

School of Built Environment, Engineering and Computing

Leeds Beckett University

**Comparison of Machine Learning Algorithms in Building Energy 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:

Date:

Student ID No: 77296873

# Acknowledgements

I would like to thank

# Abstract

To maximise energy efficiency, reduce carbon emissions, and maximise energy use, building energy forecast is a crucial component of energy management. Building energy consumption predictions using machine learning (ML) algorithms has shown encouraging results, enabling efficient energy planning and environmental initiatives. In order to help stakeholders make wise decisions, this dissertation examines and assesses various ML systems for predicting building energy.

The paper starts with a thorough analysis of the building energy forecast techniques now in use, highlighting the advantages and disadvantages of various algorithms. Then, taking into account aspects like prediction accuracy, computational efficiency, interpretability, and scalability, a group of excellent ML algorithms is chosen.

Relevant building energy consumption data is gathered and prepared, including handling missing values, outliers, and feature scaling, to aid in experimentation. The prepared dataset is used to build and train the selected machine learning algorithms, with performance-enhancing hyperparameter adjustment and optimisation.

Model interpretability and computational effectiveness are taken into account together with a number of indicators, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R2). Each ML approach for predicting building energy is explored along with its benefits and drawbacks to help with its practical implementations in real-world circumstances.

Contents

[Candidate’s Declaration ii](#_Toc142513355)

[Acknowledgements iii](#_Toc142513356)

[Abstract iv](#_Toc142513357)

[List of Figures viii](#_Toc142513358)

[List of Tables 9](#_Toc142513359)

[Chapter 1: Introduction 10](#_Toc142513360)

[1.1 Overview 10](#_Toc142513361)

[1.2 Problem Statement 10](#_Toc142513362)

[1.3 Aim and Objectives 10](#_Toc142513363)

[1.4 Outline 13](#_Toc142513364)

[This dissertation is organized as follows: 13](#_Toc142513365)

[Chapter 2 : Literature Review 15](#_Toc142513366)

[2.0 Introduction 15](#_Toc142513367)

[2.1 Background 15](#_Toc142513368)

[2.2 Machine-Learning in Energy Prediction 16](#_Toc142513369)

[2.3 Studies Performing Comparative Analysis of Machine Learning Algorithms 21](#_Toc142513370)

[2.4 Scope and Significance of Study 25](#_Toc142513371)

[Chapter 3: Methodology 27](#_Toc142513372)

[3.1 CRISP Methodology Implementation 27](#_Toc142513373)

[3.1.1 Research Design 29](#_Toc142513374)

[Data Exploration 30](#_Toc142513375)

[3.1.2 Data Collection 34](#_Toc142513376)

[3.1.3 Data Pre-processing 39](#_Toc142513377)

[3.1.4 Algorithm Selection 39](#_Toc142513378)

[3.6 Performance Metrics 40](#_Toc142513379)

[3.7 Experimental Setup 40](#_Toc142513380)

[3.8 Ethical Considerations 40](#_Toc142513381)

[Chapter 4 : Product/Research Design and Implementation 42](#_Toc142513382)

[4.1 Challenges and Limitations 44](#_Toc142513383)

[4.1.1 Data Availability and Quality 45](#_Toc142513384)

[4.1.2 Model Complexity and Interpretability 45](#_Toc142513385)

[4.1.3 Algorithm Selection and Hyperparameter Tuning 45](#_Toc142513386)

[4.1.4 Generalisation and Transferability 45](#_Toc142513387)

[4.2 Code Implementation 46](#_Toc142513388)

[Chapter 5: Research Outcomes/Results/Discussion and Evaluation 53](#_Toc142513389)

[5.1 Evaluation Metrics 53](#_Toc142513390)

[Chapter 6 : Project Management 56](#_Toc142513391)

[6.1 Project Overview 56](#_Toc142513392)

[6.2 Project Phases 56](#_Toc142513393)

[6.3 Project Resources 58](#_Toc142513394)

[6.4 Risk Management 58](#_Toc142513395)

[6.5 Communication and Collaboration 58](#_Toc142513396)

[6.6 Project Monitoring and Control 58](#_Toc142513397)

[Chapter 7: Conclusion and Future Work 60](#_Toc142513398)

[Appendices 65](#_Toc142513399)

# List of Figures

[Figure 1 32](#_Toc142682222)

[Figure 2 33](#_Toc142682223)

[Figure 3 34](#_Toc142682224)

[Figure 4 35](#_Toc142682225)

[Figure 5 44](#_Toc142682226)

[Figure 6 48](#_Toc142682227)

[Figure 7 49](#_Toc142682228)

[Figure 8 50](#_Toc142682229)

[Figure 9 50](#_Toc142682230)

[Figure 10 51](#_Toc142682231)

[Figure 11 52](#_Toc142682232)

[Figure 12 54](#_Toc142682233)

[Figure 13 55](#_Toc142682234)

[Figure 14 58](#_Toc142682235)

[Figure 15 58](#_Toc142682236)

# List of Tables

[Table 1 18](#_Toc142497061)

[Table 2 29](#_Toc142497062)

[Table 3 33](#_Toc142497063)

[Table 4 51](#_Toc142497064)

# Chapter 1: Introduction

# 1.1 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.

# 1.2 Problem Statement

A comparative analysis of machine learning methods for predicting building energy is necessary to find the most precise and trustworthy models. Researchers and practitioners can find the best strategy for predicting building energy by assessing and contrasting the performance of various algorithms.

# 1.3 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. Examine and evaluate current ML techniques: We will start by conducting a thorough assessment of the literature on machine learning methods used to estimate building energy. We may learn more about the benefits and drawbacks of different algorithms, as well as how well they work with specific building energy prediction tasks, by examining earlier research.
2. Finding Appropriate ML Algorithms: Based on the literature analysis, we will choose a few representative ML algorithms that are most suited for predicting building energy. Prediction accuracy, computational effectiveness, interpretability, scalability, and robustness to various building kinds and datasets are some of the variables we'll take into account.
3. statistics preparation and collection: For our experiment, it is essential to collect pertinent building energy usage statistics. To build a complete dataset, we will investigate openly accessible databases, energy monitoring systems, and real-time sensor data. Handling missing values, outliers, and feature scaling are all parts of the pre-processing process that make sure the dataset is ready for the ML model training.
4. Implementing the chosen ML algorithms and training them with the provided dataset will be done once the data is ready. To obtain the greatest results, this procedure will include optimising the models and fine-tuning the hyperparameters.
5. Performance Evaluation: We will use a variety of evaluation metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R2), to determine how well the ML algorithms perform. We will also take model interpretability and computational efficiency into account as additional evaluation criteria in addition to these measures.
6. Advantages and Limitations: In this chapter, we'll talk about the benefits and drawbacks of each machine learning (ML) approach for predicting building energy. By comprehending the advantages and disadvantages of each strategy, we may more effectively use them in practical situations and come to wise judgements.
7. Identifying Research Gaps: As a last step, we'll examine the findings to pinpoint any remaining questions and prospective topics for ML model-based building energy prediction research. This analysis will advance the area of energy research as a whole and direct future investigations into unresolved issues.

The outcomes of this study will be helpful for many parties concerned with sustainable development and energy efficiency. Building owners and facility managers can make use of the comparison of ML algorithms' insights to optimise energy use and cut expenditures. 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 prediction will be helpful to researchers in the field.

# 1.4 Outline

# This dissertation is organized as follows:

Chapter 1: Introduction

This chapter provides a brief overview of the study, outlining the problem statement, objectives, and significance of the research. It also presents the research questions that the study aims to answer.

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.

# Chapter 2 : 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.

### 2.1 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 HVAC system control.

Engineering-based modelling has traditionally been used to anticipate the energy consumption of buildings. 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.

In the context of building energy prediction, machine learning models can learn from previous energy consumption data and other relevant variables, such as weather conditions, occupancy, and time of day. Once trained, these models can forecast future energy use under varied scenarios with a reasonable degree of accuracy.

Building energy prediction has made use of a variety of machine learning techniques, such as artificial neural networks, decision trees, support vector machines, and linear regression models. These models each have advantages and disadvantages. The type of data available, the precise specifications of the prediction task, and the available computational resources are only a few of the variables that may influence the model selection.

It is anticipated that machine learning will become more widely used in building energy prediction as a result of the growing availability of building energy data made possible by the Internet of Things (IoT) and smart metering technology. 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.

## 2.2 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.

#### 1.2.1 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), in particular, 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).

#### 1.2.2 Support Vector Machines (SVM)

SVMs have shown promising results in building energy prediction challenges and have excellent generalisation capabilities. The main topic of this paper 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 (Dong et al., 2005).

#### 1.2.3 Random Forest (RF)

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.

#### 1.2.4 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.

#### 1.2.5 Linear Regression

(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.

#### 1.2.5 Gradient Boosting

The study focuses on commercial buildings and addresses the critical requirement for precise energy savings estimates in energy improvement programmes. The study suggests using the gradient boosting machine (GBM) technique to create baseline energy usage models, taking advantage of the growing availability of high-frequency interval data from advanced metering infrastructure (AMI). Using a dataset of 410 commercial buildings, the GBM model's performance is compared to conventional linear regression models and a random forest approach.

The study emphasises the necessity of energy efficiency programmes to reduce environmental effects in the context of the energy consumption landscape of commercial buildings in the United States. It highlights the value of measurement and verification (M&V), or M&V 2.0 approaches (also known as advanced data analytics and smart metres), in calculating energy savings.

The research examines several baseline energy modelling methods, such as artificial neural networks, support vector machines, gradient boosting machines, and linear and nonlinear regression. In particular, gradient boosting is emphasised for its promising results in raising prediction accuracy.

The article describes gradient boosting utilising the decision tree as a base learner and hyperparameter adjustment's significance in reducing overfitting. The research considers the outdoor air temperature, the day of the week, and U.S. federal holidays while applying the GBM algorithm to create baseline energy use models.

The paper's contribution—a baseline modelling strategy utilising the gradient boosting machine—is emphasised in the conclusion. It highlights how much better the GBM model performs in estimating commercial buildings' power usage compared to traditional and modern approaches. The increased interest in machine learning for energy efficiency is consistent with this study. It demonstrates how gradient-boosting algorithms can make precise forecasts.

Table 1

|  |  |  |  |
| --- | --- | --- | --- |
| 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.

### 2.3 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.

### 2.3.1 Predictive Modelling for US Bommercial 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, in particular, 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.

### 2.3.2 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.

### 2.3.3 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).

Study 2: 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, C. et al., 2017).

Study 3: 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).

The literature review, on its whole, examined the body of work on machine learning methods for forecasting building energy consumption. The advantages and uses of several algorithms, including Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest (RF), and K-Nearest Neighbours (KNN), were highlighted, and their use in predicting building energy consumption was thoroughly explored. The research also contained comparative analytical studies that evaluated the functionality of several algorithms, highlighting their potential for precise energy demand forecasts. The research underlined how crucial machine learning is to optimise energy use and lowering environmental impact in building operations. However, a more systematic methodology and a synthesis of the information from the analysed studies might help to increase the review's impact.

### 2.4 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 study's significance extends to several facets of energy management, environmental sustainability, and machine learning research, even though its focus is on using machine learning to predict building energy usage.

# Chapter 3: 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.

## 3.1 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.

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.

## 3.1.1 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

Table 2

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 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 |

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 2.

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 1

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 2

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 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 are 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.

A comparison of energy ratings

Description automatically generated

Figure 3

The distributions of the CURRENT\_ENERGY\_RATING and POTENTIAL\_ENERGY\_RATING columns in the dataset are displayed in the bar plots above:

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.

A pie chart with different colored circles

Description automatically generated

Figure 4

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 4 above.

* 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.

## 3.1.2 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 datasets for predicting building energy have been gathered for this work from various sources, including building management systems and open energy databases. A wide range of elements, including historical energy consumption data, meteorological data, building parameters (such as area and occupancy), and other pertinent variables, are present in the chosen datasets. These datasets are selected to guarantee their fit with the study's goals and the availability of the data needed for precise energy forecasting.

Definition of Data Columns

Table 3

|  |  |
| --- | --- |
| 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 |

## 3.1.3 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.

## 3.1.4 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.

## 3.6 Performance Metrics

Several performance indicators are selected to assess and contrast the performance of the machine learning algorithms. Accuracy, precision, recall, F1 score, and computing efficiency are some criteria. 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.

## 3.7 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.

## 3.8 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.

# Chapter 4 : 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.

Visual Representation of Implementation Process

A screenshot of a phone

Description automatically generated with low confidence

Figure 5

Here is a short explanation for each stage in the machine learning model development flowchart found in Figure 3:

1. Define the Problem: Clearly articulate and understand the problem you want to solve using machine learning. This involves identifying the goals, desired outcomes, and tasks the model will address.
2. Collect and Pre-process Data: Gather relevant data from various sources and ensure its quality and 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: 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: Assess the model's performance by applying it to the test dataset and evaluating the predictions against the known ground truth values. Calculate appropriate evaluation metrics such as accuracy, precision, recall, F1-score, or other domain-specific metrics.
8. Decision (Performance Satisfactory?): Determine if the model's performance meets the desired criteria or needs further improvement. Based on the evaluation results, decide whether the performance is satisfactory or unsatisfactory.
9. Modify Model Architecture or Parameters: If the model's performance is unsatisfactory, adjust the model architecture or hyperparameters. 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: 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.

## 4.1 Challenges and Limitations

This chapter explores the difficulties and constraints of estimating building energy using machine learning methods. While the significance of precise energy prediction and the study's goals were highlighted in the previous chapter, this chapter focuses on the challenges that were faced when implementing such techniques.

Producing trustworthy and beneficial outcomes depends on recognising and overcoming issues. This chapter 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.

### 4.1.1 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 restricted availability and data quality problems.

### 4.1.2 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.

### 4.1.3 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.

### 4.1.4 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.

## 4.2 Code Implementation

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 6 below the dataset is loaded using the Pandas library and an overview of the first few rows of data is show

A black screen with text on it

Description automatically generated

Figure 6

The results of the output of the code in Figure 6 can be seen in Figure 7 below. 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 need to be more helpful for our prediction purpose.

A computer screen shot of a black screen

Description automatically generated

Figure 7

Several columns in the dataset have a high percentage of missing values as shown in Figure 7. Completely absent are columns like COUNTY, SHEATING\_ENV\_EFF, and SHEATING\_ENERGY\_EFF. Many data in other columns, like FLOOR\_ENV\_EFF and ADDRESS3, must be included.

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 8

In Figure 9 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 9

As part of our cleaning operation, let us carry out some other operation 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 10. This is a straightforward and frequently successful way for addressing missing values, but there 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 10

The handling of all missing values was successful. 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.

Before the next stage of the data preparation process, the data was sampled, using 2% of which was arrived at by testing various percentages of the original dataset, as seen in Figure 11 below. This decision was ultimately arrived at due to initial runs of the code using the original dataset, where processes either timed out or crashed due to the large dataset size.

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

Description automatically generated

Figure 11

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.

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.

A screen shot of a computer program

Description automatically generated

Figure 12

In Figure 12, the one-hot encoding has been successfully applied to the dataset. After one-hot encoding, the dataset now comprises 41166 columns, as can be seen.

Let us move on to feature scalability and the division of the data into training and test sets. We will employ the StandardScaler from sklearn, which standardises features by eliminating the mean and scaling to unit variance, for feature 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 13 shows the code used to perform these operations.

A screen shot of a computer program

Description automatically generated

Figure 13

# Chapter 5: 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.

## 5.1 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), root mean square error (RMSE) have been widely employed for regression-based models such as these.

Table 4

|  |  |  |
| --- | --- | --- |
| 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 |

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

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.

A graph of blue rectangular bars

Description automatically generated

Figure 14

A graph of blue rectangular bars

Description automatically generated

Figure 15

The bar graphs above in Figures 14 and 15 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.

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.

# Chapter 6 : Project Management

## 6.1 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.

## 6.2 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.

3.3 Project Timeline

To effectively manage the project, a Gantt chart is provided below:

🡨 Insert Gantt Chart 🡪

## 6.3 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.

## 6.4 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.

## 6.5 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.

## 6.6 Project Monitoring and Control

Continuous monitoring and control of the project progress will be 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.

# Chapter 7: Conclusion and Future Work

The study 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.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.

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].

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].

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.

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