


Coursework Coversheet		 UNIVERSITY OF LEEDS	
School of Geography FACULTY OF ENVIRONMENT			
Student ID	201788996	Second Marker	
Module title/code	GEOG5405 Urban Data Science Project	Mark	
Assignment title	Practical Briefing	Less deduction (state reason)	
Supervisor	Dr Na Yan	Final Mark	

Project Title:
Incorporating EPC Data and Socioeconomic Factors for Improved Data-Driven Carbon Emission Prediction in Domestic Buildings

Feedback:
Empty space for feedback

GEOG5405M Urban Data Science Project

Incorporating EPC Data and Socioeconomic Factors for Improved
Data-Driven Carbon Emission Prediction in Domestic Buildings

Zhihao Zhang
August 21st, 2024

A dissertation submitted in partial fulfilment of the requirements of
the Master's Degree in Urban Data Science and Analytics of the
University of Leeds

Abstract

Building industry contributes around 40% of global carbon emissions (CEs), making it one of the major factors in climate change. It is important to understand the factors that influence CEs and develop accurate predictive models for reducing the environmental impact of the built environment. Energy Performance Certificates (EPCs) provide useful information for understanding building characteristics, but their effectiveness is still limited. Advances in data-driven approaches provide new possibilities for maximising the use of EPC data. Previous research mostly focused on energy consumption prediction as a method of estimating CEs rather than predicting CEs directly. This study identified three feature sets and the most effective models for predicting CEs by using different feature selection methods and ensemble learning. Our results showed that these methods can enhance the accuracy of predictive models. In particular, eXtreme Gradient Boosting (XGBoost) and Random Forest (RF) models optimised through Bayesian techniques were the most effective. This study proved the effectiveness of the proposed model in various socioeconomic contexts. By combining socioeconomic factors with building characteristics, this study addressed existing gaps and offered urban planners with data-driven support to achieve zero-carbon buildings and net-zero carbon cities.

Acknowledgements

My programme leads Dr Vikki Houlden and Dr Jiaqi Ge,
Thank you for your kind assistance this year!

My dissertation supervisor Dr Na Yan,
Thank you for your guidance in my dissertation!

My PhD supervisor Dr Sahar Mirzaie,
Thank you for your encouragement and support!

My parents,
Thank you for your unconditional love!

Thank you, Zhihao Zhang!

Table of Contents

Abstract	3
1. Introduction.....	8
1.1 Context	8
1.2 Research Aim and Objectives	8
1.3 Research Structure	8
2. Literature review	9
2.1 Energy Consumption Prediction	9
2.2 Carbon Emissions Prediction at Various Building Life Cycle Stages.....	10
2.3 Carbon Emissions Prediction at the Operation and Maintenance Stage.....	10
2.4 Research Questions.....	11
3. Methodology	12
3.1 Data Collection	12
3.2 Data Processing	12
3.3 Data Analysis	13
4. Discussions	13
4.1 Research Question 1	14
4.2 Research Question 2	15
4.3 Research Question 3	18
4.4 Limitations.....	20
5. Conclusions.....	20
5.1 Conclusions	20
5.2 Implications for Policy and Practice.....	21
Reference	23

List of Tables

Table 1: Datasets used in this study	12
Table 2: Features selected from each FS method.	17

List of Figures

Figure 1: The correlations between factors and CEs.....	14
Figure 2: The R2 values of each model.....	16
Figure 3: The CVRMSE values of each model	16
Figure 4: The importance of each feature ranked by each method.....	17
Figure 5: Growth of R2 values of each model after optimisation.....	18
Figure 6: Decline of CVRMSE values of each model after optimisation	19
Figure 7: Comparison of ensemble learning model performance.....	19

1. Introduction

1.1 Context

To address the challenges of global climate change, countries globally have implemented various policies and actions. These efforts align with the United Nations Framework Convention on Climate Change (UNFCCC) and the United Nations Sustainable Development Goals (SDGs), targeting the transition to sustainable urban development. Urban energy consumption and carbon emissions (CEs) are driven by various sectors, such as transportation and buildings, and governments address these sectors accordingly. For example, China has set a target to peak CEs by 2030 by promoting renewable energy (Xinhua, 2021); California in the United States (US) is planning to achieve zero-emission for all new passenger cars by 2035, reducing CEs in transportation (State of California, 2020); the United Kingdom (UK) is the major economy to legislate a net zero strategy by 2050, focusing on the investment in the latest zero-carbon technologies and green industries (Department for Business, Energy and Industrial Strategy, 2021).

Despite these efforts, buildings remain a major contributor to CEs due to emissions from the construction and operation of buildings (Ahmed Ali, Ahmad and Yusup, 2020). In 2022, domestic buildings accounted for one-fifth of the total CEs in the UK (Rowe and Rankl, 2024). Given their significant contribution, zero-carbon buildings are considered a critical strategy for reducing CEs (Fouly and Abdin, 2022). Recent studies have focused on the role of Energy Performance Certificates (EPCs) in mitigating CEs, highlighting their importance in providing information on CEs from buildings. However, there remain research gaps in understanding how buildings with varying characteristics influence CEs in complex urban environments. Addressing these gaps is crucial to enhancing predictive accuracy and guiding urban sustainability practices.

1.2 Research Aim and Objectives

This study aims to establish the influence of various building characteristics and socioeconomic factors on CEs during the building operation and maintenance stage. Furthermore, this study seeks to promote urban sustainability by developing advanced machine learning (ML) predictive models. It intends to provide data-driven insights for reducing CEs in diverse households and supporting zero-carbon building practices.

This study has four research objectives as follows.

- To analyse the influence of various building characteristics (e.g., size and insulation quality) and socioeconomic factors (e.g., work hours) on CEs.
- To validate the most effective feature selection method for further predicting CEs.
- To identify the best-performing CE predictive models based on various factors.
- To provide data-driven recommendations for urban planners and policymakers aimed at reducing CEs.

1.3 Research Structure

The remaining sections of this briefing are structured as follows. Chapter 2 reviews the existing literature, identifying key research gaps and outlining the research questions. Chapter 3 clarifies the methodology and data analysis process. Chapter 4 presents and discusses the key findings from the notebook. Chapter 5 concludes with the implications and limitations, followed by

proposing potential policies and practices for relevant organizations.

2. Literature review

This literature review is organized into three sections exploring current studies on CE prediction in buildings. The first section discusses research that focuses on energy consumption as a proxy for CEs, showing the limitations of this approach. The second section reviews CE predictive models across various stages of the building life cycle, comparing methodologies and performance. The third section focuses on the operation and maintenance stage to highlight the importance of building characteristics and socioeconomic factors.

2.1 Energy Consumption Prediction

Renewable energy consumption decreases CEs, whereas non-renewable energy consumption tends to increase operational CEs (Shafiei and Salim, 2014). Consequently, most previous research primarily focused on predicting energy consumption as a proxy for evaluating CEs, rather than predicting CEs directly. These studies identified some best-performing models for predicting energy consumption, which laid a solid foundation for the development and optimisation of CE predictive models. Leveraging these insights can improve the performance of CE predictive models.

Several studies have compared the performance of different ML models in various energy consumption scenarios. For example, Quevedo, Geraldi, and Melo (2023) compared multiple linear regression (MLR), support vector regression (SVR), and artificial neural network (ANN), and established a benchmark to evaluate the impact of university renovations on energy consumption in Brazilian universities. Their study found that SVR had the highest coefficient of determination (R^2) and lowest mean absolute error (MAE) among the proposed models. Similarly, Mohan et al. (2024) also confirmed that SVR could more accurately predict energy consumption in most cases by comparing multiple ML models including ANN and SVR. These results indicated SVR performed the best in energy consumption forecasting. Additionally, Pan and Zhang (2020) proposed a categorical boosting (CatBoost) model to predict building energy consumption from Seattle buildings. The R^2 values of their model were around 0.897, which was better than RF. The model they proposed can help building developers improve energy efficiency during building design and renovation. Moreover, Hosseini and Fard (2021) developed RF, Decision Tree (DT), and K-nearest neighbours (KNN) models to determine factors that affect energy consumption. The results showed that building height had a greater impact on building energy consumption than building surface area and roof area. To further improve the accuracy of energy consumption prediction, Bassi et al. (2021) conducted a case study in Chicago to examine the performance of different gradient-boosting ML models, including CatBoost, Light Gradient Boosting Machine (LightGBM), and eXtreme Gradient Boosting (XGBoost). Their results showed that XGBoost outperformed LightGBM and CatBoost.

These studies identified effective building energy consumption predictive models across different scenarios, including SVR, RF, LightGBM, CatBoost, and XGBoost. However, the performance of these models varies across these studies due to different urban socioeconomic contexts and regional case studies. Furthermore, buildings with varying energy efficiency may result in different CEs despite identical energy consumption. Hence, predicting energy

consumption alone to predict CEs is insufficient and inaccurate.

2.2 Carbon Emissions Prediction at Various Building Life Cycle Stages

The building industry generally consists of many stages, including material extraction, manufacturing, construction, operation and maintenance, and demolition (Abd Rashid and Yusoff, 2015). These interconnected stages are commonly known as the building life cycle. However, to avoid CEs shifting from one building life cycle stage to another (Verellen and Allacker, 2022), there has been recent research into predicting CEs at various stages.

In the stage of early design, Pan and Wu (2023) introduced an ensemble learning model that combines Bayesian optimization and XGBoost to predict CEs from domestic buildings in Chengdu, China. The proposed model achieved a Root Mean Square Error (RMSE) at least 40% lower than other models, which shows the effectiveness of ensemble learning models. During the construction stage, Fang, Lu, and Li (2021) investigated how CEs relate to design parameters and developed an RF model to predict CEs for 38 buildings in China. Their results showed that RF has a higher R^2 value than MLR, encouraging building designers to reconsider their design choices based on CEs. During the operation and maintenance stage, Yan et al. (2023) developed a real-time CE predictive method using a convolutional neural network (CNN) for domestic buildings in Beijing. The R^2 values of their models range from 0.91 to 0.98 and have a higher efficiency than the simulation-based method. During the demolition stage, Neelamegam and Muthusubramanian (2024) used ANN and SVR to assess CEs from recycled construction and demolition waste. Their results showed that SVR has superior performance than ANN. However, few studies have focused on CE prediction during the demolition stage due to limited datasets, making the work by Neelamegam and Muthusubramanian a valuable reference for developing CE predictive models at this stage in the future.

Although these studies offer methodological guidance for building developers in selecting building materials, they overlook how to accurately choose and minimize input parameters for simpler models. The objectives of these studies are finding suitable models to predict CEs at various stages instead of exploring the influence of several factors on CEs.

2.3 Carbon Emissions Prediction at the Operation and Maintenance Stage

The operation and maintenance stage contributes approximately 60% of a building's total CEs and covers a long period (Chen et al., 2023). Therefore, recent studies primarily focused on predicting CEs during this stage by considering various influencing factors.

Among numerous factors influencing CEs, floor area was consistently considered as one of the most important factors. For example, Kapoor et al. (2022) and Zheng et al. (2024) introduced more than ten ML models including RF, ANN, MLR, SVR, and DT, to predict CEs and their intensity in domestic buildings. Their results showed that nonlinear models were more effective than linear ones for predicting CEs, and models that included features such as floor area and heating types had higher accuracy. Additionally, Fenton et al. (2024) developed various ML models including LR and SVR to predict CEs in domestic and non-domestic buildings in France and Belgium, and their models reached around 90% prediction accuracy and emphasized the significance of floor area. However, while these studies emphasized the influence of various building characteristics on CEs, they overlooked the role of socioeconomic factors, such as

education and income levels. The daily behaviours of residents may be shaped by their socioeconomic background, which can also affect CE patterns.

In addition, other studies have also confirmed other important building characteristics. For example, Su et al. (2023) used four ML models, including XGBoost and SVR, to predict CEs from commercial buildings in Dalian, China. They identified the correlation between indoor illuminance and CEs across different seasons, showing that ensemble models performed well in predicting CEs. Similarly, Krych, Heeren, and Hertwich (2021) conducted a multivariate regression analysis on Brazilian office buildings, identifying electricity mix and cooling efficiency as the most significant factors among the 10 factors influencing CEs. However, while both studies highlighted relevant factors in commercial buildings, they do not address the applicability of their models to domestic buildings with different energy infrastructures. Domestic and commercial buildings have unique occupancy patterns and requirements; for instance, commercial buildings may have more complex ventilation systems, resulting in different CEs and usage patterns. Therefore, predictive models developed for non-domestic buildings may not be directly applicable to domestic buildings.

In the studies mentioned above, several building characteristics were identified that significantly affect CEs, mainly including floor area, installation of insulated materials, and different household systems for heating and cooling. Accurate CE prediction requires a comprehensive understanding of these factors, and the accuracy can vary with different features (Li et al., 2018). However, many researchers selected features based on previous studies or their expertise and they did not fully consider feature importance. To fill these gaps, Xikai et al. (2019) took a key step in feature selection by identifying 12 building characteristics, such as the heat transfer capacity of walls and glazing through correlation analysis. Additionally, they created four ML models including MLP and SVR, to predict CEs from domestic buildings in Tianjin, China. Their results showed that SVR had the highest R² values among the proposed models, reaching 0.8, which emphasizes the importance of feature selection in CE prediction.

Furthermore, most studies only focused on a limited number of buildings in special areas or only predicted CEs from single buildings. Therefore, these models may not be applicable to other cities with different CE patterns. Zhang et al. (2023) explained how urban morphology influences CEs by applying LightGBM at the urban-scale level. They identified urban morphology and building geometry (e.g. total floor area and natural gas) as important features for CE prediction, with an R² value improving by around 34% when these factors were considered. Their results highlight the need for adaptable models in different urban contexts. Additionally, many studies remain regional-specific with a lack of case studies in Leeds, UK.

In conclusion, the reviewed literature demonstrated significant advancements in predicting CEs by developing various predictive models and identifying key factors. However, gaps remain in accounting for the diverse factors (e.g. education levels) influencing CEs. The method to select features should also be considered to reduce model complexity and feature counts. Furthermore, there is a need for more comprehensive urban-scale studies, particularly in Leeds.

2.4 Research Questions

Based on the research background and research gaps mentioned, this study is guided by three

research questions.

- What is the correlation between building characteristics, socioeconomic factors, and CEs?
- Which feature selection method is most effective for CE predictive models?
- Based on building characteristics and socioeconomic factors, how well do ML models predict CEs perform well?

3. Methodology

The methodology section provides detailed dataset resources, explains the main steps of the study undertaken and justifies the need for the research project.

3.1 Data Collection

To fully capture the various factors influencing CEs, we collected relevant datasets from public sources, including building characteristics (e.g., glazing types, floor area) and socioeconomic factors (e.g., working hours). The unit of CEs is measured in tonnes/year, and the datasets used are summarised in Table 1.

Table 1: Datasets used in this study

Data description	Link	Period
Energy certificate data (EPC) for buildings including carbon emission data in Leeds	https://epc.opendatacommunities.org/login	2024
The Index of Multiple Deprivation scores in the Education, Skills and Training domain for each LSOA in Leeds	https://observatory.leeds.gov.uk/deprivation/map/	2019
The Index of Multiple Deprivation scores in the Income domain for each LSOA in Leeds	https://observatory.leeds.gov.uk/deprivation/map/	2019
Hours worked for different categories for each LSOA in Leeds	https://www.ons.gov.uk/datasets/TS059/editions/2021/versions/5	2024

These datasets make our models more applicable across urban areas with diverse socioeconomic contexts. Specifically, they help us examine how socioeconomic backgrounds and building design influence occupant behaviour and CEs. By incorporating these datasets, we can develop more accurate predictive models and identify the most significant factors influencing CEs in complex urban environments.

3.2 Data Processing

We began with data processing, including removing any missing values and outliers to ensure dataset completeness. Given the detailed descriptions of each category in building characteristics, we categorised them into broader groups to simplify the analysis. We applied the one-hot encoding technique to convert each categorical variable into a binary variable containing either 0 or 1 (Dahouda and Joe, 2021), as some ML models cannot handle

categorical variables directly. After data cleaning, we merged the EPC data with socioeconomic data using LSOA codes to create the final dataset. These steps provided the foundation for developing predictive models and feature selection (FS) methods.

3.3 Data Analysis

In Research Question 1, we identified the building characteristics and socioeconomic factors that are associated with CEs. Furthermore, we used the Pearson correlation coefficient to measure the strength of the linear relationship between variables (Zoran et al., 2020). This correlation analysis can fill the gap of ignoring socioeconomic data, and also support urban planners in implementing targeted policies focusing on the most significant factors.

Research question 2 focuses on determining the most effective FS method for CE predictive models. FS is a crucial process in developing best-performing models, which includes filter, wrapper, and embedded methods (Jovic, Brkic, and Bogunovic, 2015). As discussed in the literature review, researchers often select features based on their expertise. However, FS not only reduces the burden of the datasets but also avoids model overfitting (Venkatesh and Anuradha, 2019). Therefore, we applied various FS methods, including mutual information (MI), RF, and Recursive Feature Elimination (RFE), to identify three feature sets and reduced the feature count from over 40 to 13. Using different feature sets, we examined the effectiveness of FS methods by developing eight ML models as identified in the current studies. These models include LR, RF, SVR, GBR, DT, LightGBM, CatBoost, and XGBoost. The predictive accuracy of each model was evaluated using the Coefficient of Variation of the Root Mean Square Error (CVRMSE) and R2 with cross-validation. This approach can fill the gap and extend current studies into the application of FS methods in predicting CEs.

The primary objective of Research Question 3 is to optimise predictive models and identify the best-performing ones. Based on the initial training results from Research Question 2, we selected the three best-performing models for Bayesian optimisation to improve their performance and determine the best parameters. Ensemble learning can gain better prediction accuracy from the combination of weak individual models (Rimal et al., 2023). Therefore, we employed the optimal parameters to combine these models into ensemble models by using different ensemble strategies including simple, weighted, and voting averaging methods. These findings can help examine and further support the application of ensemble learning in predicting CEs more accurately, thereby helping urban planners evaluate CEs from domestic buildings.

4. Discussions

This chapter presents the key findings of our study within urban environment contexts. The correlation identified between building characteristics, socioeconomic factors, and CEs is introduced. The effectiveness of different FS methods and the performance of various ML models in predicting CEs are discussed. Finally, potential limitations and future directions are also outlined.

4.1 Research Question 1

The heat map in Figure 1 highlights the correlations between various factors and CEs.

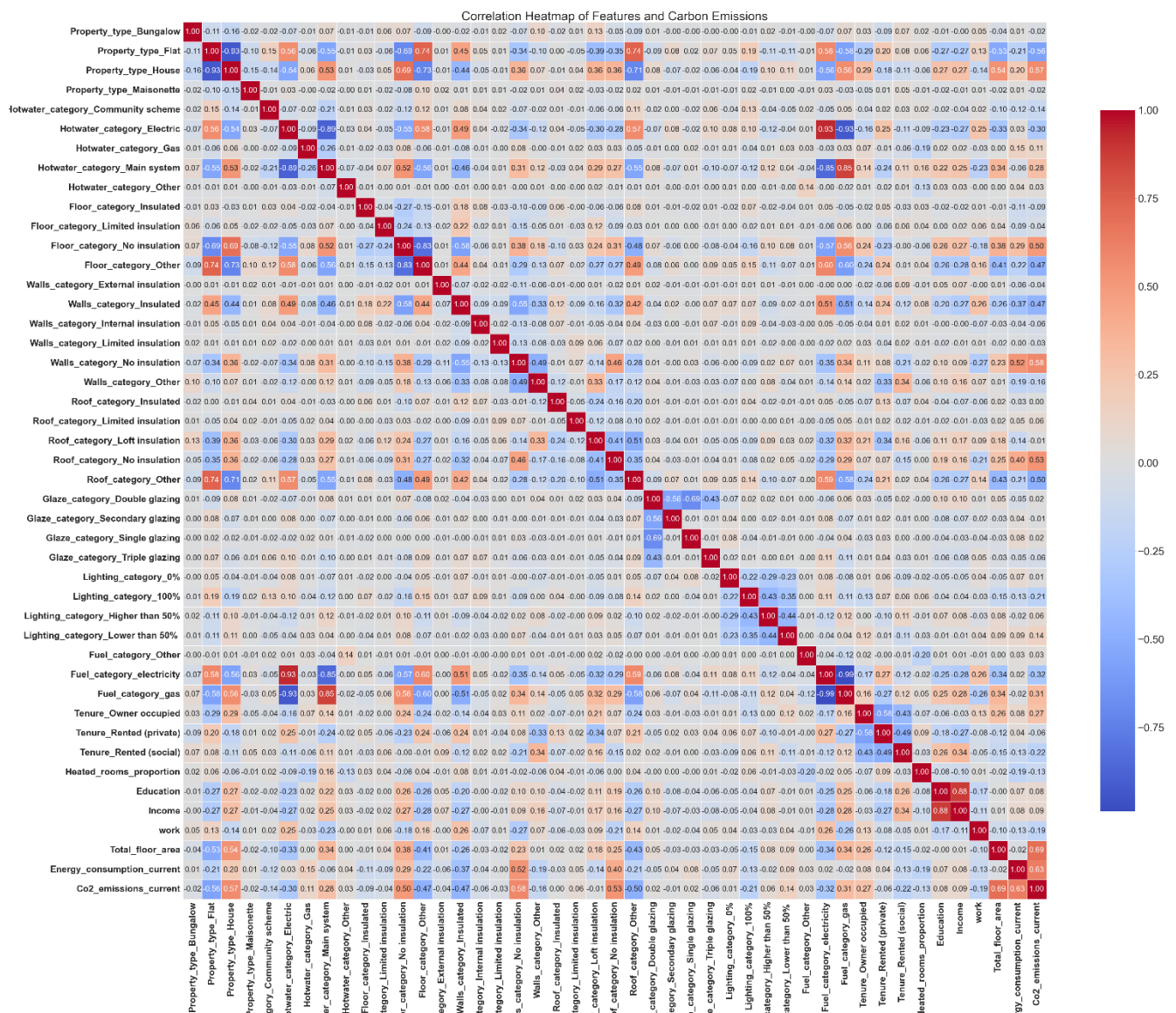


Figure 1: The correlations between factors and CEs.

This study found that buildings powered by using natural gas for heating and cooling are associated with having lower CEs than those using electricity. These results were consistent with the empirical study of Yuan et al. (2022), which found that CEs significantly increased due to the use of fossil fuels, including natural gas. Furthermore, our results indicated that larger properties with less installation of insulated materials and fewer fixed lighting outlets generally tend to have higher CEs. It suggested that improving the insulation materials of the building envelope (e.g., walls and roofs) and optimising integrated lighting design could be effective ways to reduce CEs in domestic buildings. Our findings aligned with previous research by Grazieschi, Asdrubali, and Thomas (2021), which concluded that buildings with highly insulated walls and roofs tend to have lower CEs. This is probably because of less need for lower heating and cooling energy, which results in lower CEs. Despite the advantages of insulation materials, they might need more maintenance and refurbishments during the building life cycle (Robati,

Daly, and Kokogiannakis, 2019). These findings highlighted the importance of developing zero-carbon buildings to lower CEs (Urge-Vorsatz et al., 2013; Hoang, Pham, and Nguyen, 2021).

However, building characteristics are not the only factors that affect CEs. Socioeconomic factors also play an important role in reducing CEs. Our study found that longer working hours are correlated with lower CEs, which is different from the findings of Fitzgerald, Schor, and Jorgenson (2018). They argued that longer working hours in the US have a greater environmental impact. This difference could be because their study focused on commercial or non-domestic buildings, whereas our study focused on domestic buildings. These findings supported the earlier idea that different building types have different CE patterns. Additionally, our results showed that buildings in areas with higher income and education levels are associated with lower CEs. Previous research has shown that education level and policy regulation can effectively reduce CEs (Xue, 2020), which is in line with our findings. Another possible explanation could be that residents in poorer neighbourhoods with higher CEs are likely to struggle to afford energy efficiency improvements. These findings further highlighted the need to conduct policies to address social inequality because inequality is associated with an increase in CEs (Hailemariam, Dzhumashev, and Shahbaz, 2020).

In conclusion, our findings suggested that policies should consider regional socioeconomic and cultural contexts when attempting to reduce CEs. Considering these findings into policy frameworks could help promote sustainability and reduce CEs in urban areas.

4.2 Research Question 2

To address Research Question 2, we assessed the performance of various feature selection methods by developing different ML models to predict CEs. Additionally, we confirmed the results drawn from Research Question 1.

The results showed that the choice of feature sets significantly affected the accuracy of CE prediction. Our findings were consistent with recent extensive studies, particularly those conducted by Jiang (2018) and Kong, Song, and Yang (2022). These studies emphasised the importance of selecting appropriate FS methods and identifying the most effective feature set for developing predictive models. In this case, we evaluate the performance of each FS method using R² and CVRMSE. The R² and CVRMSE values for each model with different feature sets are presented in Figure 2 and Figure 3.

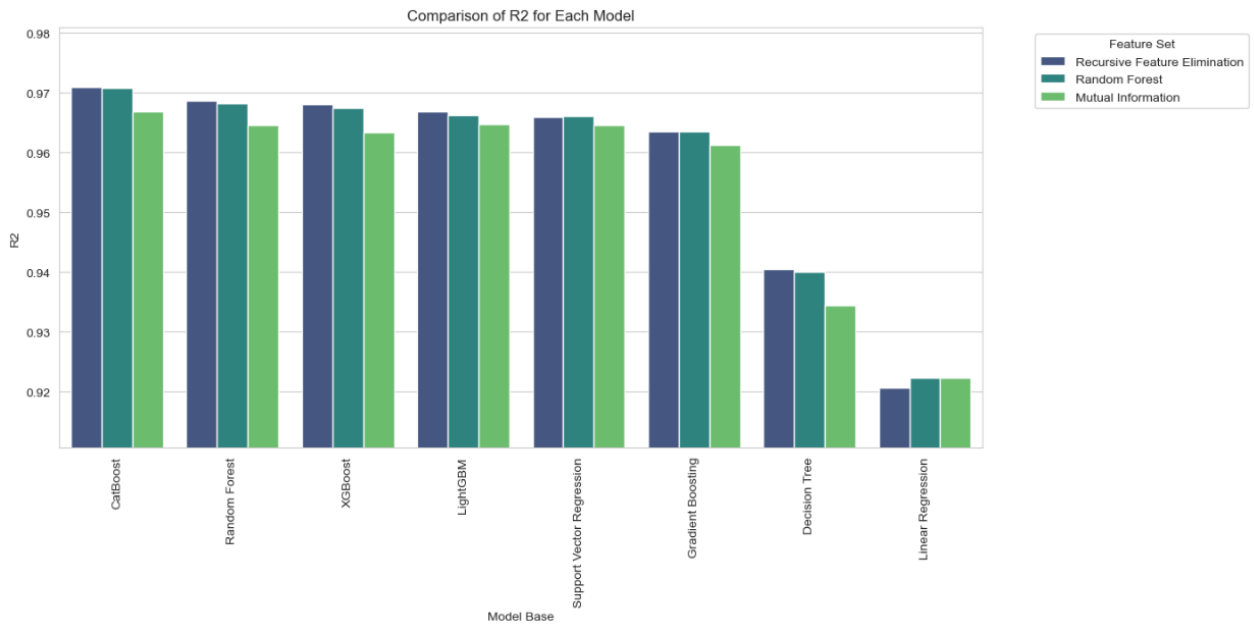


Figure 2: The R^2 values of each model

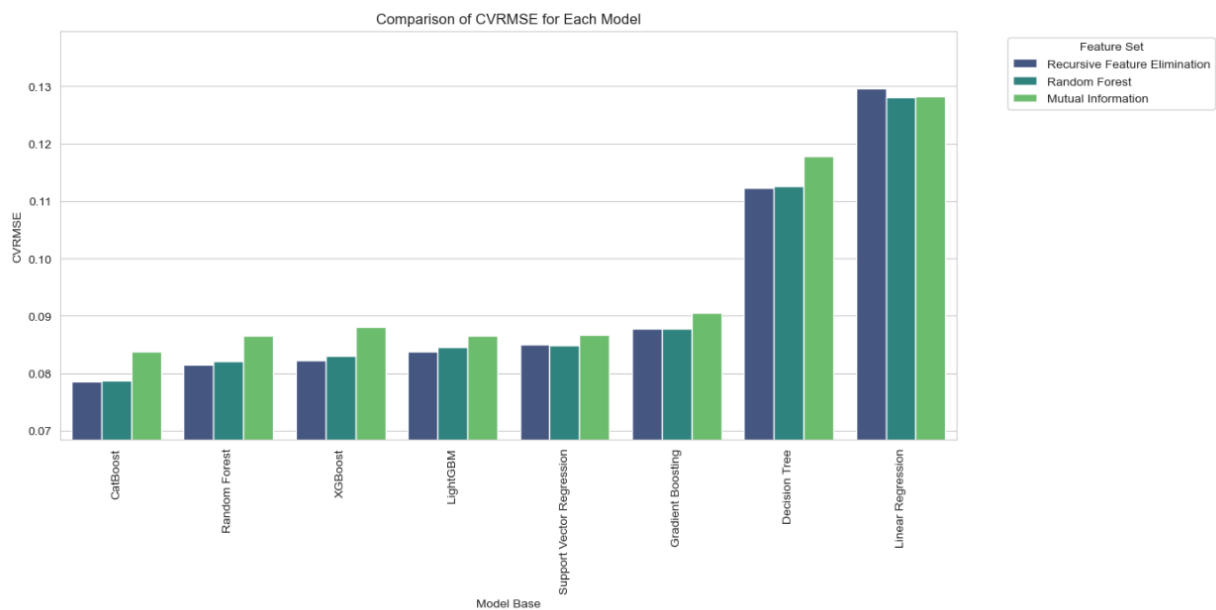


Figure 3: The CVMSE values of each model

Our findings showed that most models using the RFE method perform better than those using other FS methods, consistent with the work by Ketu (2022) and Luo et al. (2021). Luo et al. also employed three FS methods including the RFE, RF, and least absolute shrinkage and selection operator (LASSO), and found that the RFE method preserved the most informative features and outperformed other methods. Additionally, our findings showed that models using the RF method outperformed those using the MI method, which generally ranked the lowest in each model. The ineffectiveness of the MI method might be due to its sensitivity to large-scale datasets with high dimensionality (Subbiah and Chinnappan, 2021), making it less effective for datasets with numerous features. Figure 4 shows the importance of each feature ranked by different FS methods.

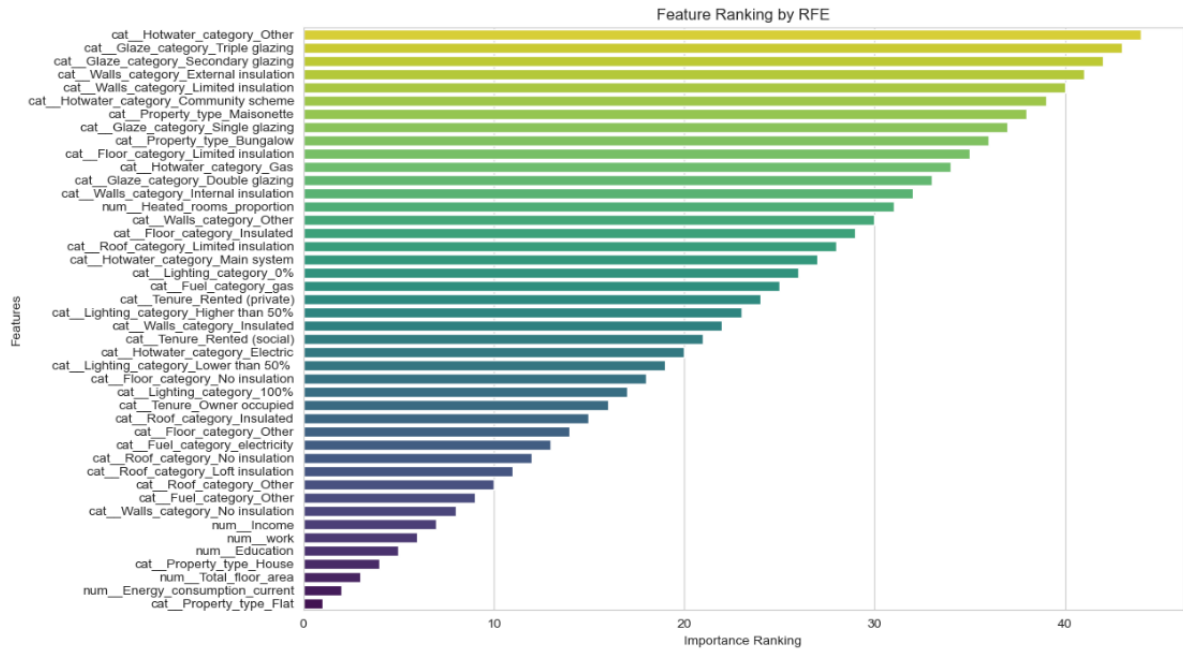


Figure 4: The importance of each feature ranked by each method

Determining the suitability of a specific method can be a challenging task due to the availability of a vast number of feature selection algorithms (Theng and Bhoyar, 2024). Similarly, the findings proved that the importance rankings of each feature differ across FS methods, leading to differences in selected feature sets. Table 2 presents the features selected by each FS method.

Table 2: Features selected from each FS method.

Rank	RFE	RF	MI
* This feature is not identified by all methods			
1	Total floor area	Energy consumption	Property type is Flat
2	Energy consumption	Property type is Flat	Energy consumption
3	IMD-Education scores	Total floor area	Total floor area
4	Work hours	Property type is House	Property type is House
5	Property type is House	Work hours	IMD-Education scores
6	Property type is Flat	IMD-Education scores	Work hours
7	IMD-Income scores	IMD-Income scores	IMD-Income scores
8	Walls without insulation	Walls without insulation	Walls without insulation
9	Roofs with "other" insulation	*Fuel is "other" type	*Fuel is "other" type
10	*Floors without insulation	*Roofs with loft insulation	*Roofs with "other" insulation
11	Roofs without insulation	Roofs without insulation	Roofs with loft insulation
12	*Floors with "other" insulation	Roofs with "other" insulation	Roofs without insulation

13	*Walls with insulation	*Floors with “other” insulation	*Fuel is electricity
----	------------------------	---------------------------------	----------------------

As shown in Table 2, ten key factors appeared consistently in the three feature sets, including property type, energy consumption, floor area, IMD-Education scores, work hours, IMD-Income scores, wall insulation, and roof insulation. These factors are also found in Research Question 1 and have a strong correlation with CEs. Furthermore, our results further explain why current studies use energy consumption as a proxy to evaluate CEs from buildings, as there is a strong positive correlation between the two (Huo et al., 2021).

In conclusion, the results highlighted the fact that the appropriate FS method for predicting CEs from buildings is the RFE method. Choosing an appropriate FS method and focusing on selected factors can not only improve CE predictive accuracy and reduce model complexity but also help urban planners develop CE reduction strategies targeted to various building types.

4.3 Research Question 3

Research Question 3 focuses on identifying the optimal model for predicting CEs using ensemble learning methods.

Although our model did not show significant improvement after optimisation (Figure 5 and Figure 6), it demonstrated that Bayesian optimisation can enhance predictive performance, with the XGBoost model showing the greatest improvement. Our findings aligned with the work of Koca Akkaya and Akkaya (2023), who developed nine ML models, including SVR and GPR, and optimised these models using Bayesian optimisation. Their results showed that the optimised GPR model has robust CE predictions. The limited improvement in CE predictive performance in this study could stem from the challenges ML models face when dealing with large feature sets and imbalanced datasets (Song et al., 2022; Wang and Sun, 2021). A similar issue was observed in the study by Wang et al. (2021), where the R2 values of their model increased by only around 3% after Bayesian optimisation due to the complexity of large feature sets.

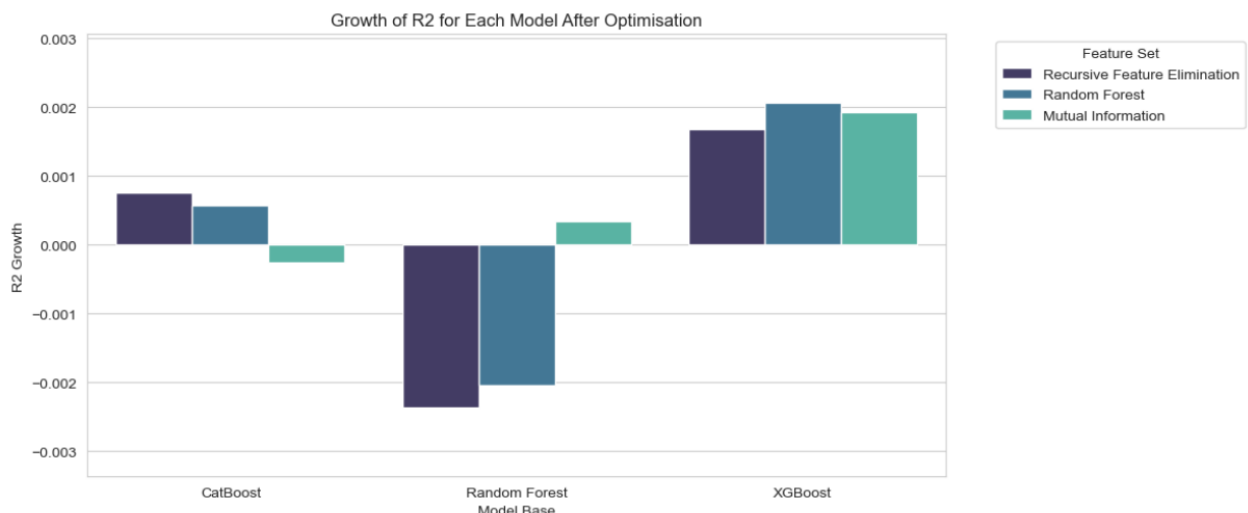


Figure 5: Growth of R2 values of each model after optimisation

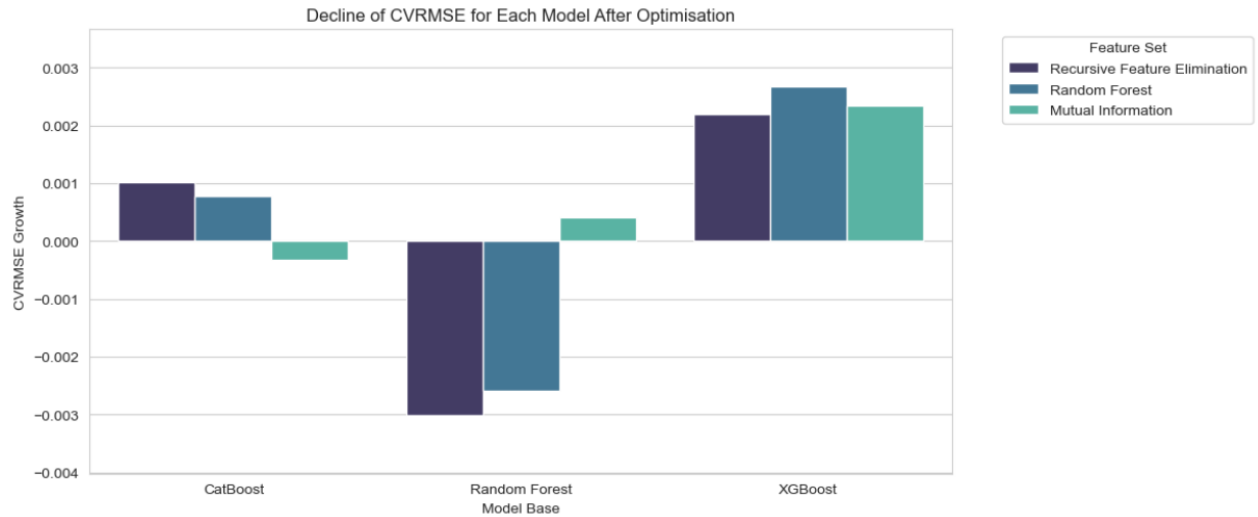


Figure 6: Decline of CVMSE values of each model after optimisation

Comparisons of model performance are shown in Figure 7. These results showed that ensemble learning can improve CE prediction accuracy, particularly through a voting ensemble approach that combines three models (the CatBoost, RF, and XGBoost models) based on the RFE method. These findings are consistent with those of Mienye and Sun (2022), who concluded that ensemble learning can enhance predictive performance by leveraging the strengths of multiple base models. Zhang et al. (2023) also confirmed the effectiveness of ensemble learning by developing a novel ensemble framework for predicting short-term CEs in three representative EU countries. Their empirical results showed that the proposed models outperformed other baseline models. Our study is also aligned with conclusions in related fields, such as enterprise CE predictions, as supported by the work of Yichao et al. (2023) and Zhou et al. (2024). Both studies highlighted the effectiveness of ensemble learning in achieving high predictive accuracy under uncertainty.

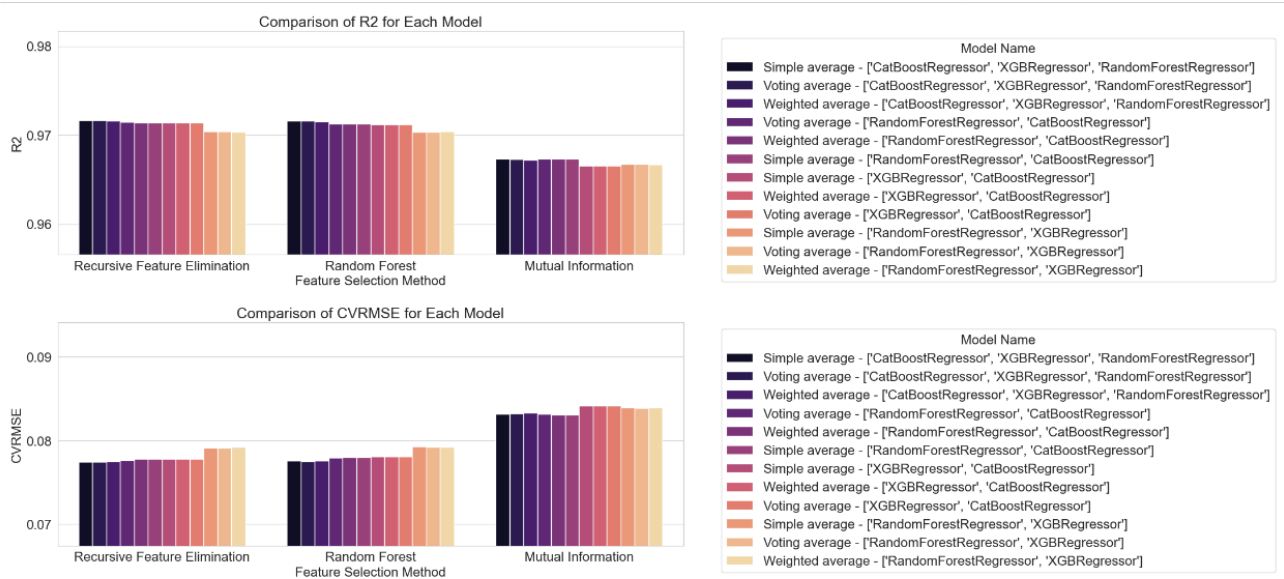


Figure 7: Comparison of ensemble learning model performance

In summary, this study suggested that utilizing ensemble learning techniques, especially in the

context of optimising CE predictions, provides a valuable framework for urban planners to estimate CEs more effectively from domestic buildings.

4.4 Limitations

While this study identified important factors affecting CEs and determined the predictive model with the best performance, several limitations should be addressed in future research.

One primary limitation is that certain factors have been overlooked. First, exploring more demographic characteristics would be beneficial, such as household age group and occupancy rate. For example, household size has a significant negative influence on household CEs per capita (Zhang, Wang, and Zhang, 2023), whereas older households may contribute to higher CEs because older households tend to have older housing equipment and more home appliances (Yagita and Iwafune, 2021). Moreover, architectural factors such as renewable energy systems, building orientation and layout should also be included. A building's orientation may impact the amount of daylighting (Silva et al., 2024), thereby influencing CEs. For instance, a building designed to effectively obtain natural light and ventilation can reduce the need for lighting systems, thereby lowering energy use and CEs. Incorporating these factors could potentially enhance the performance of the models.

Another limitation is that this study primarily focused on different building materials without accounting for variations in how these materials are applied. For example, the optimum thickness for insulated materials is different in different climate zones (Kumar et al., 2020). Future research can integrate climate data and explore how to optimize the selection and application of insulation materials based on specific climate conditions. Furthermore, researchers can investigate the influence of implementing strategies, including carbon taxes, across various areas with different socioeconomic backgrounds. Simulating different policy interventions to strengthen building regulations and exploring how policies incentivize residents to make net-zero transitions is also a beneficial choice.

Addressing these gaps and exploring these directions in future research can further improve CE predictive accuracy. This will also provide governments and professionals with a more comprehensive understanding of how different building characteristics and socioeconomic factors influence CEs and toward zero-carbon buildings.

5. Conclusions

This chapter summarises the main findings and makes clear recommendations for potential policy and practice.

5.1 Conclusions

By leveraging ML techniques and data-driven approaches, this study expanded the research on the application of ensemble learning and optimisation techniques for CE prediction and provided actionable insights for urban planners to implement targeted policies towards zero-carbon buildings.

Our research focused on three main questions. In Research Question 1, we examined the correlation between various factors and CEs, concluding that building characteristics such as floor area and insulation materials of the building envelope are closely associated with CEs. Additionally, socioeconomic factors such as education and income levels were identified as critical contributors to CEs. These findings support existing literature on the influence of building characteristics on CEs and further emphasize the importance of socioeconomic factors in CE reduction. Research Question 2 identified the most effective FS method to enhance CE predictive accuracy. The study proposed that the RFE method performed the best, which aligned with existing literature on the effectiveness of the RFE method and reinforced the conclusions drawn from Research Question 1. These findings can guide future research on selecting appropriate FE methods. Research Question 3 highlighted the advantages of ensemble learning in optimising CE predictive models. By combining three base models (the CatBoost, RF, and XGBoost models) and using a voting average ensemble strategy, ensemble learning can effectively integrate the strengths of various models. Our study has confirmed the potential of ensemble learning in CE prediction and expanded on its practical applications. Furthermore, these findings provided data-driven guidance for developers in evaluating CEs from domestic buildings.

In conclusion, this study enhanced the understanding of the correlation between building characteristics, socioeconomic factors, and CEs, while introducing new approaches for optimising CE predictive models. By employing effective FS methods and ensemble learning strategies, CE predictive accuracy was significantly improved, supporting the transition to zero-carbon buildings and net-zero cities.

5.2 Implications for Policy and Practice

Developing zero-carbon buildings and net-zero cities requires collective efforts from governments, relevant professionals, and the public. Therefore, some potential policies are proposed based on our findings.

For governments, stricter building codes and regulatory enforcement are the most effective basic measures (Zhang et al., 2021), as each building life cycle stage contributes to CEs. Additionally, reducing relevant taxes could encourage residents to upgrade their current systems to electricity or renewable energy systems. For example, they are freely replacing natural gas systems with electric systems for heating and cooling. Furthermore, socioeconomic inequality persists, and rural households potentially suffer more welfare loss in the carbon pricing policies (Jia, Wen, and Liu, 2022). Governments could adjust carbon pricing based on regional IMD levels. Moreover, governments can host educational sessions to propagate knowledge on CE reduction for shaping residents' environmentally friendly behaviours in areas where buildings with higher IMD scores tend to have higher CEs.

Building management systems are one of the critical roles in driving sustainability. Therefore, developers should implement these systems to monitor CEs dynamically and adjust operations accordingly. Building developers and house owners could implement building-integrated vegetation or plants, as green infrastructures not only influence CEs but also benefit residents' well-being (Zhu et al., 2023). Furthermore, developers should actively adopt green building standards and introduce sustainability concepts into building design (as mentioned in Research

Question 1), such as using efficient insulation for roofs and walls, triple-glazed windows, and reducing reliance on air conditioning through natural ventilation. Finally, the public should respond positively to policies by upgrading existing home energy systems. For instance, increasing the use of energy-efficient lighting outlets. They should also actively participate in environmental education activities to enhance environmental awareness.

Reference

- Abd Rashid, A. F. and Yusoff, S. (2015) 'A review of life cycle assessment method for building industry', *Renewable and Sustainable Energy Reviews*, 45, pp. 244–248. doi: 10.1016/j.rser.2015.01.043.
- Ahmed Ali, K., Ahmad, M. I. and Yusup, Y. (2020) 'Issues, impacts, and mitigations of carbon dioxide emissions in the building sector', *Sustainability*, 12(18), p. 7427. doi: 10.3390/su12187427.
- Bassi, A. *et al.* (2021) 'Building energy consumption forecasting: A comparison of gradient boosting models', in *The 12th International Conference on Advances in Information Technology*. New York, NY, USA: ACM.
- Chen, L. *et al.* (2023) 'Green construction for low-carbon cities: a review', *Environmental chemistry letters*, 21(3), pp. 1627–1657. doi: 10.1007/s10311-022-01544-4.
- Dahouda, M. K. and Joe, I. (2021) 'A deep-learned embedding technique for categorical features encoding', *IEEE access: practical innovations, open solutions*, 9, pp. 114381–114391. doi: 10.1109/access.2021.3104357.
- Department for Business, Energy and Industrial Strategy (2021) *UK's path to net zero set out in landmark strategy*, Gov.uk. Available at: <https://www.gov.uk/government/news/uk-path-to-net-zero-set-out-in-landmark-strategy> (Accessed: 12 August 2024).
- Fang, Y., Lu, X. and Li, H. (2021) 'A random forest-based model for the prediction of construction-stage carbon emissions at the early design stage', *Journal of cleaner production*, 328(129657), p. 129657. doi: 10.1016/j.jclepro.2021.129657.
- Fenton, S. K. *et al.* (2024) 'Embodied greenhouse gas emissions of buildings—Machine learning approach for early stage prediction', *Building and environment*, 257(111523), p. 111523. doi: 10.1016/j.buildenv.2024.111523.
- Fitzgerald, J. B., Schor, J. B. and Jorgenson, A. K. (2018) 'Working hours and carbon dioxide emissions in the United States, 2007–2013', *Social forces; a scientific medium of social study and interpretation*, 96(4), pp. 1851–1874. doi: 10.1093/sf/soy014.
- Fouly, S. E.-D. A. and Abdin, A. R. (2022) 'A methodology towards delivery of net zero carbon building in hot arid climate with reference to low residential buildings — the western desert in Egypt', *Journal of Engineering and Applied Science*, 69(1). doi: 10.1186/s44147-022-00084-6.
- Grazieschi, G., Asdrubali, F. and Thomas, G. (2021) 'Embodied energy and carbon of building insulating materials: A critical review', *Cleaner Environmental Systems*, 2(100032), p. 100032. doi: 10.1016/j.cesys.2021.100032.

- Hailemariam, A., Dzhumashev, R. and Shahbaz, M. (2020) 'Carbon emissions, income inequality and economic development', *Empirical economics*, 59(3), pp. 1139–1159. doi: 10.1007/s00181-019-01664-x.
- Hoang, A. T., Pham, V. V. and Nguyen, X. P. (2021) 'Integrating renewable sources into energy system for smart city as a sagacious strategy towards clean and sustainable process', *Journal of cleaner production*, 305(127161), p. 127161. doi: 10.1016/j.jclepro.2021.127161.
- Hosseini, S. and Fard, R. H. (2021) 'Machine learning algorithms for predicting electricity consumption of buildings', *Wireless personal communications*, 121(4), pp. 3329–3341. doi: 10.1007/s11277-021-08879-1.
- Huo, T. *et al.* (2021) 'Decoupling and decomposition analysis of residential building carbon emissions from residential income: Evidence from the provincial level in China', *Environmental impact assessment review*, 86(106487), p. 106487. doi: 10.1016/j.eiar.2020.106487.
- Jia, Z., Wen, S. and Liu, Y. (2022) 'China's urban-rural inequality caused by carbon neutrality: A perspective from carbon footprint and decomposed social welfare', *Energy economics*, 113(106193), p. 106193. doi: 10.1016/j.eneco.2022.106193.
- Jiang, H. (2018) 'Model forecasting based on two-stage feature selection procedure using orthogonal greedy algorithm', *Applied soft computing*, 63, pp. 110–123. doi: 10.1016/j.asoc.2017.11.047.
- Jovic, A., Brkic, K. and Bogunovic, N. (2015) 'A review of feature selection methods with applications', in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE.
- Kapoor, N. R. *et al.* (2022) 'Machine learning-based CO2 prediction for office room: A pilot study', *Wireless communications and mobile computing*, 2022, pp. 1–16. doi: 10.1155/2022/9404807.
- Ketu, S. (2022) 'Spatial Air Quality Index and Air Pollutant Concentration prediction using Linear Regression based Recursive Feature Elimination with Random Forest Regression (RFERF): a case study in India', *Natural hazards (Dordrecht, Netherlands)*, 114(2), pp. 2109–2138. doi: 10.1007/s11069-022-05463-z.
- Koca Akkaya, E. and Akkaya, A. V. (2023) 'Development and performance comparison of optimized machine learning-based regression models for predicting energy-related carbon dioxide emissions', *Environmental science and pollution research international*, 30(58), pp. 122381–122392. doi: 10.1007/s11356-023-30955-1.
- Kong, F., Song, J. and Yang, Z. (2022) 'A daily carbon emission prediction model combining two-stage feature selection and optimized extreme learning machine', *Environmental science*

and pollution research international, 29(58), pp. 87983–87997. doi: 10.1007/s11356-022-21277-9.

Krych, K., Heeren, N. and Hertwich, E. G. (2021) 'Factors influencing the life-cycle GHG emissions of Brazilian office buildings', *Buildings & cities*, 2(1), pp. 856–873. doi: 10.5334/bc.136.

Kumar, D. *et al.* (2020) 'Comparative analysis of building insulation material properties and performance', *Renewable and Sustainable Energy Reviews*, 131(110038), p. 110038. doi: 10.1016/j.rser.2020.110038.

Li, J. *et al.* (2018) 'Feature selection: A data perspective', *ACM computing surveys*, 50(6), pp. 1–45. doi: 10.1145/3136625.

Luo, M. *et al.* (2021) 'Combination of feature selection and CatBoost for prediction: The first application to the estimation of aboveground biomass', *Forests*, 12(2), p. 216. doi: 10.3390/f12020216.

Mienye, I. D. and Sun, Y. (2022) 'A survey of ensemble learning: Concepts, algorithms, applications, and prospects', *IEEE access: practical innovations, open solutions*, 10, pp. 99129–99149. doi: 10.1109/access.2022.3207287.

Mohan, R. *et al.* (2024) 'Comparative analysis of machine learning algorithms for the building energy prediction', in *2024 2nd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT)*. IEEE.

Neelamegam, P. and Muthusubramanian, B. (2024) 'Evaluating embodied energy, carbon impact, and predictive precision through machine learning for pavers manufactured with treated recycled construction and demolition waste aggregate', *Environmental research*, 248(118296), p. 118296. doi: 10.1016/j.envres.2024.118296.

Pan, H. and Wu, C. (2023) 'Bayesian optimization + XGBoost based life cycle carbon emission prediction for residential buildings—An example from Chengdu, China', *Building simulation*, 16(8), pp. 1451–1466. doi: 10.1007/s12273-023-1024-2.

Pan, Y. and Zhang, L. (2020) 'Data-driven estimation of building energy consumption with multi-source heterogeneous data', *Applied energy*, 268(114965), p. 114965. doi: 10.1016/j.apenergy.2020.114965.

Quevedo, T. C., Geraldi, M. S. and Melo, A. P. (2023) 'Applying machine learning to develop energy benchmarking for university buildings in Brazil', *Journal of building engineering*, 63(105468), p. 105468. doi: 10.1016/j.jobbe.2022.105468.

Rimal, Y. *et al.* (2023) 'Machine learning model matters its accuracy: a comparative study of ensemble learning and AutoML using heart disease prediction', *Multimedia tools and applications*, 83(12), pp. 35025–35042. doi: 10.1007/s11042-023-16380-z.

- Robati, M., Daly, D. and Kokogiannakis, G. (2019) 'A method of uncertainty analysis for whole-life embodied carbon emissions (CO₂-e) of building materials of a net-zero energy building in Australia', *Journal of cleaner production*, 225, pp. 541–553. doi: 10.1016/j.jclepro.2019.03.339.
- Rowe, G. and Rankl, F. (2024) *Housing and net zero*, Parliament.uk. Available at: <https://commonslibrary.parliament.uk/research-briefings/cbp-8830/> (Accessed: 22 August 2024).
- Shafiei, S. and Salim, R. A. (2014) 'Non-renewable and renewable energy consumption and CO₂ emissions in OECD countries: A comparative analysis', *Energy policy*, 66, pp. 547–556. doi: 10.1016/j.enpol.2013.10.064.
- Silva, B. V. F. et al. (2024) 'Sustainable, green, or smart? Pathways for energy-efficient healthcare buildings', *Sustainable cities and society*, 100(105013), p. 105013. doi: 10.1016/j.scs.2023.105013.
- Song, X.-F. et al. (2022) 'A fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for high-dimensional data', *IEEE transactions on cybernetics*, 52(9), pp. 9573–9586. doi: 10.1109/tcyb.2021.3061152.
- State of California (2020) *Governor Newsom announces California will phase out gasoline-powered cars & drastically reduce demand for fossil fuel in California's fight against climate change*, Governor of California. Available at: <https://www.gov.ca.gov/2020/09/23/governor-newsom-announces-california-will-phase-out-gasoline-powered-cars-drastically-reduce-demand-for-fossil-fuel-in-californias-fight-against-climate-change/> (Accessed: 12 August 2024).
- Su, Y. et al. (2023) 'Analysis and prediction of carbon emission in the large green commercial building: A case study in Dalian, China', *Journal of building engineering*, 68(106147), p. 106147. doi: 10.1016/j.jobbe.2023.106147.
- Subbiah, S. S. and Chinnappan, J. (2021) 'Opportunities and challenges of feature selection methods for high dimensional data: A review', *Ingénierie des systèmes d'information*, 26(1), pp. 67–77. doi: 10.18280/isi.260107.
- Theng, D. and Bhoyar, K. K. (2024) 'Feature selection techniques for machine learning: a survey of more than two decades of research', *Knowledge and information systems*, 66(3), pp. 1575–1637. doi: 10.1007/s10115-023-02010-5.
- Urge-Vorsatz, D. et al. (2013) 'Energy use in buildings in a long-term perspective', *Current opinion in environmental sustainability*, 5(2), pp. 141–151. doi: 10.1016/j.cosust.2013.05.004.
- Venkatesh, B. and Anuradha, J. (2019) 'A review of Feature Selection and its methods', *Cybernetics and Information Technologies*, 19(1), pp. 3–26. doi: 10.2478/cait-2019-0001.

- Verellen, E. and Allacker, K. (2022) 'Life cycle assessment of clustered buildings with a similar renovation potential', *The international journal of life cycle assessment*, 27(9–11), pp. 1127–1144. doi: 10.1007/s11367-022-02095-0.
- Wang, W. and Sun, D. (2021) 'The improved AdaBoost algorithms for imbalanced data classification', *Information sciences*, 563, pp. 358–374. doi: 10.1016/j.ins.2021.03.042.
- Xikai, M. *et al.* (2019) 'Comparison of regression models for estimation of carbon emissions during building's lifecycle using designing factors: a case study of residential buildings in Tianjin, China', *Energy and buildings*, 204(109519), p. 109519. doi: 10.1016/j.enbuild.2019.109519.
- Xinhua (2021) *China maps path to carbon peak, neutrality under new development philosophy*, Gov.cn. Available at: https://english.www.gov.cn/policies/latestreleases/202110/24/content_WS61755fe9c6d0df57f98e3bed.html (Accessed: 12 August 2024).
- Xue, Y. (2020) 'Empirical research on household carbon emissions characteristics and key impact factors in mining areas', *Journal of cleaner production*, 256(120470), p. 120470. doi: 10.1016/j.jclepro.2020.120470.
- Yagita, Y. and Iwafune, Y. (2021) 'Residential energy use and energy-saving of older adults: A case from Japan, the fastest-aging country', *Energy research & social science*, 75(102022), p. 102022. doi: 10.1016/j.erss.2021.102022.
- Yan, S. *et al.* (2023) 'A real-time operational carbon emission prediction method for the early design stage of residential units based on a convolutional neural network: A case study in Beijing, China', *Journal of building engineering*, 75(106994), p. 106994. doi: 10.1016/j.jobbe.2023.106994.
- Yichao, X. *et al.* (2023) 'Load-driven and energy consumption conversion-based enterprise carbon footprint estimation using stacking ensemble learning', in *2022 First International Conference on Cyber-Energy Systems and Intelligent Energy (ICCSIE)*. IEEE.
- Yuan, X. *et al.* (2022) 'The race to zero emissions: Can renewable energy be the path to carbon neutrality?', *Journal of environmental management*, 308(114648), p. 114648. doi: 10.1016/j.jenvman.2022.114648.
- Zhang, S. *et al.* (2021) 'Policy recommendations for the zero energy building promotion towards carbon neutral in Asia-Pacific Region', *Energy policy*, 159(112661), p. 112661. doi: 10.1016/j.enpol.2021.112661.
- Zhang, Y. *et al.* (2023) 'Data-driven estimation of building energy consumption and GHG emissions using explainable artificial intelligence', *Energy (Oxford, England)*, 262(125468), p. 125468. doi: 10.1016/j.energy.2022.125468.

- Zhang, Y., Wang, F. and Zhang, B. (2023) 'The impacts of household structure transitions on household carbon emissions in China', *Ecological economics: the journal of the International Society for Ecological Economics*, 206(107734), p. 107734. doi: 10.1016/j.ecolecon.2022.107734.
- Zheng, L. *et al.* (2024) 'Predicting whole-life carbon emissions for buildings using different machine learning algorithms: A case study on typical residential properties in Cornwall, UK', *Applied energy*, 357(122472), p. 122472. doi: 10.1016/j.apenergy.2023.122472.
- Zhou, W. *et al.* (2024) 'Carbon emission forecasting model for carbon asset management and power data-driven based on TPE-boosting ensemble stacking', in *2024 6th Asia Energy and Electrical Engineering Symposium (AEEES)*. IEEE.
- Zhu, S. *et al.* (2023) 'Numerical simulation to assess the impact of urban green infrastructure on building energy use: A review', *Building and environment*, 228(109832), p. 109832. doi: 10.1016/j.buildenv.2022.109832.
- Zoran, M. A. *et al.* (2020) 'Assessing the relationship between ground levels of ozone (O₃) and nitrogen dioxide (NO₂) with coronavirus (COVID-19) in Milan, Italy', *The Science of the total environment*, 740(140005), p. 140005. doi: 10.1016/j.scitotenv.2020.140005.