ARTIFICIAL INTELLIGENCE APPLICATIONS FOR IDENTIFYING KEY FEATURES TO REDUCE BUILDING ENERGY CONSUMPTION

by

Lakmini Rangana Senarathne

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Engineering Technology

West Lafayette, Indiana

August 2023

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Raji Sundararajan
School of Engineering Technology

Dr. Gaurav Nanda
School of Engineering Technology

Dr. Adel El-Shahat
School of Engineering Technology

Approved by:
Dr. Duane D. Dunlap
School of Engineering Technology

To my family, for their love and unwavering support.

To my loving husband, Charitha, for being my biggest encourager and cheerleader.

ACKNOWLEDGMENTS

I extend my heartfelt appreciation and gratitude to all those who have made invaluable contributions to the successful completion of my master's degree thesis. As a first step, I would like to thank my supervisor, Professor Raji Sundararajan, for providing expert guidance, insightful feedback, and valuable suggestions to make this a success. I am grateful for the opportunities she has provided me during my two years of education at Purdue.

I would also like to express my heartfelt gratitude to Professor Gaurav Nanda and Professor Adel El-Shahat, who guided me and strengthened my research with their knowledge and insightful discussions.

My deep appreciation goes out to all of my colleagues, co-teaching assistants, and others who have contributed to my remarkable experience at Purdue in numerous ways.

I extend my gratitude to my parents, my in-laws, and my sister for their love, support, and understanding throughout this journey. A special thanks to my husband, Charitha, for constantly encouraging me along the way. I am forever grateful for your support, love, and unwavering presence.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networks
CBECS	Commercial Building Energy Consumption Survey
CLD	Commercial Load Data
CO_2	Carbon Dioxide
HVAC	Heating, Ventilation, and Air Conditioning
IEA	International Energy Agency
IRLS	Iterated Reweighted Least Squares
MAE	Mean Absolute Error
MLP	Multi-Layer Perceptron
RF	Random Forest
RMSE	Root Mean Squared Error
UCI	University of California Irvine
US	United States
XAI	Explainable Artificial Intelligence

GLOSSARY

Accuracy	The quality of being correct (Moayedi & Mosavi, 2021)
Atmosphere	A mixture of gases surround the earth (Yoro & Daramola, 2020)
Classification	Putting into categories with the same qualities (Mao et al., 2021)
Coefficient	A numerical value that multiplies a variable (Huang & Li, 2021)
Continuous	A data that can take any value (Scanagatta et al., 2019)
Correlation	A mutual relationship between two or more things (Campagna & Fiorito, 2022)
Dependent variable	A variable whose value depends on another variable (Irfan & Ramlie, 2021)
Derivation	Developing of something from an origin (Goliatt et al., 2018)
Dry bulb temperature	Ambient air temperature (Liu et al., 2023)
Efficiency	The ratio of useful work to the total expected work (Moayedi & Mosavi, 2021)
Glazing	Windows (Prasetiyo et al., 2019)
Greenhouse Gas	Atmosphere gases that traps heat (Zhang et al., 2023)
Hypothesis	A proposed explanation made on basis of reasoning (Barros et al., 2018)
Independent Variable	A variable whose value does not depends on another variable (Irfan & Ramlie, 2021)
Linear Relationship	A straight line relationship between two variables (Kardani et al., 2021)
Machine Learning	A techniques that teaches computers to learn from experience (Tsoka et al., 2022)
Meteorological Year	A standardized period of 12 consecutive months that meteorologists used for statistical and climatological purposes (Yan & Liu, 2020)
Optimization	Making the best use of a resource (Gianniou et al., 2018)
Shapash	A python library (*Shapash*, n.d.)
Variable	Not consistent (Irfan & Ramlie, 2021)

Wilcoxon-rank-sum test	A test to compare two independent samples by comparing medians (Barros et al., 2018)

ABSTRACT

The International Energy Agency (IEA) estimates that residential and commercial buildings consume 40% of global energy and emit 24% of CO_2. A building's design parameters and location significantly impact its energy usage. Adjusting the building parameters and features in an optimum way helps to reduce energy usage and to build energy-efficient buildings. Hence, analyzing the impact of influencing factors is critical to reduce building energy usage.

Towards this, artificial intelligence applications, such as Explainable Artificial Intelligence (XAI) and machine learning (ML) identified the key building features to reduce building energy. This is done by analyzing the efficiencies of various building features that impact building energy consumption. For this, the relative importance of input features impacting commercial building energy usage is investigated. Also analyzed is the parametric analysis of the impact of input variables on residential building energy usage. Furthermore, the dependencies and relationships between the design variables of residential buildings were examined. Finally, the study analyzed the impact of location features on cooling energy usage in commercial buildings.

For the purpose of energy consumption data analysis, three datasets, named the Commercial Building Energy Consumption Survey (CBECS) datasets gathered in 2012 and 2018, University of California Irvine (UCI) energy efficiency dataset, and Commercial Load Data (CLD) were utilized. For this, Python and WEKA were used. Random Forest, Linear Regression, Bayesian Networks, and Logistic Regression predicted energy consumption using datasets. Moreover, statistical tests, such as the Wilcoxon-rank sum test were analyzed for the significant differences between specific datasets. Shapash, a Python library, created the feature important graphs.

The results indicated that cooling degree days are the most important feature in predicting cooling load with contribution values 34.29% (2018) and 19.68% (2012). Also, analyzing the impact of building parameters on energy usage indicated that 50% of overall height reduction achieves a reduction of heating load by 64.56% and cooling load by 57.47%. Also, the Wilcoxon-rank sum test indicated that the location of the building also impacts energy consumption with a 0.05 error margin. The proposed analysis is beneficial for real-world applications and energy-efficient building construction.

CHAPTER 1 . INTRODUCTION

1.1 Importance of analyzing building energy

The impact of climate change are profound and affect every living thing on the planet (Osman et al., 2022 ;Rodríguez et al., 2020). Global warming is a component of climate change, representing the elevation in the earth's surface average temperature (Rodríguez et al., 2020). Increased electrical energy consumption is one of the major causes of emitting greenhouse gases, which supports global warming (Mastrucci et al., 2021; Campagna & Fiorito, 2022). Carbon dioxide (CO_2) stands as the most notable greenhouse gas influencing global warming, with its emissions showing a steady increase over the years (Al-Ghussain, 2019; Yoro & Daramola, 2020). For example, emissions of CO_2 from the global atmosphere in 2019 were 45% higher than those between 1980 and 1990 (Yoro & Daramola, 2020). Furthermore, doubling the CO_2 amount would increase global temperatures by 3.8°C (Al-Ghussain, 2019).

The three main economic sectors are buildings, transportation, and industry. The building sector consumes considerable portion of energy among the three sectors (Chou & Bui, 2014; Prasetiyo et al., 2019). International Energy Agency (IEA) estimates, buildings consumes for 40% of total energy with 24% of CO_2 global emissions (Zhang et al., 2023; Prasetiyo et al., 2019). The building sector comprises 41% of the overall energy usage, whereas industry and transportation consume only 30% and 29%, respectively, in the United States (Choi & Bui, 2014). Therefore, reducing building energy consumption is crucial.

Buildings have two different types. The types are residential and commercial. Commercial buildings include types of offices, malls, hospitals, hotels, and many other buildings. As for the US Department of Energy, both residential and commercial buildings 40% emitted 40% greenhouse gasses by 2010 (Ahmad & Zhang, 2020). Moreover, as of 2016, commercial buildings have used over 60% of the energy only for electricity (Lokhandwala & Nateghi, 2018). Identifying the factors that influence commercial building cooling energy consumption is therefore important to achieve higher sustainability and less environmental impact.

1.2 Building energy usage impacting factors

Many factors impact building energy usage (Tsanas & Xifara, 2012; Aqlan et al., 2014; Invidiata et al., 2018). A few impacting factors are heating, ventilation, and air conditioning (HVAC), population growth, longer time spent in buildings, and climate. Furthermore, weather conditions, i.e., dry, cold, and seasonal weather, dry bulb temperature, thermal properties of building materials and number of floors also impact the building energy usage (Araújo et al., 2023; Bekkouche et al., 2017). Also, building characteristics play a major role in controlling building energy usage, and proper design strategies could reduce the building energy demand (Tsanas & Xifara, 2012).

Investigating building characterizing impacts on building energy efficiency is vital, because inadequate building design and structure have led to a 40% increase in CO_2 emissions from building energy usage (Xu et al., 2012). Heating load (HL) and cooling load (CL) are the two suitable parameters for analyzing building energy usage (Tsanas & Xifara, 2012; Irfan & Ramlie, 2021).

Climate location is another factor that impacts building energy consumption (Timmons et al., 2016; Renuka et al., 2022). The climates have different conditions such as hot, cold, humid, and dry (Phan & Lin, 2014). Therefore, the requirements of building energy also vary with the climate conditions in specific locations. For example, buildings use less cooling energy in cold areas than in hot climates (Phan & Lin, 2014).

1.3 Explainable Artificial Intelligence and its advantages

Artificial Intelligence (AI) is computers simulating human knowledge and training to take decisions based on human behaviors (Zhang & Lu, 2021). Image processing and intelligent robots are applications of AI. Explainable Artificial Intelligence (XAI) is a subset of Artificial Intelligence which is used to explain the dataset in detail using visualization techniques such as feature importance graphs (Zhang & Lu, 2021). XAI is useful technique to understand the decision or prediction made by the AI and identify which variables in the dataset have more impact on the predictive outcomes by supervised machine learning models (Kim et al., 2020; Angelov et al., 2021). Shapash is one of the popular XAI methods, which is a Python library that helps to make machine learning easily understandable and interpretable. The Shapash library

generates a visualization dashboard to implement machine learning model outputs. (Shapash, n.d.; Molnar, n.d.; Amin et al., 2022). Shapash displays precise results using plots to help users understand the models using a web app that allows them to switch between global and local explainability, without difficulty. Model explainability at the global level focuses on the features that have the most significant impact on outcomes, while it focuses on individual decisions at the local level. Shapash outputs include graphs depicting feature importance, contribution, and local explanations (Saboni et al., 2022)

1.4 Machine Learning and its advantages

Machine learning is a viable approach for analyzing with a high performing speed and easy implementation (Dogan & Birant, 2021). Machine Learning is a subfield of AI, which uses algorithms to make predictions using a training dataset (Xie et al., 2022). Machine learning is a popular model because of its easy identification of trend and patterns. It can review complex and large data and identify trends and patterns that are difficult for humans if do manually (Khanzode, 2020). Furthermore, ML is capable of handling multidimensional data in dynamic environments. The study used ML to identify patterns of energy related data, but due to its wide applications, many areas can use ML, such as healthcare (Dahiya et al., 2022). Specifically, supervised machine learning uses a labelled dataset for the predictions (Wang et al., 2021). The models use two testing conditions: train-test split and k-fold cross validation. Partitioning the data into k subsets or folds is the method of k-fold cross validation, with each fold taking its turn as the test data where other folds serves as the training data, and it is a reliable testing method (Chou & Bui, 2014).

Tsanas and Xifara (2012) used machine learning to model random forest, and iteratively reweighted least squares to predict HL and CL of a building. Moreover, Aqlan et al., (2014) predicted HL and CL using artificial neural networks and cluster analysis.

1.5 Statement of the Problem

Increasing building energy efficiency plays a significant part in reducing energy consumption, which, in turn, leads to reduced greenhouse gas emissions. One of the basic measurements to calculate the energy usage of a building is to measure heating load and cooling

load. Studies developed methods to identify the best machine learning algorithm with high accuracy and impact from input variables for predicting HL and CL. The problem was the need to identify the input variable's positive and negative impact on output variables and the relationship between design parameters. Also, the study was needed to analyze the relative importance of building characteristics and to analyze how the location feature impact building energy usage.

1.6 Objectives

The residential and commercial building sector consumed 41.10% of total energy consumption and emitted 40% of greenhouse gasses (Ahmad & Zhang, 2020). Reduced energy consumption helps to increase building energy efficiency. Therefore, the study analyzed methods to increase building energy efficiency using machine learning models.

Towards this, the objectives of the research are:

1. To study the positive and negative impacts of building design variables on energy usage.
2. To study building energy efficiency using the association of input variables.
3. To identify most important input features on building energy usage.
4. To analyze the impact of building features on energy consumption.
5. To reduce building energy usage and hence to reduce CO_2 emissions.

1.7 Research Questions

The following research questions guide this study.

1. What is the relative importance of input features/ variables impacting commercial buildings energy usage?
2. What design input variables impact the residential building heating and cooling energy usage positively and negatively?
3. What are the relationships and dependencies between design variables concerning residential building energy efficiency?
4. How impactful is the location feature to building cooling energy usage?

1.8 Importance of the Study

The results of this study will significantly contribute to improving living conditions for everyone. By reducing building energy usage, it is possible to reduce CO_2 emissions, which play a crucial role in today's global environmental challenges. The goal of the study is to identify the direction, magnitude of the impact, and relationships between building design parameters that can reduce energy consumption. Moreover, a clear understanding of the relative importance of features makes it easier to decide which one to prioritize. An architect or building designer can incorporate energy-saving features into the design of a new building by understanding how location impacts energy consumption. The practical value of the findings is significant since they can apply to real-world building projects, resulting in the creation of more energy-efficient structures. The study, in turn, can help reduce the world's energy demand and promote better living conditions for people everywhere.

1.9 Assumptions

The following assumptions are being made:

1. The UCI dataset (*UCI Machine Learning Repository*, n.d.) used is an adequate representation of real world data.
2. Parameters i.e., climate, number of occupants, and comfort levels are constant when analyzing impact from building parameters.
3. The weather data of Indianapolis and Lafayette, both located in the state of Indiana, are the same.
4. The building cooling energy usage of the city represents the whole state cooling energy usage.

1.10 Limitations

The following limitations are being made:

1. The UCI dataset considers only 12 building shapes.

2. Machine Learning model Bayesian networks used are applicable for smaller number of variables.
3. Random Forest produces ineffective predictions with larger number of trees.

1.11 Delimitations

The following delimitations are being made:

1. The data collection was not performed as part of the research process.

1.12 Chapter Summary

The introduction indicates that it is important to analyze building energy usage to reduce greenhouse gas emissions. The building design parameters are the most controllable factor in building energy-efficient structures. Identifying the most impacting features and the influence of building location is also important in analyzing energy efficiency. Due to pattern recognition capabilities, machine learning is a viable approach to analyzing energy data. The study includes four research questions to help with improving energy efficiency.

CHAPTER 2 . REVIEW OF LITERATURE

Chapter two provides a summary of recent literature in the research area of analyzing building energy efficiency using statistical techniques.

2.1 Overview of Building Energy Efficiency

Global warming, the increasing in average surface temperature, is a major challenge and a serious issue around the world caused by human activity (Al-Ghussain, 2019). The issue affects not only the human population but every living thing as well. Global warming is mainly due to greenhouse gas emissions, such as water vapor, methane, carbon dioxide, and nitrous oxide, that cause global warming. Among greenhouse gasses, carbon dioxide (CO_2) is the most prominent influencing gas, contributing 76% of emissions (Al-Ghussain, 2019).

Residential and commercial buildings accounts for more than 41% of energy consumption in US (Chou & Bui, 2014). The energy consumption of buildings is impacted by factors, such as design parameters, population density, and urbanization (Aqlan et al., 2014). Moreover, building energy consumption has been on the rise in recent years, driven by several factors, such as population growth, increased demand for building services, climate, and building characteristics (Zhang et al., 2023; Aqlan et al., 2014). Proper design of architectural parameters is a prominent energy- saving technique in buildings (Kim & Suh, 2021; Sarkar & Bardhan, 2020). By fine-tuning and enhancing these design parameters, it becomes feasible to lower energy consumption. The building design is one of the primary factors influencing energy usage (Chung & Rhee, 2014). Location is also a major factor impacting building energy consumption, since the weather and climate impacting the building depends on its location. By building structures with optimized energy design depending on its locations can increase building energy efficiency (Ascione et al., 2019).

Among many factors influencing building energy usage, it is important to understand which features are most important. Identifying the most influential features or factors that affect building energy usage allows designers to prioritize their efforts and investments in energy efficiency (Medal et al., 2021).

2.2 Impact from Building design variables

Heating and cooling loads are the basic parameters that define the environmental condition, representing the energy per unit time that needs be added or removed from the building (Chou & Bui, 2014; Shanthi & Srihari, 2018). Calculating building HL and CL is crucial in determining the necessary equipment to maintain a controllable indoor temperature, while ensuring economic and environmental viability (Abediniangerabi et al., 2022; Gong et al., 2020). Studies have used UCI energy efficiency dataset to analyze the building energy usage using eight input design variables i.e., relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution. The two output variables in UCI dataset are heating load and cooling load (*UCI Machine Learning Repository*, n.d.).

Several studies have used these input variables to forecast heating and cooling loads. Aqlan et al., (2014) identified RC, wall area, surface area, roof area, overall height, and glazing area as the most important factors for predicting HL and CL. Aqlan et al., identified that overall height significantly impacts the building heating and cooling energy. The study by Tsanas & Xifara, (2012) to identify influential variables of HL and CL, RC, wall area, and roof area have more influence on energy loads than other input variables.

Furthermore, Irfan and Ramlie (2021) conducted a study on the role of input variables concerning two output variables: HL and CL. Their research demonstrated that orientation does not significantly affect changes in HL and CL. However, overall height, wall area, and surface area have a notable impact on both energy loads. The study by Nazir et al., (2020) identified that the most important factors influencing heating and cooling load prediction are RC, overall height, wall area, glazing area, surface area, and roof area. The study did not delve into the negative or positive impact of independent variables or explore linear relationships between independent and dependent variables.

2.3 Explainable Artificial Intelligence

Artificial Intelligence (AI) is the simulation of human intelligence and the training of computers to learn human behaviors, including learning, judgment, and decision-making (Zhang & Lu, 2021). Artificial intelligence is used in many applications such as image processing, and intelligent robots (Zhang & Lu, 2021; van der Velden et al., 2022).

Explainable Artificial Intelligence (XAI) is a technique that improves the explainability of machine learning models. The aim of XAI is to give a better understanding of machine learning outputs (Machlev et al., 2022). The XAI tools gives a better understand on influence of input variables on the output (Tsoka et al., 2022). XAI also explains the procedure of decision are made by Artificial models increasing the confidence in the model (Ersoz et al., 2022). XAI helps researchers to understand the workings of ML models with high accuracy and performance (Machlev et al., 2022).

Previous studies have used XAI to give accurate predictions with better understanding. For example, a study by Zhang et al., (2023) used XAI using Light Gradient Boosting Machine integrated with SHapley Additive exPlanations (SHAP). Zhang et al., predicted the influence form different characteristics on building energy consumption quantitatively. A study by Tsoka et al., (2022) analyzed whether the building can achieve an energy performance certificate (EPC) using an artificial neural network classification model. However, Tsoka et al., used XAI tools such as Local Interpretable Model-Agnostic Explanation (LIME) and SHAP for the classification. According to the results of XAI, it is possible to remove not important input features without significantly affecting the accuracy of the ANN classification models.

2.4 Machine Learning

The use of machine learning is a viable approach to predicting energy efficiency data to get accurate results (Fathi et al., 2020). Machine learning is a subset of artificial intelligence (Pruneski et al., 2022). Machine learning is a computational method that uses available training data to make accurate predictions (Guo & Li, 2023). Machine learning divides into different categories, such as supervised, unsupervised, and semi supervised (Xie et al., 2022; Yu et al., 2023). The category depends on labeled and unlabeled training data (Guo & Li, 2023).

Supervised learning uses a labeled dataset (Karatzas & Katsifarakis, 2018). Supervised learning enables future predictions and classifications (Pruneski et al., 2022). Supervised machine learning includes algorithms such as regression, support vector machine, random forest, and neural networks (Guo & Li, 2023; Pruneski et al., 2022). Supervised learning has two models: regression and classification. Regression is predicting a continuous value using an existing training dataset. Some of the regression algorithms are Linear regression, Bayesian Ridge Regression, and Support Vector Regression (SVR) (Liapikos et al., 2022). Classification

algorithms predict a discrete or binary output (Matheus). Random Forest, Logistic Regression, and Support Vector Machine are a few of classification machine learning algorithms(Hassan et al., 2022). Unsupervised learning uses unlabeled input data to predict associations and patterns between them. Clustering methods help to predict the patterns (Hernandez-Matheus et al., 2022; Z. Huang et al., 2022). Semi supervised learning methods also work on analyzing high-dimensional data such as clustering and dimensionality reduction (Guo & Li, 2023). Because of these advantages, many studies have used machine learning to analyze building energy usage and efficiencies (Goliatt et al., 2018; Mokeev, 2019).

The two ways to divide data for machine learning predictions are test- train split and k-fold cross-validation. The ML model acquires knowledge about variables by training on a labeled dataset (Mishra et al., 2022). Subsequently, the model is put to the test and validated using the results obtained from test data. Test-train split divides the data randomly into two sets considering proportions (Boudjella & Boudjella, 2021a). A common percentage is training 75% and test 25% (Boudjella & Boudjella, 2021b). However, in k-fold cross validation, the dataset is divided into k folds (Chou & Bui, 2014). k-fold cross-validation is a reliable data-dividing method compared to train-test split (Abediniangerabi et al., 2022).

2.5 Prediction Algorithms

Studies have used many machine learning algorithms to predict building HL and CL using UCI dataset.

- Aqlan et al., (2014) used K-means clustering method and artificial neural networks (ANN) to analyze UCI HL and CL data. The results indicated combination of ANN and cluster analysis effectively predicted HL and CL with high accuracy.
- Tsanas & Xifara, (2012) utilized Random Forest (RF) and iteratively reweighted least square (IRLS) methods to estimate the association between variables. Results indicated RF out-performed IRLS.
- Goliatt et al., (2018) employed four regression models, which are Support Vector Regression (SVR) Gaussian process, Multi-Layer Perceptron, Neural Networks (MLP), and RF to predict building energy efficiency. They evaluated the performance using five metrics: Mean Absolute Error (MAE), R-squared (R2), Root Mean Square Error

(RMSE), Mean Absolute Percentage Error (MAPE), and Synthetic Index (SI). The results indicate that the Gaussian process is a viable and effective method for prediction of heating and cooling loads in buildings.

- Moayedi et al., (2021) utilized genetic algorithm (GA) and imperialist competition algorithm (ICA) to enhance the artificial neural network performance to predict HL and CL. The study's findings revealed that incorporating an optimization algorithm significantly improved the model's performance, with ICA outperforming GA in this context.
- Chou & Bui, (2014) developed predictions of heating load (HL) and cooling load (CL) using five single models: artificial neural networks (ANNs), support vector regression (SVR), classification and regression tree (CART), chi-squared automatic interaction detector (CHAID), and general linear regression (GLR), along with ensemble models. They conducted a comparison of these models' performance and identified that SVR had predicted HL with the best performance, while SVR+ANN ensemble model precited CL with the best performance.
- Permai & Tanty (2018) conducted a study on the impact of the Frequentist method and Bayesian approach on linear regression for predicting HL and CL. The results indicated that the Bayesian approach combined with linear regression yielded better results for Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Deviance (MAD) compared to the Frequentist method.
- Nazir et al., (2020) analyzed heating and cooling loads predictions using artificial neural networks. Results indicated that overall height, surface area, relative compactness, wall area, roof area, and glazing area are the most impacting factors for heating and cooling loads.
- Boudjella & Boudjella (2021a) compared the cooling load prediction result with the actual class result using the K Nearest Neighbor model.

Previous studies have primarily focused on predicting heating load (HL) and cooling load (CL) using supervised and unsupervised machine learning models. However, none of these studies have specifically analyzed the linear model coefficients to explore the relationship

between building parameters and various levels of energy efficiency for HL and CL. Our study aims to bridge this gap by analyzing model coefficients and establishing the relationship between input building parameters and five energy efficiency levels for heating and cooling loads. To achieve this, the study predicted model coefficients using linear regression and logistic regression, and Bayesian network was utilized to model the relationships.

2.6 Feature Importance

The CBECS 2018 and 2012 datasets analyzed the feature importance in the study. However, only a few limited studies used the CBECS 2018 data since the microdata in year 2018 was released in December 2022 (Administration, n.d.).

However, many studies used the CBECS 2012 dataset. The study by Lokhandwala & Nateghi (2018) identified the main predictors for predicting cooling load intensity of CBECS 2012. According to their results, the key predictors are climate and building types. Lokhandwala and Nateghi (2018) used Random Forest as the machine learning model for prediction since it obtained the highest R^2 of 0.86 and the lowest in-sample mean standard error (MSE) of 15.27%. They analyzed the importance of variables based on the percentage reduction of prediction accuracy by excluding each variable from the dataset. The results identified that cooling degree days, percentage of building cooled, principle building activity and the census division as the topmost important variables in both 2003 and 2012.

The study conducted by Deng et al., (2018) delved into the statistical relationship between continuous and categorical building characteristics. They found that SVM and RF exhibited superior performance in predicting energy usage intensity.

Indeed, numerous studies have been dedicated to identifying important variables within datasets for predicting the target variable. Casalicchio et al., (2019) introduced a local feature importance measure to idenitfy individual observations. They proposed visual tools, such as partial importance (PI) and individual conditional importance (ICI) plots, to demonstrate the impact from each feature on the model's performance.

Similarly, A study by Yan & Liu, (2020) developed a model for predicting cooling energy usage in residential buildings using air conditioners. They underwent a feature selection engineering process to select the most correlated and significant input features for energy usage

prediction. After evaluating different models, they concluded that XGBoost performed the best in predicting cooling energy use due to its reduced complexity and reduced risk of overfitting.

From the above it can be seen that no studies have been performed using explainable AI to identify the important features to predict energy usage using CBECS 2018 and 2012 data. Towards this, Shapash, a recent explainable ML library identified important features related to commercial building energy usage.

2.6 Impact of the location

The weather differences significantly impact building energy consumption (Santamouris et al., 2015; Delgarm et al., 2016). The climate and weather influencing the building's energy consumption depend on its location. Therefore, studies analyzed the impact of weather and climate parameters such as humidity, solar radiation, and wind speed on energy consumption (Santamouris et al., 2015).

The climate has different types. Timmons et al., (2016) studied the relationship between urban location and carbon emissions from residential buildings. The results of Timmons' study indicated a significant difference in residential energy between urban and nonurban areas due to factors such as population density.

A study by Santamouris et al., (2015) noted that the temperature elasticity of electricity demand is different for countries with warm, mild, and cold climates. Santamouris et al., indicated that the parameter considered varies by climate and location: 1.7% for warm climates, 0.54% for mild climates, and 0.51% for cold climates.

Furthermore, a study by Cao et al., (2016) explained that energy-saving techniques differ for different climate conditions based on their location. For example, thermal insulation and passive solar heat gain are well suited to cold climates, while solar shading and ground cooling are best suited to tropical climates.

The above studies focused on the climate parameter mostly, not the location. Therefore, it is important to identify the relationship between location on building energy consumption to building energy-efficient structures suitable for the location.

2.7 Chapter Summary

The review of literature indicated that factors such as population growth, climate, and building characteristics increase building energy consumption. Explainable Artificial Intelligence is a useful method to give accurate prediction results with more details. Machine Learning is a viable approach to predict building energy consumption, using many algorithms categorized into regression and classification. The studies mentioned in the review of literature focused on improving the accuracy of machine learning algorithm predictions. Therefore, identifying relationships between building input features to energy consumption is important. Furthermore, analyzing the most important features is needed to increase building energy efficiency. Analyzing the impact from the location on building energy consumption is also important to increase energy efficiency.

CHAPTER 3 . RESEARCH METHODOLOGY

3.1 Explainable Artificial Intelligence

3.1.1 Explainable Artificial Intelligence – Shapash

Explainable Artificial Intelligence (XAI) is a useful technique to understand the decision or prediction made by the AI and identify which variables in the dataset have more impact on the predictive outcomes by supervised machine learning models (Kim et al., 2020; Angelov et al., 2021). Shapash is one of the popular XAI methods, a Python library that helps make machine learning easily understandable and interpretable. The study uses Shapash to analyze the most important input features impacting cooling energy usage of buildings.

The Shapash library generates a visualization dashboard to implement machine learning model outputs. Visualization outputs display explicit labels to easily comprehend a model's summary (Shapash, n.d.; Molnar, n.d.). Shapash displays precise results using plots to help users understand the models using a web app that allows them to switch between global and local explainability easily. Model explainability at the global level focuses on the features that have the most significant impact on outcomes, while it focuses on individual decisions at the local level. The library results in three graphs showing the feature importance, feature contributions, and local explanation graph for each data ID (Ghosh et al., 2022). Shapash library supports predictions with many machine learning models, including Catboost, XGboost, Random Forest, LightGBM, Support vector machine (Ghosh & Sanyal, 2021).

The feature importance graph identifies the relative importance of input variables. The feature importance graph is a bar graph that gives results of the top 20 features identified based on their impact (Amin et al., 2022). The length of the bar represents the contribution of each feature. Shapash calculates the contribution using Shapley value (Molnar, n.d.). The fundamental concept involves quantifying the feature contribution value to the model's prediction relative to the average prediction of the target variable. By understanding how much each feature influences the model's output compared to the overall average prediction, it is possible to assess the relative importance of different features in making accurate predictions. For that, Shapash uses linear regression for simple models. For more complex models, Shapash uses other methods, such as cooperative game theory, to calculate Shapley values. Equation 3.1 shows the mathematical

formula to calculate Shapley value, as proposed by Štrumbelj & Kononenko (2014) using Monte-Carlo sampling. Here, let M represent the number of iterations. $\hat{f}(x_{+j}^{m})$ refers to the prediction for x, where a random data point z replaces the random number of values of features, excluding feature j value. Similarly, x_{-j}^{m} is the same as x_{+j}^{m}, but with the replaced values of j feature (Molnar, n.d.).

$$\widehat{\emptyset}_j = \frac{1}{M} \sum_{m=1}^{M} (\hat{f}(x_{+j}^{m}) - \hat{f}(x_{-j}^{m})) \qquad (3.1)$$

Equation 3.2 is a derivation of equation 3.1, assigning that $\emptyset_j^m = \hat{f}(x_{+j}^{m}) - \hat{f}(x_{-j}^{m})$. Equation 3.2 gives the averaging process of making the samples a probability distribution of x. Repeating the procedure can obtain the Shapley values for all input features (Molnar, n.d.).

$$\emptyset_j(x) = \frac{1}{M} \sum_{m=1}^{M} \emptyset_j^m \qquad (3.2)$$

The feature importance graph uses the SHAP, based on the game theory-based optimal Shapley values. The feature importance graph considers all the features; therefore, it is a global explainability. SHAP calculates the average absolute Shapley values per feature across the data to calculate the global importance of features, as shown in equation 3.3 (Molnar, n.d.).

$$I_j = \frac{1}{n} \sum_{i=1}^{n} |\emptyset_j^{(i)}| \qquad (3.3)$$

The value of I_j is the x axis value or the contribution of a feature. Therefore, SHAP feature importance (feature contribution) measures the mean absolute Shapley value (Molnar, n.d.; Lundberg & Lee, 2017). The interpretation of the Shapley values is that feature X changes the predicted absolute Y probability on an average by k×100% points, assuming the input feature X obtained a mean absolute Shapley value of k for predicting Y.

Random Forest implemented XAI to analyze the feature importance impacting building cooling energy consumption.

3.1.2 Random Forest Regressor

A Random Forest trains several decision trees which is an ensemble method (Biau & Scornet, 2016). A decision tree is a data classification process based on decision nodes, branches, and leaf nodes. A decision node passes a test, and a branch contains the outcome. A class label appears at the end of each leaf node (Li et al., 2020). The decision tree stops if the tree reaches the maximum number of levels or contains fewer observations than a predefined number.

Random Forest generates multiple decorrelate decision trees by randomly sampling a subset of features for each tree. When constructing each tree, the training sample set contains 75% randomly selected training samples. The other 25% uses as the test dataset to calculate the prediction error (Xie et al., 2021). Random Forest calculates the average values of predictions from all the trees as the final decision, and it is the decision with majority votes for classifications (Li et al., 2018; Singh et al., 2016). Random forest predictions are fast, scalable, robust to noise, have no overfitting, and have easy interpretation. However, the prediction speed decreases with the ensemble model's increasing number of decision trees (Abediniangerabi et al., 2022).

Two Commercial Buildings Energy Consumption Survey (CBECS) datasets interpreted the results using Random Forest in this study.

3.1.3 CBECS Datasets

The Commercial Buildings Energy Consumption Survey (CBECS) is dataset that gathers information on energy consumption and consumption patterns of commercial buildings in the United States (Administration, n.d.). The CBECS datasets include a wide range of building types, such as schools, hospitals, correctional facilities, religious facilities. These buildings are located in all 50 states and the District of Columbia.

CBECS 2018 and CBECS 2012 datasets are named according to the year collected in the study. CBECS 2018 dataset includes the data of 6436 buildings, representing 5.9 million US commercial buildings, which consumed 6.8 quadrillion of BTU energy in 2018 (Administration, n.d.). The CBECS 2012 dataset contains 6720 records, where each data is associated with 1119 attributes (Tian et al., 2019).

Cooling energy used in buildings is the output variable in research question 1. Energy used for cooling a building depends on many factors, and among them, this study used 21 input variables from CBECS 2018 and 2012 (Lokhandwala & Nateghi, 2018). The 21 input variables include five categories. The categories are climate, cooling features, building usage, construction, appliances and other features. Table 3.1 shows the summary of input variables. Table 3.1 includes the categories, name, label, and variable type. According to Table 3.1, the value ranges and categories are the same for CBECS 2018 and 2012 for some input features, such as CENDIV, MAINT, WKHRS, and MONUSE. However, input features like CDD65, MAINCL, and HWRDCL have different value ranges and categories in the two datasets.

Table 3.1. Summary of CBECS 2018 and 2012 input variables

Category	**Input variable**	**Input variable label**	**Variable type**
Climate	Cooling degree days (base 65)	CDD65	Number (days)
	Census division	CENDIV	Character
Cooling features	Percent cooled	COOLP	Number (%)
	Main cooling equipment	MAINCL	Character
	Regular maintenance for HVAC system	MAINT	Characters
	How cooling reduced in a 24-hour cycle	HWRDCL	Character
Building usage	Total hours open per week	WKHRS	Number (Hours)
	Months of year the building was in use	MONUSE	Number (Months)
	Total occupancy percent	TOTOCPP	Number (%)
	Number of businesses	NOCC	Number
	Number of workers	NWKER	Number
	Principle building activity	PBA	Character
	Percent lit when open	LTOHRP	Number (%)
Building construction	Floor to ceiling height	FLCEILHT	Number (feet)
	Glass percent in building	GLSSPC	Character
Appliances and other features	Laboratory equipment	LABEQP	Character
	Area of data center or server farm	DCNTRSFS	Character
	Linear accelerators	LINACC	Character
	Number of Xray machines	XRAYN	Number
	Cost of Electricity	ELCOST	Number ($\$.k^{-1}$ Btu^{-1})
	Number of computers	PCTERMN	Number

Total occupancy percent (TOTOCPP) and cost of electricity (ELCOST) variables are derivatives of the original CBECS datasets. Equations 3.4 and 3.5 show the formulas for generating TOTOCPP and ELCOST. To calculate the two variables, percentage occupancy (OCCUPYP), lodging room percent occupancy (LODOCCP), annual electricity expenditures (ELEXP), and annual electricity consumption (ELBTU) were considered.

$$\text{TOTOCPP (\%)} = \text{OCCUPYP (\%)} + \text{LODOCCP (\%)} \qquad (3.4)$$

$$\text{ELCOST } (\$k^{-1}Btu^{-1}) = \text{ELEXP (\$)} / \text{ELBTU (kBtu)} \qquad (3.5)$$

The output variable of CBECS datasets is the electrical energy usage intensity (EUI), labeled as ELCLPERSQFT (electricity used for cooling per square feet). Equation 3.6 shows the formula to calculate EUI (Lokhandwala & Nateghi, 2018).

$$\text{EUI } (\text{kBtu.ft}^{-2}) = \text{ELCLPERSQFT} = \frac{\text{Eelectricity used for Cooling (kBTU)}}{\text{Building area (square feet)}} = \frac{\text{ELCLBTU}}{\text{SQFT}} \qquad (3.6)$$

Table 3.2 summarizes the cooling EUI of CBECS 2018 and 2012 datasets. According to Table 3.2, the cooling EUI of 2018 buildings ranges from 0 to 139.6 kBtu, and it was 0 to 787.5 kBtu in 2012. Additionally, the mean and median also have a lower value in 2018 (7.9 and 4.7, respectively) than in 2012 (9.8 and 4.9, respectively). Table 3.2 indicates that buildings in 2018 used less cooling EUI than 2012 buildings.

Table 3.2. Statistics of cooling EUI in CBECS 2018 and 2012 datasets

	CBECS 2018 / kBtu.ft^{-2}	CBECS 2012/ kBtu.ft^{-2}
Minimum	0.0	0.0
Maximum	136.6	787.5
Q1	2.1	1.9
Median	4.7	4.9
Q3	9.5	11.5
Inter quartile range (IQR)	7.4	9.6
Mean	7.9	9.8

3.1.4 Workflow for identifying feature importance.

Figure 3.1 shows the workflow of research question 1, which is to identify the most important features to predict commercial buildings cooling energy usage. Random Forest predicted cooling EUI using 21 input variables from CBECS 2018 and 2012 datasets separately. Random Forest ran the predictions ten times to get the results using the train (75%)-test (25%) split. The Random Forest prediction created feature importance graphs for CBECS 2018 and 2012 using Shapash. Then, the study identified the most important features of the two datasets separately and compared them in terms of each feature's contribution and importance order.

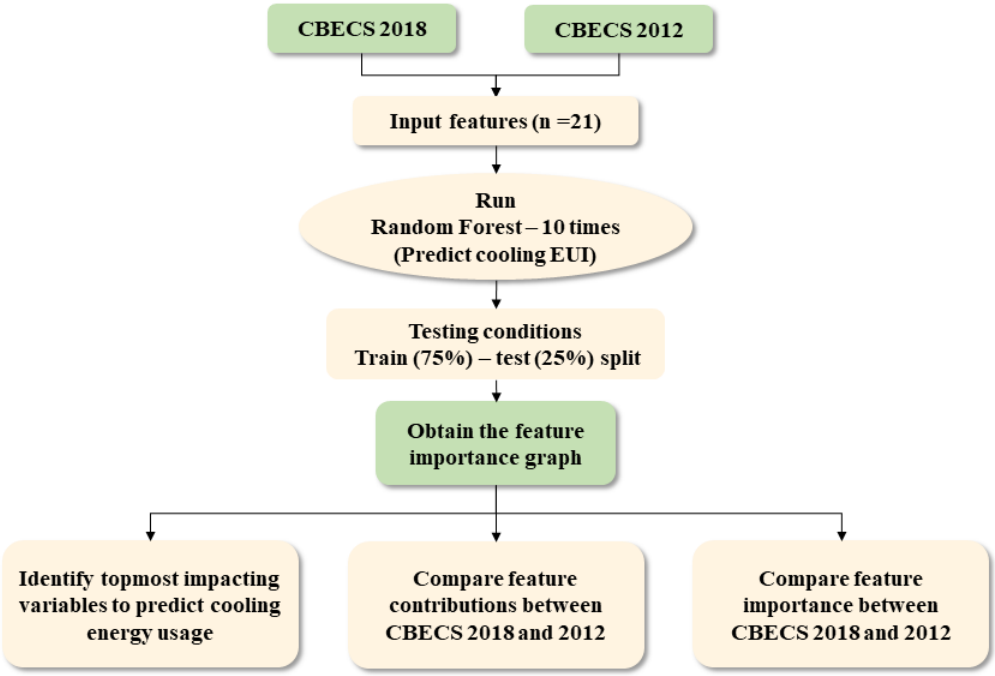


Figure 3.1. Workflow for identifying feature importance.

3.2 Machine Learning Applications

3.2.1 Machine Learning Algorithms

The two types of machine learning algorithms are Regression and classification algorithms. One regression model, i.e., Linear Regression and two classification models: Logistic Regression and Bayesian Networks are used in the study.

3.2.1.1 Linear Regression

Linear regression is a regression machine learning model which explains the linear relationship between one dependent (output/ response) variable and independent (input/explanatory) variables (Deng et al., 2018; Xie et al., 2021). Linear regression predicts the continuous variables (Srihari, 2018). Equation 3.7 shows the general Linear regression model, where β_0 = intercept, β_i = model coefficients, and ε is the random error (Pandit et al., 2021). Y is the dependent variable, where X_i are independent variables. The model coefficient of an input variable is proportional to the input feature contribution. For example, an input feature with a higher coefficient contributes more and have a higher influence on the output variable (Srihari, 2018).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_k X_k + \varepsilon \qquad (3.7)$$

Linear regression has two types: Simple Linear Regression (SLR) and Multiple Linear Regression (MLR) (Pandit et al., 2021; Goyal et al., 2020). If the model contains one independent variable (i =1), then it is a simple linear regression model. Therefore, SLR model includes one output variable (Y) and one input variable (X), and finds a correlation between input and output (Gianey & Choudhary, 2018). When there are multiple input variables ($i \geq 2$), it is a MLR model (X. Xie et al., 2021).

The Linear regression minimizes the residual sum of squares between the observation (y_i) and predicted output (y_{i_pred}) (Deng et al., 2018). Equation 3.8 shows the Linear regression objective function, where n is the number of data points and p is the number of input variables (Deng et al., 2018).

$$\text{Min} \sum_{i=1}^{n} \left(y_i - y_{i\_pred}\right)^2 \quad \text{where } y_{i\_pred} = \sum_{j=1}^{p} (x_{ij}\beta_j - \beta_0) \qquad (3.8)$$

3.2.1.2 Logistic Regression

Logistic regression solves classification problems where the output variable is dichotomous (Gianey & Choudhary, 2018). Logistic regression predicts the output by fitting a logistic function/curve to the dataset (Sala et al., 2021; Singh et al., 2016).

The classification process of Logistic regression calculates z as the summation of input data X ($x_0, x_1, x_2, \dots, x_n$) where each feature is multiplied by a regression coefficient W ($w_0, w_1, w_2, \dots, w_n$). Equation 3.9 shows the formula to calculate z (Zou et al., 2019).

$$z = w_0x_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n \quad (3.9)$$

Then, Logistic regression model calculates probabilities based on z, since a conditional probability distribution represents the fitted model. Equation 3.10 shows the conditional probability distribution of Logistic regression (Li et al., 2020). P(Y=1|X) is the probability of the output variable Y taking the value 1 (or success) and given the independent variables X and a set of coefficients W. (Y=0|X) is the probability of Y taking the value 0 (or failure).

$$P(Y = 1|X) = \frac{\exp(Z)}{1+\exp(Z)} \text{ and } P(Y = 0|X) = \frac{1}{1+\exp(Z)} \quad (3.10)$$

The objective of logistic regression is estimating coefficients (W) that best fit the observed data, and then use these coefficients to predict the probability.

When dealing with more than two classes, logistic regression transforms into Multiclass Logistic Regression (MLR) (De Loera & Hogan, 2020). The most common approach for MLR is the one-vs-rest method. Each classifier determines whether an example belongs to a specific class (De Loera & Hogan, 2020). The MLR model can handle multiple classes and make predictions accordingly using this one-vs-rest method.

3.2.1.3 Bayesian Networks

The Bayesian network (BN) is based upon Bayes' theorem, which is a supervised ML model (Tian et al., 2019). Bayesian network utilizes a directed acyclic graphical model to classify data, taking into account the dependencies among different attributes. Bayesian networks identify relationships between variables by analyzing their dependencies and independence, or they represent the causality of variables. This study focused on the two aspects of Bayesian Network (1) Bayesian network components and (2) model hyperparameters.

The Bayesian network consists of (a) the graph structure and (b) the table of numerical conditional probabilities (S. Huang et al., 2018).

3.2.1.3. a Graph structure

Bayesian network graph's nodes represent variables. The two types of nodes are parent nodes and child nodes. Parent nodes have an impact on child nodes. The arrows in the graph represent the relationships between connected nodes, pointing from the parent node to the child node, indicating the direction of influence.

Figure 3.2 illustrates the Bayesian network model with six nodes (S. Huang et al., 2018). Here, X_f is a parent node since it influences X_b, as indicated by the arrow direction from X_f to X_b. However, X_b acts as both parent and child nodes: it is the parent of X_c and the child of Xf. This configuration represents the dependencies and influences among the variables in the Bayesian network.

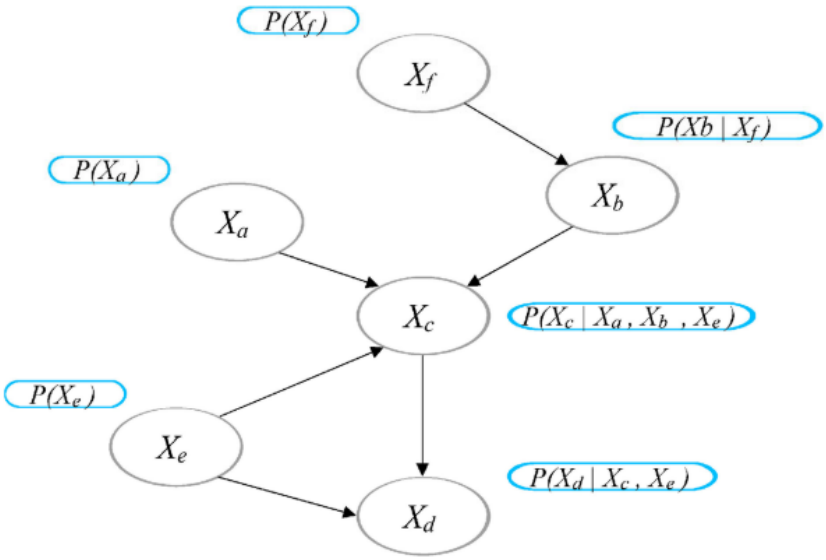


Figure 3.2. Bayesian network model (Huang et al., 2018)

Below equations show that a node can represent as a function of its parents (Huang et al., 2018).

$$X_b = f(X_f)$$

$$X_d = f(X_c, X_e)$$

$$X_c = f(X_a, X_b, X_e)$$

3.2.1.3. b Conditional probability tables

Each node in the Bayesian network graph consists of a conditional probability table. The conditional probability table defines how each variable behaves in the presence of its parent nodes. The conditional table includes semi ranges of each variable. Table 3.3 provides prototypes of these split sections for variables X_c and X_e in Figure 3.2. The assumed ranges of X_c and X_e span between $[X_{c,L}, X_{c,H}]$ and $[X_{e,L}, X_{e,H}]$, where $X_{c,L}$ is the minimum and $X_{c,H}$ is the maximum value for variable X_c, and similarly, $X_{e,L}$ is the minimum and $X_{e,H}$ is the maximum value for variable X_e. Variables X_e and X_c have m and n sections, respectively, reflecting the partitioning of their respective total ranges. These sections help define the conditional probabilities and dependencies of each variable with its parents in the Bayesian network.

Table 3.3. Semi ranges in a conditional probability table

Variable	Range	Number of Sections	Sections
X_e	$[X_{e,L}, X_{e,H}]$	m	$E_1 = [X_{e,L}, X_{e,1}]$ . . $E_m = [X_{e,m-1}, X_{e,H}]$
X_c	$[X_{c,L}, X_{c,H}]$	n	$C_1 = [X_{b,L}, X_{b,1}]$ . . $C_n = [X_{c,n-1}, X_{c,H}]$

Equation 3.11 expresses the conditional probability of node Xd (parents: Xc and Xe) in Figure 3.2 (Huang et al., 2018).

$$P\left(\frac{Xd=xd,i}{Xc \in Cj \cap Xe \in Ek}\right) = P\left(X_d = x_{d,i} \cap X_c \in C_j \cap X_e \in E_k\right) / P\left(X_c \in C_j \cap X_e \in E_k\right) \quad (3.11)$$

$P(X_d = x_{d,i} \cap X_c \in C_j \cap X_e \in E_k)$ is the probability of $X_d = x_{d,i}$ when the of X_c and X_e is within $\{X_c \in C_j, X_e \in E_k\}$ set and $P(X_c \in C_j \cap X_e \in E_k)$ is probability of X_c and X_e is within $\{X_c \in C_j, X_e \in E_k\}$.

Combining these conditional probabilities results in the creation of conditional probability tables for each node in the Bayesian network. The tables play a crucial role in

identifying the relationship between child nodes and their parent nodes. The table is organized to consider all possible combinations of variable ranges for the parent nodes and the specific child node. By accounting for all these combinations, the conditional probability table quantifies how the child node's probability is affected by its parent nodes across various scenarios.

The model hyperparameters used in the study are (c) search algorithm and (d) number of parents.

3.2.1.3. c Search algorithm

Various types of Hillclimbing and general-purpose search algorithms were employed for structure learning of the Bayesian network in this study (Scanagatta et al., 2019). The Hillclimbing network search algorithms used included Hillclimber, LAGD Hillclimber, K2 and Repeated Hillclimber. On the other hand, the general-purpose network search algorithms considered were Tabu Search, Simulated Annealing and Tree Augmented Naïve (TAN) classifier.

Each algorithm follows a distinct approach in ordering variables for network structure learning. For instance, the Hillclimber algorithm changes arrows without an order until it reaches a good network (Bouckaert, 2004; Scanagatta et al., 2019). Tabu Search, on the other hand, follows the principle of using Hillclimbing until a local optimum is achieved and then selecting the most suitable candidate in the neighborhood (Bouckaert, 2004). These different search algorithms contribute to the exploration of various network structures during the learning process of the Bayesian network.

3.2.1.3. d Number of parents

The number of parents (NP) serves as an upper bound on the number of parents of each node in the learned Bayesian network. NP plays a crucial role in influencing the prediction accuracy of the Bayesian network model.

For the purpose of identifying the impact of input variables on output variables, the study utilized results from Linear Regression and Logistic Regression. These regression analyses allowed the researchers to understand how the input variables contribute to the variation in the output variables. Then, results from the Bayesian Network prediction were analyzed to identify

the relationships between input and output variables. For the visualization of the model coefficients obtained, the study employed Word-Clouds.

3.2.2 Visualization

Word-Cloud is a visualization method commonly used to provide a visual summary of the main themes and recurring terms within a text document. Word-Cloud includes the distinct words of a text. Generally, the font size of each word is linearly proportional to its occurring frequency within the document (Heimerl et al., 2014). Larger font sizes indicate more frequent occurrences, while smaller font sizes represent less frequent ones. The different colors can highlight the other aspects of Word-Cloud.

Figure 3.3 displays the Word-Cloud of the SCOPUS database. The database includes the keywords from 893 publications (Cristino et al., 2022). The Word-Cloud highlights that words like "building," "energy," and "efficiency" have larger font sizes compared to other words, indicating their higher frequency of occurrence among the keywords. Notably, "building" stands out with the largest font size, suggesting it appears most frequently.

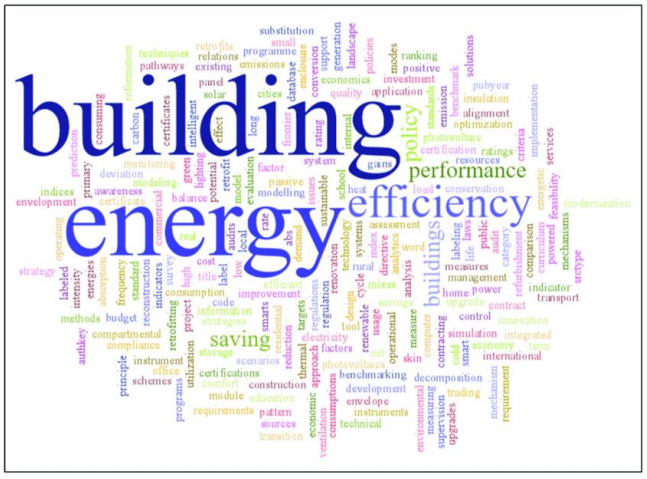


Figure 3.3. Word-Cloud of the SCOPUS database (Cristino et al., 2022)

University of California Irvine (UCI) dataset has been used to interpret data using machine learning algorithms and visualization.

3.2.3 UCI Dataset

Dataset from the University of California Irvine (UCI) Machine Learning Repository contains the energy loads of buildings (UCI Machine Learning Repository, n.d.). UCI dataset includes the energy loads of 12 different building shapes. Combining 18 cubes of 3.5 × 3.5 × 3.5 m^3 dimension generated the simulated building shapes using Ecotect software (Tsanas & Xifara, 2012). All buildings have the same volume of 771.75m^3, but consist of distinct surface areas and dimensions. The material used for all 18 cubes are the same and selected maintaining the lowest U-value. The U-values for building characteristics are 1.780 (walls), 0.860 (floors), 0.500 (roofs), and 2.260 (windows) – these were the default values used by Tsanas and Xifara (2012) to create the dataset.

The location of the buildings is in Athens, Greece. The simulation assumed that each building had seven residents and 70 W of sedentary activity. According to the design specifications, the interior conditions were as follows: clothing: 0.6 Clo, humidity: 60%, air speed: 0.30 m/s, and lighting level: 300 Lux. The buildings had set the internal gain to sensible (5) and latent (2 W/m^2). The infiltration rate set value was 0.5 for a 0.25 air change per hour. The thermal properties of the buildings were mixed mode (95% efficiency), 19-24 ˚C (66.2 -75.2 ˚F) thermostat range, and 15-20 hours of operation (Tsanas & Xifara, 2012).

UCI dataset includes data of 768 buildings with eight input variables and two output variables. The input variables and their ranges are X1: Relative Compactness (0.62 - 0.98), X2: Surface area (514.5 - 808.6) m^2, X3: Wall are (245 - 416.5) m^2, X4: Roof area (110.25 - 220.5) m^2, X5: Overall height (3.5 -7) m, X6: Orientation (2 -5), X7: Glazing are (0 - 0.4), and X8: Glazing are distribution (0 - 5). The two output variables are Y1: Heating Load (6.01 - 43.1) kWh/m^2, and Y2: Cooling Load (10.9 – 48.03) kWh/m^2.

The dataset of 768 buildings used by Tsans & Xifara (2012) comprises a combination of input and output variables that contribute to the characterization of various building scenarios. Firstly, there are 12 distinct types of building shapes, each representing different architectural designs and layouts. Secondly, the dataset includes three different glazing area options, specifically 10%, 25%, and 40% of the floor area, which reflects variations in the amount of glass used in the building's exterior. Additionally, for each glazing area, there are five distribution scenarios that determine how the glazing is distributed on each side of the building. These scenarios include uniform distribution (25% on each side), as well as specific distributions

on the north, east, south, and west sides (55% on one side and 15% on each of the other sides). Finally, the dataset accounts for the rotation of buildings to align with four cardinal points, namely North, East, South, and West.

The data mentioned yields a total of 720 distinct buildings, which is the result of multiplying 12 buildings for three sections (m) and five sections (n) in the data (12 × 3 × 5 × 4). Additionally, considering 12 buildings for four orientations without glazing created 48 buildings (12 × 4). Therefore, the total number of buildings is 768.

The 12 building shapes have different RC values. Equation 3.12 is the mathematical formula for RC (Chou & Bui,2014), including volume (V m^3) and surface area (A m^2) of a building. Surface area is the addition of roof area, wall area, and floor area.

$$RC = \frac{6\,V^{\frac{2}{3}}}{A} \qquad (3.12)$$

Figure 3.4 shows the building shapes and their RC values (Chou & Bui, 2014). According to Figure 3.4, RC values of the building shapes are 0.98, 0.90, 0.86, 0.82, 0.79, 0.76, 0.74, 0.714, 0.69, 0.66, 0.64, and 0.62.

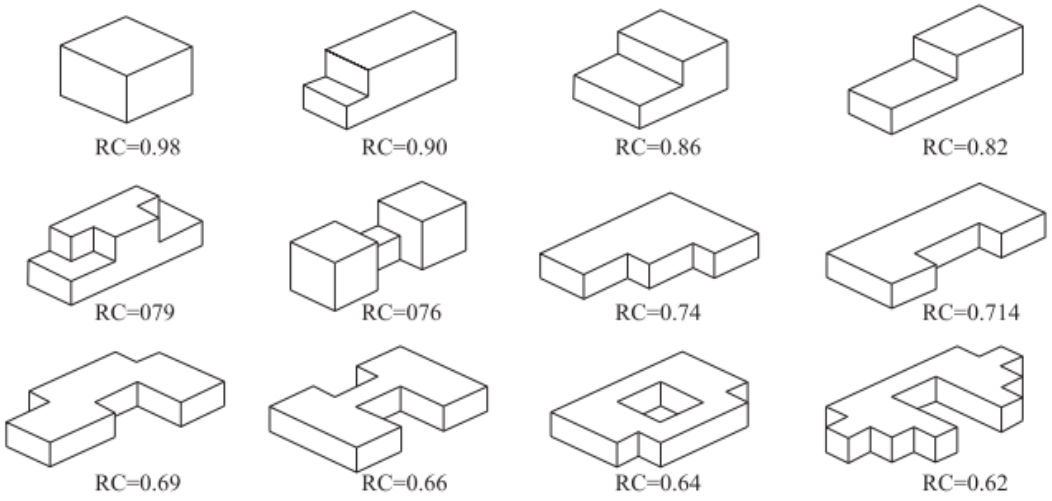


Figure 3.4. Building shapes of UCI dataset according to their Relative Compactness values (Chou & Bui, 2014)

To create the energy efficiency classes needed for classification algorithms, the discretization process was used.

3.2.4 Data Discretization

The output variables of research questions 2 and 3 are energy efficiency classes of buildings. Therefore, discretizing UCI dataset created energy efficiency classes. The discretization process involves two main steps, as described below.

1) Arrange the data in ascending order

2) Divide the data into five classes with equal ranges (Aqlan et al., 2014)

Since the UCI dataset has two outputs: HL and CL, discretization process applied for both outputs separately. The purpose of dividing output loads into five classes is to increase the scope of energy efficiency.

The same procedure was used to create three energy efficiency classes for the results comparison.

3.2.5 Workflow

3.2.5.1 Identifying the impact of input variables on output variables

Figure 3.5 shows the workflow of research question 2 to identify the impact of input variables on HL and CL. Figure 3.5 shows that Linear regression predicted the HL and CL, analyzed using RMSE and MAE. Then the Logistic regression predicted the HL and CL energy efficiency classes using the discretized data. Accuracy was the evaluation metric for Logistic regression predictions. Testing condition were10-fold cross-validation. Finally, model coefficients obtained from Linear and Logistic regression were analyzed and visualized.

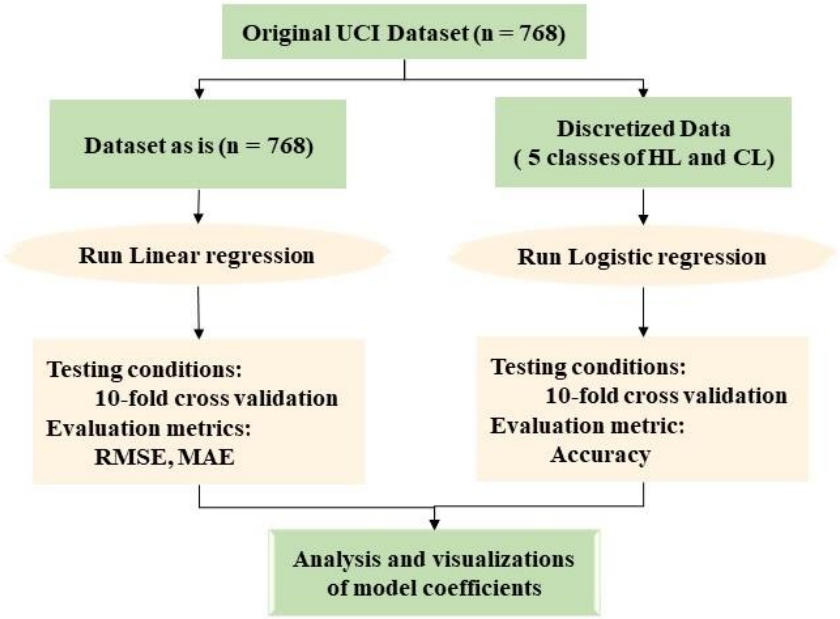


Figure 3.5. Workflow for identifying the impact of input variables on output variables

3.2.5.2 Analyzing relationship between input and output variables

Figure 3.6 shows the workflow of the procedures in research question 3, which is analyzing the relationship between input and output variables. The Bayesian network classifier used the discretized UCI dataset with five energy efficiency classes to test predictions with 10-fold cross-validation. Accuracy evaluated the prediction results. Research question 3 analyzed the Bayesian network graphical structures by analyzing the impact of model hyperparameters. Finally, the study analyzed the conditional probability tables of nodes. Finally, the same procedure used three discretized classes for the results comparison.

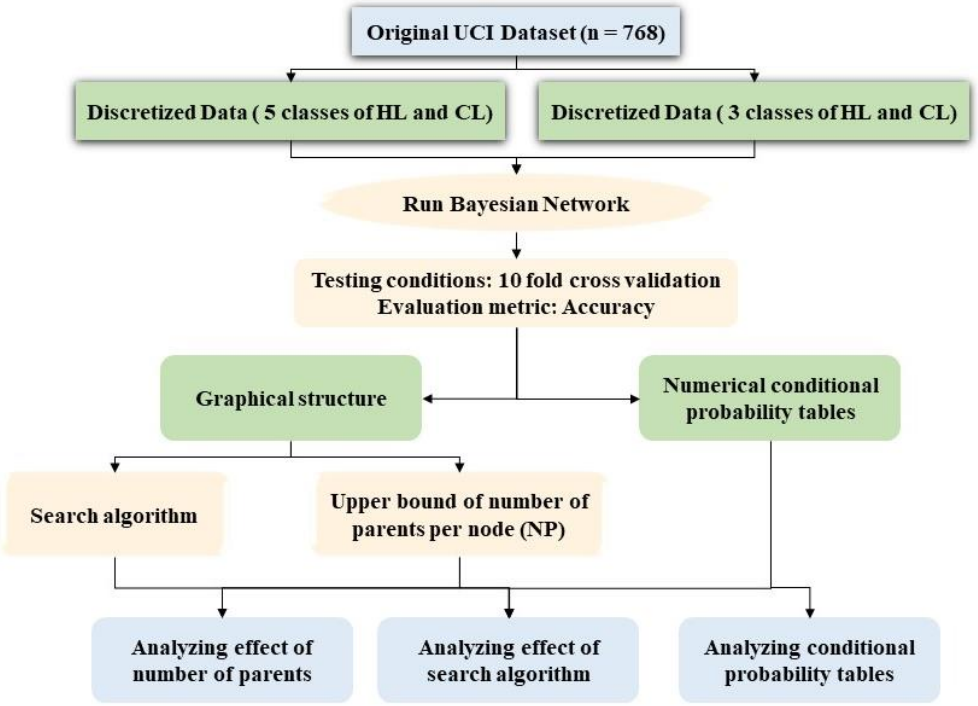


Figure 3.6. Workflow for Analyzing relationship between input and output variables

3.3 Impact of location on energy consumption

3.3.1 Mann-Whitney-Wilcoxon rank sum test

The Wilcoxon rank sum test is a statistical test that does not assume data is normally distributed (nonparametric test) (Barros et al., 2018). The test uses to determine whether the two independent samples come from the same distributed populations by comparing dataset medians. Therefore, Wilcoxon rank sum test is equivalent to two-sample t-test (Doorn et al., 2020; Jiang et al., 2020).

Independent samples with continuous, non-normal distribution can apply the Wilcoxon-rank sum test (Barros et al., 2018). The null and alternative hypothesis of the test are as below (Barros et al., 2018).

- Null hypothesis: Two samples have the same distribution.
- Alternative hypothesis: Two samples do not have the same distribution.

The test statistic of the Wilcoxon test is U and calculate using summing r^x or r^y and subtracting $\frac{n_x(n_x+1)}{2}$ or $\frac{n_y(n_y+1)}{2}$ respectively, where r_i^x is the rank of x_i, and r_i^y is the rank of y_i. The test checks the difference between two groups by comparing U with the values correspond to no difference (Perolat et al., 2015).

Wilcoxon- rank sum test checks whether there is a difference between building cooling energy usage between states cities and Lafayette. The U value considered to check no difference is 0.05. Commercial Load Data (CLD) was used to interpret the results using Wilcoxon rank-sum test.

3.3.2 CLD Dataset

Commercial Load Data (CLD) (Ong & Clark, 2022) contains hourly load profiles for both commercial and residential buildings across all Typical Meteorological Year 3 (TMY3) locations in the United States. The dataset was initially created around 2012 to support various solar water heating and photovoltaic analyses.

The CLD dataset comprises weather data, commercial load profile data, and residential load profile data. However, in the study being referred to, only the commercial load profile data was utilized.

Weather data from 10 states, representing one each from nine census divisions and Indiana as the reference state were selected. Building cooling usage for electricity (kWh) was the weather data variable selected in two different building types: midrise apartments and supermarkets.

Table 3.4 summarizes the CDL dataset of midrise apartments. Table 3.4 includes the states and their symbols, cities considered in each city, minimum, maximum, median, and mean values of building cooling usage for electricity.

Table 3.4. Summary of CLD dataset – Midrise apartments

State	Symbol	City	Min (kWh)	Max (kWh)	Median (kWh)	Mean (kWh)
Indiana	IN	Lafayette	0	532.722	19.746	80.792
		Indianapolis	0	543.634	22.424	88.433
Illinois	IL	Willard	0	583.947	17.405	85.409
California	CA	San Francisco	1.995	177.380	42.794	49.149
Arizona	AZ	Tucson	0	548.267	128.500	180.678
Texas	TX	Austin	0	579.475	150.794	204.081
South Dakota	SD	Brookings	0	445.307	8.083	54.675
Tennessee	TN	Knoxville	0	460.722	45.125	113.209
North Carolina	NC	Durham	0	463.705	62.086	119.996
New York	NY	Rochester	0	403.692	7.604	63.451
Massachusetts	MA	Boston	0	478.348	11.898	62.159

Similarly, Table 3.5 shows the summary of supermarkets in CDL dataset. Table 3.5 contains the state, symbol, city, minimum, maximum, median, and mean values of building cooling usage for electricity.

Table 3.5. Summary of CLD dataset – Supermarkets

State	Symbol	City	Min	Max	Median	Mean
Indiana	IN	Lafayette	0	85.391	0	2.967
		Indianapolis	0	77.235	0	3.158
Illinois	IL	Willard	0	86.596	0	3.329
California	CA	San Francisco	0	17.712	0	0.259
Arizona	AZ	Tucson	0	84.271	0	8.946
Texas	TX	Austin	0	85.897	1.458	10.246
South Dakota	SD	Brookings	0	63.487	0	1.788
Tennessee	TN	Knoxville	0	64.272	0.047	4.463
North Carolina	NC	Durham	0	68.892	0.115	4.783
New York	NY	Rochester	0	56.788	0	2.054
Massachusetts	MA	Boston	0	51.983	0	1.943

3.3.3 Workflow for analyzing impact of location on energy consumption

Research question 4 follows the following steps for analysis.

1. Selected 11 different cities in the US in ten states.
2. Divided data according to four seasons.
3. Performed Wilcoxon- rank-sum test for all the data and for four seasons separately.
4. Compare the U value with the error margin.

3.4 Data Split

Two different data splitting techniques for predictions were used in this study. The g techniques are train-test split and k-fold cross validation.

Research questions 2 and 3 used10-fold cross-validation. For research question 1, the data splitting technique was train-test split. The training data included 75% of the total data, and 25% was testing data.

3.5 Data Evaluation Metrics

The evaluation metrics are accuracy, mean absolute error (MAE), and root mean squared error (RMSE). Equations 3.13, 3.14, and 3.15 show the mathematical formulas for evaluation metrics (Li et al., 2021; Kabir et al., 2017; Roostaei et al., 2021).

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ predictions} \times 100\% \qquad (3.13)$$

$$MAE = \sum \frac{|Predicted\ output - Actual\ output|}{Total\ predictions} \qquad (3.14)$$

$$RMSE = Sqrt\ (\sum \frac{|Predicted\ output - Actual\ output|^2}{Total\ predictions}) \qquad (3.15)$$

3.6 Software Implementation

The study used Waikato Environment for Knowledge Analysis (WEKA) workbench (Holmes et al., 1994) for data analysis. The workbench is a popular and powerful data mining and machine learning software that provides a comprehensive set of tools and algorithms for data analysis. The WEKA workbench encompasses a large number of machine learning models, making it a versatile tool for various data analysis tasks. These algorithms include classification, regression, clustering, and association rule mining, among others. Researchers can use these algorithms to build predictive models, discover patterns in the data, and gain valuable insights.

Python is the programming language used to predict and get explainable artificial intelligence outputs using the Shapash library and to run Wilcoxon-rank-sum test.

3.7 Chapter Summary

Three datasets were used to analyze the impact of building features on energy consumption. Linear Regression, Random Forest, Logistic Regression, and Bayesian Networks predicted energy consumption using input features. Data discretization and statistical tests such as Mann-Whitney- Wilcoxon rank sum test made it possible to obtain accurate results. Python and WEKA implemented the data analysis, and finally, the chapter includes the workflow for four research questions.

CHAPTER 4 . RESULTS

4.1 Explainable Artificial Intelligence

4.1.1 Feature Importance

Running the Random Forest regressor resulted in a feature importance graph that visualized the order of the feature's importance and their contributions. Figure 4.1 shows the feature importance graph and contributions from each feature for predicting cooling EUI using CBECS 2018 dataset. All ten runs with Random Forest obtained the same feature importance graphs as Figure 4.1.

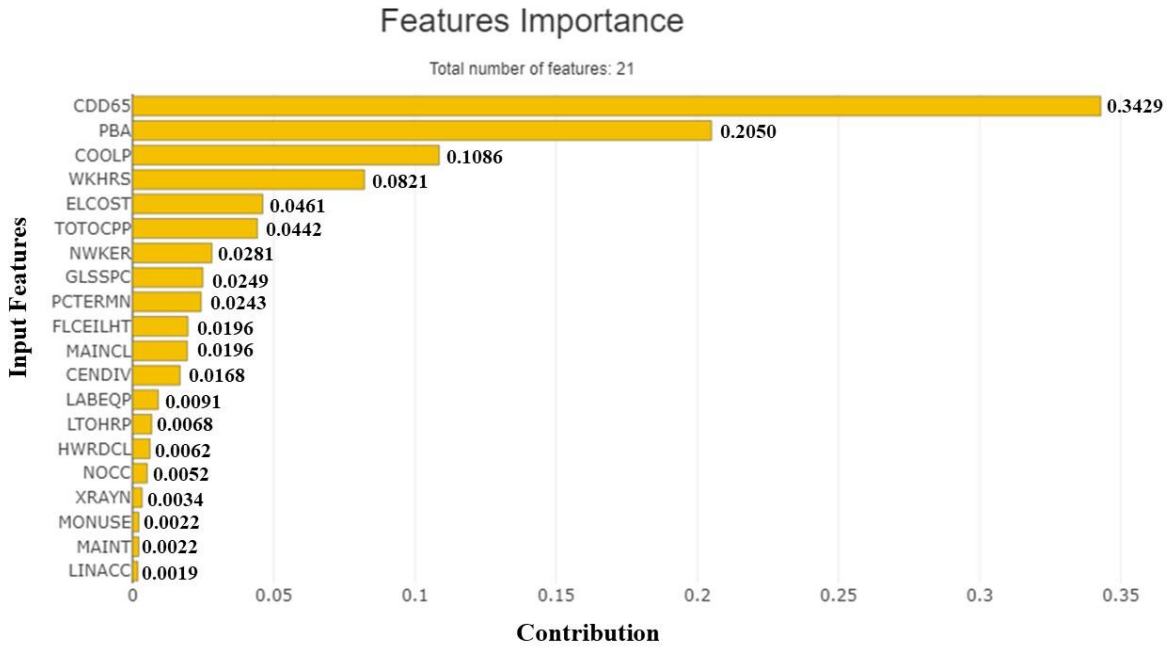


Figure 4.1. Shapash CBECS 2018 dataset feature importance graph and contributions for predicting cooling EUI

The results indicate that for the CBECS 2018 dataset, cooling degree days (CDD65) is the topmost important feature, with a contribution of 0.3429, which means that cooling degree days change the predicted values of absolute cooling EUI probability by 34.29% on average. The second most important feature is principle building activities (PBA), and the third is cooling percentage (COOLP). Features WKHRS, ELCOST, TOTCPP, NWKER, GLSSPC, PCTERMN,

FLCEILHT, MAINCL, CENDIV, LABEQP, LTOHRP, HWRDCL, NOCC, XRAYN, MONUSE, MAINT, and LINACC have the importance orders from 4th to 20th respectively. Since Shapash shows only 20 inputs, the least important feature to predict cooling EUI using CBECS 2018 data is the area of the data center or servers (DCNTRSFC).

The first two important features, cooling degree days and principle building activity contribute to 54.79% of the impact to predict cooling EUI in the CBECS 2018. When considering the first five features, cooling degree days, principle building activity, cooling percentage, cost of electricity, and total occupancy, made an 82.89% contribution.

Figure 4.2 shows the feature importance graph and contributions from each feature for predicting cooling EUI using CBECS 2012 dataset. The feature importance graphs are the same for all ten runs with Random Forest.

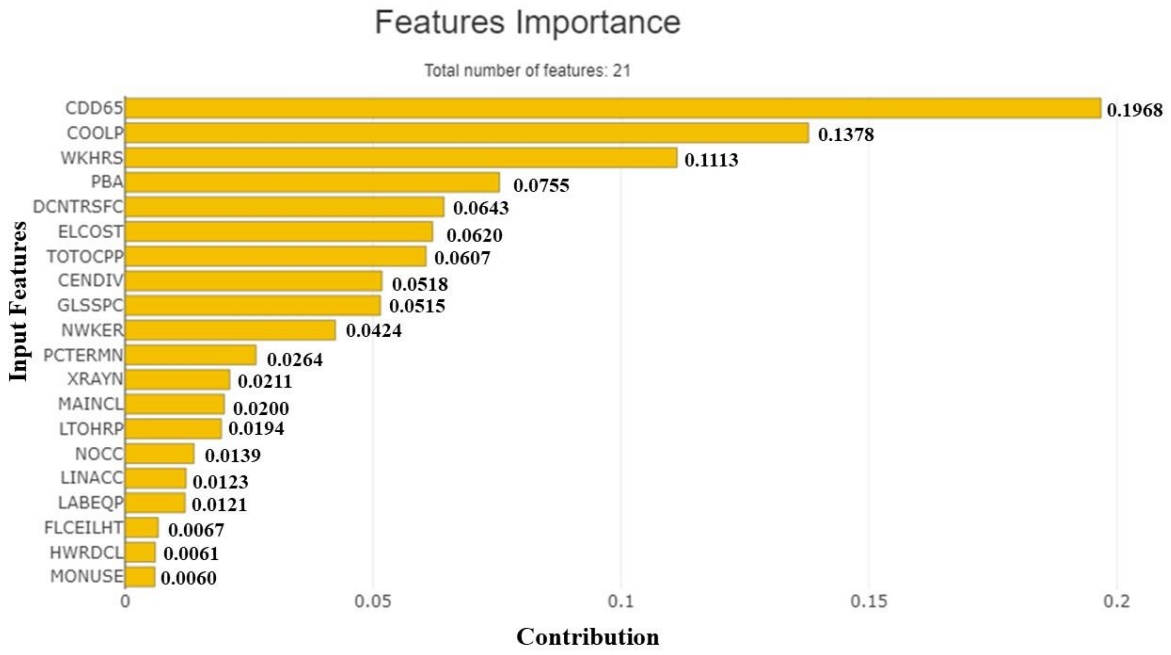


Figure 4.2. Shapash CBECS 2012 dataset feature importance graph and contributions for predicting cooling EUI

According to Figure 4.2, cooling degree days (CDD65) is the topmost important feature with a contribution of 0.1968. The results indicate that cooling degree days change the absolute cooling EUI probability predicted on average by 19.68%. The second important feature is the percent cooled (COOLP) with a 0.1378 contribution, and the third is the total hours open per

week (WKHRS) with a contribution value of 0.1113. These results indicate that cooling degree days, percent cooled, and total hours open per week are the top three important features for predicting cooling EUI. DCNTRSFC, ELCOST, TOTOCPP, CENDIV, GLSSPC, NWKER, PCTERMN, XRAYN, MAINCL, LTOHRP, NOCC, LINACC, LABEQP, FLCEILHT, HWRDCL have the feature importance order numbers from 4th to 20th respectively. The feature importance order indicates that the importance of predicting cooling EUI decreases from the area of data centers or servers (DCNTRSFC) to how cooling is reduced in a 24-hour cycle (HWRDCL). The 20th feature in the order is the months of the year the building was in use (MONUSE) with a 0.006 contribution. The results indicate that MONUSE can change the absolute cooling EUI probability predicted on average by 0.6%. However, Shapash shows only the top 20 input features in the feature importance graph. Since Figure 4.2 does not include the variable MAINT and the dataset has only 21 input features, regular maintenance of HVAC system (MAINT) has the least order number, which is 21.

However, adding cooling degree days, cooling percentage, total hours open per week, and principle building activity contributes 0.5214, indicating that four features contribute to almost 50% of cooling EUI. The first two most important features only contribute to 33.46%. Moreover, the first nine features contribute to 81.17% of cooling EUI.

4.1.2 Comparison of Feature Importance

Figure 4.3 shows the heatmap comparing each feature's importance order obtained with CBECS 2018 and 2012. The yellow color represents the obtained order number for each feature for CBECS 2018 dataset, and it is in blue for CBECS 2012. If the importance order is the same for both years, points are in red on the heat map.

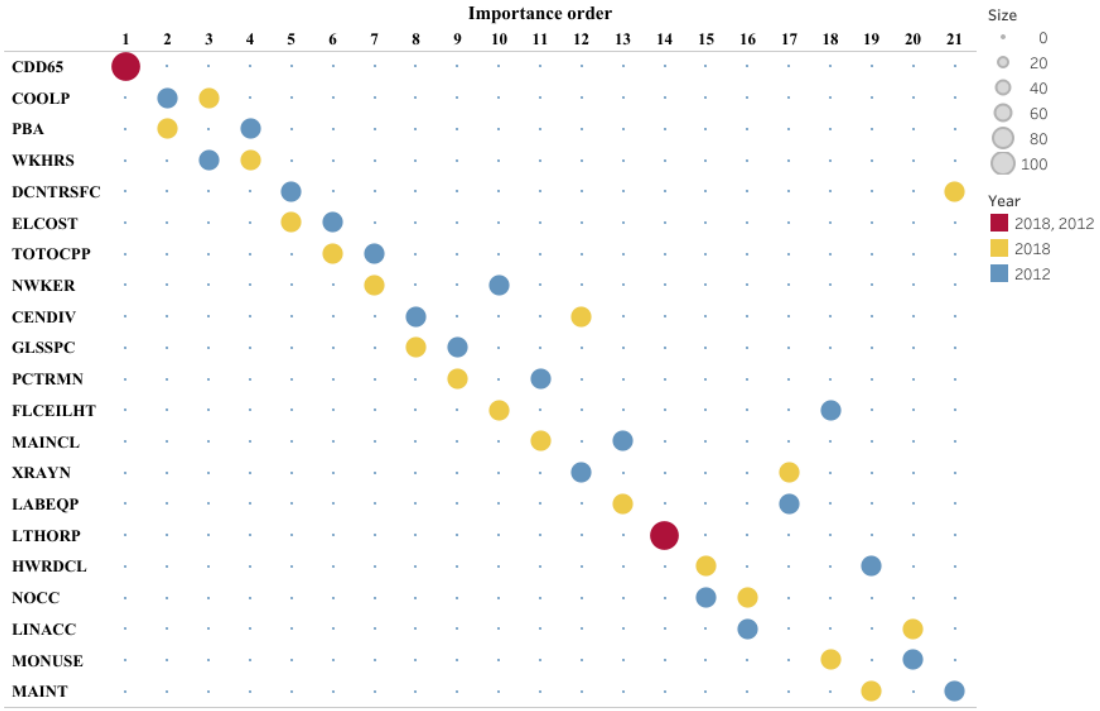


Figure 4.3. Heatmap of feature importance order from CBECS 2018 and 2012

As shown in Figure 4.3, cooling degree days (CDD65) and LTHORP (percent lit when open) have the same order numbers, 1 and 14, in 2018 and 2012. Cooling degree days have the importance number 1, indicating that it is the most important feature in predicting cooling EUI in both 2018 and 2012. CDD65 estimates the energy required for cooling during a warm season. The cooling degree days directly reflect a building's cooling load. A building's cooling system has to work harder when the outside temperature is high to maintain a comfortable indoor temperature (Santamouris, 2016). Other factors, such as the cooling equipment (MAINCL), months in use (MONUSE), and number of businesses (NOCC) can also impact energy usage. However, these factors are more static and typically don't vary as much daily as CDD65. Therefore, CDD65 is the most important feature in predicting cooling EUI in both 2018 and 2012. Furthermore, the percentage of buildings that are illuminated when they are open has an impact on building cooling usage, but it has the importance number 14th in both years.

However, cooling percentage (COOLP), total hours open per week (WKHRS), area of data center or server farm (DCNTRSFC), census divisions (CENDIV), number of Xray machines (XRAYN), number of businesses (NOCC), linear accelerators (LINACC) obtained a lower importance in 2018 compared to 2012, in predicting cooling EUI.

Cooling percentage has the importance numbers 3 in 2018 and 2 in 2012. The change in importance can be attributed to the increased usage of energy-efficient windows and improved building insulation, which reduces heat transfers (Amirifard et al., 2019; Khatibi et al., 2023). Total hours open per week also has importance numbers 4 in 2018 and 3 in 2012. As remote work has increased and automated systems have developed, such as turning off lights in unoccupied buildings, this could contribute to reducing the importance in 2018 (Aste et al., 2017). Furthermore, the area of data centers or servers achieved 21st in 2018 and 5th in 2018 as importance. Data centers or servers had a less significant impact on cooling EUI in 2018, possibly due to the reduced server density and the increased use of cloud computing. Because data centers needed more physical space, cloud computing recently helped reduce cooling EUI by using off-site data centers (Pradhan et al., 2016; Riahi, 2015). Census divisions may have had different climates during times due to climate change. Also, building codes and standards in census divisions have been updated to reduce energy usage (Levinson, 2016). Therefore, possibly due to these factors, census division was the 12th important feature to predict cooling EUI in 2018 and 8th in 2012. Number of Xray machines has the importance of 17 and 12 in 2018 and 2012. The change in XRAYN can be due to recent efficient X-ray machines and shifting to digital X-ray systems which reduces the number of machines needed. The number of businesses has an importance of 16 in 2018 and 15 in 2012. Since there could be more commercial activities in the buildings which produce more heat in 2012, buildings might have used more cooling energy. As for linear accelerators, their importance number in 2018 is 20, down from 16 in 2012, might be due to the development of energy-efficient linear accelerators, which reduce energy usage.

According to Figure 4.3, principle building activity (PBA), cost of electricity (ELCOST), total occupancy percent (TOTOCPP), number of workers (NWKER), glass percentage (GLSSPC), number of computers (PCTERMN), floor-to-ceiling height (FLCEILHT), main cooling equipment (MAINCL), laboratory equipment (LABEQP), how cooling reduced in a 24-hour cycle (HWRDCL), months of the year the building was in use (MONUSE), regular maintenance for HVAC systems (MAINT) obtained a higher importance number in 2018 than 2012 .

The increased complexity of buildings, such as their size, number of occupants, and urbanization, can be reasons to increase the principle building activity importance to 2 in 2018

from 4 in 2012 (Delzendeh et al., 2017; Cao et al., 2016). Also, the increasing temperature with climate change created a rise in air conditioning which can lead to higher electricity consumption and cooling cost of electricity (Andrić et al., 2021). Total occupancy has more importance in predicting cooling EUI in 2018 than 2012. Increasing population growth may have contributed to this by increasing the use of space in modern buildings, resulting in more people per square foot (Cao et al., 2016). Moreover, with the increase in flexible working hours, many people are working longer than in the past, increasing the cooling time period.

With the growth of the service sector, the number of people working in buildings has increased. For example, in the CBECS 2018 dataset, the NWKER range is 0-7500, and 0-6500 in CBECS 2012. Compared to 2012, NWKER has a greater importance in 2018 due to an increase in the number of workers requiring more space and, therefore, a greater need for cooling. Modern buildings increasingly feature extensive glass facades, with the increased use of glass. Glass allows natural light into the building and increases its openness. However, glasses also allow more heat to enter, increasing the building's cooling load (Alam & Islam, 2017). Compared to 2012, people used more computers in 2018, and they generate heat. Not just the computers installed in buildings generate heat; personal laptops also increase cooling energy consumption. Aside from this, today's computers are more likely to generate heat than their predecessors, especially high-performance computers used in data centers.

Considering the building sizes, generally, buildings in 2018 could be taller than in 2012 and often feature higher ceilings to create a more open and spacious feel. Maintaining a comfortable indoor temperature needs more space to cool and more air to cool (Senarathne et al., 2022). Therefore, higher FLCEILHT increases the cooling energy. Therefore, FLCEILHT is the 10th most important feature in 2018 and 18th in 2012.

Figure 4.3 shows that cooling reduced in a 24-hour cycle obtained an importance of 19 and 15 in 2018 and 2012, respectively. The result is that earlier, the buildings controlled its cooling systems by turning them on and off using a simple switch. However, modern buildings may use more advanced systems that adjust cooling levels throughout the day based on occupancy and external conditions (Amirifard et al., 2019). The advancements allow for precise control and reduce cooling energy usage. Main cooling equipment has more options in 2018 than in 2012. For example, MAINCL did not include modern cooling equipment, such as split systems and fuel/oil/kerosene thrillers, in CBECS 2012. Also, with the global temperature rise,

cooling equipment needs to supply more energy to maintain the required temperature increasing its impact on cooling EUI. The increased impact from laboratory equipment to cooling EUI in 2018 can be with the use of more laboratory machines and research facilities with the development of science. Additionally, modern laboratory equipment, such as freezers and incubators require significant energy to operate, which increases the impact of cooling EUI. The number of months the building was in use also has a higher importance in 2018 than in 2012, because with the temperature changes, buildings need more cooling energy to maintain requirements (Campagna & Fiorito, 2022). Therefore, even if the building used the same number of months in both years, more cooling energy is needed in 2018 with the high energy requirement.

Regular maintenance is essential to ensure optimal equipment performance. However, when it comes to building cooling energy usage, regular maintenance of equipment may have the least impact because there are several other factors, such as building layout that have a more significant effect on energy usage. Even with well-maintained equipment, the cooling system is inefficient if the building is not properly insulated. Due to the development of well-insulated buildings, MAINT has a lower impact on cooling EUI in 2018 than in 2012. Therefore, MAINT has important numbers 21 and 19 in 2018 and 2012.

4.1.3 Comparison of feature contributions

Figure 4.4 compares contributions from each input feature to predict cooling EUI, in 2018 and 2012. Here, the features are arranged from left to right using the feature importance in 2018. The reason for arranging it as 2018 is that CBECS 2018 is the most recent dataset, which could be more accurate. Of all the 21 inputs, four features have contributed more than 5% in both years. The features are CDD65, PBA, COOLP, and WKHRS. According to the results, overall, the most important features impacting cooling EUI in commercial buildings are cooling degree days, principle building activity, cooling percentage, and total hours open per week.

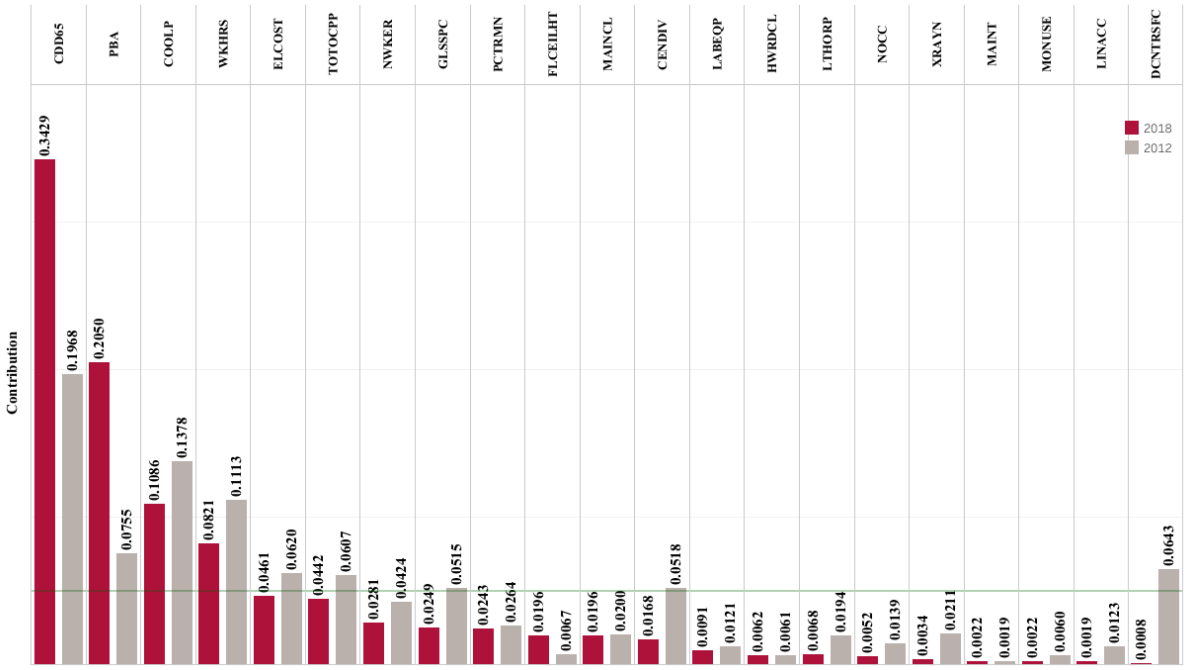


Figure 4.4. Contributions from features in CBECS 2018 and 2012 to predict cooling EUI

However, according to a contribution comparison of features, such as cost of electricity, total occupancy percentage, number of workers, glass percentage, number of computers, main cooling equipment, laboratory equipment, and months of usage, they contributed less in 2018 than in 2012. However, they have a higher importance in 2018 compared to 2012. The observation is because of the comparison of two independent datasets, and their calculated contribution is relative to the specific dataset.

4.2 Machine Learning Applications

4.2.1 Discretized UCI dataset

Figure 4.5 shows the histogram created for heating load by the discretization process with five classes using UCI dataset. In Figure 4.5, class 1 includes 207 buildings data and 192 for class 2. Classes 3, 4, 5 include 98, 166, and 105 building data respectively.

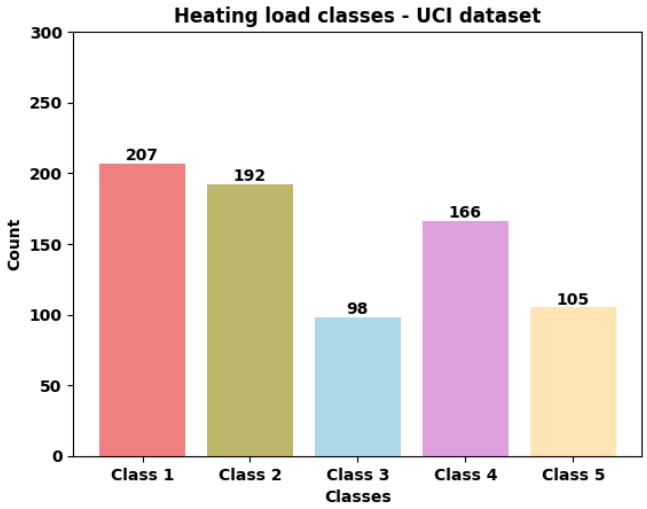


Figure 4.5. Heating load classes for UCI dataset

Figure 4.6 shows the histogram for cooling load with five classes. As of Figure 4.6, class 1 to 5 in cooling load has 322, 90, 167, 154, and 35 building data respectively.

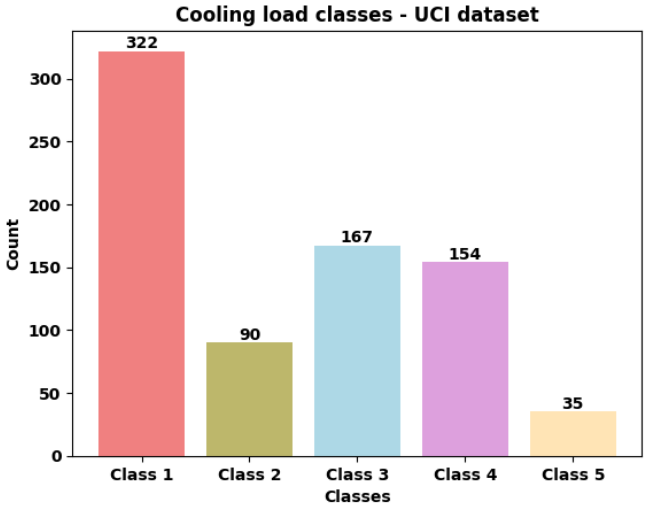


Figure 4.6. Cooling load classes for UCI dataset

Table 4.1 shows the names of the five classes, their ranges, and the count of building data in each class for heating and cooling.

For example, Class 1 is named "very high-energy efficiency" and includes the lowest range of heating (6.01 - 13.428 kWh/m^2) and cooling loads (10.9 - 18.326 kWh/m^2). Buildings in this class require a relatively small amount of energy to achieve the desired indoor temperature,

making them highly efficient in their energy usage. On the other hand, Class 5 is named "very low-energy efficiency" and comprises the highest ranges of heating (35.682 – 43.1 kWh/m^2) and cooling loads (40.604 – 48.03 kWh/m^2). Buildings in this class need a higher amount of energy to regulate temperature when there are outdoor temperature changes, indicating low efficiency in their energy consumption.

Table 4.1. Summary of discretized data - UCI dataset (five classes)

Class	HL class range kWh/m^2	HL count	CL class range kWh/m^2	CL count
Very high energy efficiency	6.01 - 13.428	207	10.9 - 18.326	322
High energy efficiency	13.428 - 20.846	192	18.326 - 25.752	90
Medium energy efficiency	20.846 - 28.264	98	25.752 - 33.178	167
Low energy efficiency	28.264 - 35.682	166	33.178 - 40.604	154
Very low energy efficiency	35.682 - 43.1	105	40.604 - 48.03	35

Table 4.2 shows the discretized UCI data for three classes and their ranges. The purpose of creating 3 classes is to compare the results.

Table 4.2. Summary of discretized data - UCI dataset (three classes)

Class	HL class range (kWh/m^2)	CL class range (kWh/m^2)
Very high energy efficiency	6.01 – 18.373	10.9 – 23.276
Medium energy efficiency	18.373- 30.736	23.276 – 35.653
Very low energy efficiency	30.736 - 43.1	35.653 - 48.03

4.2.2 Impact of building design variables on energy usage

4.2.2.1 Linear Regression

Equations 4.1 and 4.2 represent the linear equations derived from Linear Regression. Each equation consists of input variables along with their corresponding model coefficients, which determine the impact of each variable on the output (HL or CL). The magnitude of each coefficient represents its impact on the output variable HL. Larger coefficients indicate a higher impact, while smaller coefficients indicate a smaller impact. For example, the highest coefficient

is for Relative Compactness (64.77), which is negative, indicating that increasing Relative Compactness by one unit will decrease the heating load by 64.77 units. Conversely, a decrease in Relative Compactness by one unit will increase the heating load by 64.77 units.

X2 and X4 have negative coefficients, where X3,X5,X7,and X8 have positive coefficients for predicting HL. Orientation (X6) is not included in the equation, suggesting that it does not impact the HL prediction.

$$Y1 = -64.77X_1 - 0.043X_2 + 0.016X_3 - 0.09X_4 + 4.17X_5 + 19.93X_7 + 20.38X_8 + 83.93 \quad (4.1)$$

Equation 4.2 shows that X1, X2, and X4 have negative coefficients, indicating that increasing these variables decreases the predicted cooling load. On the other hand, X5 and X7 have positive coefficients, meaning that an increase in these variables will result in an increase in the predicted cooling load. However, X3, X6, and X8 have coefficients of zero, indicating that these variables do not have any impact on the prediction of cooling load. Relative compactness has the highest (70.79) coefficient for predicting cooling load in a negative direction. If relative compactness increases by one unit, the cooling load will decrease by 70.79 units.

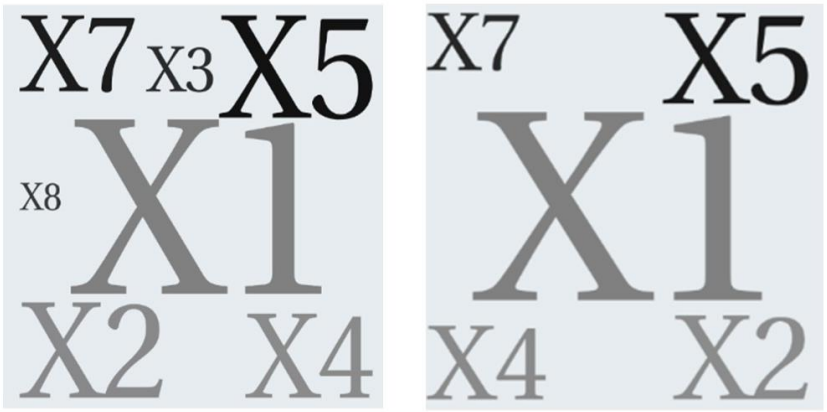
$$Y2 = -70.79X_1 - 0.04X_2 - 0.09X_4 + 4.28X_5 + 14.82X_7 + 93.76 \quad (4.2)$$

Figure 4.7(a) represents the Word-Cloud visualization of the coefficients obtained from predicting the heating load using Linear Regression. The visualization highlights the positive coefficients (X3, X5, X6, X8) using a dark color and the negative coefficients (X1, X2, X4) in a light color. The largest magnitude among the negative coefficients in Figure 4.7 (a) is for X1 (Relative Compactness), indicated by the largest font size in light color. The results indicate that Relative Compactness impacts the HL most compared to other variables.

Figure 4.7(b) shows the Word-Cloud of the Linear Regression model coefficients for predicting the cooling load, with negative (X1, X2, X4) and positive (X5, X7) coefficients.

The comparison between Figures 4.7(a) and 4.7(b) shows that increased values of X1 (Relative Compactness), X2 (Surface Area), and X4 (Roof Area), as well as decreased values of

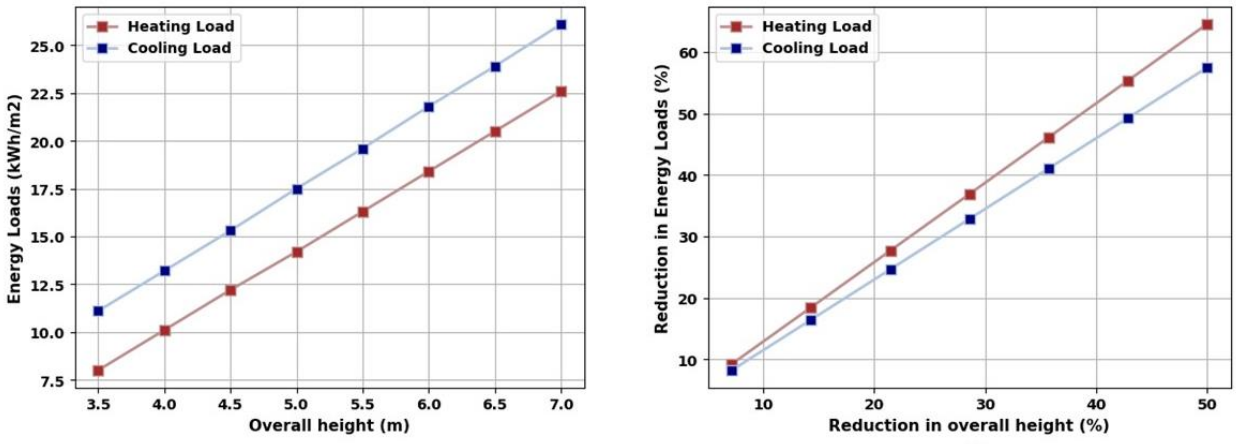
X5 (Overall Height) and X7 (Glazing Area Ratio), result in reduced energy usage for both heating and cooling in buildings.



(a) Heating Load / Y1 (b) Cooling Load / Y2

Figure 4.7. Word-Cloud for Linear regression model coefficients

Figure 4.8 (a) shows the changes in HL and CL with overall height. According to Figure 4.8 (a), HL and CL increase with the overall height. For example, for a fixed dataset, HL and CL are 22.6 and 26.08 for X5 =7m and 8.00 and 11.09 for X5= 3.5m. Figure 4.8 (b) shows the energy reduction percentage with the reduction percentage of overall height. Figure 4.8 (b) shows that a 50% overall height reduction results in 64.56% HL reduction and 57.47% CL reduction.



(a) (b)

Figure 4.8. Changes of heating and cooling loads with overall height (a) Overall height vs Energy load (b) Overall height percentage reduction vs Energy load percentage reduction

Similarly, Figure 4.9 (a) shows the energy loads change with Glazing area ratio, where HL and CL are proportional to glazing area ratio. For example, X7 values of 0.4 and 0.1 obtained 30.57, 24.59 (HL), and 32.01, 26.08 (CL). Figure 4.9 (b) shows the percentages of energy reduction to glazing area ratio reductions. As of Figure 4.9 (b), 75% of glazing area ratio creates a 19.5% reduction in HL and 13.88% in CL.

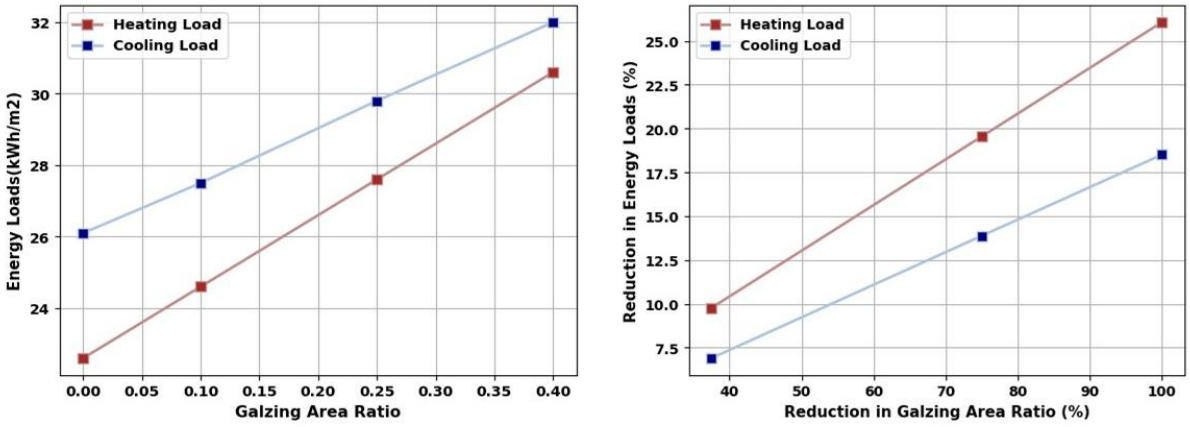


Figure 4.9. Changes of heating and cooling loads with glazing area ratio (a) Glazing area ratio vs Energy load (b) Glazing area ratio percentage reduction vs Energy load percentage reduction

Figure 4.10 (a) shows the impact of relative compactness on HL and CL. However, unlike overall height and glazing area ratio, energy loads decrease with the increase of relative compactness. Figure 4.10 (b) shows how much HL and CL increase with the reduction of RC. According to the figures, a 36.73% reduction of RC creates a 103.16% increase in HL and 97.68% in CL.

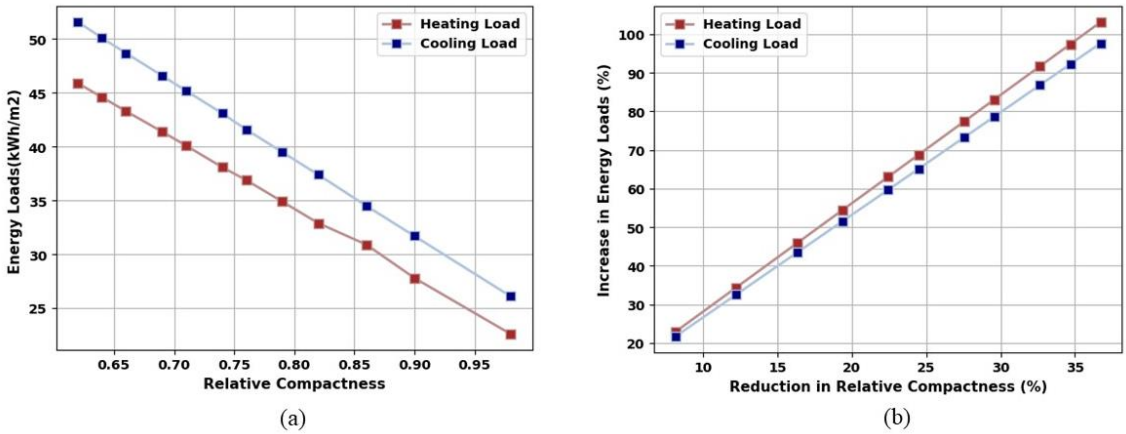


Figure 4.10. Changes of heating and cooling loads with relative compactness (a) RC vs Energy load (b) RC percentage reduction vs Energy load percentage reduction

Table 4.3 shows the evaluation metrics (RMSE and MAE) of Linear regression predictions. According to Table 4.3, input variables predicted heating load with 2.95 (RMSE) and 2.09 (MAE) values. For cooling load predictions, metrics values were 3.23 (RMSE) and 2.27 (MAE).

Table 4.3. Evaluation metrics for Linear Regression predictions for HL and CL

	RMSE/ kWh/m^2	MAE / kWh/m^2
Heating load	2.95	2.09
Cooling load	3.23	2.27

4.2.2.2 Logistic Regression

Table 4.4 shows the model coefficients from Logistic Regression for predicting HL. Table 4.4 includes coefficients for five energy efficiency classes.

Table 4.4. Model Coefficients for Logistic Regression HL classes predictions

Y1	X1	X2	X3	X4	X5	X6	X7	X8	Constant
Very high efficiency	0	0	0.01	-0.01	-1.31	0	8.33	-0.22	0
High efficiency	-0.63	-0.01	-0.02	0.01	1.52	0	-8.61	0.17	0
Medium efficiency	0	-0.01	-0.01	-0.02	1.01	-0.02	2.89	0.09	0
Low efficiency	0	0.01	-0.04	0.08	-0.76	-0.01	-18.14	0.04	0
Very low efficiency	-17.09	-0	0.01	-0.02	1.6	-0.01	6.68	-0.02	0

Table 4.5 shows the model coefficients for CL predictions and includes coefficients for five energy efficiency classes.

Table 4.5. Model Coefficients for Logistic Regression CL classes predictions

Y2	X1	X2	X3	X4	X5	X6	X7	X8	Constant
Very high efficiency	0	0	0.01	-0.01	-0.56	-0.05	-1.87	-0.07	0
High efficiency	0	-0.01	-0.03	0.01	1.92	-0.05	-1.85	0.04	0
Medium efficiency	-5.13	0	0.02	-0.05	0.53	0.02	4.56	0.09	-1.21
Low efficiency	0	0	-0.03	0.06	-1.01	-0	-0.13	-0.1	0
Very low efficiency	-3.39	-0	0	-0.01	0.5	0	0.24	-0.01	-1.55

Table 4.6 is the Word-Cloud visualization for coefficients from Tables 4.4 and 4.5. Table 4.6 includes the model coefficients for five energy efficiency classes predicting HL and CL.

According to Table 4.6 the common observations compatible with linear regression results are X5, and X1. Considering X5, in very high energy efficiency class, it acts as a negative variable. However, X5 acts as a positive impacting variable and X1 (RC) as a negative impacting variable in very low energy efficiency class.

Table 4.6. Word-Cloud for Logistic Regression model coefficients

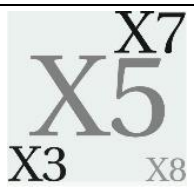
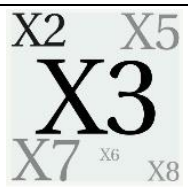
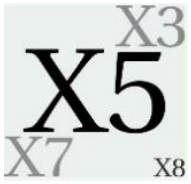
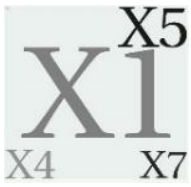
Range	HL	CL
Very high energy efficiency		
High energy efficiency		
Medium energy efficiency		
Low energy efficiency		
Very low energy efficiency		

Figure 4.11 (a) shows the changes of the probability of success in very high energy efficiency class with overall height. According to Figure 4.11 (a), the probability of success in very high energy class decreases with the overall height. For example, X5 =3.5 obtained 0.012 of probability and 0.00013 for X5=7 for HL. Figure 4.11 (b)shows the increase of probability percentage with a reduction percentage of overall height. According to Figure 4.11 (b), a 50% overall height reduction resulted in a 9.5×10^3 (HL) and 5.17×10^2 (CL) probability percentage increase.

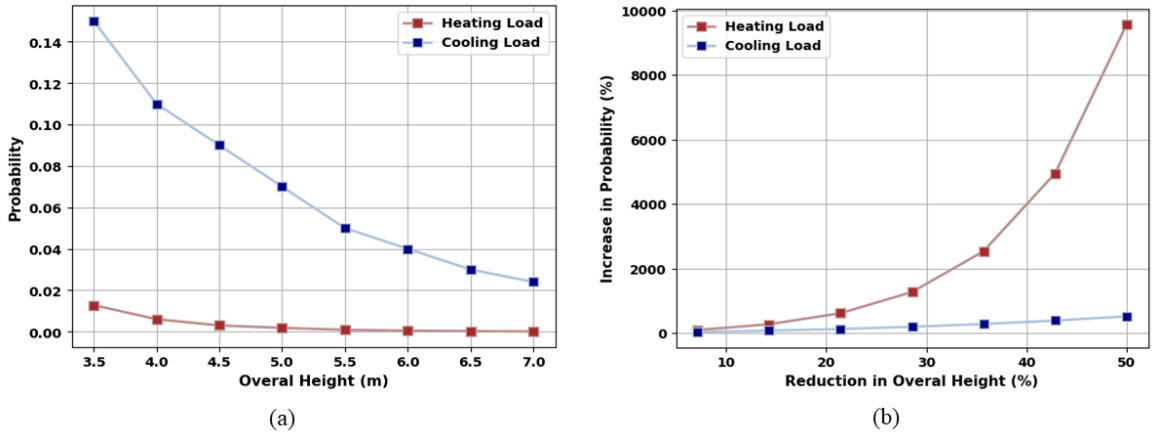


Figure 4.11. Changes of probability with overall height - Very high energy efficiency class (a) Overall height vs Probability (b) Overall height percentage reduction vs probability percentage increase

Figure 4.12 (a) shows the changes of probability of success in very low energy efficiency class with the overall height. The aim within the class is to reduce the probability of success. Therefore, according to Figure 4.12 (a), increasing overall height decreases the probability of success. Figure 4.12 (b) shows the percentage reduction in the probability of success with the percentage reduction of overall height. According to Figure 4.12.(b), 50% of overall height reduction reduces the probability by 99.51% (HL) and 81.04% (CL).

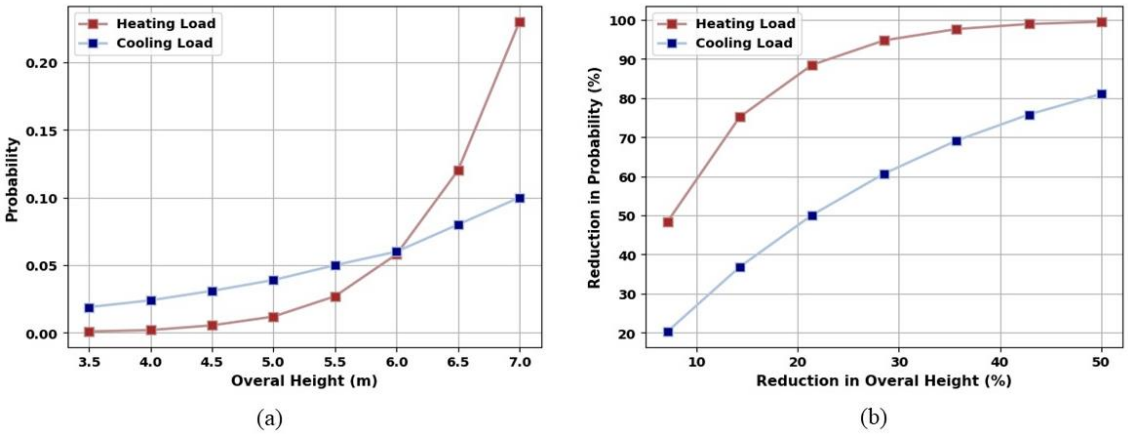


Figure 4.12. Changes of probability with overall height -Very low energy efficiency class (a) Overall height vs Probability (b) Overall height percentage reduction vs probability percentage reduction

However, RC performs differently than overall height in very low energy efficiency class. Figure 4.13 (a) shows the variation in the probability of success with the RC. According to Figure 4.13 (a), probability in very low energy efficiency class decreases with RC. Figure 4.13 (b) shows the reduction of probability with the increased percentage of RC. According to Figure 4.9.(b), 36.73% increase in RC reduces the probability of success by 7.1×10^{3}%(HL) and 1.98×10^{2} %(CL).

Therefore, Logistic Regression predictions results that decreased overall height and increased RC, increases the probability of including in a high energy efficacy class.

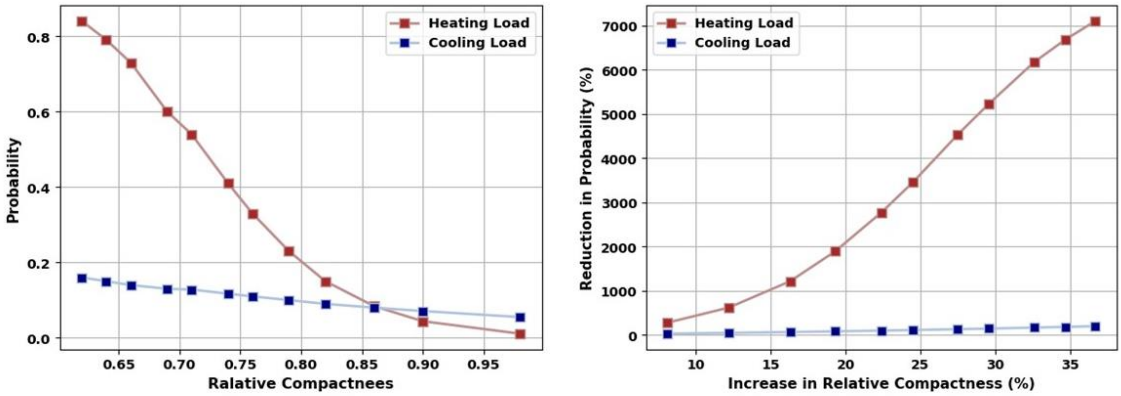


Figure 4.13. Changes of probability with RC- Very low energy efficiency class (a) RC vs Probability (b) RC percentage increase vs probability percentage reduction

Table 4.7 shows the evaluation metric (Accuracy) to analyze Logistic regression. Table 4.7 includes the accuracy values for predicting HL and CL. A Logistic Regression model predicted heating energy efficiency classes with an accuracy of 76.30% and cooling energy efficiency classes with an accuracy of 73.17% according to Table 4.7.

Table 4.7. Evaluation metrics for Logistic Regression.

	Accuracy
Heating Load (Y1)	76.30%
Cooling Load (Y2)	73.17%

4.2.3 Relationships and dependencies between design variable

4.2.3.1 Effect of Bayesian Network Number of Parents

Figures 4.14(a) and 4.14(b) show the graphical structures of Bayesian networks from Hillclimber algorithm for heating load (HL) and cooling load (CL), using discretized data with five classes. The number of parents (NP) is set to 1.

Figure 4.14(a) shows that the variable Y1 acts as a parent for eight input nodes (X1 to X8). Each input node has only one arrow directed toward it from Y1, adhering to the upper bound of the number of parents. The absence of arrows between the X input nodes indicates that they are independent of each other, or there is no causal relationship between them. For example, the structure of Figure 4.14 (a) shows that Y1 can influence the selection of X1 and X2 individually, but X1 and X2 are not dependent on each other.

Similarly, in Figure 4.14(b), the variable Y2 serves as the parent for all the input nodes (X1 to X8). Each input node has only one arrow coming from Y2, again satisfying the NP value of 1. As with Figure 4.14(a), there are no arrows connecting the X input nodes in Figure 4.14(b), indicating their independence from each other. For instance, Figure 4.14(b) demonstrates that Y2 influences X3 and X4 individually, without any connection between X3 and X4.

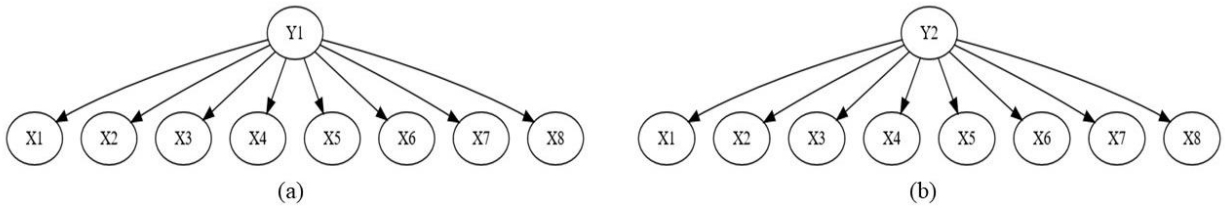


Figure 4.14. Bayesian network for Hillclimber algorithm (NP = 1)

Figures 4.15 (a, b) show the Bayesian networks from Hillclimber algorithm, for predicting HL and CL, where NP =3.

Figure 4.15(a) shows that the nodes have parents where three parents is the maximum, satisfying the NP value. Below shows the relationships between nodes.

1. Y1 is the only parent for Nodes X5, X6, and X8, indicating that these variables are independent of each other since there are no arrows between them.

2. The arrows between X1, X2, X3, and X4 indicate their dependencies, showing the relationship between relative compactness (X1) and surface area (X2) as defined in Equation 1.
3. X4 and X7 have three parents, indicating that these variables are affected by three other variables.

The relationships between the nodes are similar in Figure 4.15(b) satisfying the relationship of Equation 3.1.

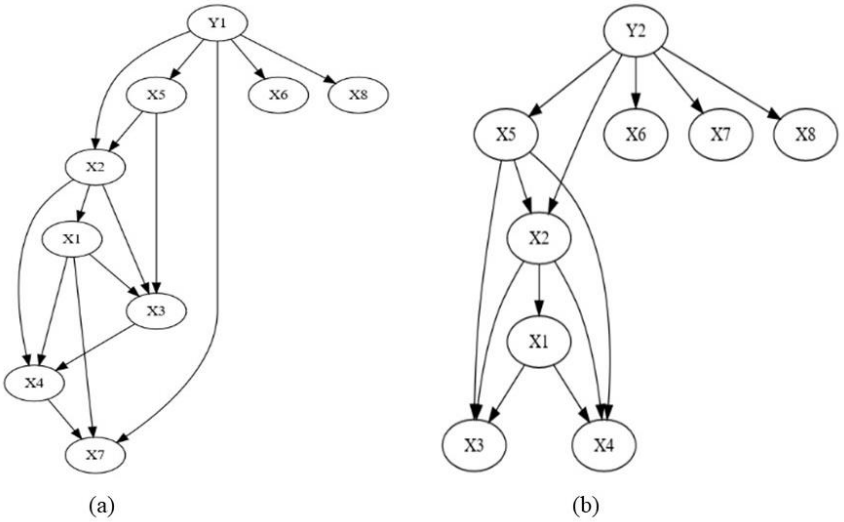


Figure 4.15. Bayesian network for Hillclimber Algorithm (NP = 3)

The comparison of Figures 4.14 and 4.15 highlights the differences in the Bayesian network structures when using different search algorithms and the same NP, where NP = 1. For example, the number of parents for variable X2 varies between the two figures. X2 has only one parent (Y1) in Figure 4.14 (a), while in Figure 4.15 (a), it has two parents (Y1 and X5). Similarly, the number of parents for variable X3 is different between the two figures, with X3 having one parent in Figure 4.14 (a) and three parents in Figure 4.15 (a).

The same observation applies to the comparison of Figures 4.15 and 4.16, where the Bayesian network structures generated by the Tabu search algorithm (NP = 3) are compared. The connections of some variables, such as Y1, X1, X2, X5, X6, X7, and X8, are similar between Figure 4.15 (a) and Figure 4.16 (a). Similarly, X4 has three parents in Figure 4.16 (a) and two

parents in Figure 4.15 (a). The same differences can be observed when comparing Figure 4.15 (b) and Figure 4.16 (b) for variable X3.

The comparisons indicate that different search algorithms can lead to distinct graphical structures in Bayesian networks, even when the upper bound of number of parents (NP) is the same. This variation is due to the different learning methods and strategies employed by the search algorithms to explore and identify the relationships and dependencies between the variables in the dataset.

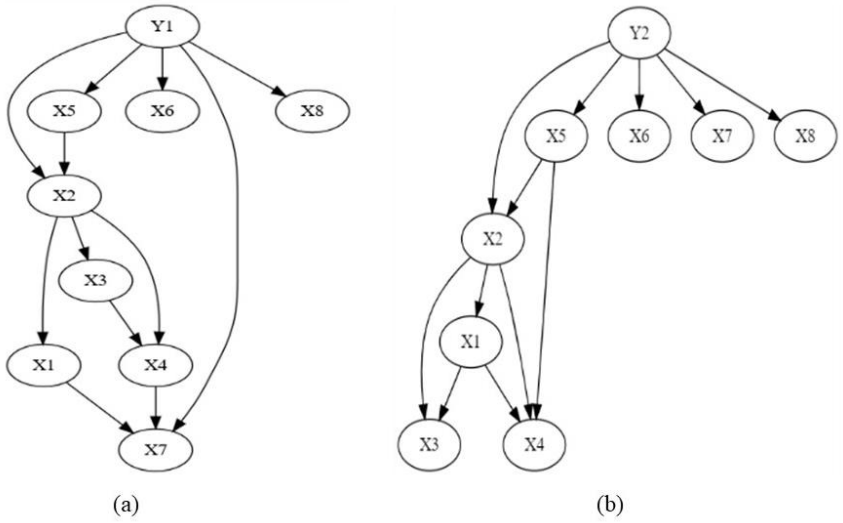


Figure 4.16. Bayesian network for Tabu Search (NP = 3)

4.2.3.2 Effect of Bayesian Network Search Algorithm

Figure 4.17 (a-g) presents the Bayesian network graphical structures generated by seven search algorithms for predicting HL, with NP = 3. Each subfigure (a-g) corresponds to a specific search algorithm. Figures 4.17 (a-e) represent the networks from Hillclimber, K2, LAGD Hillclimber, Repeated Hillclimber, and Tabu search algorithms, respectively, while Figures 4.17 (f, g) depict the networks generated by Simulated annealing and TAN algorithms. Simulated annealing and TAN generate general structures and do not have an NP parameter.

The seven networks demonstrate varying structures due to the differences in the connections between nodes in each network. For example, in Figures 4.17 (a, c, d), variable X4 has variables X1, X2, and X3 as its parents. However, in Figures 4.17 (b, d, e, f), the parents of variable X4 are different. Additionally, Figures 4.17 (a, c) show similar connections for the X4 node but have different parents for the X1 node.

Figure 4.17 suggests that each search algorithm uses a different learning approach and method to build the model. The choice of search algorithm significantly influences the resulting Bayesian network, impacting its predictive performance and the relationships it captures between the input variables.

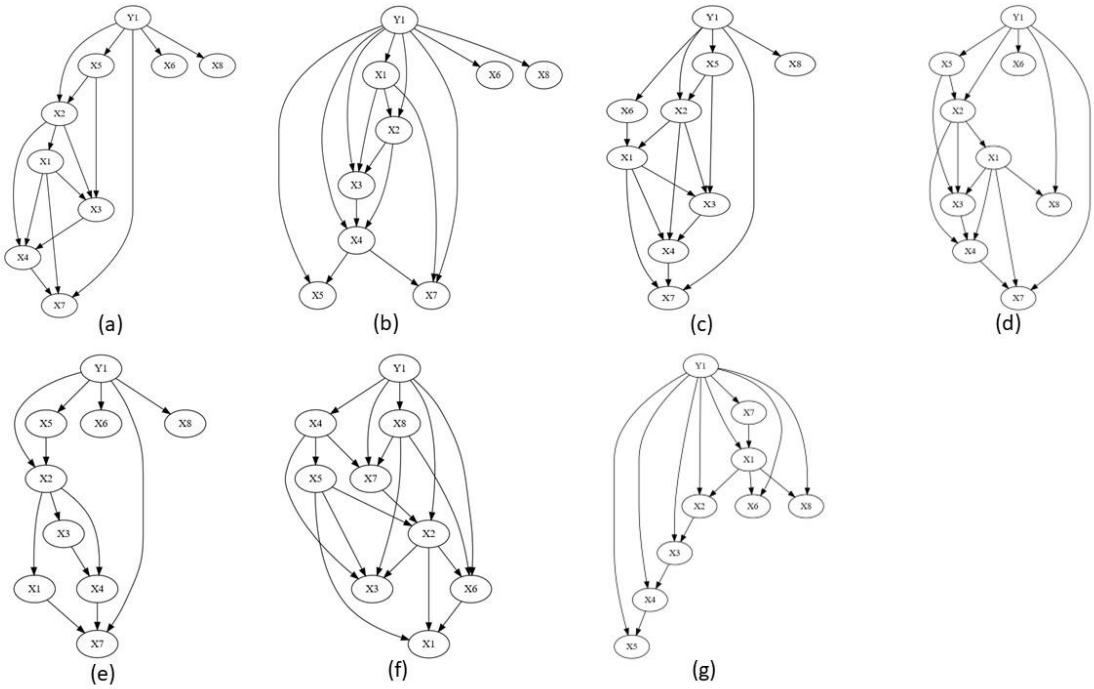


Figure 4.17. Bayesian networks from search algorithms for predicting HL (NP = 3)

Figure 4.18 shows the graphical structures for predicting cooling load using search algorithms. Each subfigure (a-g) corresponds to a specific search algorithm. Figures 4.18 (a-e) represent the networks generated by Hillclimber, K2, LAGD Hillclimber, Repeated Hillclimber, and Tabu search algorithms, respectively, with NP = 3. On the other hand, Figures 4.18 (f, g) illustrate the networks generated by Simulated annealing and TAN algorithms.

The figure clearly demonstrates that different search algorithms result in distinct learning methods. For instance, Figures 4.18 (a, c, d, e) indicate that variables X1, X2, and X5 act as parents for the X4 node, while in Figure 4.18 (b), X1, X2, and Y2 are the parents for the same node. Furthermore, in Figures 4.18 (f) and (g), the parents of node X4 are Y2 and X2, respectively.Similarly, the number of parents for node X2 varies across Figures 4.18 (a, d). In one network, X5 and Y2 are parents of X2, while in the other, Y2 is the only parent.

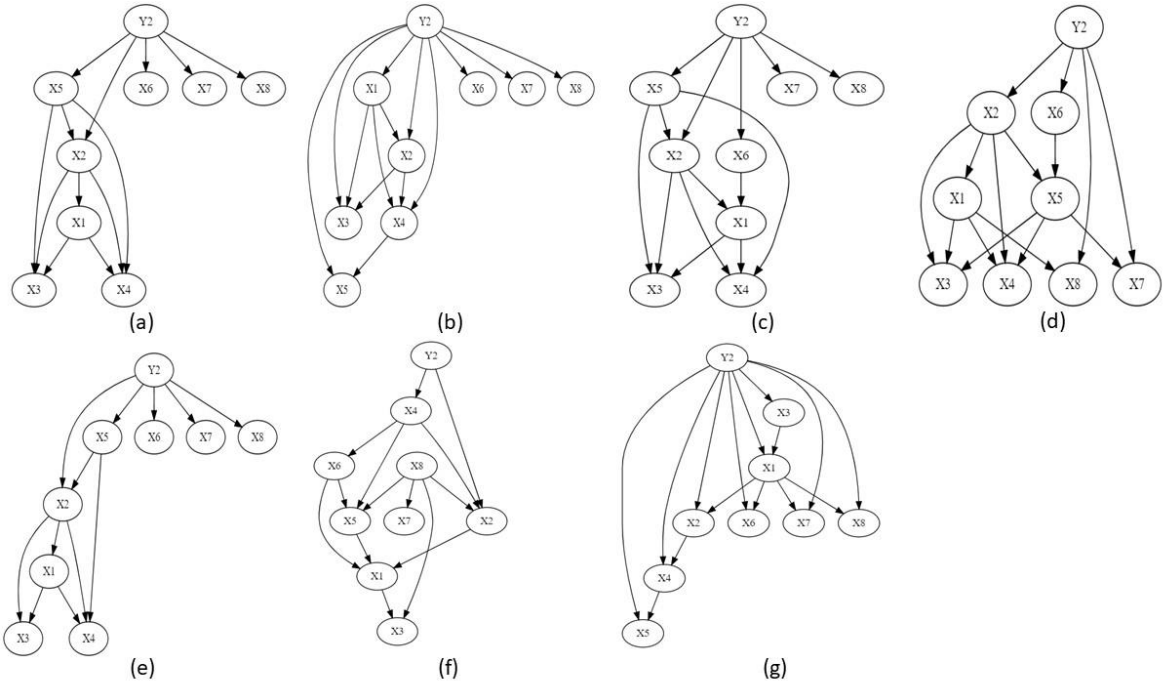


Figure 4.18. Bayesian networks from search algorithms for predicting CL (NP = 3)

Figure 4.19 presents a stacked bar chart summarizing the prediction accuracies (%) of Bayesian network search algorithms. The chart includes results for Hillclimber, K2, LAGD Hillclimber, Repeated Hillclimber, and Tabu search algorithms for NP values of 1, 2, and 3. The maximum NP value recorded is 3 because, for these algorithms, the accuracies (%) obtained didn't change after NP = 3.

For visualizing purposes, the counted maximum number of parents in Simulated annealing is represented as NP = 3 in Figure 4.17, whereas the actual value is 4 and NP= 3 in Figure 4.18. Similarly, in both Figures 7 and .18, TAN is shown to have NP = 2.

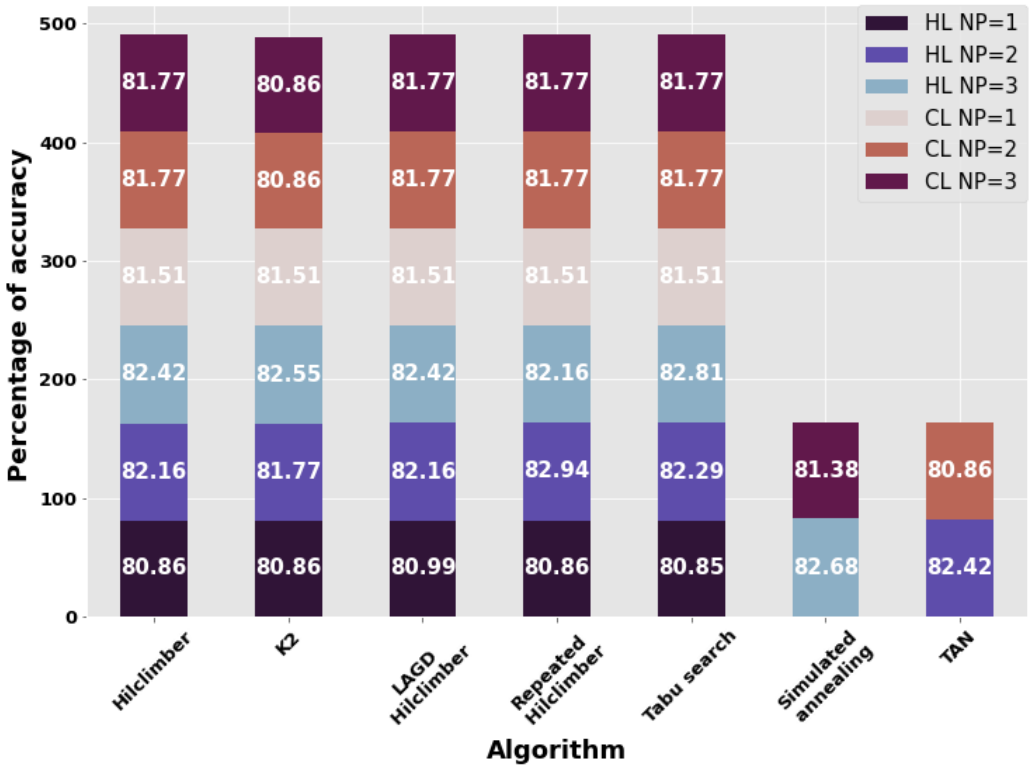


Figure 4.19. Bayesian networks accuracies for search algorithm with NP values

According to the results from Figure 4.19, accuracy tends to increase with a NP value, indicating that more complex models with more dependencies among variables lead to better predictions. The trend holds true, except for CL predictions form K2 and HL predictions from Repeated Hillclimber.

Comparing the predictions, Tabu search achieved the highest accuracy for HL (82.81%) and CL (81.77%) predictions.

Comparing the algorithms Simulated annealing and TAN, Simulated annealing outperformed TAN for both HL and CL predictions with 82.68% and 81.38% accuracy respectively.

4.2.3.3 Analyzing Bayesian Network Conditional Probability Tables

Generally, conditional probability table includes the semi ranges of input variables to create the table. For example, below are the few semi ranges of the input nodes of Tabu search Bayesian network. For X1 the ranges were 0.62-0.63, 0.63-0.75, 0.65-0.75, 0.75-0.805, 0.805-

0.84, 0.84-0.98, and for X5 it was 3.5-5.25, 5.25-7. However, the ranges might vary depending on the specific search algorithm used in the analysis.

Table 4.8 presents the conditional probability table for node X2 using the Tabu search algorithm, which is used for predicting HL. The table shows the probabilities associated with different combinations of values for X2 (surface area) and Y1 (energy efficiency class) in the UCI dataset.

The observation form the table is that the highest conditional probability of 0.855 is obtained for the Y1 in (6.01-13.428) kWh/m^2, when X2 falls within the range of (673.75–771.75) m^2 and X5 (overall height) falls within the range of (3.5–5.25) m. This indicates that to achieve very high-energy efficiency in heating load, the surface area (X2) of the building should be in the range of (673.75–771.75) m^2, and the overall height (X5) should be in the range of (3.5–5.25) m for the given dataset.

Similarly, the highest probability of 0.875 is obtained for the Y1 in (35.682-43.1) kWh/m^2 class, when X2 falls within the range of (624.75–673.75) m^2 and X5 falls within the range of (5.25–7) m. This indicates that for a building to have very low-energy efficiency in heating load, the surface area (X2) should be in the range of (624.75–673.75) m^2, while having overall height (X5) in the range of (5.25–7) m for the given dataset.

Table 4.8. Conditional probability table of X2 for predicting HL (Five classes)

Class	Y1 (kWh/m^2)	X5 (m)	Ranges of X2 (m^2)					
			514.5-600.25	600.25-624.75	624.75-673.75	673.75-771.75	771.75-796.25	796.25-808.6
Very high energy efficiency	6.01-13.428	3.5-5.25	0.002	0.002	0.002	0.855	0.021	0.117
	6.01-13.428	5.25-7	0.167	0.167	0.167	0.167	0.167	0.167
High energy efficiency	13.428-20.846	3.5-5.25	0.003	0.003	0.003	0.431	0.336	0.225
	13.428-20.846	5.25-7	0.639	0.25	0.028	0.028	0.028	0.028
Medium energy efficiency	20.846-28.264	3.5-5.25	0.167	0.167	0.167	0.167	0.167	0.167
	20.846-28.264	5.25-7	0.52	0.421	0.045	0.005	0.005	0.005
Low energy efficiency	28.264-35.682	3.5-5.25	0.167	0.167	0.167	0.167	0.167	0.167
	28.264-35.682	5.25-7	0.701	0.109	0.18	0.003	0.003	0.003
Very low energy efficiency	35.682-43.1	3.5-5.25	0.167	0.167	0.167	0.167	0.167	0.167
	35.682-43.1	5.25-7	0.106	0.005	0.875	0.005	0.005	0.005

Table 4.9 presents the conditional probability table for node X2 in the Tabu search algorithm, which is used for predicting CL (cooling load).

From the table, it can be observed that the surface area (X2) of the building should be in the range of (673.75–771.75) m^2, and the overall height (X5) should be in the range of (3.5–5.25) m for the given dataset to achieve very high energy efficiency. Similarly, for a building to obtain very low energy efficiency in cooling load, the surface area (X2) should be in the range of (624.75–649.25) m^2, while overall height (X5) in the range of (5.25–7) m for the given dataset.

Table 4.9. Conditional probability table of X2 for predicting CL (Five classes)

Class	Y2 (kWh/m^2)	X5 (m)	Ranges of X2 (m^2)						
			514.5-600.25	600.25-624.75	624.75-649.25	649.25-673.75	673.75-771.75	771.75-796.25	796.25-808.6
Very high energy efficiency	10.9-18.326	3.5-5.25	0.002	0.002	0.002	0.002	0.782	0.014	0.198
	10.9-18.326	5.25-7	0.143	0.143	0.143	0.143	0.143	0.143	0.143
High energy efficiency	18.326-25.752	3.5-5.25	0.008	0.008	0.008	0.008	0.038	0.924	0.008
	18.326-25.752	5.25-7	0.46	0.46	0.016	0.016	0.016	0.016	0.016
Medium energy efficiency	25.752-33.178	3.5-5.25	0.143	0.143	0.143	0.143	0.143	0.143	0.143
	25.752-33.178	5.25-7	0.666	0.284	0.015	0.026	0.003	0.003	0.003
Low energy efficiency	33.178-40.604	3.5-5.25	0.143	0.143	0.143	0.143	0.143	0.143	0.143
	33.178-40.604	5.25-7	0.403	0.016	0.206	0.365	0.003	0.003	0.003
Very low energy efficiency	40.604-48.03	3.5-5.25	0.143	0.143	0.143	0.143	0.143	0.143	0.143
	40.604-48.03	5.25-7	0.065	0.013	0.792	0.091	0.013	0.013	0.013

To decide which node's conditional probability tables to analyze, the nodes directly connected to the output nodes Y1, and Y2 in the Bayesian network obtained from the search algorithms were analyzed. The process repeated for both HL and CL predictions. The purpose of selecting nodes that were connected directly to the outputs is to understand the influence of input parameters on the output variables HL and CL. The most connected nodes to Y1 (HL) and Y2 (CL) were X2 (surface area), X6 (orientation), X5 (overall height), X7 (glazing area ratio), and X8 (glazing area distribution), therefore, the study analyzed selected nodes.

Table 4.10 presents the ranges of input nodes (X2, X5, and X7) that obtained the highest conditional probabilities for each class. The table includes ranges for predicting HL using five energy efficiency classes.

The results indicate that X2 achieved very high-energy efficiency when X2 in (673.75–771.75) m2 for HL. Similarly, dataset achieved very low energy efficiency when X2 in (624.75–673.75) m2, which was consistent across all algorithms.

Regarding node X5 (overall height), the highest conditional probabilities for very high-energy and high-energy efficiency classes were achieved when X5 in (3.5–5.25) m, which was true for all algorithms. X5 in (5.25–7) m resulted in the highest conditional probabilities for medium-energy efficiency, low-energy efficiency, and very low-energy efficiency classes, across all search algorithms. The results indicate that increased overall heights are associated with lower energy efficiency of a building.

For node X7 (glazing area ratio), X7 in (0–0.175) achieved the highest conditional probabilities for predicting very high-energy efficiency of HL, while using X7 in (0.325–0.4) led to very low-energy efficiency, consistently across all algorithms. These results suggest that lower glazing areas increase building energy efficiency, as lower glazing areas reduce the heat transfer through buildings. The findings align with a study by Alwetaishi (2019) that recommends a glazing to wall ratio of 10% for both hot and dry and hot and humid climates. The selected value of 10% as the recommended glazing area corresponds to the lower percentages of 5%, 10%, 20%, 40%, and 50%.

Table 4.10. Highest ranges of nodes X2, X5 and X7 for HL (Five classes)

Node	Y1 range (kWh/m^2)	Algorithm						
		Hillclimber	K2	LAGD Hillclimber	Repeated Hillclimber	Tabu Search	Simulated Annealing	TAN
X2 (m^2)	Very high energy efficiency	673.75-771.75	673.75-771.75	673.75-771.75	673.75-771.75	673.75-771.75	673.75-771.75	673.75-771.75
	High energy efficiency	514.5-600.25	673.75-771.75	514.5-600.25	514.5-600.25	514.5-600.25	771.75-796.25	673.75-771.75
	Medium energy efficiency	514.5-600.25	514.5-600.25	514.5-600.25	514.5-600.25	514.5-600.25	514.5-600.25	514.5-600.25
	Low energy efficiency	514.5-600.25	514.5-600.25	514.5-600.25	514.5-600.25	514.5-600.25	514.5-600.25	514.5-600.25
	Very low energy efficiency	624.75-673.75	624.75-673.75	624.75-673.75	624.75-673.75	624.75-673.75	624.75-673.75	624.75-673.75
X5 (m)	Very high energy efficiency	3.5-5.25	3.5-5.25	3.5-5.25	-	3.5-5.25	-	3.5-5.25
	High energy efficiency	3.5-5.25	3.5-5.25	3.5-5.25	-	3.5-5.25	-	3.5-5.25
	Medium energy efficiency	5.25-7	5.25-7	5.25-7	-	5.25-7	-	5.25-7
	Low energy efficiency	5.25-7	5.25-7	5.25-7	-	5.25-7	-	5.25-7
	Very low energy efficiency	5.25-7	5.25-7	5.25-7	-	5.25-7	-	5.25-7
X7	Very high energy efficiency	0-0.175	0-0.175	0-0.175	0-0.175	0-0.175	0-0.175	0-0.175
	High energy efficiency	0.325-0.4	0-0.175	0.325-0.4	0.325-0.4	0.325-0.4	0-0.175	0.325-0.4
	Medium energy efficiency	0-0.175	0-0.175	0-0.175	0-0.175	0-0.175	0-0.175	0-0.175
	Low energy efficiency	0-0.175	0.375-0.4	0.325-0.4	0-0.175	0-0.175	0.325-0.4	0.325-0.4
	Very low energy efficiency	0.325-0.4	0.325-0.4	0.325-0.4	0.325-0.4	0.325-0.4	0.325-0.4	0.325-0.4

The procedure was repeated for CL predictions and obtained the ranges for nodes X2 and X5, which achieved the highest probabilities for predicting cooling load (CL) using different search algorithms. The results indicate that the highest conditional probability is achieved when X2 is within (673.75–771.75) m2, and X2 range is (624.75–649.25) m2, for very high and very low energy efficiency classes respectively. The results are consistently observed across all algorithms. The results obtained for X2 matches for both HL and CL.

Analyzing the node X5 obtained for very high-energy and high-energy efficiency classes of CL, the best range for X5 is (3.5–5.25) m, and the result holds true for all algorithms. For the remaining three classes, the best range for X5 is in (5.25–7) m, consistent across all search algorithms.

The results consistently conclude that reducing the overall height (X5) of a building can led to higher energy efficiency. These results align with the findings of Aqlan et al., (2014) which identified X5 (overall height) as the most crucial factor for reducing heating and cooling requirements, suggesting that engineers should focus on reducing the overall height of buildings to achieve higher energy efficiency.

Regarding nodes X7 (glazing area) for CL and X6 (orientation) and X8 (glazing area distribution) for both HL and CL, the obtained probability values are consistent across all ranges. Therefore, it is difficult to draw specific conclusions about the range that leads to certain conditional probabilities for these three parameters.

The results obtained for the three classes correlated with the results obtained with five classes. The results obtained only for the K2 and TAN algorithms, as they are the only ones which connected directly with the class and input node.

For the K2 and TAN algorithms, the highest conditional probabilities were achieved when X5 in (3.5–5.25) m for the very high-energy efficiency class. For both the medium and very low-energy efficiency classes X5 was in (5.25–7) m.

Similarly, when analyzing the results for node X2 with three classes, the findings also correlate well with the results from the five-class analysis for most of the search algorithms. For example, the very high-energy efficiency class obtained the best range as X2 in (673.75–771.75) m2, while the very low-energy efficiency class had the highest probability when X2 in (624.75–673.75) m2. These ranges for X2 are consistent with the five-class analysis for both predictions.

However, there were slight changes in the results for the K2 and Simulated annealing algorithms when considering X2 ranges for HL. The X7 ranges obtained from the three-class analysis correlated well only for the very low-energy efficiency class and the best range was X7 within (0.325-0.4), and this result was consistent across all algorithms.

Similarly, the highest conditional probability ranges were obtained for CL predictions using three classes, for X5 and X2 nodes. The results for X5 in predicting CL are similar to the results observed for HL classes. The very high-energy efficiency class for CL obtained the highest conditional probability when X5 is in (3.5–5.25) m and X2 are (673.75–771.75) m^2, while the medium and very low-energy efficiency classes had the highest probabilities when X5 in (5.25–7) m and X2 in (624.75–673.75) m^2.

4.3 Impact of location on energy consumption

Table 4.11 shows the Wilcoxon rank sum test U values for mid-rise apartments of ten states. The U values are calculated compared to Lafayette, Indiana. According to Table 4.11, Indianapolis in Indiana and Illinois obtained U values larger than 0.05 in all cases (all year, spring, summer, fall, and winter).

The results indicate that the two samples have the same distribution, meaning cooling usage for electricity in mid-rise apartments has the same distribution in Indianapolis and Illinois. However, other states, i.e., CA, AZ, TX, SD, TN, NC, NY, and MA, obtained that the two samples have a different distribution. The results indicate that cooling usage for electricity in mid-rise apartments does not have the same distribution in the mentioned states. The results obtained 100% accuracy for states except CA (80%), SD (80%), NY (60%), and MA (40%). Therefore, the results of Table 4.11 indicate that location factor impacts the cooling usage for electricity in mid-rise apartments.

Table 4.11. Wilcoxon rank sum test results for mid-rise apartments relative to Indiana -Lafayette

City/State	All year	Spring	Summer	Fall	Winter
Indiana (Indianapolis)	0.678	0.0915	0.327	0.787	0.792
Illinois	0.802	0.790	0.316	0.822	0.678
California	9.700×10^{-5}	0.122	2.480×10^{-24}	1.17×10^{-11}	4.057×10^{-30}
Arizona	2.216×10^{-26}	3.604×10^{-10}	1.184×10^{-22}	1.857×10^{-16}	2.904×10^{-27}
Texas	2.237×10^{-26}	2.738×10^{-17}	1.025×10^{-25}	2.744×10^{-17}	3.071×10^{-16}
South Dakota	0.0033	0.030	1.577×10^{-7}	0.040	0.091
Tennessee	1.611×10^{-5}	0.0009	3.409×10^{-5}	7.814×10^{-6}	0.0006
North Carolina	1.953×10^{-8}	1.895×10^{-6}	0.0003	1.056×10^{-6}	3.927×10^{-11}
New York	0.021	0.285	4.773×10^{-5}	0.109	0.0275
Massachusetts	0.069	0.000	0.006	0.840	0.271

Table 4.12 shows the U values of Wilcoxon rank sum test for supermarkets in the ten states. For supermarkets, Indianapolis and Illinois obtained U values larger than 0.05, indicating the two samples have the same distribution. Other eight states obtained that the cooling usage for electricity is different relative to Lafayette, Indiana. The results' accuracy is 100% for CA, AZ, TX, TN, and NC. The accuracy for SD, NY, and MA is 60%.

Table 4.12. Wilcoxon rank sum test results for supermarkets relative to Indiana -Lafayette

City/State	All year	Spring	Summer	Fall	Winter
Indiana (Indianapolis)	0.739	0.078	0.612	0.94	0.362
Illinois	0.642	0.8366	0.338	0.728	0.667
California	5.55×10^{-28}	9.16×10^{-19}	1.49×10^{-30}	4.59×10^{-7}	0.0495
Arizona	1.03×10^{-13}	3.76×10^{-10}	1.95×10^{-22}	9.16×10^{-6}	0.002
Texas	1.14×10^{-35}	1.68×10^{-21}	8.16×10^{-26}	7.466×10^{-22}	1.22×10^{-19}
South Dakota	0.007	0.241	3.99×10^{-7}	0.045	0.119
Tennessee	4.32×10^{-8}	2.05×10^{-6}	3.614×10^{-5}	5.705×10^{-7}	3.902×10^{-7}
North Carolina	8.45×10^{-11}	7.17×10^{-10}	0.00016	4.607×10^{-8}	1.59×10^{-11}
New York	0.027	0.432	5.66×10^{-5}	0.385	0.0048
Massachusetts	0.043	0.00165	0.00051	0.925	0.318

Figure 4.20 (a- e) shows the climate data of the ten states over the 2012 year. Figure 4.20 (a- e) shows the precipitation, heating degree days, cooling degree days, solar radiation intensity, and average temperature, respectively. According to Figure 4.20, Indiana and Illinois have a closer relationship than the other eight states. Therefore, Indiana and Illinois have a similar climate, while the other eight states have a different climate compared to Indiana.

The climate of the state changes with the location. Therefore, the results indicate that location features impact the cooling usage for electricity in commercial buildings.

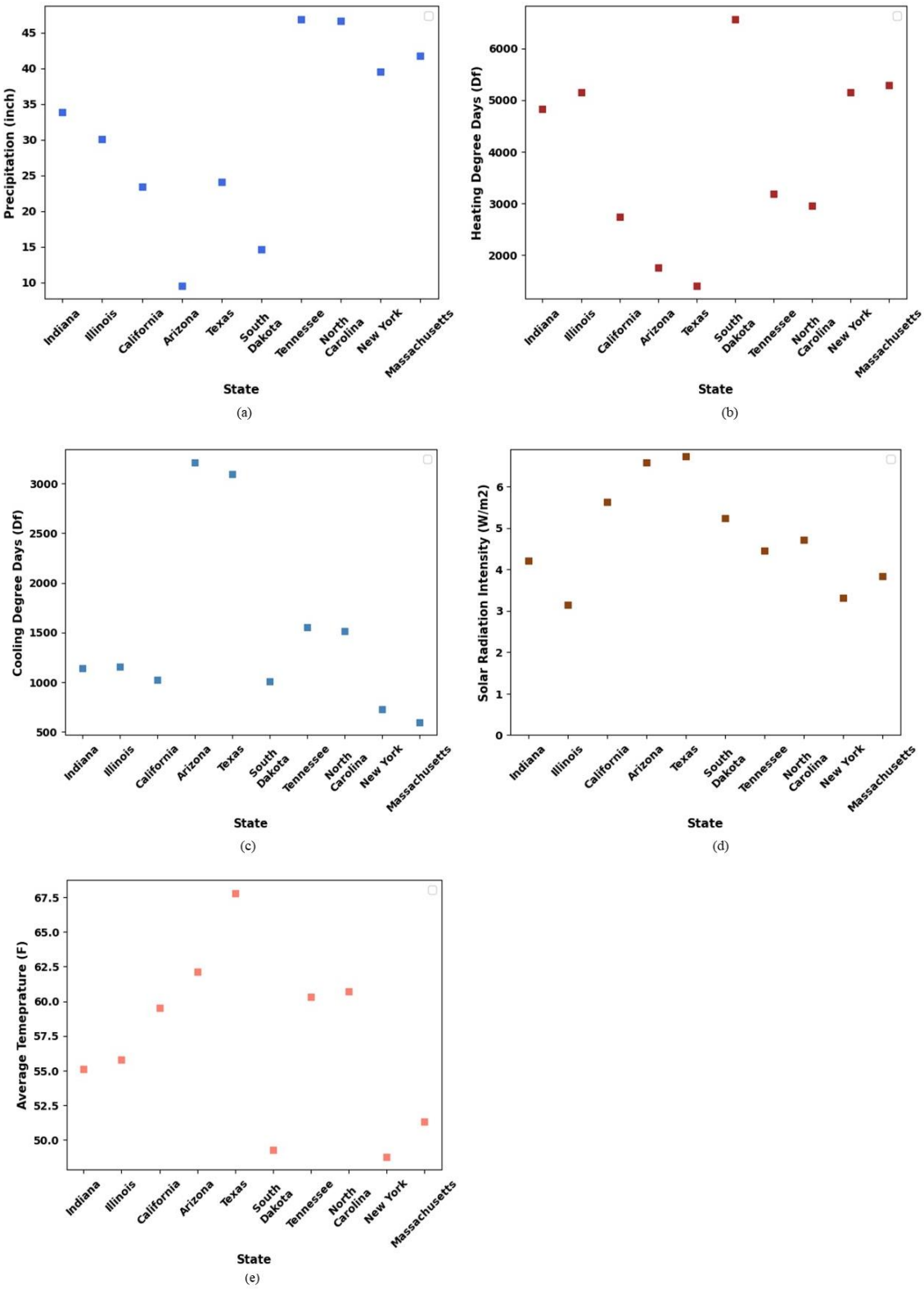


Figure 4.20. Climate data of US states over 2012

4.4 Chapter Summary

Discretized UCI dataset created five and three energy efficiency classes to analyze the input feature relation with the energy efficiency. Shapash, created feature important graphs to identify the most important feature impacting commercial buildings cooling energy usage. Feature importance graph for CBECS 2018 dataset indicated that the first two features, i.e., cooling degree days and principle building activity contribute to 54.79% to predict cooling EUI. The four topmost important features to predict cooling EUI are cooling degree days, cooling percentage, total hours open per week, and principle building activity.

Linear Regression and Logistic Regression analyzed the direction of the input feature's impact on building energy consumption. Predictions indicated that reduced overall heights and increased relative compactness reduce heating and cooling loads. The reduction of overall height by 50% reduces the heating load by 64.56% and cooling load by 57.47%. Logistic Regression predictions identified that reduction of overall height by 50% in very high energy class increases the probability of success percentage by 9.5×10^{3}% (HL) and 5.17×10^{2}% (CL).

Bayesian networks identified the relationships and dependencies between input and out features using the structure of the networks and conditional probability tables. Tabu search with NP =3 obtained the highest accuracy results for heating (82.81%) and cooling load (81.77%).

Finally, Wilcoxon -rank sum test analyzed that the location of the building impacts its cooling energy usage for electricity. The results indicated that two cities in Indiana, i.e., Indianapolis, and Illinois, did not differ significantly with Lafayette - Indiana with a 0.05 error margin. However, California, Arizona, South Dakota, Texas, Tennessee, North Carolina, New York, and Massachusetts obtained a significant difference in distributions relative to Lafayette-Indiana.

CHAPTER 5 . SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

5.1 Conclusions

The aim of this study is to increase residential and commercial building energy efficiency by analyzing impacting factors on heating and cooling energy usage. The chapter presents the conclusions of this research.

1. The study analyzed the residential and commercial buildings' heating and cooling energy usage using machine learning algorithms and statistical techniques. Identifying the impact of building parameters on its energy consumption is important to increase the building's energy efficiency by reducing energy usage.
2. The importance of features impacting commercial buildings, their importance orders, and contributions to cooling energy usage intensity (EUI) of commercial buildings were analyzed.
3. Random Forest predicted the cooling EUI using CBECS 2018 and 2012 datasets with a 25% test and 75% training data. Explainable ML Shapash, generated the feature importance graphs for datasets included with feature importance order and contributions from each input feature.
4. Feature importance graph for CBECS 2018 dataset indicated that the first two features, i.e., cooling degree days and principle building activity contribute to 54.79% to predict cooling EUI. The four topmost important features in CBECS 2012 to predict cooling EUI are cooling degree days, cooling percentage, total hours open per week, and principle building activity. The four features contribute to 52.14% of predictions.
5. Considering each feature contribution to predicting cooling EUI in both 2018 and 2012, four out of 21 inputs, i.e., cooling degree days, principle building activity, cooling percentage, and total hours open per week, each contributed more than 5% in both years. Cooling degree days obtained the highest contribution in both 2018 and 2012, being the topmost important feature to predict cooling EUI.

6. The findings can be used in the real world for identifying the most important features and to decide which feature should be given more priority before and after building construction.
7. The common observations of Linear Regression predictions are that the overall height and glazing ratio has positive coefficients, whereas relative compactness has a negative coefficient. The reduction of overall height by 50% reduces the HL by 64.56% and CL by 57.47%. The decrease of glazing area ratio by 75% reduces HL and CL by 19.5% and 13.88%. However, a 36.73% reduction of RC increases HL by 103.16% and CL by 97.68%.
8. The common observations of Logistic Regression predictions that were compatible with Linear regression results were overall height and relative compactness. The reduction of overall height by 50% in very high energy class increases the probability of success percentage by 9.5×10^{3}% (HL) and 5.17×10^{2}% (CL). Overall height reduction by 50% decreases the probability of success by 99.51%(HL) and 81.04%(CL) in very low energy class. The probability of success in very low energy efficiency class decreases with RC such that 36.73% increase in RC reduces the probability of success by 7.1×103%(HL) and 1.98×102 %(CL).
9. The Logistic Regression prediction accuracy of discretized data are 76.30% (HL) and 73.17% (CL).
10. Different Bayesian network search algorithms generate different networks. The upper bound of the number of parents impacts the network design.
11. The prediction accuracy increased with the increase of number of allowed parents per node. Tabu search with NP =3 gave the best accuracy results for heating (82.81%) and cooling load (81.77%).
12. The results indicate that reduced overall height and glazing are ratio has a high energy efficiency for energy loads. The results obtained using five and three energy classes were the same.
13. Identifying whether the location impacts the building's energy consumption is important to design suitable buildings optimized for the specific climate.

14. The Wilcoxon rank sum test calculated the U values to identify whether commercial building cooling usage for electricity impacted from the location. The Wilcoxon rank sum test analyzed the cooling energy used for electricity for ten states relative to Indiana for two commercial building types of mid-rise apartments and supermarkets.
15. The results indicated that two cities in Indiana, i.e., Indianapolis, and Illinois, did not differ significantly with Lafayette - Indiana with a 0.05 error margin. However, California, Arizona, South Dakota, Texas, Tennessee, North Carolina, New York, and Massachusetts obtained a significant difference in distributions relative to Lafayette-Indiana. Therefore, according to the results, the location feature impacts the building energy usage.

5.2 Recommendations

1. Using Shapash, running the Random Forest identified the relative importance of features impacting cooling energy usage of commercial buildings. But, using other algorithms which support Shapash, such as XGBoost, Catboost, and Support Vector Machine could be done to compare the results for different algorithms.
2. Linear Regression and Logistic Regression identified the impact from input design variable on heating load and cooling load of residential buildings in machine learning applications. However, using other algorithms, such as Artificial Neural Networks could provide more insights.
3. Only one test i.e., Wilcoxon- rank sum test identified whether the location impacts the cooling energy usage of commercial buildings. However, using the other state-of-the-art, sophisticated statistical tests, such as anova test, t-test could provide additional insights.

LIST OF REFERENCES

Abediniangerabi, B., Makhmalbaf, A., & Shahandashti, M. (2022). Estimating energy savings of ultra-high-performance fibre-reinforced concrete facade panels at the early design stage of buildings using gradient boosting machines. *Advances in Building Energy Research*, *16*(4), 542–567. https://doi.org/10.1080/17512549.2021.2011410

Administration, U. S. E. I. (n.d.). *Commercial Buildings Energy Consumption Survey (CBECS)*. Retrieved January 31, 2023, from https://www.eia.gov/consumption/commercial/

Ahmad, T., & Zhang, D. (2020). A critical review of comparative global historical energy consumption and future demand: The story told so far. *Energy Reports*, *6*, 1973–1991. https://doi.org/10.1016/j.egyr.2020.07.020

Al-Ghussain, L. (2019). Global warming: review on driving forces and mitigation. *Environmental Progress and Sustainable Energy*, *38*(1), 13–21. https://doi.org/10.1002/ep.13041

Alam, M. J., & Islam, M. A. (2017). Effect of external shading and window glazing on energy consumption of buildings in Bangladesh. *Advances in Building Energy Research*, *11*(2), 180–192. https://doi.org/10.1080/17512549.2016.1190788

Amin, M. N., Salami, B. A., Zahid, M., Iqbal, M., Khan, K., Abu-Arab, A. M., Alabdullah, A. A., & Jalal, F. E. (2022). Investigating the Bond Strength of FRP Laminates with Concrete Using LIGHT GBM and SHAPASH Analysis. *Polymers*, *14*(21), 4717. https://doi.org/10.3390/polym14214717

Andrić, I., Le Corre, O., Lacarrière, B., Ferrão, P., & Al-Ghamdi, S. G. (2021). Initial approximation of the implications for architecture due to climate change. *Advances in Building Energy Research*, *15*(3), 337–367. https://doi.org/10.1080/17512549.2018.1562980

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), 1–13. https://doi.org/10.1002/widm.1424

Aqlan, F., Ahmed, A., Srihari, K., & Khasawneh, M. T. (2014). Integrating artificial neural networks and cluster analysis to assess energy efficiency of buildings. *IIE Annual Conference and Expo 2014*, 3936–3943.

Araújo, G. R., Teixeira, H., Gomes, M. G., & Rodrigues, A. M. (2023). Multi-objective optimization of thermochromic glazing properties to enhance building energy performance. *Solar Energy*, *249*(October 2022), 446–456. https://doi.org/10.1016/j.solener.2022.11.043

Ascione, F., Bianco, N., Maria Mauro, G., & Napolitano, D. F. (2019). Building envelope design: Multi-objective optimization to minimize energy consumption, global cost and thermal discomfort. Application to different Italian climatic zones. *Energy*, *174*, 359–374. https://doi.org/10.1016/j.energy.2019.02.182

Aste, N., Manfren, M., & Marenzi, G. (2017). Building Automation and Control Systems and performance optimization: A framework for analysis. *Renewable and Sustainable Energy Reviews*, *75*(October 2016), 313–330. https://doi.org/10.1016/j.rser.2016.10.072

Barros, R. S. M. de, Hidalgo, J. I. G., & Cabral, D. R. de L. (2018). Wilcoxon Rank Sum Test Drift Detector. *Neurocomputing*, *275*, 1954–1963. https://doi.org/10.1016/j.neucom.2017.10.051

Bekkouche, S. M. A., Benouaz, T., Hamdani, M., Cherier, M. K., Yaiche, M. R., & Benamrane, N. (2017). Diagnosis and comprehensive quantification of energy needs for existing residential buildings under Sahara weather conditions. *Advances in Building Energy Research*, *11*(1), 37–51. https://doi.org/10.1080/17512549.2015.1119059

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Bouckaert, R. R. (2004). *Bayesian Network Classifiers in Weka*. https://researchcommons.waikato.ac.nz/bitstream/handle/10289/85/%0Acontent.pdf?sequence=1

Boudjella, A., & Boudjella, M. Y. (2021a). Cooling Load Energy Performance of Residential Building: Machine Learning-Cluster K-Nearest Neighbor CKNN (Part I). In *Lecture Notes in Networks and Systems* (Vol. 174, Issue Part I). Springer International Publishing. https://doi.org/10.1007/978-3-030-63846-7_41

Boudjella, A., & Boudjella, M. Y. (2021b). Heating Load Energy Performance of Residential Building: Machine Learning-Cluster K-Nearest Neighbor CKNN (Part I). In *Lecture Notes in Networks and Systems* (Vol. 174, Issue Part I). Springer International Publishing. https://doi.org/10.1007/978-3-030-63846-7_41

Campagna, L. M., & Fiorito, F. (2022). On the Impact of Climate Change on Building Energy Consumptions: A Meta-Analysis. *Energies*, *15*(1). https://doi.org/10.3390/en15010354

Cao, X., Dai, X., & Liu, J. (2016). Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade. *Energy and Buildings*, *128*, 198–213. https://doi.org/10.1016/j.enbuild.2016.06.089

Casalicchio, G., Molnar, C., & Bischl, B. (2019). Visualizing the Feature Importance for Black Box Models. *Springer, Cham*, *11051*. https://doi.org/https://doi.org/10.1007/978-3-030-10925-7_40

Chou, J. S., & Bui, D. K. (2014). Modeling heating and cooling loads by artificial intelligence for energy-efficient building design. *Energy and Buildings*, *82*(2014), 437–446. https://doi.org/10.1016/j.enbuild.2014.07.036

Chung, M. H., & Rhee, E. K. (2014). Potential opportunities for energy conservation in existing buildings on university campus: A field survey in Korea. *Energy and Buildings*, *78*, 176–182. https://doi.org/10.1016/j.enbuild.2014.04.018

Cristino, T. M., Neto, A. F., Wurtz, F., & Delinchant, B. (2022). The Evolution of Knowledge and Trends within the Building Energy Efficiency Field of Knowledge. *Energies*, *15*(3). https://doi.org/10.3390/en15030691

Daeung Danny Kim, & Suh, H. S. (2021). Heating and cooling energy consumption prediction model for high-rise apartment buildings considering design parameters. *Energy for Sustainable Development*, *61*, 1–14. https://doi.org/https://doi.org/10.1016/j.esd.2021.01.001

Dahiya, N., Gupta, S., & Singh, S. (2022). A Review Paper on Machine Learning Applications, Advantages, and Techniques. *ECS - The Electrochemical Society*, *107*(1). https://doi.org/10.1149/10701.6137ecst

De Loera, J. A., & Hogan, T. (2020). Stochastic Tverberg Theorems With Applications in Multiclass Logistic Regression, Separability, and Centerpoints of Data. *SIAM Journal on Mathematics of Data Science*, *2*(4), 1151–1166. https://doi.org/10.1137/19m1277102

Delgarm, N., Sajadi, B., Kowsary, F., & Delgarm, S. (2016). Multi-objective optimization of the building energy performance: A simulation-based approach by means of particle swarm optimization (PSO). *Applied Energy*, *170*, 293–303. https://doi.org/10.1016/j.apenergy.2016.02.141

Delzendeh, E., Wu, S., Lee, A., & Zhou, Y. (2017). The impact of occupants' behaviours on building energy analysis: A research review. *Renewable and Sustainable Energy Reviews*, *80*(September 2016), 1061–1071. https://doi.org/10.1016/j.rser.2017.05.264

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*, 34–43. https://doi.org/10.1016/j.enbuild.2017.12.031

Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications*, *166*(September 2020), 114060. https://doi.org/10.1016/j.eswa.2020.114060

Ersoz, B., Sagiroglu, S., & Bulbul, H. I. (2022). A Short Review on Explainable Artificial Intelligence in Renewable Energy and Resources. *11th IEEE International Conference on Renewable Energy Research and Applications, ICRERA 2022*, 247–252. https://doi.org/10.1109/ICRERA55966.2022.9922870

Farhad Amirifard, Sharif, S. A., & Nasiri, F. (2019). Application of passive measures for energy conservation in buildings – a review. *Advances in Building Energy Reserach*, *13*(2), 282–315. https://doi.org/https://doi.org/10.1080/17512549.2018.1488617

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*(September), 110287. https://doi.org/10.1016/j.rser.2020.110287

Ghosh, I., Chaudhuri, T. D., Alfaro-Cortés, E., Gámez, M., & García, N. (2022). A hybrid approach to forecasting futures prices with simultaneous consideration of optimality in ensemble feature selection and advanced artificial intelligence. *Technological Forecasting and Social Change*, *181*(May). https://doi.org/10.1016/j.techfore.2022.121757

Ghosh, I., & Sanyal, M. K. (2021). Introspecting predictability of market fear in Indian context during COVID-19 pandemic: An integrated approach of applied predictive modelling and explainable AI. *International Journal of Information Management Data Insights*, *1*(2), 100039. https://doi.org/10.1016/j.jjimei.2021.100039

Gianey, H. K., & Choudhary, R. (2018). Comprehensive Review On Supervised Machine Learning Algorithms. *Proceedings - 2017 International Conference on Machine Learning and Data Science, MLDS 2017*, *2018-Janua*, 38–43. https://doi.org/10.1109/MLDS.2017.11

Gianniou, P., Liu, X., Heller, A., Nielsen, P. S., & Rode, C. (2018). Clustering-based analysis for residential district heating data. *Energy Conversion and Management*, *165*(December 2017), 840–850. https://doi.org/10.1016/j.enconman.2018.03.015

Goliatt, L., Capriles, P. V. Z., & Duarte, G. R. (2018). Modeling Heating and Cooling Loads in Buildings Using Gaussian Processes. *2018 IEEE Congress on Evolutionary Computation, CEC 2018 - Proceedings*. https://doi.org/10.1109/CEC.2018.8477767

Gong, M., Zhou, H., Wang, Q., Wang, S., & Yang, P. (2020). District heating systems load forecasting: a deep neural networks model based on similar day approach. *Advances in Building Energy Research*, *14*(3), 372–388. https://doi.org/10.1080/17512549.2019.1607777

Goyal, M., Pandey, M., & Thakur, R. (2020). Exploratory Analysis of Machine Learning Techniques to predict Energy Efficiency in Buildings. *ICRITO 2020 - IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, 1033–1037. https://doi.org/10.1109/ICRITO48877.2020.9197976

Guo, T., & Li, X. (2023). Machine learning for predicting phenotype from genotype and environment. *Current Opinion in Biotechnology*, *79*, 102853. https://doi.org/10.1016/j.copbio.2022.102853

Hassan, S. U., Ahamed, J., & Ahmad, K. (2022). Analytics of machine learning-based algorithms for text classification. *Sustainable Operations and Computers*, *3*(July 2021), 238–248. https://doi.org/10.1016/j.susoc.2022.03.001

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 1833–1842. https://doi.org/10.1109/HICSS.2014.231

Hernandez-Matheus, A., Löschenbrand, M., Berg, K., Fuchs, I., Aragüés-Peñalba, M., Bullich-Massagué, E., & Sumper, A. (2022). A systematic review of machine learning techniques related to local energy communities. *Renewable and Sustainable Energy Reviews*, *170*(October), 112651. https://doi.org/10.1016/j.rser.2022.112651

Holmes, G., Donkin, A., & Witten, I. H. (1994). WEKA: A machine learning workbench. *Australian and New Zealand Conference on Intelligent Information Systems - Proceedings*, 357–361. https://doi.org/10.1109/anziis.1994.396988

Huang, S., Zuo, W., & Sohn, M. D. (2018). A Bayesian Network model for predicting cooling load of commercial buildings. *Building Simulation*, *11*(1), 87–101. https://doi.org/10.1007/s12273-017-0382-z

Huang, Y., & Li, C. (2021). Accurate heating, ventilation and air conditioning system load prediction for residential buildings using improved ant colony optimization and wavelet neural network. *Journal of Building Engineering*, *35*(September 2020), 101972. https://doi.org/10.1016/j.jobe.2020.101972

Huang, Z., Wu, Y., Tempini, N., Lin, H., & Yin, H. (2022). An Energy-efficient And Trustworthy Unsupervised Anomaly Detection Framework (EATU) for IIoT. *ACM Transactions on Sensor Networks*, *18*(4), 1–18. https://doi.org/10.1145/3543855

Invidiata, A., Lavagna, M., & Ghisi, E. (2018). Selecting design strategies using multi-criteria decision making to improve the sustainability of buildings. *Building and Environment*, *139*(November 2017), 58–68. https://doi.org/10.1016/j.buildenv.2018.04.041

Jiang, Y., He, X., Lee, M. L. T., Rosner, B., & Yan, J. (2020). Wilcoxon rank-based tests for clustered data with r package clusrank. *Journal of Statistical Software*, *96*, 1–26. https://doi.org/10.18637/jss.v096.i06

Kabir, R., Rahman, A., & Samad, T. (2017). A Network Intrusion Detection Framework based on Bayesian Network using Wrapper Approach. *International Journal of Computer Applications*, *166*(4), 13–17. https://doi.org/10.5120/ijca2017913992

Karatzas, K., & Katsifarakis, N. (2018). Modelling of household electricity consumption with the aid of computational intelligence methods. *Advances in Building Energy Research*, *12*(1), 84–96. https://doi.org/10.1080/17512549.2017.1314831

Kardani, N., Bardhan, A., Kim, D., Samui, P., & Zhou, A. (2021). Modelling the energy performance of residential buildings using advanced computational frameworks based on RVM, GMDH, ANFIS-BBO and ANFIS-IPSO. *Journal of Building Engineering*, *35*(July 2020), 102105. https://doi.org/10.1016/j.jobe.2020.102105

Khanzode, K. C. A. (2020). *ADVANTAGES AND DISADVANTAGES OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING : A LITERATURE REVIEW*. *9*(1), 30–36.

Khatibi, A., Jahangir, M. H., & Astaraei, F. R. (2023). Developing an IoT-based electrochromic windows for smart buildings. *Advances in Building Energy Research*. https://doi.org/10.1080/17512549.2023.2175371

Kim, M., Jun, J. A., Song, Y. J., & Pyo, C. S. (2020). Explanation for building energy prediction. *International Conference on ICT Convergence*, *2020-Octob*, 1168–1170. https://doi.org/10.1109/ICTC49870.2020.9289340

Levinson, A. (2016). How much energy do building energy codes save? Evidence from California houses. *American Economic Review*, *106*(10), 2867–2894. https://doi.org/10.1257/aer.20150102

Li, M., Nanda, G., Chhajedss, S. S., & Sundararajan, R. (2020). Machine learning-based decision support system for early detection of breast cancer. *Indian Journal of Pharmaceutical Education and Research*, *54*(3), S705–S715. https://doi.org/10.5530/ijper.54.3s.171

Li, R. A., McDonald, J. A., Sathasivan, A., & Khan, S. J. (2021). A multivariate Bayesian network analysis of water quality factors influencing trihalomethanes formation in drinking water distribution systems. *Water Research*, *190*, 116712. https://doi.org/10.1016/j.watres.2020.116712

Li, Y., Zou, C., Berecibar, M., Nanini-Maury, E., Chan, J. C. W., van den Bossche, P., Van Mierlo, J., & Omar, N. (2018). Random forest regression for online capacity estimation of lithium-ion batteries. *Applied Energy*, *232*(February), 197–210. https://doi.org/10.1016/j.apenergy.2018.09.182

Liapikos, T., Zisi, C., Kodra, D., Kademoglou, K., Diamantidou, D., Begou, O., Pappa-Louisi, & Theodoridis, G. (2022). Quantitative structure retention relationship (QSRR) modelling for Analytes' retention prediction in LC-HRMS by applying different Machine Learning algorithms and evaluating their performance. *Journal of Chromatography B*, *1191*. https://doi.org/https://doi.org/10.1016/j.jchromb.2022.123132

Liu, H., Liang, J., Liu, Y., & Wu, H. (2023). A Review of Data-Driven Building Energy Prediction. *Buildings*, *13*(2). https://doi.org/10.3390/buildings13020532

Lokhandwala, M., & Nateghi, R. (2018). Leveraging advanced predictive analytics to assess commercial cooling load in the U.S. *Sustainable Production and Consumption*, *14*, 66–81. https://doi.org/10.1016/j.spc.2018.01.001

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *2017-Decem*(Section 2), 4766–4775.

Machlev, R., Heistrene, L., Perl, M., Levy, K. Y., Belikov, J., Mannor, S., & Levron, Y. (2022). Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI*, *9*(March), 100169. https://doi.org/10.1016/j.egyai.2022.100169

Mao, Y., Yang, H., Sheng, Y., Wang, J., Ouyang, R., Ye, C., Yang, J., & Zhang, W. (2021). Prediction and Classification of Formation Energies of Binary Compounds by Machine Learning: An Approach without Crystal Structure Information. *ACS Omega*, *6*(22), 14533–14541. https://doi.org/10.1021/acsomega.1c01517

Mastrucci, A., van Ruijven, B., Byers, E., Poblete-Cazenave, M., & Pachauri, S. (2021). Global scenarios of residential heating and cooling energy demand and CO2 emissions. *Climatic Change*, *168*(3–4), 1–26. https://doi.org/10.1007/s10584-021-03229-3

Medal, L. A., Sunitiyoso, Y., & Kim, A. A. (2021). Prioritizing Decision Factors of Energy Efficiency Retrofit for Facilities Portfolio Management. *Journal of Management in Engineering*, *37*(2), 1–12. https://doi.org/10.1061/(asce)me.1943-5479.0000878

Mishra, P., Swain, B. R., & Swetapadma, A. (2022). A Review of Cancer Detection and Prediction Based on Supervised and Unsupervised Learning Techniques. In P. K. Pattnaik, A. Vaidya, S. Mohanty, S. Mohanty, & A. Hol (Eds.), *Smart Healthcare Analytics: State of the Art* (pp. 21–30). Springer Singapore. https://doi.org/10.1007/978-981-16-5304-9_3

Moayedi, H., & Mosavi, A. (2021). Suggesting a stochastic fractal search paradigm in combination with artificial neural network for early prediction of cooling load in residential buildings. *Energies*, *14*(6). https://doi.org/10.3390/en14061649

Moayedi, H., Nguyen, H., & Kok Foong, L. (2021). Nonlinear evolutionary swarm intelligence of grasshopper optimization algorithm and gray wolf optimization for weight adjustment of neural network. *Engineering with Computers*, *37*(2), 1265–1275. https://doi.org/10.1007/s00366-019-00882-2

Mokeev, V. V. (2019). Prediction of heating load and cooling load of buildings using neural network. *Proceedings - 2019 International Ural Conference on Electrical Power Engineering, UralCon 2019*, 417–421. https://doi.org/10.1109/URALCON.2019.8877655

Molnar, C. (n.d.). *Interpretable Machine Learning*. Retrieved January 22, 2023, from https://christophm.github.io/interpretable-ml-book/

Muhammad Irfan, & Faizir Ramlie. (2021). Analysis of Parameters which Affects Prediction of Energy Consumption in Buildings using Partial Least Square (PLS) Approach. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *25*(1), 61–68. https://doi.org/10.37934/araset.25.1.6168

Nazir, A., Wajahat, A., Akhtar, F., Ullah, F., Qureshi, S., Malik, S. A., & Shakeel, A. (2020). Evaluating Energy Efficiency of Buildings using Artificial Neural Networks and K-means Clustering Techniques. *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies: Idea to Innovation for Building the Knowledge Economy, ICoMET 2020*. https://doi.org/10.1109/iCoMET48670.2020.9073816

Ong, S., & Clark, N. (2022). *Open Energy Data nitiative (OEDI)*. https://data.openei.org/submissions/153

Osman, A. I., Chen, L., Yang, M., Msigwa, G., Farghali, M., Fawzy, S., Rooney, D. W., & Yap, P. S. (2022). Cost, environmental impact, and resilience of renewable energy under a changing climate: a review. *Environmental Chemistry Letters*, *21*(2), 741–764. https://doi.org/10.1007/s10311-022-01532-8

Pandit, P., Dey, P., & Krishnamurthy, K. N. (2021). Comparative Assessment of Multiple Linear Regression and Fuzzy Linear Regression Models. *SN Computer Science*, *2*(2), 1–8. https://doi.org/10.1007/s42979-021-00473-3

Permai, S. D., & Tanty, H. (2018). Linear regression model using bayesian approach for energy performance of residential building. *Procedia Computer Science*, *135*, 671–677. https://doi.org/10.1016/j.procs.2018.08.219

Perolat, J., Couso, I., Loquin, K., & Strauss, O. (2015). Generalizing the Wilcoxon rank-sum test for interval data. *International Journal of Approximate Reasoning*, *56*(PA), 108–121. https://doi.org/10.1016/j.ijar.2014.08.001

Phan, L., & Lin, C. X. (2014). A multi-zone building energy simulation of a data center model with hot and cold aisles. *Energy and Buildings*, *77*, 364–376. https://doi.org/10.1016/j.enbuild.2014.03.060

Pradhan, P., Behera, P. K., & Ray, B. N. B. (2016). Modified Round Robin Algorithm for Resource Allocation in Cloud Computing. *Procedia Computer Science*, *85*(Cms), 878–890. https://doi.org/10.1016/j.procs.2016.05.278

Prasetiyo, B., Alamsyah, & Muslim, M. A. (2019). Analysis of building energy efficiency dataset using naive bayes classification classifier. *Journal of Physics: Conference Series*, *1321*(3). https://doi.org/10.1088/1742-6596/1321/3/032016

Pruneski, J. A., Pareek, A., Kunze, K. N., Martin, R. K., Karlsson, J., Oeding, J. F., Kiapour, A. M., Nwachukwu, B. U., & Williams, R. J. (2022). Supervised machine learning and associated algorithms: applications in orthopedic surgery. *Knee Surgery, Sports Traumatology, Arthroscopy*, *31*(4), 1196–1202. https://doi.org/10.1007/s00167-022-07181-2

Renuka, S. M., Maharani, C. M., Nagasudha, S., & Raveena Priya, R. (2022). Optimization of energy consumption based on orientation and location of the building. *Materials Today: Proceedings*, *65*, 527–536. https://doi.org/10.1016/j.matpr.2022.03.081

Riahi, G. (2015). E-learning systems based on cloud computing: A review. *Procedia Computer Science*, *62*(Scse), 352–359. https://doi.org/10.1016/j.procs.2015.08.415

Rodríguez, M. V., Cordero, A. S., Melgar, S. G., & Andújar Márquez, J. M. (2020). Impact of global warming in subtropical climate buildings: Future trends and mitigation strategies. *Energies*, *13*(23), 1–22. https://doi.org/10.3390/en13236188

Roostaei, J., Colley, S., Mulhern, R., & May, A. A. (2021). Predicting the risk of GenX contamination in private well water using a machine-learned Bayesian network model. *Journal of Hazardous Materials*, *411*(October 2020), 125075. https://doi.org/10.1016/j.jhazmat.2021.125075

Saboni, A., Ouamane, M. R., Bennis, O., & Kratz, F. (2022). Model Reports, a Supervision Tool for Machine Learning Engineers and Users. *International Journal of Education and Information Technologies*, *16*(February), 50–54. https://doi.org/10.46300/9109.2022.16.5

Sala, J., Li, R., & Christensen, M. H. (2021). Clustering and classification of energy meter data: A comparison analysis of data from individual homes and the aggregated data from multiple homes. *Building Simulation*, *14*(1), 103–117. https://doi.org/10.1007/s12273-019-0587-4

Santamouris, M. (2016). Cooling the buildings – past, present and future. *Energy and Buildings*, *128*, 617–638. https://doi.org/10.1016/j.enbuild.2016.07.034

Santamouris, M., Cartalis, C., Synnefa, A., & Kolokotsa, D. (2015). On the impact of urban heat island and global warming on the power demand and electricity consumption of buildings - A review. *Energy and Buildings*, *98*, 119–124. https://doi.org/10.1016/j.enbuild.2014.09.052

Sarkar, A., & Bardhan, R. (2020). Optimal interior design for naturally ventilated low-income housing: a design-route for environmental quality and cooling energy saving. *Advances in Building Energy Research*, *14*(4), 494–526. https://doi.org/https://doi.org/10.1080/17512549.2019.1626764

Scanagatta, M., Salmerón, A., & Stella, F. (2019). A survey on Bayesian network structure learning from data. *Progress in Artificial Intelligence*, *8*(4), 425–439. https://doi.org/10.1007/s13748-019-00194-y

Senarathne, L. R., Nanda, G., & Sundararajan, R. (2022). Influence of building parameters on energy efficiency levels: a Bayesian network study. *Advances in Building Energy Research*, *16*(6), 780–805. https://doi.org/10.1080/17512549.2022.2108142

Shanthi, J ;Srihari, B. (2018). Prediction of Heating and Cooling Load to improve Energy Efficiency of Buildings Using Machine Learning Techniques. *Journal of Mechanics of Continu and Mathematical Sciences*, *13*(5). https://doi.org/https://doi.org/10.26782/jmcms.2018.12.00008

Shapash. (n.d.). Retrieved November 30, 2022, from https://shapash.readthedocs.io/en/latest/index.html%0A

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. *Proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development, INDIACom 2016*, 1310–1315.

Srihari, J. (2018). Prediction of Heating and Cooling Load to improve Energy Efficiency of Buildings Using Machine Learning Techniques. *Journal of Mechanics of Continua and Mathematical Sciences*, *13*(5), 97–113. https://doi.org/10.26782/jmcms.2018.12.00008

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x

Tian, Z., Si, B., Shi, X., & Fang, Z. (2019). An application of Bayesian Network approach for selecting energy efficient HVAC systems. *Journal of Building Engineering*, *25*(November 2018), 100796. https://doi.org/10.1016/j.jobe.2019.100796

Timmons, D., Zirogiannis, N., & Lutz, M. (2016). Location matters: Population density and carbon emissions from residential building energy use in the United States. *Energy Research and Social Science*, *22*, 137–146. https://doi.org/10.1016/j.erss.2016.08.011

Tsanas, A., & Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, *49*, 560–567. https://doi.org/10.1016/j.enbuild.2012.03.003

Tsoka, T., Ye, X., Chen, Y. Q., Gong, D., & Xia, X. (2022). Explainable artificial intelligence for building energy performance certificate labelling classification. *Journal of Cleaner Production*, *355*(December 2021), 131626. https://doi.org/10.1016/j.jclepro.2022.131626

UCI Machine Learning Repository. (n.d.). Energy Efficiency Data Set. Retrieved January 1, 2022, from https://archive.ics.uci.edu/ml/datasets/Energy%2Befficiency

van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, *79*, 102470. https://doi.org/10.1016/j.media.2022.102470

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E. J. (2020). Bayesian rank-based hypothesis testing for the rank sum test, the signed rank test, and Spearman's ρ. *Journal of Applied Statistics*, *47*(16), 2984–3006. https://doi.org/10.1080/02664763.2019.1709053

Wang, Z., Liu, J., Zhang, Y., Yuan, H., Zhang, R., & Srinivasan, R. S. (2021). Practical issues in implementing machine-learning models for building energy efficiency: Moving beyond obstacles. *Renewable and Sustainable Energy Reviews*, *143*(August 2020), 110929. https://doi.org/10.1016/j.rser.2021.110929

Xie, X., Wu, T., Zhu, M., Jiang, G., Xu, Y., Wang, X., & Pu, L. (2021). Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land. *Ecological Indicators*, *120*, 106925. https://doi.org/10.1016/j.ecolind.2020.106925

Xie, Y., Wu, D., Dong, B., & Li, Q. (2022). *Trained Model in Supervised Deep Learning is a Conditional Risk Minimizer*. 1–24. http://arxiv.org/abs/2202.03674

Xu, X., Taylor, J. E., Pisello, A. L., & Culligan, P. J. (2012). The impact of place-based affiliation networks on energy conservation: An holistic model that integrates the influence of buildings, residents and the neighborhood context. *Energy and Buildings*, *55*, 637–646. https://doi.org/10.1016/j.enbuild.2012.09.013

Yan, L., & Liu, M. (2020). A simplified prediction model for energy use of air conditioner in residential buildings based on monitoring data from the cloud platform. *Sustainable Cities and Society*, *60*(June 2019), 102194. https://doi.org/10.1016/j.scs.2020.102194

Yoro, K. O., & Daramola, M. O. (2020). CO2 emission sources, greenhouse gases, and the global warming effect. In *Advances in Carbon Capture*. Elsevier Inc. https://doi.org/10.1016/b978-0-12-819657-1.00001-3

Yu, T., Boob, A. G., Volk, M. J., Liu, X., Cui, H., & Zhao, H. (2023). Machine learning-enabled retrobiosynthesis of molecules. *Nature Catalysis*, *6*(2), 137–151. https://doi.org/10.1038/s41929-022-00909-w

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration*, *23*(March), 100224. https://doi.org/10.1016/j.jii.2021.100224

Zhang, Y., Teoh, B. K., Wu, M., Chen, J., & Zhang, L. (2023). Data-driven estimation of building energy consumption and GHG emissions using explainable artificial intelligence. *Energy*, *262*(PA), 125468. https://doi.org/10.1016/j.energy.2022.125468

Zou, X., Hu, Y., Tian, Z., & Shen, K. (2019). Logistic Regression Model Optimization and Case Analysis. *Proceedings of IEEE 7th International Conference on Computer Science and Network Technology, ICCSNT 2019*, 135–139. https://doi.org/10.1109/ICCSNT47585.2019.8962457

PUBLICATIONS

a) Journal Papers:

1. Senarathne, L.R., Nanda, G., Sundararajan, R. (2022). Influence of Building parameters on Energy Efficiency Levels: A Bayesian Network Study. Advances in Building Energy Research, 16:6,780-805, doi: 10.1080/17512549.2022.2108142

2. Senarathne, L.R., Nanda, G., Sundararajan, R. (2023). Identifying Salient Features of Cooling Energy Usage of Commercial Buildings using Explainable Machine Learning. Advances in Building Energy Research (In Revision)

b) Conference Paper:

Senarathne, L.R., Nanda, G., Sundararajan, R. "Supervised Machine Learning Models to Assess Impact of Building Parameters on Energy Efficiency,"*2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)*, Bangalore, India, 2022, pp. 1-7, doi: 10.1109/GCAT55367.2022.9971834.