consumption in Tropical Region” (Dong et al., 2005) is the application of support vector machines (SVM), a novel neural network algorithm, for forecasting building energy consumption in tropical regions. The study aims to evaluate the viability and efficacy of SVM in predicting building load. As case studies, four randomly chosen commercial buildings in Singapore are examined. The input features for the case studies are monthly mean outdoor dry-bulb temperature, relative humidity, and global solar radiation. Utility invoices from landlords are gathered for model testing and development. Using a stepwise searching technique and a radial basis function (RBF) kernel, the study additionally examines the effects of two SVM parameters, C and. With coefficients of variation (CV) less than 3% and percentage error (%error) within 4%; the prediction results demonstrate remarkable accuracy.

### Random Forest (RF)

As a result of its capacity to manage intricate interconnections and nonlinear relationships between many parameters, Random Forest is a reliable method for predicting the energy consumption of buildings. It excels at identifying complex patterns in various datasets frequently used in building energy research. Random Forest's ensemble structure minimises overfitting and improves forecast accuracy. At the same time, its feature importance analysis sheds light on key factors influencing how much energy is consumed. Random Forest is a trustworthy option for precisely estimating building energy demand due to its adaptability to diverse data types and robustness against outliers.

According to (Wang, Z. et al., 2018), buildings are the most prominent global energy consumers, and there has been considerable growth in global energy consumption. Engineering-based building energy modelling provides a thorough understanding of building energy behaviours but is constrained by the challenge of obtaining complete building data for existing structures. Due to its usefulness and precision in making predictions, empirical modelling, which uses machine learning techniques, has grown in popularity. However, some algorithms employed in empirical modelling may experience instability problems, which may affect the accuracy of the findings of the predictions. Ensemble learning techniques like Random Forest (RF), which increase prediction accuracy by mixing many models, have been created to address this. The research compares RF with other methods and seeks to prove the viability of RF in short-term building energy forecasts.

### Decision Trees

Decision trees are intuitive and straightforward machine learning models. They partition the data into subsets based on attribute values, resulting in a tree-like model of decisions. These models are simple to understand and visualize, can handle both numerical and categorical data, and require relatively little data pre-processing.

In the context of building energy prediction, decision trees can provide interpretable rules indicating how different features contribute to energy consumption. However, individual decision trees can be prone to overfitting. To overcome this, ensemble methods like Random Forests and Gradient Boosting, which combine the predictions of multiple decision trees, are often used. These ensemble methods typically perform better and are more robust than individual decision trees.

The decision tree approach, which is excellent at classifying and forecasting categorical variables, was used in the study "A decision tree method for building energy demand modelling" (Yu et al., 2010) to design a building energy demand predictive model. According to the study, this strategy has an advantage over regression and ANN since it produces precise prediction models with comprehensible tree architectures. When used to analyse the energy performance of residential buildings, it achieved good accuracy (93% for training and 92% for test data), quickly identified key energy intensity factors, and delivered useful information for enhancing energy efficiency. This simple method can be used for a variety of applications in energy consumption prediction and efficiency improvement because it requires little computational expertise.

### Linear Regression

Because it is straightforward to understand, linear regression is a good approach for predicting building energy usage. Since it successfully captures linear correlations between independent variables and energy usage, it is suitable for superficial relationships. In addition, linear regression sheds light on the size and direction of feature contributions, assisting in identifying essential variables affecting energy use. It is an efficient option for early analysis. It serves as a benchmark model for more advanced algorithms thanks to its simplicity of development and ability to handle massive datasets.

(Ciulla & D’Amico, 2019) in “Building energy performance forecasting: A multiple linear regression approach” explains the difficulties in selecting the best strategies while taking complexity, user skill, and accuracy into account. The authors create a linear regression model as a workaround for the shortcomings of previous approaches. They stress the requirement for a simple, trustworthy model for a first energy assessment. The research entails building a comprehensive energy database and developing correlations between energy demand and pertinent factors using the multiple linear regression (MLR) method. The outcomes demonstrate the accuracy with which the MLR model predicts the energy requirements of buildings. The study's conclusion highlights the MLR method's generality and simplicity as well as its ability to assist users of all skill levels in making decisions related to energy planning. An overview of the study's history, goals, methods, and conclusions is provided in this section.

### Gradient Boosting

Due to its ensemble nature and capacity for capturing complex interactions, gradient boosting emerges as the perfect algorithm for developing energy consumption prediction. It excels at simulating nonlinear interactions between variables affecting energy use and reveals crucial insights about feature relevance. It provides excellent predicted accuracy, which is essential for efficient energy management. It is resilient against overfitting and has the flexibility to handle various data types. Its flexibility through parameter modification and incremental learning further strengthens its capacity to predict building energy consumption trends accurately.

In “Gradient boosting machine for modelling the energy consumption of commercial buildings” by (Touzani et al., 2018) Gradient Boosting proved advantageous in accurate savings estimations for energy efficiency projects in commercial buildings. With the availability of high-frequency interval data from advanced metering infrastructure (AMI), the algorithm capitalized on the vast dataset. This data abundance facilitated the application of advanced statistical learning models, including gradient boosting, for precise predictions of building baseline energy consumption. The method's effectiveness was demonstrated through a comparison with industry standards like piecewise linear regression and random forest algorithms, showcasing that gradient boosting consistently enhanced prediction accuracy metrics, surpassing 80% of cases and affirming its potential as a powerful tool for energy efficiency applications

Table ‑ Strengths, Weaknesses and Performance of various ML Algorithms

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Strengths | Weaknesses | Performance in Predicting Building Energy |
| Gradient Boosting | - High predictive accuracy.  - Handles complex non-linear relationships.  - Robust to outliers. | - Prone to overfitting if not tuned properly.  - Not easily interpretable.  - Computationally expensive during training. | - Good performance if hyperparameters are tuned appropriately.  - Can capture complex patterns in data.  - May struggle with large datasets due to high computational requirements. |
| Linear Regression | - Low computational cost  - Low computational cost  - Provides insights into variable importance  - Easily scalable | - Assumes a linear relationship in the data  - Sensitive to outliers  - Limited flexibility in handling complex data | - Suitable for situations where the relationship between variables is approximately linear.  - Might not capture complex non-linear patterns in the data.  - Quick and easy to implement. |
| Neural Networks | - Ability to capture complex relationships  - State-of-the-art performance in many tasks  - Good at feature learning  - Versatile for different data types | - Computationally intensive, especially in training  - Computationally intensive, especially in training  - Prone to overfitting if not properly regularized  - Lack of interpretability | - High potential for accurate predictions, especially with large datasets.  - Can handle non-linear and high-dimensional data.  - Can be time-consuming to train.  - Might need extensive hyperparameter tuning. |
| Random Forest | - High predictive accuracy  - Robust to overfitting  - Handles non-linear relationships  - Robust to outliers | - Lack of interpretability  - Slower to train than some algorithms  - Not suitable for real-time applications | - Good performance even with high-dimensional and noisy data.  - Can handle large datasets.  - Provides insights into feature importance. |
| Decision Trees | - Simple to understand and interpret  - Fast to train and predict  - Handles non-linear relationships  - Can handle both numerical and categorical data | - Prone to overfitting  - Instability with small variations in data  - Limited generalization ability | - Easy to interpret and visualize.  - Suitable for smaller datasets.  - May not be the best choice for very complex relationships in data. |

It should be noted that the specific dataset and data quality can affect how well these algorithms perform. To achieve the most excellent results when estimating building energy, it is crucial to thoroughly pre-process the data, manage missing values, and choose the suitable characteristics. Due to their proficiency in managing complex relationships and outliers, ensemble approaches such as Random Forest and Gradient Boosting frequently deliver favourable results in real-world applications for this kind of regression job. Neural networks may also perform well, mainly when there is a requirement to capture complicated patterns in vast amounts of data. When simplicity and ease of interpretation are essential considerations, linear regression and decision trees can be useful beginning points.

## Scope and Significance of Study

This study's main objective is to apply machine learning techniques to forecast building energy usage. The information utilised in this study includes a number of building characteristics that affect energy usage, including energy ratings, age, CO2 emissions, and others.

In this study, we examine and contrast how well various machine learning algorithms forecast the energy use of buildings. This comparison broadens the scope of our understanding in this area by exploring more sophisticated models in addition to more conventional linear regression ones.

The significance of this study is multi-faceted:

1. Increasing Energy Efficiency: This study can help with the larger objective of increasing energy efficiency by anticipating energy usage. These forecasts can be used by building owners, managers, and policymakers to spot energy inefficiencies and take appropriate action, which will lead to decreased energy use and running costs.
2. Informing Design and Construction: The study may provide useful information to architects, engineers, and builders as they plan and construct new structures. They can integrate architectural features and construction techniques that make the building more energy-efficient by comprehending the aspects that influence energy use.
3. Supporting Energy Planning: Accurate energy consumption forecasts can support energy planning at the building, neighbourhood, and city scales. The likelihood of energy shortages or overproduction can be decreased by using these forecasts by energy providers to better manage energy production and delivery.
4. Supporting Environmental Sustainability: This study indirectly supports environmental sustainability by encouraging energy efficiency and lowering energy usage. Fossil fuels are frequently burned in the process of producing energy, which releases greenhouse gases into the atmosphere. We can lessen these emissions and thus the effects of climate change by consuming less energy.
5. Research Advancement in Machine Learning: By offering a comparison of several algorithms, this work advances the field of machine learning. This can help guide current and future work in this area, resulting in the development of algorithms that are more precise and effective.

The literature on machine learning methods for forecasting building energy demand is thoroughly reviewed in this chapter. It explores numerous strategies, algorithms, and their uses in energy prediction, offering insights into their advantages and disadvantages. The chapter emphasises the value of precise energy forecasting for sustainable building management, energy efficiency, and the environment. It examines the benefits and downsides of machine learning methods such as neural networks, SVMs, RFMs, DTDs, linear regression, and gradient boosting. This chapter thoroughly explains the performance and applicability of different algorithms by examining studies that compare them. It also lays the groundwork for current research and provides insightful information for academics and industry professionals looking to apply machine learning for accurate building energy consumption prediction.

# Methodology

This chapter presents the research approach for comparing ML algorithms for building energy prediction. It describes the procedures for gathering, pre-processing, and preparing the building energy consumption data. The chapter also discusses the chosen machine learning (ML) algorithms and how they were set up for the experimental evaluation.

## CRISP Methodology Implementation

The CRISP methodology governs the entire research process. The CRISP process is meticulously carried out at each stage to provide a methodical and thorough investigation.

Most methodologies (both for data mining and data science process models) have developed from CRISP-DM, which can be seen as the canonical approach. The six processes are business understanding, data understanding, data preparation, modelling, evaluation, and deployment. It expands and multiplies the steps in the original KDD plan (Martinez-Plumed et al., 2021).

Understanding the problem domain and determining the study objectives are both parts of business understanding. Exploring the dataset, discovering relationships, and picking up domain knowledge are the primary goals of data understanding. Data integration, transformation, and cleaning are all included in data preparation. Machine learning algorithms must be chosen and trained before being used in modelling. Evaluation rates the effectiveness of the models, and deployment entails using the selected model in real-world settings.

To better understand systems and/or their processes, data mining (DM) refers to the process of looking for hidden patterns or relationships in data. These patterns could be used, for instance, to assess the degree of a relationship between variables or forecast future results (Gibert et al., 2018).

The six phases of the process model are scope definition, business understanding, data understanding, data preparation, modelling, evaluation, and deployment. The first stage integrates data and business insight while recognising the influence both have on project viability. Tasks like data preparation, modelling, evaluation, and deployment are covered in the following phases (Studer et al., 2021).

1. Business Understanding:

The goals and specifications of the building energy prediction task are outlined in this phase. Consideration is given to the problem's unique properties, such as its linearity, nonlinearity, high dimensionality, and interpretability requirements (Lee, I. & Shin, 2020). Real-time predictions, integration with control systems, and decision-making interpretability are all considered task requirements.

1. Data Understanding:

In this stage, the data collection procedure begins. Building attributes, meteorological information, historic energy usage information, and occupancy information are all gathered as relevant data (Roh et al., 2021). Public datasets, utility companies, or on-site sensors and monitoring systems can all be used to get data. In order to capture seasonal and long-term fluctuations, the data is thoroughly analysed to ensure its representativeness, correctness, and coverage over a considerable period.

1. Data Preparation:

Data preparation is done during this stage to convert the raw data into a format appropriate for analysis. Data cleansing, addressing missing values, removing outliers, and normalising or scaling features are all included in this (Nasrul Aziz et al., 2019). Extracting pertinent characteristics or changing the data using feature engineering approaches to increase its prediction value is possible. Predictions are reliable and accurate because the pre-processed dataset is prepared for algorithm training.

1. Modelling:

Various machine learning methods are chosen in this phase for building energy prediction. Support vector machines, decision tree-based algorithms (like random forests), ensemble methods (like gradient boosting), regression-based algorithms (like linear regression), and neural network algorithms are all taken into consideration. Using a training-test split, the algorithms are trained on the pre-processed dataset. To improve prediction performance, model parameters are adjusted using methods like grid search, random search, or Bayesian optimisation. To avoid overfitting, methods like regularisation and cross-validation are used.

1. Evaluation:

Using appropriate assessment metrics for regression tasks, the trained models are assessed in this phase. The accuracy and dependability of the model's predictions are evaluated using metrics like mean squared error (MSE), root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R-squared). Additional analysis such as residual analysis and visualisations may be carried out to understand model performance better.

1. Deployment:

Comparing the effectiveness of the trained models is the final step. The predicting skills of the models are evaluated and ranked using the assessment criteria established in the preceding phase. It is possible to use statistical tests to determine whether observed performance differences are statistically significant, such as paired t-tests or ANOVA. The comparison focuses on interpretability, resilience, accuracy, and computational efficiency measures. Based on the performance comparison findings and considering the unique needs of the work and the actual implementation of the models, the best technique for predicting building energy is supported.

## Research Design

In order to assess and compare the effectiveness of several machine learning algorithms for building energy prediction using the CRISP methodology, this study uses a comparative analysis research design. The comparative analysis design enables a systematic assessment of the algorithms, determining the best method for predicting building energy. This study sheds light on the advantages and disadvantages of various algorithms in forecasting building energy consumption.

## Data Exploration

Count (the number of non-null values), mean (the average value), std (the standard deviation, which measures variability), min (the smallest value), 25% (the 25th percentile), 50% (the median or 50th percentile), 75% (the 75th percentile), and max (the enormous value) are some of the statistics of a few of the columns included in Table 3-1.

Table ‑ Sample Description of Data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | BUILDING\_REFERENCE\_NUMBER | CURRENT\_ENERGY\_EFFICIENCY | POTENTIAL\_ENERGY\_EFFICIENCY | FIXED\_LIGHTING\_OUTLETS\_COUNT | LOW\_ENERGY\_FIXED\_LIGHT\_COUNT | UPRN |
| count | 1.499000e+03 | 1499.000000 | 1499.000000 | 649.000000 | 541.000000 | 1.485000e+03 |
| mean | 5.434052e+09 | 62.841227 | 77.260841 | 10.674884 | 5.717190 | 7.228184e+07 |
| std | 3.045439e+09 | 13.705714 | 11.005734 | 6.053058 | 5.284101 | 5.208312e+05 |
| min | 1.966018e+07 | 1.000000 | 8.000000 | 0.000000 | 0.000000 | 6.318120e+07 |
| 25% | 2.876931e+09 | 55.000000 | 73.000000 | 7.000000 | 2.000000 | 7.212900e+07 |
| 50% | 5.531148e+09 | 65.000000 | 79.000000 | 10.000000 | 5.000000 | 7.225558e+07 |
| 75% | 8.149844e+09 | 72.000000 | 85.000000 | 12.000000 | 8.000000 | 7.250052e+07 |
| max | 1.000381e+10 | 94.000000 | 117.000000 | 48.000000 | 40.000000 | 7.278870e+07 |

From these data, we may learn more about the range and distribution of values in each numerical column. With a mean of roughly 62.84 and a standard deviation of roughly 13.71, the CURRENT\_ENERGY\_EFFICIENCY values, for instance, range from 1 to 94, demonstrating a wide range of energy efficiency ratings in the dataset.

A comparison of a graph

Description automatically generated with medium confidence

Figure ‑ Energy Efficiency

1. Current Energy Efficiency: The first histogram reveals that the current energy efficiency ratings are distributed in a generally bell-shaped manner but are biased to the right, with a peak around the value of 60. This indicates that while there are quite a few buildings with ratings higher or lower than 60, most buildings in the dataset currently have ratings in the range of 60 to 70.
2. Potential Energy Efficiency: The second histogram demonstrates that the potential energy efficiency rating distribution is broadly bell-shaped but biased to the left, with a peak around 80. This means that if specific modifications were implemented, many of the buildings in the dataset could reach an energy efficiency rating of approximately 80.

A screenshot of a graph

Description automatically generated

Figure ‑ Correlation Matrix of Data

The correlation matrix, which gauges the linear correlations between pairs of numerical variables in the dataset, is represented visually in the heatmap above labelled Figure 3-2.

Each square in the heatmap represents the correlation coefficient between two variables, and its colour denotes the strength and direction of the link:

* Dark red squares represent a significant positive correlation, meaning that as one variable rises, so does the likelihood that the other will also.
* The presence of dark blue squares denotes a strong negative connection, meaning that as one variable rises, the other tends to fall.
* Lighter squares show a weaker association.

The heatmap is symmetrical along the diagonal because there is always a 1:1 correlation between a variable and itself.

We can see from the heatmap that several pairs of variables are highly associated. For instance, it is assumed that CURRENT\_ENERGY\_EFFICIENCY and POTENTIAL\_ENERGY\_EFFICIENCY is positively correlated since structures with high current energy efficiency ratings are likely to have high potential energy efficiency ratings as well.

The heatmap also has a large number of squares that are pretty light in colour, indicating that there is little association between those particular pairings of variables.

The distributions of the CURRENT\_ENERGY\_RATING and POTENTIAL\_ENERGY\_RATING columns in the dataset are displayed in the bar plots below in Figure 3-3:

A comparison of energy ratings

Description automatically generated

Figure ‑ Distribution of Current and Potential Energy Ratings

1. Current Energy Ratings: According to the first bar plot, the most prevalent current energy rating is "D," which is followed by "E" and "C." Few structures currently have an "A" or "B" energy rating.
2. Potential Energy Ratings: According to the second bar plot, the potential energy rating of "B" is the most prevalent, followed by "C" and "A." This shows that many of the buildings in the dataset could obtain a higher energy rating with specific adjustments.

Let us move on to the TENURE column, which indicates whether a building is owned by its current owner or is rented out. To illustrate the percentage of buildings in each group, we will use a pie chart which can be seen in Figure 3-4 below.

A pie chart with different colored circles

Description automatically generated

Figure ‑ Distribution of Tenure/Building Occupancy

* About 47.0% of all buildings are occupied by their owners.
* Private rentals make up about 25.0% of all properties.
* 21.1% of all buildings are rented out on a social basis.

This provides information about the buildings' ownership status in the dataset. However, the precise ratios could change based on the particulars of the dataset.

Our examination of the exploratory data is now complete. Our comprehension of the data has improved as a result of looking at the distributions, summary statistics, and correlations of the dataset's variables. We have also discovered some patterns and relationships that may be relevant for future research or modelling.

### Data Collection

“Data collection is the process of gathering data for use in business decision-making, strategic planning, research and other purposes” (Stedman & McLaughlin, 2022). Relevant dataset for predicting building energy have been obtained for this work from trusted sources. A wide range of elements, including historical energy consumption data, building parameters (such as area and occupancy), and other pertinent variables, are present in the chosen dataset. The dataset is selected to guarantee their fit with the research's goals and the availability of the data needed for precise energy forecasting.

The dataset was obtained from the uk.gov website and is publicly available, the data collected is related to buildings found in Leeds. Table 3-2 gives a description of each of the columns found in the dataset.

Table ‑ Description of Columns in Dataset

|  |  |
| --- | --- |
| Column Header | Description |
| LMK\_KEY | Unique identifier for each building record |
| ADDRESS1 | First line of the building's address |
| ADDRESS2 | Second line of the building's address |
| ADDRESS3 | Third line of the building's address |
| POSTCODE | Postal code of the building's address |
| BUILDING\_REFERENCE\_NUMBER | Reference number assigned to the building |
| CURRENT\_ENERGY\_RATING | Current energy rating of the building |
| POTENTIAL\_ENERGY\_RATING | Potential energy rating of the building |
| CURRENT\_ENERGY\_EFFICIENCY | Current energy efficiency of the building |
| POTENTIAL\_ENERGY\_EFFICIENCY | Potential energy efficiency of the building |
| PROPERTY\_TYPE | Type of property |
| BUILT\_FORM | Form of construction of the building |
| INSPECTION\_DATE | Date when the building was inspected for energy assessment |
| LOCAL\_AUTHORITY | Local authority responsible for the building's location |
| CONSTITUENCY | Constituency where the building is located |
| COUNTY | County where the building is located |
| LODGEMENT\_DATE | Date when the building's energy assessment was lodged |
| TRANSACTION\_TYPE | Type of transaction associated with the building |
| ENVIRONMENT\_IMPACT\_CURRENT | Current environmental impact rating of the building |
| ENVIRONMENT\_IMPACT\_POTENTIAL | Potential environmental impact rating of the building |
| ENERGY\_CONSUMPTION\_CURRENT | Current energy consumption of the building |
| ENERGY\_CONSUMPTION\_POTENTIAL | Potential energy consumption of the building |
| CO2\_EMISSIONS\_CURRENT | Current CO2 emissions of the building |
| CO2\_EMISS\_CURR\_PER\_FLOOR\_AREA | Current CO2 emissions per floor area of the building |
| CO2\_EMISSIONS\_POTENTIAL | Potential CO2 emissions of the building |
| LIGHTING\_COST\_CURRENT | Current lighting cost of the building |
| LIGHTING\_COST\_POTENTIAL | Potential lighting cost of the building |
| HEATING\_COST\_CURRENT | Current heating cost of the building |
| HEATING\_COST\_POTENTIAL | Potential heating cost of the building |
| HOT\_WATER\_COST\_CURRENT | Current hot water cost of the building |
| HOT\_WATER\_COST\_POTENTIAL | Potential hot water cost of the building |
| TOTAL\_FLOOR\_AREA | Total floor area of the building |
| ENERGY\_TARIFF | Energy tariff associated with the building |
| MAINS\_GAS\_FLAG | Indicates whether the building has mains gas |
| FLOOR\_LEVEL | Level of the building's floor |
| FLAT\_TOP\_STOREY | Indicates whether the building has a flat top storey |
| FLAT\_STOREY\_COUNT | Number of storeys in the building |
| MAIN\_HEATING\_CONTROLS | Type of main heating controls in the building |
| MULTI\_GLAZE\_PROPORTION | Proportion of multi-glazed windows in the building |
| GLAZED\_TYPE | Type of glazing in the building |
| GLAZED\_AREA | Area covered by glazed windows in the building |
| EXTENSION\_COUNT | Number of extensions in the building |
| NUMBER\_HABITABLE\_ROOMS | Number of habitable rooms in the building |
| NUMBER\_HEATED\_ROOMS | Number of heated rooms in the building |
| LOW\_ENERGY\_LIGHTING | Indicates whether the building has low-energy lighting |
| NUMBER\_OPEN\_FIREPLACES | Number of open fireplaces in the building |
| HOTWATER\_DESCRIPTION | Description of the hot water system in the building |
| HOT\_WATER\_ENERGY\_EFF | Energy efficiency of the hot water system |
| HOT\_WATER\_ENV\_EFF | Environmental efficiency of the hot water system |
| FLOOR\_DESCRIPTION | Description of the building's floor |
| FLOOR\_ENERGY\_EFF | Energy efficiency of the building's floor |
| FLOOR\_ENV\_EFF | Environmental efficiency of the building's floor |
| WINDOWS\_DESCRIPTION | Description of the windows in the building |
| WINDOWS\_ENERGY\_EFF | Energy efficiency of the windows |
| WINDOWS\_ENV\_EFF | Environmental efficiency of the windows |
| WALLS\_DESCRIPTION | Description of the walls in the building |
| WALLS\_ENERGY\_EFF | Energy efficiency of the walls |
| WALLS\_ENV\_EFF | Environmental efficiency of the walls |
| SECONDHEAT\_DESCRIPTION | Description of the secondary heating system in the building |
| SHEATING\_ENERGY\_EFF | Energy efficiency of the secondary heating system |
| SHEATING\_ENV\_EFF | Environmental efficiency of the secondary heating system |
| ROOF\_DESCRIPTION | Description of the building's roof |
| ROOF\_ENERGY\_EFF | Energy efficiency of the roof |
| ROOF\_ENV\_EFF | Environmental efficiency of the roof |
| MAINHEAT\_DESCRIPTION | Description of the main heating system in the building |
| MAINHEAT\_ENERGY\_EFF | Energy efficiency of the main heating system |
| MAINHEAT\_ENV\_EFF | Environmental efficiency of the main heating system |
| MAINHEATCONT\_DESCRIPTION | Description of the main heating controls in the building |
| MAINHEATC\_ENERGY\_EFF | Energy efficiency of the main heating controls |
| MAINHEATC\_ENV\_EFF | Environmental efficiency of the main heating controls |
| LIGHTING\_DESCRIPTION | Description of the lighting system in the building |
| LIGHTING\_ENERGY\_EFF | Energy efficiency of the lighting system |
| LIGHTING\_ENV\_EFF | Environmental efficiency of the lighting system |
| MAIN\_FUEL | Main fuel used in the building |
| WIND\_TURBINE\_COUNT | Number of wind turbines installed in the building |
| HEAT\_LOSS\_CORRIDOR | Indicates whether the building has a heat loss corridor |
| UNHEATED\_CORRIDOR\_LENGTH | Length of unheated corridor in the building |
| FLOOR\_HEIGHT | Height of the building's floor |
| PHOTO\_SUPPLY | Indicates whether the building has photo supply |
| SOLAR\_WATER\_HEATING\_FLAG | Indicates whether the building has solar water heating |
| MECHANICAL\_VENTILATION | Indicates whether the building has mechanical ventilation |
| ADDRESS | Complete address of the building |
| LOCAL\_AUTHORITY\_LABEL | Label for the local authority responsible for the building |
| CONSTITUENCY\_LABEL | Label for the constituency where the building is located |
| POSTTOWN | Town associated with the building's address |
| CONSTRUCTION\_AGE\_BAND | Age band of the building's construction |
| LODGEMENT\_DATETIME | Date and time when the building's energy assessment was lodged |
| TENURE | Tenure type of the building |
| FIXED\_LIGHTING\_OUTLETS\_COUNT | Number of fixed lighting outlets in the building |
| LOW\_ENERGY\_FIXED\_LIGHT\_COUNT | Number of low-energy fixed lights in the building |
| UPRN | Unique Property Reference Number (UPRN) of the building |
| UPRN\_SOURCE | Source of the UPRN |

### Data Pre-processing

Several pre-processing steps are carried out to guarantee data uniformity and quality. In order to eliminate any missing values, outliers, or inconsistent entries, data cleaning procedures are first used. Any required data transformations, such as scaling or normalisation, are executed to achieve consistency in the dataset. Feature extraction and selection are also made during pre-processing to determine the most pertinent features for predicting building energy. This process is essential for enhancing the effectiveness of machine learning algorithms.

### Algorithm Selection

For the comparison analysis, various machine learning algorithms that are appropriate for predicting building energy are used. These algorithms include Gradient Boosting, Random Forest, Linear Regression, Support Vector Machines, and Neural Networks. The decision is made based on the algorithms' applicability to the prediction of building energy, their standing in the industry, and other studies that have successfully used them and returned positive results.

## Performance Metrics

Several performance indicators are available to assess and contrast the performance of the machine learning algorithms. For this research the main performance evaluation metrics used are the Root Mean Square Error and Mean Absolute Error. Accuracy, precision, recall, F1 score, and computing efficiency are some of the other criteria available. Precision and recall give information about how well the algorithm can categorise different energy usage levels, while accuracy indicates how accurate the forecasts are overall. The F1 score provides a fair measurement by combining recall and precision. The time and resources needed for model training and prediction are considered in computational efficiency.

## Experimental Setup

The experiments are conducted using a computer system with appropriate hardware and software configurations. The machine learning algorithms are implemented using widely used libraries and frameworks such as scikit-learn and pandas. The training and evaluation processes are carried out using a cross-validation approach to ensure robustness and minimize bias. The hyperparameters of the algorithms are tuned using techniques such as grid search or random search to optimize their performance. The experiments are performed multiple times to account for any randomness and ensure reliable results.

## Ethical Considerations

1. **Data security and privacy:** The dataset may include private information about particular properties or buildings, such as addresses and building identification numbers. This data must be handled to respect privacy rights and comply with data protection regulations. Always ensure that data is sent and kept securely; where it is practical, think about pseudonymizing or anonymizing personally identifiable information.
2. **Fairness and Bias:** Machine learning algorithms can unintentionally reinforce or even exacerbate pre-existing biases in the data. For instance, the models might inaccurately anticipate higher energy consumption for some populations if energy efficiency ratings have historically been lower in particular geographic areas or for particular types of houses. Identifying and correcting any biases in the data and model projections is critical.
3. **Transparency and Explainability:** Complex machine learning models can be opaque and challenging to decipher. Because of this, it may be challenging for stakeholders to comprehend how predictions are created, which may result in mistrust or inappropriate use of the model. In order to achieve model transparency and explainability, approaches like feature significance and SHAP values can be used to reveal the model's inner workings.
4. **Accountability and Responsibility:** Choices based on the model's predictions may affect the real world, such as spending money to increase energy efficiency or deciding how much energy to use. It is crucial to set responsibilities for the model's predictions and any actions performed as a result of them.
5. Energy efficiency and consumption are intimately related to sustainability and climate change. While forecasting and improving building energy efficiency is typically a good thing for the environment, it is vital to think about any unforeseen repercussions that can arise. For instance, may emphasising energy efficiency result in abandoning other crucial sustainable building design elements?

These are only some of the possible ethical issues. The dataset details, the model's intended application, the legal and cultural setting, and other elements can all influence the specific problems and suitable actions.

The methodical comparison of machine learning methods for predicting building energy is described in this chapter. The CRISP technique includes modelling, assessment, deployment, business knowledge, data comprehension, data preparation, and modelling. The choice and configuration of machine learning methods like Gradient Boosting, Random Forest, Linear Regression, Support Vector Machines, and Neural Networks are covered in depth. Additionally, the chapter emphasises the significance of experimental design, performance indicators, and ethical issues. The extensive data gathering, pre-processing, and exploration provide an understanding of the steps to produce accurate forecast models for building energy consumption. This chapter is a thorough manual that gives readers a clear road map for the research strategy and how to implement it.

# Product/Research Design and Implementation

This chapter focuses on the product/research design and implementation process for building energy prediction using machine learning (ML) models. Here the data is loaded, and various pre-processing activities are carried out to prepare the data for model building. Building energy prediction is critical to energy management to achieve energy efficiency, reduce carbon emissions, and optimise energy utilisation. ML algorithms have shown great promise in this domain, providing valuable insights, and enabling data-driven decision-making for stakeholders.

Figure 4-1 shows a visual representation of the implementation process, used to develop the various models for evaluation.

A screenshot of a phone

Description automatically generated with low confidence

Figure ‑ Flowchart of ML Model Implementation

Here is a short explanation for each stage in the machine learning model development flowchart which is used as a guide for implementation later in section 4.3 of this chapter:

1. Define the Problem: In this stage I clearly articulate and understand the problem to solve using machine learning. This involves identifying the goals, desired outcomes, and tasks the model will address.
2. Collect and Pre-process Data: Here relevant data is gathered from relevant and trusted sources to ensure its integrity. Pre-process the data by cleaning, transforming, and encoding it into a suitable format for machine learning algorithms.
3. Split Data into Training and Test Sets: Divide the data into two separate sets: the training set, used to train the model, and the test set, used to evaluate the model's performance on unseen data.
4. Choose a Machine Learning Algorithm: Select an appropriate machine learning algorithm based on the nature of the problem, data characteristics, and desired outcomes. Consider classification, regression, clustering, or other specialized algorithms.
5. Define Model Architecture and Parameters: Here the I design the architecture of the machine learning model, including the number and type of layers, activation functions, and other specific configurations. Set the model's hyperparameters, such as learning rate, batch size, and regularization parameters.
6. Train the model using Training Data: Feed the training data into the model and optimize the model's parameters through an iterative process, such as gradient descent. This step involves adjusting the model's weights and biases to minimize the error between predicted and actual outcomes.
7. Evaluate Model Performance using Test Data: To assess the model evaluation metrics such as RMSE and MAE are used to determine how well the models perform in their predictions.
8. Decision (Performance Satisfactory?): Here based on the scores obtained from the evaluation of the models a decision is made on whether to retrain the model or keep the model as is.
9. Modify Model Architecture or Parameters: If the model's performance is unsatisfactory, the model architecture or hyperparameters is adjusted. This may involve changing the number of layers, activation functions, regularization techniques, or other model configurations.
10. Retrain the Model: Re-run the training process with the modified model architecture or parameters using the training data. Continue iterating and refining the model until the desired performance is achieved.
11. Deploy the Model for Predictions: Once the model meets the performance criteria, deploy it to make predictions on new, unseen data. This step involves integrating the model into a production environment, such as a web application or an automated system.
12. Monitor and Update the Model as Needed: Continuously monitor the model's performance in real-world scenarios and gather feedback. Update the model periodically by retraining it on new data or making necessary adjustments to ensure its accuracy and relevance over time.
13. End: This the flowchart's final stage indicates the completion of the machine learning model development process.

These stages represent the sequential steps in developing a machine learning model, from problem definition to model deployment and ongoing monitoring.

## Software and Hardware Infrastructure

The software and hardware infrastructure utilised for this research using machine learning with an emphasis on predicting building energy is described in this part. Predictive models' effectiveness and accuracy primarily depend on the processing power of the underlying software tools and the hardware configuration employed. The successful installation and assessment of machine-learning algorithms for predicting building energy use depends on a properly prepared environment.

### Software Framework

A robust software framework that supports several machine-learning libraries, data pre-processing tools, and visualisation utilities was necessary to construct prediction models for building energy usage. Python's broad ecosystem and widespread use in the data science community led to its selection as the principal programming language. The following computer programs were used:

1. Python: Python served as the core programming language for implementing machine-learning algorithms, data manipulation, and visualization tasks. The availability of libraries such as NumPy, pandas, and Matplotlib facilitated efficient data handling and exploration.
2. Scikit-Learn: Scikit-Learn, a widely used machine-learning library in Python, provided an array of tools for classification, regression, and model evaluation. It streamlined the implementation of decision trees, random forests, support vector machines, and neural networks for building energy prediction.
3. Jupyter Notebooks: Jupyter Notebooks were utilized for code development, data visualization, and result analysis. These interactive notebooks allowed for seamless integration of code, visualizations, and textual explanations, enhancing the reproducibility and interpretability of the project.

### Hardware Configuration

The successful execution of machine-learning algorithms for building energy prediction demands a capable hardware setup to handle the computational complexity and large datasets involved. The following hardware configuration was utilized for the project:

1. CPU and RAM: A workstation equipped with a multi-core CPU (Central Processing Unit) and ample RAM (Random Access Memory) ensured efficient processing and manipulation of datasets during training and evaluation of machine-learning models. Specific numbers are 16 core CPU and 32 gigabytes of RAM.
2. Storage: Sufficient storage capacity was necessary to accommodate large datasets, trained models, and intermediate results. The workstation used was equipped with an SSD (Solid State Drive) for faster read/write operations and with a size of 1 Terabyte (TB).

## Challenges and Limitations

This section of the chapter explores the difficulties and constraints of estimating building energy using machine learning methods. While the significance of precise energy consumption prediction and the research’s goals were highlighted in the previous chapter, here we will focus on the challenges that were faced when implementing such techniques.

Producing trustworthy and beneficial outcomes depends on recognising and overcoming issues. This section offers a realistic view of the difficulties of estimating building energy use using machine learning by understanding the constraints and potential hazards. These revelations increase the research's transparency and provide direction for creating successful mitigation solutions.

Throughout this chapter, we will delve into the complexities of data collection, model complexity, interpretability, overfitting, and generalisation. Each of these difficulties can affect the precision and viability of energy forecast models. We'll also go into the effects of using real-world data, which is frequently noisy, incomplete, or prone to unforeseen changes.

This chapter prepares the reader for the upcoming portions of the dissertation by addressing these difficulties head-on. It emphasises the need for methodological rigour and careful thought when using machine learning to predict building energy. Readers will get a sophisticated grasp of the potential obstacles and complexities that must be negotiated to develop accurate and effective energy projections through a thorough review of problems and constraints.

### Data Availability and Quality

For precise prediction, extensive and high-quality building energy datasets must be available. However, building energy forecast models is difficult due to data's scarce availability and data quality problems.

### Model Complexity and Interpretability

Given the complexity of some machine learning algorithms, such as ANN, it can be challenging to understand and interpret the underlying prediction mechanisms. In practice, interpreting models is frequently favoured for obtaining understanding and establishing credibility.

### Algorithm Selection and Hyperparameter Tuning

Which machine learning technique is most appropriate for a particular building energy prediction task depends on various factors, including dataset features, computational requirements, and forecast accuracy. Correct hyperparameter tweaking is necessary to optimise algorithm performance.

### Generalisation and Transferability

When used on new structures or in other geographic regions, machine learning models trained on specific datasets may encounter difficulties. The actual implementation of models needs to guarantee their generalizability and transferability.

## Code Implementation

This section delves into the practical realisation of the machine-learning models and algorithms discussed in earlier chapters. This segment showcases the translation of theoretical concepts and methodologies into tangible code, elucidating the step-by-step process of preparing data, configuring models, training, and evaluating their performance. This section bridges theory and application, shedding light on the technical intricacies and practical challenges encountered during the project's execution. This section offers insights into the complexities of the machine-learning models' deployment for accurate energy consumption forecasting through illustrative code snippets, explanations, and visualisations.

### Data Cleaning: Handling Missing Values

The first step in this study was to explore the data to understand the state it is in and decide on what needs to be done to prepare the data for modelling.

In Figure 4-2 below the dataset is loaded using the Pandas library and an overview of the first few rows of data is shown in figure 4-2.

A black screen with text on it

Description automatically generated

Figure ‑ Load dataset into Dataframe using Pandas

In figure 4-3 it is observed that the dataset is reasonably vast, with 92 columns. Looking at these columns to see what kind of information they hold; we can determine which columns we should ignore, and which may be relevant for predicting building energy. For example, some columns may have IDs or other data exclusive to each row and will not be helpful for our prediction purpose.

A computer screen shot of a black screen

Description automatically generated

Figure ‑ Output of first 5 rows from dataset

Several columns in the dataset have a high percentage of missing values as shown in Figure 4-4. Completely absent are columns like COUNTY, SHEATING\_ENV\_EFF, SHEATING\_ENERGY\_EFF, FLOOR\_ENV\_EFF and ADDRESS3, therefore removed as they will not have any positive impact on the modelling process.

The columns with more than 50% missing values were removed because the models will not likely benefit from their contents.

Identifier columns, such as "LMK\_KEY," "BUILDING\_REFERENCE\_NUMBER," "UPRN," "ADDRESS1," "ADDRESS2," "ADDRESS3," and "POSTCODE," were also removed because they are specific to each building and will not provide any generalizable data for our models.

Next is choosing an acceptable handling method for the remaining missing information. This entails using a more straightforward approach, such as the mean or median of the column, to fill in any missing values or using a more complicated technique, such as regression or K-Nearest Neighbours, to forecast the missing values based on the other columns. The type of data and the column typically determine the best course of action.

A computer screen with white text

Description automatically generated

Figure ‑ Percentage of missing values for Columns

In Figure 4-5 we find all the columns with null values and sum it up and find the percentage of these null values against all the number of rows in the dataframe, we then go ahead to update the dataframe dropping all the columns with more than 50% missing.

A computer screen shot of a black screen

Description automatically generated

Figure ‑ Drop columns with 50% or more missing

As part of our data cleaning, let us carry out some other operations on the remaining missing values in the dataset. We will look at the remaining columns with missing values to determine the most effective course of action for each.

We can substitute the column's median for missing values in numerical columns and substitute the most popular category for missing values in categorical columns, as shown in Figure 4-6. This is a straightforward and frequently successful way for addressing missing values, but it is important to take note that for more complex values more advanced techniques such as K-Nearest Neighbour are available but since the values for our dataset is not too diverse and straight forward this technique will suffice.

A computer screen with text

Description automatically generated

Figure ‑ Fill in columns with missing values

After the successful handling of missing data value, the missing values in categorical columns were replaced with the most prevalent category and in numerical columns with the column's median leaving the dataset with no missing values that could negatively affect out models.

### Data Sampling

When working with huge datasets, sampling data has a number of benefits, including increased effectiveness and reduced resource usage, as well as quicker model training and simpler troubleshooting. Large datasets can place a strain on the system, especially when utilising StandardScaler with one-hot encoding, which can cause problems with memory, computation, and scalability. These problems are lessened through sampling, which also makes model development processes more controllable and effective.

Here is some further insight to the strain on the system using large datasets in the modelling of predictive models:

1. Memory Consumption: Large datasets use a lot of memory, which could result in memory overflow problems, slow down operations, or even cause the system to crash. This is particularly problematic when carrying out memory-intensive operations like one-hot encoding, which converts the dataset into a large binary matrix.
2. Computation Time: When working with huge datasets, operations like one-hot encoding require repeatedly iterating through the complete collection of data. This hinders production by delaying model training and other procedures.
3. Scaling Issues: Large datasets put a burden on the scalability of algorithms and preparation procedures. For each feature, StandardScaler, for example, computes the mean and standard deviation, which is computationally intensive for large datasets.
4. I/O bottlenecks: This can be caused by loading and storing big datasets from storage devices, which can impede data access and preprocessing operations.
5. Challenges of Parallelization: Parallelizing processes on huge datasets can be difficult and require optimised hardware, which isn't always available.

Here are also the advantages of sampling, which are as follows:

1. Enhanced Efficiency: Because they require more processing and memory, large datasets can make model inference and training take longer. Faster experimentation, model iteration, and testing are possible thanks to sampling a portion of the data. When experimenting with various methods, hyperparameters, or model designs, this is especially helpful.
2. Reduced Resource Consumption: Processing and memory are two major computing resources needed when working with massive datasets. If you're working on a system with limited resources, sampling can be extremely helpful because it lowers the demand on these resources and allows you to work with smaller batches of data.
3. Faster Model Training: Training a model often takes less time when using a smaller sample of data as opposed to the complete dataset. This is especially helpful when you need to quickly evaluate the model's performance and make the necessary adjustments throughout the development phase.
4. Easier Debugging and Testing: Debugging and testing code can be made simpler by using smaller datasets, which are also simpler to handle. This enables you to spot problems and make changes more quickly.
5. Mitigating Overfitting: Overfitting is when a model learns noise from the data rather than useful patterns, which can be caused by large datasets. This risk can be reduced by using a smaller sample by simplifying the model and enhancing generalisation.
6. Exploratory Data Analysis (EDA): Sampling can speed up exploratory data analysis (EDA), allowing you to learn more about the traits and connections in the dataset more quickly. This can direct the choice of features and the pre-processing procedure.

In this research the decision to sample data from the dataset was taken. This is as a result of the code run timing out or crashing due to the large size of the data especially after scaling the data which is addressed in the next sub-section. A sample size of 2% of which the code can be found in figure 4-7 below. This sample size was arrived at by testing various percentages of the original dataset, and this value was the one with the most favourable outcome.

A computer screen shot of a black rectangular with green and white text

Description automatically generated

Figure ‑ Data Sampling

### Feature Engineering: Data Scaling

Now to get the data ready for modelling. This comprises:

1. Encoding categorical variables: Many machine learning methods demand numerical input data. As a result, we must transform categorical variables into a format that the algorithms can use. In order to accomplish this, we will use one-hot encoding, which generates additional binary columns for each category or label contained in the original column.
2. Scaling of features: The scale of the features affects how some machine learning algorithms perform. So that each characteristic has a similar range, the data must be scaled. To do this, the StandardScaler from sklearn is utilized, which scales the features to have a mean of 0 and a variance of 1.
3. Dividing the data into training and test sets: To assess how well the models perform on new data, the dataset is divided into a training set and a test set so that the models may be trained and tested separately.

In Figure 4-8, the one-hot encoding has been successfully applied to the dataset. After one-hot encoding, feature scalabiling was then carried out. Using the OneHotEncoder to encode the various categorical columns and StandardScaler from sklearn, which standardises features by eliminating the mean and scaling to unit variance, for feature scaling.

A screen shot of a computer program

Description automatically generated

Figure ‑ Data Encoding and Scaling

We will divide the data using the sklearn train\_test\_split function, leaving 20% of the data for testing.

Our sample has 1450 features after one-hot encoding, scaling, and other operations.

The mutual information score for each feature concerning the target variable is then determined. For this, the use of the sklearn mutual\_info\_regression function was implemented. After computing the scores, the top 30 features with the highest scores were chosen.

It should be noted that mutual information measures the dependency between two variables. Greater levels indicate greater dependence. Therefore, the features that depend most heavily on the target variable are chosen based on those with the highest mutual information scores. Figure 4-9 shows the code used to perform these operations.

A screen shot of a computer program

Description automatically generated

Figure ‑ Feature Selection

A crucial stage of this research's development took place within the confines of this chapter, where theoretical concepts and practical application of knowledge converged. The essence of this chapter was to go beyond the theoretical bounds and delve into the actual orchestration of code implementation that gave the theoretical framework life. This chapter clarified the precise plan that crystallised theoretical ideas into practical reality as the centre of synergy between insight and action.

This chapter set out on an instructive journey into the crux of data preparation. It shed light on the complex network of choices that supported the transformation of unprocessed data into a finely woven tapestry, weaved to mesh with the complexities of prediction algorithms. It gave academics and practitioners a compass by outlining each careful step.

# Research Outcomes/Results/Discussion and Evaluation

The findings of the study are discussed in this chapter, along with the effects of applying several machine learning models to the problem of predicting building energy use. It serves as the central theme of our research, tying together the threads of the data analysis, machine learning, and research methodology we used.

This chapter aims to give readers a thorough grasp of the effectiveness and ramifications of the machine-learning models. We will go into the experimental findings, assessing each model's performance objectively and discussing the emerging trends and insights. These results will be assessed based on our original research goals and queries.

## Evaluation Metrics

To compare the performance of different machine learning algorithms, various evaluation metrics are available, but our focus will be on mean absolute error (MAE) and root mean square error (RMSE) which have been widely employed for regression-based models such as these.

In statistics and machine learning, the Mean Absolute Error (MAE) metric is frequently used to assess the effectiveness of a prediction model, notably in regression assignments. It calculates the average absolute difference between the actual observed values in the dataset and the anticipated values produced by the model. The MAE is determined mathematically as follows:

MAE = (1/n) \* Σ|i=1 to n| |Yi - Ŷi|

Where:

* The total number of data points is n.
* Yi denotes the actual value seen at the i-th data point.
* The projected value for the i-th data point generated by the model is represented by Ŷi.
* Differences are guaranteed to be positive since || represents the absolute value.

Another often used metric in statistics and machine learning to assess the effectiveness of prediction models, particularly in regression tasks, is root mean square error (RMSE). It establishes the typical size of the discrepancies between the actual observed values in the dataset and the model's anticipated values. Although Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are closely linked, RMSE offers a more comprehensible evaluation by calculating the square root of MSE. The following is the RMSE mathematical formula:

RMSE = √((1/n) \* Σ(i=1 to n) (Yi - Ŷi)^2)

Where:

* The total number of data points is n.
* Yi denotes the actual value seen at the i-th data point.
* The predicted value for the i-th data point generated by the model is represented by Ŷi.

Using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to evaluate models for predicting building energy consumption is a practical and insightful approach for several reasons:

* Interpretability: In a real-world setting, both RMSE and MAE provide measures that are directly related to the energy consumption prediction errors. These indicators are helpful for communicating with stakeholders and making decisions since they are simple for stakeholders, such building managers, and owners, to grasp.
* Sensitivity to Errors: RMSE and MAE penalise greater errors more severely because they are sensitive to prediction errors. Large prediction errors in the context of building energy prediction can result in significant cost and energy loss, making the elimination of such mistakes a top priority.
* Robustness to Outliers: Because RMSE squares the errors, it is more susceptible to outliers. This may be advantageous when outliers indicate serious problems, such as abrupt increases in energy use or equipment failure. When you need a more reliable measure of error, MAE, on the other hand, may be selected because it is less impacted by outliers.
* Model Comparison: Using RMSE and MAE, a straightforward model comparison is possible. Using the corresponding error measures, the precision of each model's forecast of building energy use may be contrasted. For the selection and creation of models, this is essential.
* Bias Assessment: MAE provides a clear measure of the average prediction bias, helping you identify whether your model tends to systematically underpredict or overpredict energy consumption. Understanding bias is crucial for making informed adjustments and improvements to the predictive model.
* Practical Utility: RMSE and MAE are commonly used in the field of building energy prediction. This means that stakeholders and industry professionals are familiar with these metrics, making it easier to communicate model performance and results within the context of building energy management.
* Model Monitoring: After deploying a predictive model for building energy consumption, ongoing monitoring is essential. RMSE and MAE can be used to assess model performance over time and trigger maintenance or updates when performance deteriorates beyond acceptable thresholds.

It's crucial to bear in mind that while evaluating building energy prediction models, there are other metrics to consider than RMSE and MAE. Additional metrics, including coefficient of determination (R-squared), may provide more details on the model's efficacy. Additionally, they should be driven by domain-specific knowledge and specific project goals to ensure that the assessment criteria are in line with the objectives of the building energy prediction assignment.

Based on the provided root mean square error (RMSE) and mean absolute error (MAE) values in Table 4-5, we can compare and discuss the performance of Linear Regression, Decision Tree, Gradient Boosting and Random Forest models.

Table 4‑5 RMSE and MAE values of Models

|  |  |  |
| --- | --- | --- |
| Model | Sample Data RMSE | Sample Data MAE |
| Linear Regression | 16.518992 | 9.939798 |
| Decision Tree | 20.214058 | 11.166551 |
| Random Forest | 15.863960 | 8.779496 |
| Gradient Boosting | 16.354889 | 10.166565 |

1. Linear Regression: In the case of linear regression, it seeks to match a linear relationship between the input data and the desired variable. The RMSE and MAE values represent the model's average prediction error. In this instance, the model's RMSE of 16.518992 shows that the predictions are, on average, 16.52 units off the actual data. Similarly, the MAE of 9.939798 shows that the forecasts differ on average by 9.94 units.
2. Decision Tree: Based on the input attributes, the Decision Tree model creates a tree-like structure of decisions. The forecasts appear to have a greater average error than Linear Regression, according to the RMSE of 20.214058. Similar to this, the MAE of 11.166551 shows that the forecasts vary by 11.17 units on average. Decision trees have a tendency to overfit the training set, which increases error on unobserved data.
3. Random Forest: An ensemble model called Random Forest combines various decision trees to enhance generalisation. In general, Random Forest offers more accurate predictions than both Linear Regression and Decision Tree, as evidenced by the lower RMSE of 15.863960. The lower MAE of 8.779496 shows that the average forecast deviation is likewise lower.
4. Gradient Boosting: Gradient Boosting is another ensemble method that systematically creates numerous models, with each model aiming to fix the flaws of the one before it. Despite having an RMSE of 16.354889, higher than Random Forest's, it is still better than the Decision Tree. The average variance of the model is approximately 10.17 units, according to the MAE of 10.16656.

The fact that these findings are based on a smaller sample of the data and a subset of the attributes must be noted. It can be advantageous to train the models on the entire dataset and/or use more features for a more accurate comparison, depending on the computational resources available.

The selection of hyperparameters can also have an impact on model performance.

In conclusion, even though the Random Forest model performed the best in our investigation, it is crucial to consider each model's assumptions and constraints, the type of data being used, and the precise specifications of the task when selecting a model for a machine learning project.

The bar graphs above in Figures 5-1 and 5-2 visually represents the MAE and RMSE respectively, the values of the four models. The length of the bars indicates the relative magnitude of the values for each model.

A graph of blue rectangular bars

Description automatically generated

Figure ‑ Model MAE Comparison

A graph of blue rectangular bars

Description automatically generated

Figure ‑ Model RMSE Comparison

In summary, among the models, Random Forest has the lowest RMSE and MAE, which offers the best overall predictions for this dataset. Following closely behind with relatively low RMSE and MAE values is gradient boosting. Although it has a moderate RMSE, linear regression has the greatest MAE. Due to overfitting, Decision Tree may not be doing as well as the other models because it has the highest RMSE and MAE.

Remember that while RMSE and MAE offer some insight into the model's prediction ability, they need to reveal the whole picture. To choose the best model and enhance it, more research is required, such as cross-validation and understanding the context of the data.

# Project Management

## Project Overview

This chapter outlines the project management approach for implementing the comparative analysis of machine learning algorithms for building energy prediction. It provides an overview of the project's scope, objectives, and deliverables. The project will follow a structured and systematic approach to ensure efficient execution and successful completion.

## Project Phases

The project will be divided into several phases to facilitate better planning, execution, and monitoring. The following phases will be undertaken:

Project Initiation:

* Establish the goals, parameters, and deliverables for the project.
* Identify the crucial parties and create efficient communication channels.
* Create the project team and distribute roles and duties.

Literature Review and Research:

* Conduct a thorough literature review to assemble pertinent data and insights.
* Use machine learning methods to find previous studies on building energy prediction.
* To create a strong foundation for the research, analyse and synthesise the results.

Data Acquisition and Pre-processing:

* Verify and gather the required datasets for predicting building energy.
* Prepare the collected data by cleaning, transforming, and extracting features.
* Verify the accuracy of the data and its suitability for the chosen machine learning techniques.

Algorithm Selection and Implementation:

* Analyse various machine learning techniques that are appropriate for predicting building energy.
* Choose the algorithms to be compared in the analysis.
* Utilise the proper tools and programming languages to implement the chosen algorithms.

Model Training and Evaluation:

* Utilise the provided datasets to train the machine learning models.
* To enhance the performance of the models, fine-tune their hyperparameters.
* Utilise the right metrics and validation methods to assess the models' performance.

Comparative Analysis and Results:

* The effectiveness of the various machine learning algorithms can be compared.
* Draw insightful conclusions from the analysis of the results.
* Find the building energy forecast algorithms with the best performance.

Discussion, Conclusion, and Recommendations:

* Discuss the results' implications and how they relate to the study's goals.
* Write a summary of the comparative analysis's main conclusions and contributions.
* Offer suggestions for additional study and beneficial uses.

## Project Timeline

To effectively manage the project, the following approach was used:

Month 1: Project Initiation and Data Collection

* Define research objectives and scope.
* Select and refine the research topic.
* Create a project plan and timeline.
* Begin a literature review.
* Gather relevant datasets for building energy prediction.

Month 2: Data Pre-processing and Algorithm Selection

* Explore and clean the collected data.
* Define relevant features.
* Research and select machine learning algorithms.
* Start implementing chosen algorithms.

Month 3: Model Training and Evaluation

* Train machine-learning models on pre-processed data.
* Evaluate model performance using appropriate metrics.
* Begin comparative analysis.

Month 4: Comparative Analysis, Dissertation Writing, and Preparation of Deliverables

* Complete the comparative analysis of algorithms.
* Draft the Introduction, Literature Review, and Methodology chapters of the dissertation.
* Create a poster and video presentation.
* Begin proofreading and editing.

## Project Resources

The project will require the following resources:

* Project management: Oversees the entire coordination and administration of the project.
* Domain experts: People with knowledge of machine learning methods and building energy that can offer direction and insight throughout the project.
* Data sources: Information on weather patterns, building attributes, and energy usage.
* Computing Resources: Enough processing power and software programmes to put machine learning models into use and train them.

## Risk Management

A risk management plan will be set up to identify and reduce potential risks and uncertainties throughout the project. Data quality problems, algorithmic restrictions, resource shortages, and timetable delays are only a few examples of risks. Regular risk assessments will be carried out to lessen the effects of identified hazards, and appropriate contingency plans will be created.

## Communication and Collaboration

For the project to be successfully completed, I, in collaboration with my supervisor, set up an effective means of communication. Frequent meetings, progress reports, and documentation will be used to guarantee lucid and open communication. Collaboration solutions like project management software and version control systems will be used to ensure smooth cooperation and work tracking.

## Project Monitoring and Control

Continuous monitoring and control of the project progress is conducted to ensure adherence to the planned timeline, quality standards, and objectives. Regular project status updates, progress tracking, and performance evaluation will be carried out to identify any deviations from the plan and take corrective actions promptly.

# Conclusion and Future Work

The research wraps up in this chapter, summarising the results, discussing the contributions of the predictive machine learning models for building energy and discussing the project's goals. The chapter also identifies the study's shortcomings. It proposes directions for further research, like investigating cutting-edge machine learning approaches, including real-time data streams, or integrating the models into energy management systems for real-world applications.

1. Understanding the Data: There are a variety of numerical and categorical variables in the building energy dataset, each with a different range and distribution. Building energy use is significantly influenced by factors like CURRENT\_ENERGY\_EFFICIENCY and POTENTIAL\_ENERGY\_EFFICIENCY.
2. Performance of the models: The dataset was used to train and test various models, and in terms of RMSE and MAE, the Random Forest and Gradient Boosting models performed the best.
3. Importance of Feature: Building energy consumption was found to be significantly predicted by features like CURRENT\_ENERGY\_EFFICIENCY, POTENTIAL\_ENERGY\_EFFICIENCY, and TENURE. This is consistent with the intuitive knowledge that a building's ownership status and energy efficiency rating may significantly influence how much energy it uses.

To further improve the performance of the predictive models these steps can be iterated over until a desired performance metric is attained, it is also important to note that there is a limit to how much the model can improve on a given dataset.

1. **Engineering of Features:** More advanced feature engineering may enhance model performance. This can entail adding new features, altering numerical features, or encoding categorical variables differently.
2. **Advanced Models:** We might investigate more sophisticated machine learning models like neural networks or ensemble techniques that combine several different models.
3. **Hyperparameter tweaking:** For the models we have chosen, we could spend more time on hyperparameter tweaking. This can be done manually or automatically using grid or random search tools.
4. **Time Series Analysis:** We could investigate time series analysis techniques if the dataset contains a time component (for instance, the date the energy consumption measurement was made).

In “Strategies for minimizing building energy performance gaps between the design intend and the reality” by (Zou et al., 2019) :

1. Energy Performance Gap: The disparity between predicted energy consumption during the design stage of buildings and the actual energy use during operation, which can be significantly higher than predicted.
2. Lack of Accuracy in Design Parameters: Inaccurate design parameters used during the design stage can lead to unreliable estimates of energy consumption.
3. Failure to Account for Uncertainties: The failure to consider uncertainties, such as occupant behaviour and weather fluctuations, can result in deviations between predicted and actual energy consumption.
4. Lack of Accountability: The absence of clear accountability for energy performance in building projects may lead to complacency in achieving energy efficiency goals.
5. Poor Communication: Inadequate communication between stakeholders during the design process can lead to misunderstandings and discrepancies in energy performance expectations.
6. Lack of Knowledge and Experience: Insufficient knowledge and experience of building energy professionals can hinder accurate energy modelling and performance prediction.
7. Inefficient and Over-complicated Design: Complex and inefficient building designs may lead to higher energy consumption than anticipated.
8. Lack of Post-construction Testing: The absence of post-construction testing and validation of energy performance can lead to unforeseen discrepancies.
9. Lack of Feedback: The lack of feedback mechanisms and learning from past projects can hinder the improvement of energy performance in future projects.
10. Inadequate Building Codes and Regulations: Current construction codes and regulations may not provide enough specificity and detail to ensure actual energy performance aligns with design intentions.
11. Absence of Specific Requirements for Building Simulation Software: There are no specific requirements regarding the use of building simulation software, which can impact the accuracy of energy consumption calculations.
12. Limited Coping Strategies: Despite the existence of the energy performance gap, there is a lack of comprehensive strategies to address and minimize the gap effectively.
13. Rapidly Evolving Industry Challenges: The building energy industry faces constant changes and challenges due to technological advancements, and coping strategies need to adapt accordingly.

To improve the accuracy of design stage energy consumption estimates and achieve significant reductions in overall energy usage and emissions in the building sector, these gaps must be filled.

To conclude this chapter of this research, an encapsulation of a comprehensive overview of the findings derived from predictive machine learning models for building energy consumption. The chapter underscores the significance of these models in advancing our understanding of energy usage patterns within buildings by effectively summarizing the attained results and elucidating the pivotal role played by factors such as CURRENT\_ENERGY\_EFFICIENCY, POTENTIAL\_ENERGY\_EFFICIENCY, and TENURE in predicting energy consumption and making contributions to the existing knowledge in the field. Moreover, the chapter remains cognizant of the study's limitations, presenting a balanced perspective on the research outcomes.

References

Ciulla, G. & D’Amico, A. (2019) Building Energy Performance Forecasting: A Multiple Linear Regression Approach. *Applied Energy*, 253 November, p. 113500.

Deng, H., Fannon, D. & 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 March, pp. 34–43.

Dong, B., Cao, C. & Lee, S. E. (2005) Applying Support Vector Machines to Predict Building Energy Consumption in Tropical Region. *Energy and Buildings*, 37 (5) May, pp. 545–553.

Fathi, S., Srinivasan, R., Fenner, A. & Fathi, S. (2020) Machine Learning Applications in Urban Building Energy Performance Forecasting: A Systematic Review. *Renewable and Sustainable Energy Reviews*, 133 November, p. 110287.

Gibert, K., Izquierdo, J., Sànchez-Marrè, M., Hamilton, S. H., Rodríguez-Roda, I. & Holmes, G. (2018) Which Method to Use? An Assessment of Data Mining Methods in Environmental Data Science. *Environmental Modelling & Software*, 110 December, pp. 3–27.

Jain, R. K., Smith, K. M., Culligan, P. J. & Taylor, J. E. (2014) Forecasting Energy Consumption of Multi-Family Residential Buildings Using Support Vector Regression: Investigating the Impact of Temporal and Spatial Monitoring Granularity on Performance Accuracy. *Applied Energy*, 123 June, pp. 168–178.

Khalil, M., McGough, A. S., Pourmirza, Z., Pazhoohesh, M. & Walker, S. (2022) Machine Learning, Deep Learning and Statistical Analysis for Forecasting Building Energy Consumption — A Systematic Review. *Engineering Applications of Artificial Intelligence*, 115 October, p. 105287.

Lee, I. & Shin, Y. J. (2020) Machine Learning for Enterprises: Applications, Algorithm Selection, and Challenges. *Business Horizons*, 63 (2) March, pp. 157–170.

Li, C., Ding, Z., Zhao, D., Yi, J. & Zhang, G. (2017) Building Energy Consumption Prediction: An Extreme Deep Learning Approach. *Energies 2017, Vol. 10, Page 1525* [Online], 10 (10) October, p. 1525. Available from: <https://www.mdpi.com/1996-1073/10/10/1525/htm> [Accessed 29 June 2023].

Lu, C., Li, S. & Lu, Z. (2022) Building Energy Prediction Using Artificial Neural Networks: A Literature Survey. *Energy and Buildings*, 262 May, p. 111718.

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J. & Flach, P. (2021) CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33 (8) August, pp. 3048–3061.

Nasrul Aziz, M., Ahmad, T. & Nasrul Aziz, M. (2019) ICIC Express Letters ICIC International c ⃝2019 ISSN. [Online], 13 (2), pp. 93–101. Available from: <https://www.researchgate.net/publication/331730342> [Accessed 24 May 2023].

Nutkiewicz, A., Yang, Z. & Jain, R. K. (2018) Data-Driven Urban Energy Simulation (DUE-S): A Framework for Integrating Engineering Simulation and Machine Learning Methods in a Multi-Scale Urban Energy Modeling Workflow. *Applied Energy*, 225 September, pp. 1176–1189.

Ohueri, C. C., Liew, S. C., Bamgbade, J. A., Kumar Goyal, L., Singh Rai -, H., Yasser, M., Abdel Rashid, I., Mohamed Nagy, A., Long, R. & Li, Y. (n.d.) IOP Conference Series: Earth and Environmental Science Research on Energy-Efficiency Building Design Based on BIM and Artificial Intelligence You May Also like The Development of BIM-MyCREST Process Model for Concept Design of Green Buildings-BIM Approach for Sustainable Design of Flat Slab Buildings: A Review Integrated Model for BIM and Risk Data in Construction Projects Research on Energy-Efficiency Building Design Based on BIM and Artificial Intelligence.

Roh, Y., Heo, G. & Whang, S. E. (2021) A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33 (4) April, pp. 1328–1347.

Seyedzadeh, S., Pour Rahimian, F., Rastogi, P. & Glesk, I. (2019) Tuning Machine Learning Models for Prediction of Building Energy Loads. *Sustainable Cities and Society*, 47 May, p. 101484.

Somu, N., Raman M R, G. & Ramamritham, K. (2021) A Deep Learning Framework for Building Energy Consumption Forecast. *Renewable and Sustainable Energy Reviews*, 137 March, p. 110591.

Stedman, C. & McLaughlin, E. (2022) *Data Collection: Methods, Challenges and Key Steps* [Online]. Available from: <https://www.techtarget.com/searchcio/definition/data-collection> [Accessed 29 June 2023].

Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S. & Müller, K. R. (2021) Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. *Machine Learning and Knowledge Extraction 2021, Vol. 3, Pages 392-413* [Online], 3 (2) April, pp. 392–413. Available from: <https://www.mdpi.com/2504-4990/3/2/20/htm> [Accessed 28 June 2023].

Touzani, S., Granderson, J. & Fernandes, S. (2018) Gradient Boosting Machine for Modeling the Energy Consumption of Commercial Buildings. *Energy and Buildings*, 158 January, pp. 1533–1543.

Wang, Z., Wang, Y., Zeng, R., Srinivasan, R. S. & Ahrentzen, S. (2018) Random Forest Based Hourly Building Energy Prediction. *Energy and Buildings*, 171 July, pp. 11–25.

Yu, Z., Haghighat, F., Fung, B. C. M. & Yoshino, H. (2010) A Decision Tree Method for Building Energy Demand Modeling. *Energy and Buildings*, 42 (10) October, pp. 1637–1646.

Zou, P. X. W., Wagle, D. & Alam, M. (2019) Strategies for Minimizing Building Energy Performance Gaps between the Design Intend and the Reality. *Energy and Buildings*, 191 May, pp. 31–41.

# Appendices

Meeting Minutes- Dissertation Meetings  
Date: 17th April 2023  
Time: 14:00  
Topic: Machine learning-based building energy prediction

Attendees:

* Kobina Folson, Student
* Hadeel Jazaa, Supervisor

Agenda Items:

1. Identifying the Area of Focus:

* We discussed the importance of identifying an area of focus for the dissertation.
* based on your strengths.
* The supervisor suggested exploring a specific area of machine learning-based in building energy prediction.

1. Weekly Meetings and Meeting Minutes:

* Weekly meetings to discuss progress and prepare meeting minutes.
* Preparation of minutes after each session.

1. Milestones for Dissertation:

* Discuss the milestones for the dissertation and formulate a plan to achieve them.
* The milestones include:
  + Ethical clearance
  + Preparation of proposal
  + Drafting the first chapter, with a focus on the literature review
  + Final work
  + Preparation of presentation
  + Create a Gantt chart to track milestone progress.

1. Schedule for Submission:

* We discussed the plan for submission and decided to work towards submitting.
* the dissertation by the deadline.

5. Resources:

* All information for the dissertation was provided on the school's virtual learning environment (VLE).

6. Expectations for Next Meeting:

* Prepare and share minutes of the meeting with the supervisor.
* Design and share an initial plan in the form of a Gantt chart for the next meeting.

Action Items:

* Focus on machine learning-based building energy prediction for the dissertation.
* Prepare meeting minutes after each meeting.
* Create a Gantt chart to track progress towards the milestones.
* Send the meeting notes and an initial Gantt chart plan to the supervisor before the next meeting.

Next Meeting:

* The next meeting day is **on 24th April 2023**.
* The agenda will include reviewing the initial plan, progress updates towards the milestones, and any challenges faced.

Meeting Minutes: Session 2

Date: 24th April 2023

Time: 14:00

Topic: Machine learning-based building energy prediction

Attendees:

* Hadeel Jazzaa, Supervisor
* Kobina Folson, Student

Agenda:

* Form the aim and objectives of the project.
* Draft Project Proposal.
* Discuss progress with Project Research.
* Where we can look for prospective datasets.

Minutes:

* We discussed and agreed upon the aims and objectives of the project, which are expected to be achieved within the given timeline.
* The team reviewed the progress made on the project research and identified areas that need more attention and resources.
* We discussed potential data sources for the project and agreed to investigate open datasets and explore other possibilities.
* Draft the project proposal, including a detailed explanation of the problem statement, research questions, methodology, and expected outcomes.
* The next meeting is scheduled for [date and time], and the following are expected to be submitted:
  + Aims and objectives of the project.
  + Meeting minutes
  + Draft Proposal

Actions:

* Finalize the aims and objectives of the project and submit them before the next meeting.
* Send out the meeting minutes to all attendees.
* Draft and submit the project proposal before the next meeting.

Next Meeting:

* Date and Time: 2nd May, 11:00
* Agenda:
  + Review and finalize project aim and objectives.
  + Review and provide feedback on the draft proposal.
  + Identify potential challenges and risks associated with the project.