Predicting Energy Consumption in Commercial Buildings using its Property features and Machine Learning Algorithms

MSc Research Project
Data Analytics

Digvijay Rai
Student ID: x18134645

School of Computing
National College of Ireland

Supervisor: Dr. Muhammad Iqbal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Digvijay Rai
Student ID:	x18134645
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Muhammad Iqba
Submission Due Date:	12/12/2018
Project Title:	Predicting Energy Consumption in Commercial Buildings using property features and Machine Learning Algorithms
Word Count:	7918
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	11th December 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Energy Consumption in Commercial Buildings using its Property features and Machine Learning Algorithms

Digvijay Rai
x18134645

Abstract

Population growth is a very crucial factor that leads to an upsurge in demand for residential services and lavishness, which is triggering the depletion of energy assets. A stable surge has been noticed from 20% to 40% worldwide towards energy consumption collectively from both residential and commercial buildings. In the past few years, many regression algorithms have been used for predicting energy consumption in buildings. The state-of-the-art for energy consumption prediction in commercial buildings using its property features was 82% (r^2 score) based on the topical studies. This study focuses on building a novel multi-class classifier that will categorize energy consumption prediction in multinomial classes using the Commercial Buildings Energy Consumption Survey (CBECS) dataset. This research compares four classification algorithms, namely Gaussian Naïve Bayes, Random Forest, K-Nearest Neighbour and Logistic Regression using analysis of variance (ANOVA) and principal component analysis(PCA). Accuracy, precision, recall, and f1 score are considered as evaluation metrics for this study. Amongst all the classification algorithms, K-Nearest Neighbor achieved the best efficiency of 97% for both ANOVA and PCA, followed by Random forest and Gaussian Naive Bayes . The accuracy of each model was evaluated using the 10-Fold Cross-Validation technique.

Keywords: *Energy consumption, Commercial Buildings, Property Features, Buildings Energy Consumption Survey*

1 Introduction

1.1 Background

Power is obtained using physical and chemical resources, which is called Energy, primarily to offer light and heat. Soon, shortage of energy can be noticed due to the massive growth in the population and increase in development of commercial lavishness and services In the year 2015, U.S. Energy Information Administration (EIA) released a data which provides information on energy consumed by commercial and residential buildings is 39 quadrillion Btus. (British Thermal Unit), that is 40% of overall energy consumed in the United States. Similarly, as per the European Commission, commercial and residential buildings consumed 40% of the overall energy consumed in the European Union (EU). Also, the future design of commercial and residential buildings may have an effect on energy consumption. International Energy Agency has gathered frightening data on trends of energy consumption (Pérez-Lombard et al.; 2008). In the earlier two decades (1984-2004), growth in energy by 49% had been noticed with an upsurge of 2% in an average annual increase.

1.2 Motivation

As per the studies and predictions did in recent times in the field of energy consumption in buildings. Energy consumption continues to grow in a significant way in emerging financial countries (Africa, Middle East, America, and Southeast Asia) will regularly increase with a standard annual proportion of 3.2% and this will exceed in the year 2020 with average growth rate of 1.1% by the technologically advanced nations (Australia, Western Europe, New Zealand, North America, and Japan). Enhancement of energy conservation has become very critical due to the rapid increase of the world population, ensuing growth in energy consumption, depletion of the ozone layer, and decreasing natural resources.

Since 2003, there is a 14% of the increase in total buildings and a 22% of the increase in total floor space and an upsurge of 7% has been noticed in energy consumed by 5.6 million commercial buildings in the United States according to the survey conducted by CBECS in the year 2012 [1]. As per the Figure 1, of the survey demonstration that's in 2012, commercial buildings in U.S have consumed total energy of 6,963 trillion of British thermal units (Btu): 4,241 trillion Btu consumed by electricity, 2,248 trillion Btu consumed by natural gas, 341 trillion Btu consumed by distinct heat and 134 trillion Btu consumed by fuel oil. The usage of natural gas and electricity has been increased by 7% and 19% respectively. Total energy usage in commercial buildings has doubled due to the massive increase in the use of electricity since CBECS began to have a record of it. Electricity usage augmented from 2,200 trillion Btu in the year 1979 to 4,241 trillion Btu in the year 2012. Due to the successive rise in consumption of energy, depletion of the ozone layer, and shrinking of natural resources, the necessity to improve conservation of energy has become critical.

Energy consumption in recent years has become a serious issue because of the rate in which energy demand is increasing, and it is predicted to increase at a very high rate. Also, it is essential to investigate the factors contributing to high energy consumption. This investigation will be helpful for stakeholders to generate various alternative measures in order to control the usage of energy in the country. This research is based on the Atlanta dataset of the United States, and once it is successfully completed, then the same method can be applied to other cities of the United States to classify the energy consumption in both commercials as well as residential buildings. This study considers a single city of United states grounded on the available data, but similar kinds of data can be gathered from other cities of the United States or from different countries, and this can answer real-world problems regardless of the country.

1.3 Research Question:

" Can energy consumption be accurately predicted as very-low, low, medium, and high classes using the commercial buildings property features and machine learning algorithms"?

[1] https://www.eia.gov/consumption/commercial/reports/2012/energyusage/

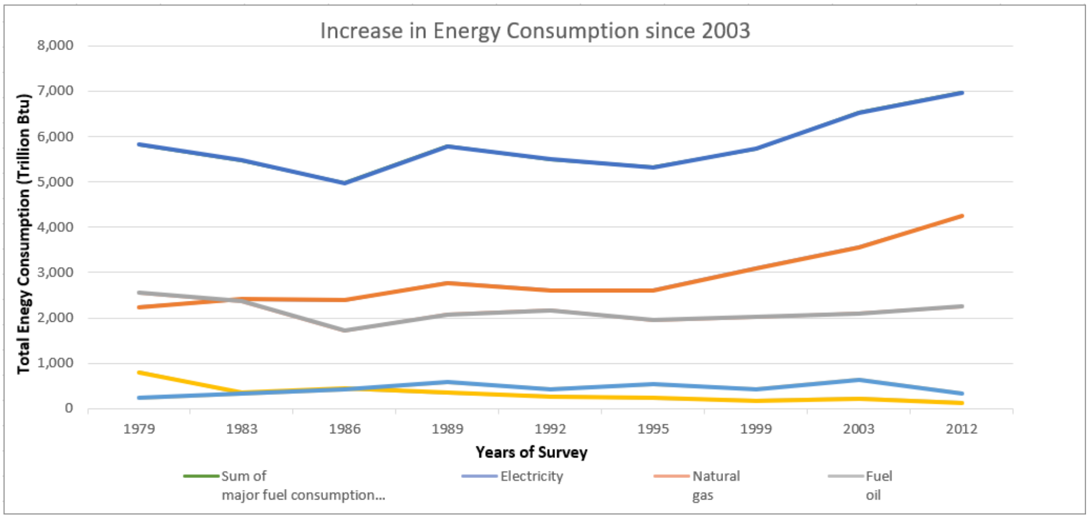

Figure 1: Increase in Energy Consumption since 2003

Property features of the commercial buildings are taken into consideration to predict energy consumption into classes of four, namely: very-low,low, medium, and high, using four different classification machine learning algorithms. These property features will be advantageous in predicting the accuracy of energy consumed in commercial buildings. This will help government officials and city developers can make improved decisions for saving energy in the future.

1.4 Proposed Objectives and Deliverable's:

1. To study the significance of the CBECS dataset and perform pre-processing.
2. Binning the energy consumption data into four classes, namely: Very-Low, Low, Medium, and High.
3. To reduce the variables affecting energy consumption prediction using the Dimensionality Reduction technique PCA and feature selection technique ANOVA.
4. To predict energy consumption in commercial buildings using property features and four classification algorithms like Random Forest, Gaussian Naïve Bayes, Logistic Regression and K-Nearest Neighbour.
5. Evaluation and Comparison of the best classification algorithm for predicting energy consumption.

The structure of this paper is arranged as follows: Section 2 analyzes the related work performed previously in this domain, Section 3 explains the KDD methodology used in the research, Section 4 discusses the two-tier design used for this research, Section 5 discusses implementation steps of this study, Section 6 analyzes and discusses the results of all the four classifiers and Section 7 concludes and provides future scope of the research.

2 Related Work

2.1 Predicting Energy consumption in Commercial buildings using CBECS dataset:

The importance of the CBECS dataset can be seen as different researchers used it for their respective projects. CBECS dataset had been used in research papers of (Deng et al.; 2018), (Robinson et al.; 2017),(Bin; 2009) [Nateghi]to predict energy consumption in commercial buildings by applying different machine learning algorithms. (Bin; 2009) used ANN and multi-linear regression model to predict energy use intensity. As a result, ANN helped to provide better accuracy when compared to the multi-linear regression model when diverse climatic zones grouped census data. (Robinson et al.; 2017) used many regression models, namely SVR, Linear regression, Ridge regression, XGBoost, etc., to predict energy consumption in which XGBoost outperformed every other regression model in all the metrics by providing best accuracy amongst all the regression models. As per researcher, the two most influential factors in commercial buildings are density and Floor-area-ratios (FAR), which tend to increase energy consumption. (Deng et al.; 2018) had used both statistical and machine learning methods and made a comparison between them to check which one is providing better results of energy consumption in commercial buildings. In the statistical and machine learning model, the researcher used principal component analysis and random forest respectively for dimensionality reduction. As a result, machine learning models like support vector machines and random forest showed better performance than the statistical models.

2.2 Predicting Energy Consumption in Commercial Buildings using Simulation Tools:

Researchers had used a lot of simulation tools to generate data for both commercial and residential buildings. Researchers (Seyedzadeh et al.; 2019) and (Tsanas and Xifara; 2012) had used a simulation tool known as Building Performance Simulation (BPS) to create simulated data for 786 diverse kinds of buildings including eight essential input variables. Apart from the Random Forest model (Tsanas and Xifara; 2012)), researcher-developed iteratively reweighted the least-squares (IRLS) model to predict two output variables like heating load (HL) and cooling load (CL) for 768 diverse residential buildings. As per the result, the Random forest outperformed the IRLS model in predicting the heating load and cooling load of residential buildings. Researchers (Seyedzadeh et al.; 2019) simulated two different, namely Eco test and Energyplus datasets using the BPS tool. Eco Test dataset has eight essential input variables, and EnergyPlus has twenty-eight input variables to predict accuracy, training time, tuning, and response time using Random Forest, Artificial Neural Network, Support Vector Machine (SVM), Gaussian Process Regression and Gradient Boosted Regression Tree (GBRT). Based on the result, for the GBRT model, the value of Root Mean Squared Error was very accurate, and SVM outperforms every other model for the simplicity and speed of calculation. The Neural Network model tends to perform faster for an enormous amount of energy simulation data than the other models of machine learning.

In these research papers, (Salakij et al.; 2016) and (Shen et al.; 2019) developed Building Energy Analysis Model (BEAM) and SimBldPy models to compare and validate with simulated results of EnergyPlus. (Shen et al.; 2019) Used SimBldPy, which a python-based tool for simulation which is capable of analyzing the content from text-based files. Random Forest technique was applied to find the effect of climate change on energy consumption of building on an hourly basis. Both energy consumption and hourly climate datasets were trained using the Random Forest technique. Several over-fitting and optimization problems were resolved using the SimBldPy modeling tool and random forest technique. The researcher (Salakij et al.; 2016) used Model-Based Predictive Control (MBPC) to develop BEAM for predicting heat and moisture transfer of the buildings. A model was developed by implementing Linear Quadratic Tracking (LQT) method to update the set point of temperature based on the occupant's activities. According to the outcomes, the MBPC method aided to preserve 43% of thermal energy utilization when compared with the traditional method.

Researchers (Rahman and Smith; 2017) and (Amasyali and El-Gohary; 2018) used the EnergyPlus simulation tool, which fits in the capabilities of BLAST and DOE-2, which are other simulation tools. Real climatic data of the years 2013 and 2014 along with schedule factors which are default values generated from EnergyPlus to predict future fuel consumption of different types of buildings (Rahman and Smith; 2017). Updated variables were utilized to estimate the building's heating and cooling loads, which were reliant upon heat and mass balances. The EnergyPlus generated hourly fuel consumption data was utilized as the dependent variable, and model variables like weather data and schedule variables were utilized as an independent variable. Four models, namely Neural network, Gaussian process regressor Neural auto-regressive with eXogenous input (NARX), and ridge regression, were used to train and test the datasets. As per the outcomes, the Neural network and Gaussian process had less error prediction (which were within 5%) compared to NARX and ridge regression. Along with using the EnergyPlus simulation tool for generating, training, and testing data, the researcher. (Amasyali and El-Gohary; 2018) also used various data-based dependent algorithms, including deep neural network, random forest, linear regression, and random forest, were implemented and tested based on the accuracy to predict hourly cooling energy utilization of the building. As a result, best perform was shown by SVM in terms of coefficient of variance (CV) with 8.59% and coefficient of determination (R squared)with 96.63% followed by deep neural network with CV of 8.88% and R square of 96.11%, random forest with CV of 9.35% and R square of 95.59% and linear regression with CV of 19.99% and R square of 79.11%. Amongst all models, the linear regression model took 1-sec of training time. However, the support vector machine model was an excellent fit to predict the consumption of cooling energy in the building.

The researcher (García-Martín et al.; 2019) used the Massive Online Analysis tool to generate data synthetically and run the algorithms in the same tool. The researcher used prequential evaluation, where models were tested and then trained, and it provided online accuracy measurements. Three different models of a convolutional neural network, namely, Inception-V3, MobileNET and DenseNet, and obtained 68.91% of average relative accuracy over the three-test convolutional neural network model. The forecasted and calculated energy consumption for the MobileNet convolution layer is least, and it represents 95% computation inside the deep convolutional neural network. Amongst the three-convolutional neural network that MobileNet was the best energy-saving convolutional neural network under the same implementation environments.

2.3 Using historical data for predicting consumption of energy in commercial Building:

Many researchers have used historical data in their study to predict energy consumption in commercial buildings. Researcher (Ahmad et al.; 2017) have used machine learning algorithms such as Articial Neural Network (ANN) and Random Forest to make a prediction of energy consumption in buildings. Climatic condition data was gathered from the closest meteorological station at a Madrid airport along with historical data of electricity utilization from the hotel building. In this study, A comparison based on performance has been made between a random forest and ANN. As per the results, ANN performed marginally better than random forest, and the researcher also noticed that ANN is providing high accuracy for a completely new testing dataset. Researcher (Xu et al.; 2019) and his team had combined ANN algorithm with an SNA (social network analysis) to predict energy consumption in multiple buildings. For this study, the researcher has done a correlation test on historical data of energy utilization in buildings. 17 building's historical information had been considered from the campus of Southeast University, China, for validating the SNA-ANN method. Four different types of groups, namely office, laboratory, residential, and education, had been created once the month electricity used data was gathered from the buildings. Then the gathered data was trained and validated with the combined SNA-ANN method, in which all the four groups of buildings have shown high prediction accuracy.

A non-linear correlation was noticed by the researcher (Liu and Chen; 2013) between light energy consumption and variables of office buildings. The researcher developed an SVR (support vector regressor) model due to the excellent performance in the past forecast of light energy consumption using historical data of office building. The predicted outcomes showed that the ANN model performed better than the SVR model. Train and test dataset of SVR model showed prediction accuracy of 0.9845 R square and 0.1233 MSE and 0.9273 R square, 0.6571 MSE respectively when compared to train and test dataset of ANN, 0.9920 R square, 0.0623 MSE, and 0.6980 R square, 3.1403 MSE respectively. (Nateghi and Mukherjee; 2017) had used Representative Concentration Pathways (RCP), which was appropriated from the IPCC's 5th Assessment Report to evaluate the climatic change for an extended period in commercial and residential buildings in the districts of Indiana. Two climate scenarios, namely a high level of greenhouse concentration (RCP8.5) and stabilized scenario (RCP4.5), were used in this study. The Bayesian predictive model was used to train and test the historical data of energy demand and climatic conditions. Based on the research outcome, the researchers concluded that the energy consumption of an average building under both climate condition scenarios is projected to surge by 5.4% for RCP 8.5 and 5.1% for RCP 4.5. Additionally, during winter heating, energy demand will reduce by 3.5% in an average household. Because of the anticipated daytime temperature, there is a down surge in the utilization of cooling energy in residential sectors when compared to in commercial sectors. Researcher (Wahid and Kim; 2016) aim was to ease the energy providers to conclude the distribution of energy to several flats based on their demand. The researcher used (K-Nearest Neighbors) KNN classifier for regular prediction of energy consumption based on classification. The historical data of 520 flats of Seoul, the Republic of Korea, which contains hourly energy consumption data, were utilized and divided into 60% of training and 40% of testing ratio, providing 95.96% of accurate results. The model efficiency was validated using 5-fold and 10-fold cross-validation. The researcher Casteleiro-Roca et al. (2019) presented a new method to forecast energy load for a hotel based on a cross intelligent topology executed with a combined technique of intelligent regression and clustering (Support Vector Regression and Artificial Neural Network). The historical data of a 5-star hotel located in Canary Island, Spain, was used, and the data contains information of its own energy demand, temperature, and occupancy rate as inputs. Prediction attained was satisfactory, demonstrating an encouraging potential for its utilization in managing the energy systems in hotels.

2.4 Predicting Energy Consumption using Transfer Learning Models, Extreme Deep Learning and Artificial Intelligence Techniques:

A method known as Hephaestus was proposed by the author (Ribeiro et al.; 2018). It was a method for predicting energy usage between the cross-building, which was usually utilized for transfer learning with time-series data. More extended timestamp data was gathered from other buildings to improve the prediction accuracy of the selected building, which had a lesser amount of information in the dataset. Hephaestus permits the utilization of machine learning techniques, which helps in data pre-processing and post-processing phases. The unwanted time series data was removed from many buildings and types it in time-independent data by adjusting the non-temporal section. The Hephaestus method was validated using a case study of energy consumption prediction for multiple schools. Additionally, numerous school's data were used for predicting the accuracy of the chosen school data, and results show 11.2% of increment in prediction accuracy.

Support Vector Regression (SVR) method was used by the researcher (Zhong et al.; 2019) to establish a model for energy consumption in residential buildings. The environmental (indoor and outdoor) attributes were used as input variables, which were achieved by data sampling and pre-processing. Fifty days of data with a time interval of one hour were gathered and merged (indoor environment data with meteorological data) to predict the building's cooling load (CL). The researcher divided the dataset into 70% of training data and 30% of testing data. Model creation and accuracy prediction was done using training data and tested the generalization ability by using testing data. Two state-of-the-art models, namely EDL (Extreme deep learning) and GBR (gradient boosting regression), were used for comparison. The RAE (relative absolute error), RMSE (root mean squared error), MAE (mean absolute error), and RRSE (root relative squared error) of the GBR and EDL models displayed the same performance and predicted models were higher than the R values. As concluded by the researcher by using EDL and GBR methods, predictive models can be built for elevated generalization ability and energy consumption. The researcher (Platon et al.; 2015) used institutional building of Canada data along with weather forecast data to predict the hourly energy consumption of the institutional building. CBR (case-based reasoning) and ANN (artificial neural network) were the artificial intelligence techniques that were utilized to develop the analytical model. PCA (Principal component analysis) was used for dimensionality reduction. As per the results, the ANN model always outperformed the CBR model by achieving a 7.3%

error rate.

2.5 Predicting Energy Consumption using Advance Machine Learning and Robust Data Mining Models:

A method of component-based was used by the author (Geyer and Singaravel; 2018) to develop models of advance machine learning for entire building design and describing a component's design of the entire building. Construction level components and zone-level components are the two decomposition levels that were examined. The Artifical neural network was used to display the behavior of the component due to the high-level flexibility of whole data regression. Dataset was trained using Bayesian regularization and Levenberg-Marquardt algorithms, which were executed in MATLAB. The complex dataset was trained using Bayesian regularization, whereas a less complicated dataset was trained using Levenberg-Marquardt algorithms. Almost every model achieved accuracy of a high level for R square. As per this research outcome, a high-quality prediction may be attained with 3.9% of errors for heating and 3.7% of errors for cooling based on the deviation of the training case from its data.

The author (Naganathan et al.; 2016) gives importance to launching a robust data-mining approach to determine complex and massive data formed by the technologies of tracking and sensing. This method gives a new approach for improved understanding of energy loss and prevention techniques during transmission. Two new concepts were presented in this paper: First, to set an algorithm for clustering, which would give information of different cluster substations, and second, build up a cluster alongside a semi-supervised machine learning method to mechanize the procedure to find the variables answerable for energy loss. The dataset was gathered from the official buildings of Arizona universities, and it had five million observations. The Semi-supervised energy model (SSEM) presents information on actual demand for energy data, which helps to integrate unlabeled data with labeled data to exceptional the value of energy loss. Data cleansing and pre-processing were done using the k-means algorithm. The data was divided into two parts i.e., 80% data for training and 20% for testing. In the anticipated model, semi-supervised learning will foresee the factors of unlabeled information by using labeled information as it quantifies the information confidence level before using them. This was the way to prevent energy loss by improving the accuracy of the output.

2.6 Predicting Energy Consumption using Sensor data, power transmitter data and electricity metered data:

In recent years many researchers had used different types of a tool like a sensor, power transmitter, and electricity to generate the electricity consumption data. A device known as FIESTA-IoT was used by the researcher (Smpokos et al.; 2018) to do a correlation check between the weather condition and the energy consumption in Data Center. The RealDC testbed was utilized to collect real-time data using sensors. A model was created with the correlation between the variables of energy consumption and the weather conditions using multivariable linear regression orderly by using weather attributes to predict energy consumption. As per the outcomes of this research, it was noticed that there were limited weather attributes which were affecting the energy consumption and developed model accomplishes to forecast the energy consumption based on climatic condition with the proper accuracy. The commercial buildings of Washington, D.C., and California, Seattle, were taken into consideration by the researcher (Touzani et al.; 2018) to gather the data from electricity meter and to do model performance evaluation. The weather condition dataset was obtained from the closest meteorological station. On a large data of energy consumption gradient boosting machine (GBM) algorithm was executed, and performance comparison was made with the available benchmark model and state-of-the-art model known as Time-of-Weekend Temperature (TOWT) and Random forest respectively. This paper results displayed that the GBM model helped to increase the energy reduction estimation accuracy of the buildings.

The researcher (Ahmad et al.; 2018) acquired limited data of energy consumption from the power transmitter for predicting energy demand using a supervised machine learning technique. Compact Regression Gaussian process (CRGP), Binary Decision Tree, generalized linear regression, and Stepwise Gaussian Processes Regression models were used for energy prediction. The CV (Coefficient of variation) and MAPE (mean absolute percentage error) were employed for the forecasting performance evaluation. Model of Binary Decision tree showed high accuracy, and it was precise in calculating the overall energy consumption, whereas the generalized linear regression model performance was lesser to some extent when compared to the other models. The Stepwise-GLR model faced unmanageable difficulties due to the huge number of noises which got implanted into the algorithm. The CRGP model usually allows any kind of fitting and forecasting, but in some situations, it might not be possible to analyze the deviations in the predicted outcome, and involving the precise method may be costly due to the size of the input variable set. Researcher (Ahmad et al.; 2017) used the data which was taken from the meter, which was getting transmitted by utilizing a cable RS232 and was collected by serial ports of the server. An application utilizing the visual basic was set up on the server to store every 15 seconds data. The researcher used logistic regression to monitor, predict and visualize electricity consumption data. The researcher (Dong et al.; 2018) used an electric water heater dataset captured by the smart meter called Ecotope. The dataset includes the data on regular hot water draws utilized with a calibrated 6-node prototype of a classic 50-gallon electric water boiler to produce profiles of electrical consumption for 20 homes. The researcher used three machine-learning, namely Gaussian Naïve Bayes, Random Forest, and Support Vector Machine, to predict energy consumption for 20 homes. As per the result, the random forest model performed the worst out to three machine learning algorithms by achieving 94% accuracy, and the model of support vector machine performed the best by achieving 96% accuracy.

3 Methodology

3.1 Introduction

This study aims to establish the best classification algorithm for classifying energy consumption in very-low, low, medium, and high categories for the CBECS dataset using Random forest, Gaussian Naïve Bayes, Logistic Regression and K-Nearest Neighbour. A method known as KDD (Knowledge discovery in datasets) is applied to complete this research project and retreat the knowledge from diverse sets of data (Naganathan et al.; 2016). The process of KDD consists of data storage, accessing the data, executing algorithms on massive datasets, and knowledge of the outcomes as seen in [2]Figure 2. Feature selection (Analysis of Variance) and dimensionality reduction (Principal component analysis) method are used in this project for selecting the best variables which contribute more to classifying the energy consumption in commercial buildings.

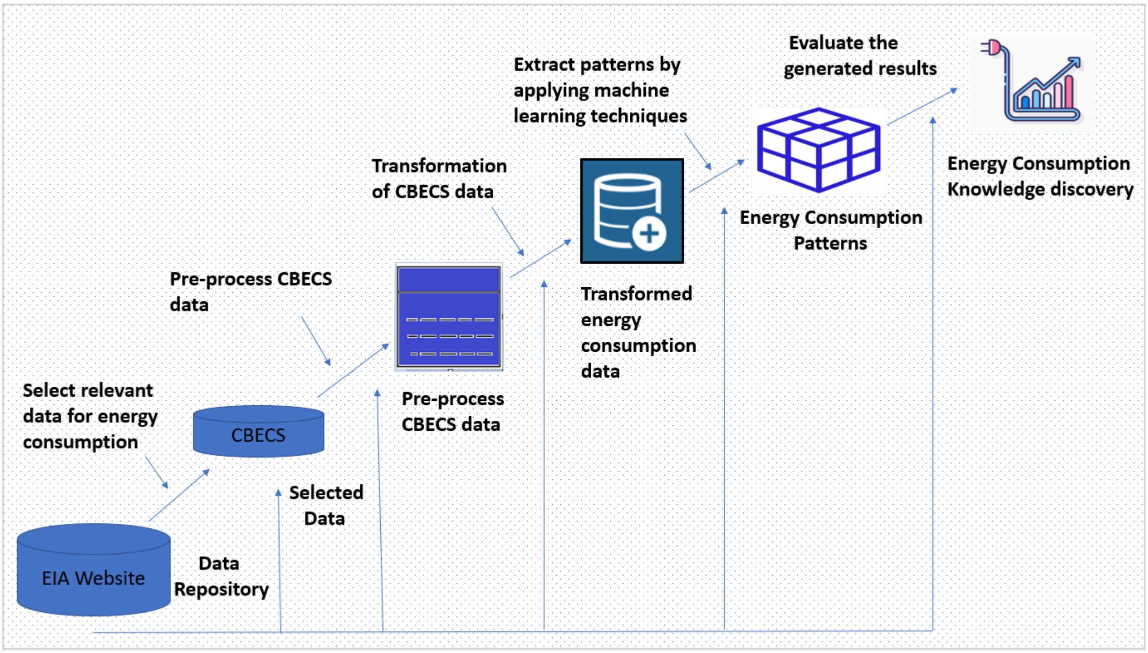
Evaluate the generated results
Extract patterns by applying machine learning techniques
Transformation of CBECS data
Energy Consumption Knowledge discovery
Energy Consumption Patterns
Pre-process CBECS data
Transformed energy consumption data
Select relevant data for energy consumption
CBECS
Pre-process CBECS data
Selected Data
EIA Website
Data Repository

Figure 2: KDD Methodology

3.2 Dataset Selection

The selection of the dataset is a very crucial part of any research project. In this phase, the dataset is selected, organized, and extracted to only include the data which is required for the research project. It is a selection of what parts and pieces of the data that is relevant to the problem which the researcher is trying to solve. The 2012 CBECS dataset is selected and downloaded from the U.S. Energy Information Administration (EIA) website.[3] The 2012 CBECS dataset is publicly available on the EIA website, and there is no ethical implication associated with this dataset.

3.3 Data pre-processing

The real-world dataset will always contain a considerable number of errors such as missing values, data inconsistency, and incomplete data which required to be handled to carry-out any research project smoothly. Before doing data pre-processing, it is indispensable to get data understanding (exploratory data analysis). Plotting a correlation graph between

[2] http://www2.cs.uregina.ca/ dbd/cs831/notes/kdd/1_k*dd.html*

[3] https://www.eia.gov/consumption/commercial/reports/2012/energyusage/

independent and dependent variables will help the researcher to get a better understanding of the dataset, and independent variables which are having high correlation had to be removed. In exploratory data analysis, data should be cleansed to provide the desired outputs. The data cleansing part includes the removal of missing values, unnecessary data, and not applicable (NA) values.

3.4 Data Transformation

It is a method of converting data from one structure to another structure. Data transformation is very crucial for activities like data management and data integration. For machine learning algorithms, categorical values are hidden text, and it is necessary to encode the data correctly in advance. There are two approaches for encoding the categorical variables, namely label encoding approach and one hot encoder approach. In this project label, the encoding approach has been followed using python's sci-kit learn library. Data Binning was performed first on the dependent variable of the dataset, which is "Major Fuel Consumption in BTU" (MFBTU) before applying a label encoding approach on it. A technique to generate four defined bins (based on this project) of a continuous variable is called Data Binning.

3.5 Data Mining

Usually, the dataset that feeds into the data mining is in a flat form, and it means there are rows and columns into a single entity in which columns are representing variables and rows are representing the samples. Then the pre-processed data is transformed as per different data mining algorithms, and then transformed and finalized data is pushed into different data mining algorithms specific to different data mining patterns and tasks. As per this research, four classification machine learning algorithms had been applied to the encoded energy consumption dataset to get the desired knowledge.

3.6 Evaluation

The knowledge obtained after applying four classification machine learning algorithms is then evaluated in this phase to get clear insights about the desired outcome. The knowledge obtained is then visualized in order to present it in front of the end-users to get smooth and useful insight into the energy consumption data. Evaluation is performed based on four metrics, namely classification accuracy, F1 score, precision, and recall.

3.6.1 Classification Accuracy

It is an accurate numeral of forecasts made out of the over-all expectations made. Its tasks fine when samples are similarly separated into each class. It will be sufficient to find model precision once the issue of data imbalance is appropriately managed.

3.6.2 F1 Score

F1 Score is required to enable stability betwixt precision and recall. The balanced mean of the precision and recall is known as F1 Score, where 0 signifies worst value, and q signifies the best value of the F1 Score. It provides knowledge of the classifier's robustness and accurateness.

3.6.3 Precision

It gives statistics on how precise the model is out of the entire anticipated positive and how many of them are really positive. It contains data about the model accuracy and information on actually positive out of anticipated positive.

3.6.4 Recall

The over-all numeral of True positive divided by Total genuine positive. It calculates the over-all numeral of genuine positive the model predicts by classifying it as genuine positive. Just by a similar understanding, Model metric alike recall can be utilized to select the most elegant model where false negative is allied by the high cost.

4 Design Specification

The design architecture of this project is split into two parts, as seen in Figure 3. The section in the top is the data layer, and the below section combines both the Application Layer and Business Logic Layer. The energy consumption data in the data layer will be extracted from the U.S. Energy Information Administration website and cleansed by utilizing python programming language[4]. Exploratory data analysis (EDA) is carried out on the cleansed data to get more understand of the distinct data features. Post EDA, data encoding had been performed by utilizing Python's Scikit learn library.

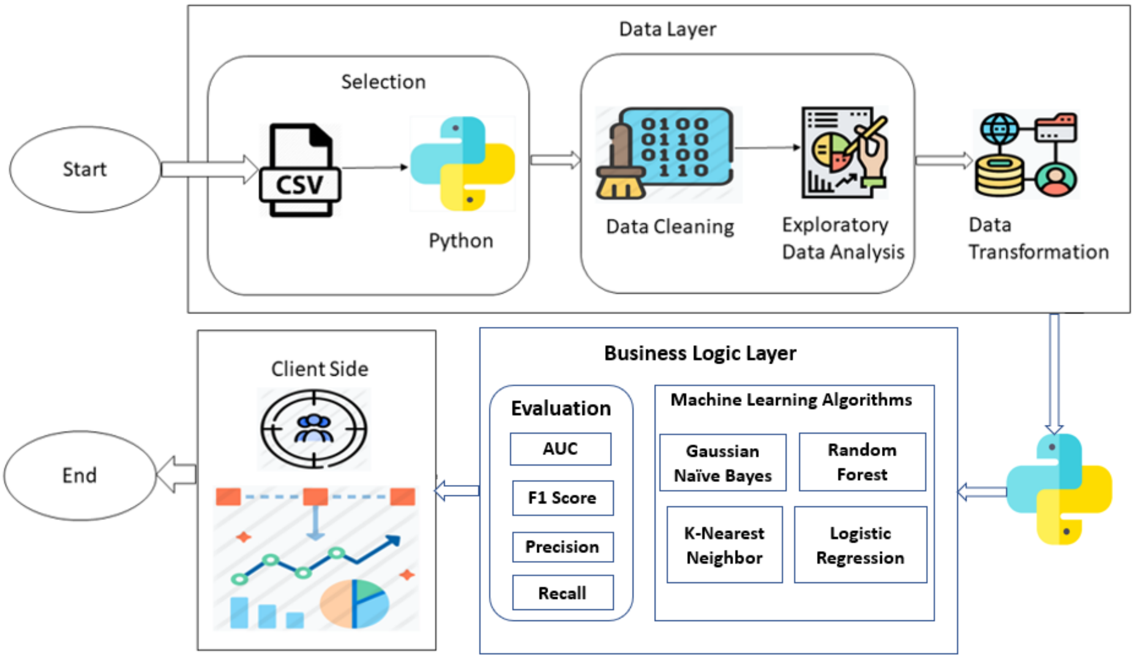

Figure 3: Two-Tier Architecture

In the Business logic layer, four machine learning algorithms had been applied on the encoded/transformed data, and their performance outcomes had been evaluated using

[4] https://www.eia.gov/consumption/commercial/reports/2012/energyusage/

classification accuracy, F1 score, precision, and recall. In the application layer, performance outcomes had been visualized to deliver a clear understanding of outcomes to the end-users.

5 Implementation

5.1 Data Selection

The U.S. Energy Information Administration (EIA) website was used to download the 2012 CBECS dataset[5], which is available in the .csv file, and it holds untabulated data of individual buildings. In July 2015 CBECS dataset was released, and due to a mistake in the dataset, it was revised by the EIA in August 2016. The CBECS dataset file consists of 6,720 observations and 1119 variables, which was gathered from 50 States of Columbia. Totally 5.6 million buildings data from the United States were represented in this sample. This dataset is not significant as it is real-time data of Atlanta city, and also each observation is tied in with a single responding, in-range sampled building. As per the researcher [Robinson], principal building activity (PBA), number of floors, cooling degree day, heating degree days, and square footage are the most influential factors which contribute more towards energy consumption.

5.2 Data Pre-processing

The first step in data preparation is to check missing values in the data and remove them wisely to avoid significant data loss. Below are the steps of data pre-processing.

1. The total size of the dataset 6720 observations and 1119 variables, and for this project, columns having more than 20 percent of missing values are removed.

2. After removing 20% of missing values, 6720 observations, and 812 variables are left.

5.2.1 Data Binning

After removing 20% of missing values, Binning has performed on “MFBTU” (Major Fuel consumption in BTU) column, and it was divided into very-low, low, medium, and high classes and the binned data will store in a newly created column known as EC (Electricity consumption).

Energy consumption data had been binned into four categories, namely:

1. Energy consumption values between (0 - 605000) kWh/ft^2 (kilowatt-hours per square feet) is in “Very-Low” category.

2. Energy consumption values between (605000 - 3100000) kWh/ft^2 are in the “Low” category.

3. Energy consumption values between (3100000 - 15200000) kWh/ft^2 are in the “Medium” category.

[5] https://www.eia.gov/consumption/commercial/data/2012/index.php?view=microdata

4. Energy consumption values between (15200000- 1418866360) kWh/ft^2 are in the “High” category.

After a binning number of variables, the count had been increased by one (EC column) to store the binning data of “MFBTU” column

1. Then the remaining missing values were removed from the entire dataset, and the dataset was left with 2758 observations and 813 variables.
2. The binning approach was applied to the dependent variable, which is Major fuel consumption in BTU’s (MFBTU) to differentiate the energy consumption data in very-low, low, medium, and high categories followed by handling the class imbalance issue.

5.3 Handling Class Imbalance

In almost every real-world classification data, there always a little amount of class imbalance issue[6]. This issue arises when the categories in the dataset are not evenly divided. The main aim of the desired goal is to adjust the metrics and methods based on the purpose. Meaningless results will be obtained if the class imbalance issue was not resolved during the data pre-processing.

The class imbalance issue is handled by dividing an equal portion of data in each class. The random down-sampling method was used in this project to resolve the class imbalance issue. It assists in downgrade the majority class sample to minority class samples to sustain the right stability in the dataset.

1. The data in all four categories should be equal in order for machine learning algorithms to learn appropriately.

   Number of observations in “Very-Low” class = 684
   Number of observations in “Low” class = 678
   Number of observations in “Medium” class = 696
   Number of observations in “High” class = 700

2. To solve class imbalance issue, the random down-sampling method is used instead of random up-sampling which will bring down majority class data to minority class, and this is to avoid unnecessary creation of noise data which happen in the case of random up-sampling method. Below are the number of observations in each class.

   Number of observations in “Very-Low” class = 678
   Number of observations in “Low” class = 678
   Number of observations in “Medium” class = 678
   Number of observations in “High” class = 678

3. After data pre-processing, the final dataset is having 2,712 observations and 813 variables.

[6] https://towardsdatascience.com/dealing-with-imbalanced-classes-in-machine-learning-d43d6fa19d2

5.4 Feature Engineering

Feature engineering is a very crucial part of any research project as it helps to improve the functioning of machine learning models. Basically, some input variables are required by machine learning algorithms to predict the output variable, and in order for machine learning algorithms to work correctly requires features with some specific characteristics. Doing a correlation check on this data was not possible due to a considerable number of variables in the dataset, which is why feature selection and feature reduction methods, i.e., analysis of variance (ANOVA) and principal component analysis (PCA), are used in this project to select the best number of variables for predicting energy consumption.

5.4.1 ANOVA

ANOVA is analysis of variance which basically helps to finish the job of choosing the best features[7]. It is a statistical technique, utilized to examine the means for multiple groups which are significantly diverse from each other. ANOVA performs F-tet check to find if any significant diversities are there between the groups. The outcome of ANOVA's F-ratio will be near to 1 if there are no significant diversities between the groups than that means all the variance are equal.

5.4.2 Principal Component Analysis

A technique that was utilized to reduce the dimensions from a vast dataset into a small-scale dataset which continues to hold nearly all data variables of the enormous dataset (Li et al.; 2017). The machine learning model's accuracy tends to degrade if the features of the dataset are reduced[8]. The main benefit of dimensionality reduction is to save unlimited data as possible, which basically helps to provide better outcomes with the small-scale dataset, which are generally simple to interpret and visualize.

5.5 Classification Algorithms

In this project, a total of four classification algorithms are being used, namely Gaussian Naïve Bayes, Random Forest, K-Nearest Neighbour, and Logistic Regression to classify energy consumption in very-low, low, medium and high. At this stage, the encoded data from the previous step is utilized to build and train the four different classifiers. Python language was utilized as the language of choice for executing this research. The available libraries in python were used to perform data pre-processing. Once the data pre-processing is completed, cleansed data was then divided into the train (80%) and test (20%), and its performance was evaluated by utilizing the different classifiers. Below are the four classification machine learning algorithms which were applied to complete this research project.

5.5.1 Gaussian Naïve Bayes

A secure but exceptionally powerful algorithm for model prediction is known as Naïve Bayes[9]. It can be prolonged to actual-valued parameters, most frequently by presuming

[7] https://towardsdatascience.com/anova-for-feature-selection-in-machine-learning-d9305e228476

[8] https://towardsdatascience.com/a-step-by-step-explanation-of-principal-component-analysisb836fb9c97e2

[9] https://machinelearningmastery.com/naive-bayes-for-machine-learning/

a gaussian distribution. This add-on to Naïve Bayes is known as Gaussian Naïve Bayes. Other parameters can be utilized to evaluate the division of the data. However, the normal distribution (or Gaussian) is the simplest to work as it only needs to evaluate the standard deviation and the mean of the training data. Researcher (Dong et al.; 2018) used the Gaussian Naïve Bayes algorithm on the water heater dataset to predict energy consumed by electricity for 20 homes and stated that the performance of gaussian naïve Bayes was the second-best amongst the three applied machine learning algorithms.

5.5.2 Random Forest

Random Forest comprises of too many individual decision trees that work as a group. Each separate tree in random forest divulges a category prediction, and the decision of model prediction is dependent on the votes. Researcher (Deng et al.; 2018) considered and applied random forest on medium-size school, and office buildings from the subset of CBECS 2003 dataset (Tsanas and Xifara; 2012) also used random forest on simulated data to predict heating load (HL) and cooling load (CL) of residential buildings. In various domains, a random forest classifier is famous for its fantastic performance and computational efficiency.

5.5.3 K-Nearest Neighbor

KNN is a supervised machine learning technique that is simple and easy to implement, and it can be utilized to resolve both regression and classification problems[10]. KNN works on the assumption that similar things exist in nearby proximity. For KNN, there is no prerequisite of tuning several parameters, build a model, or assemble extra assumptions. For classification, KNN usually works by detecting the gap between the numerous data examples and a query, choosing the definite number of examples (K) which is nearby to the query and then selects the most common label. Researcher (Wahid and Kim; 2016) used the KNN classifier to make a prediction of daily energy consumption on 520 flats of Seoul, Korea.

5.5.4 Logistic Regression

Logistic Regression is utilized to allocate annotation to a distinct set of classes[11]. Unlike linear regression, which outcomes continuous numerical values, logistic regression converts its outcomes using the sigmoid function to provides a value that can be depicted to multiple distinct classes. It is a technique that was borrowed from the domain of statistics by machine learning. An S-shaped curve which can lay hold of any genuine-valued integer and depict it within a value betwixt 0 and 1. Researcher (Ahmad et al.; 2018) had used logistic regression in his research for electricity monitoring, prediction, and visualization.

6 Evaluation

For any research, machine learning model evaluation is a vital part, and there are numerous model evaluation methods. Evaluation of below four classifier is performed using accuracy, precision, recall and F1 Score.

[10] https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[11] https://ml-cheatsheet.readthedocs.io/en/latest/logistic$_regression.htmlintroduction$

6.1 Experiment 1: ANOVA Feature Selection Method

ANOVA feature selection method is used to select the best features, which helped to get the best accuracy for all the four classifiers. In Table 1, several machine learning classifiers are stated with their accuracy, precision, recall, and f1 score. Amongst all the machine learning classifier algorithms, the K-Nearest Neighbor algorithm provided better performance for all the four evaluation metrics. Accuracy, Precision, Recall, and F1 Score provided by the K-Nearest Neighbor algorithm are 97.05%, 97%, 97%, and 97%, respectively. Followed by Random Forest and Gaussian Naive Bayes has accuracy, precision, recall and f1 score of 97.05%, 97%, 97%, 97% and 88.95%, 89%, 89%, 89% respectively, Logistic Regression obtain accuracy, precision, recall and f1 score of 79.74%, 80%, 80%, 79% respectively. Also, each machine learning model's accuracy is evaluated using k10-fold cross validation. Amongst all, K-Nearest Neighbor algorithm provided accuracy of 97.41% followed by Random Forest, and Gaussian Naïve Bayes has 93.95%, 88.23%, respectively. Logistic regression achieves a k10-fold cross validation accuracy of 79.75%.

Table 1: Results of ANOVA Methodology

ANOVA Feature Selection Method				
Machine Learning Algorithms	Accuracy	Precision	Recall	F1 Score
Gaussian Naive Bayes	88.95%	89%	89%	89%
Random Forest Classifier	96.31%	96%	96%	96%
K-Nearest Neighbor	97.05%	97%	97%	97%
Logistic Regression	79.74%	80%	80%	79%

6.2 Experiment 2: Principal Component Analysis

The principal component analysis is a dimensionality reduction method used to scale down variables from a vast dataset. In Table 2, the results of all the classifiers have been shown on which dimensionality reduction method principal component analysis has been applied by keeping n value to the dimension value on which the machine learning models are getting the best accuracy. That means it will reduce and merge the variables in different columns. As per the results shown in Table 2, K-Nearest Neighbor outperformed all the other machine learning algorithms in terms of all the four-evaluation metrics by achieving 98.15% of accuracy and 97% of precision, recall and f1 score followed by Random forest has an accuracy of 96.50%, and 97% of precision, recall and f1 score. Gaussian Naïve Bayes and Logistic Regression obtain accuracy, precision, recall and f1 score of 93.37%, 94%, 93%, 93% and 70.97%, 71%, 72%, 69% respectively. Also, K10-fold cross validation is used to evaluate each machine learning model's accuracy. Amongst all K-Nearest Neighbor provided accuracy of 98.38%, followed by Random Forest and Gaussian Naïve Bayes has an accuracy of 97.59% and 93.30%, respectively. The Logistic Regression model achieved a K10-fold cross validation accuracy of 67.78%.

Table 2: Results of PCA Methodology

PCA Dimensionality Reduction Method				
Machine Learning Algorithms	Accuracy	Precision	Recall	F1 Score
Gaussian Naive Bayes	93.37%	94%	93%	93%
Random Forest Classifier	96.50%	97%	97%	97%
K-Nearest Neighbor	98.15%	97%	97%	97%
Logistic Regression	70.97%	71%	72%	69%

6.3 Experiment 3: Deep Analysis on K-Nearest Neighbor

This in-depth analysis is performed on all the machine learning models. However, it is necessary to showcase the analysis of the K-Nearest Neighbor algorithm since it has achieved an outstanding accuracy of 98.15% for PCA and 97.05% for ANOVA. A ”for loop” is written for both ANOVA and principal component method to select the number of variables on which the machine learning algorithms are providing the best accuracy.

6.3.1 ANOVA

The preprocessed data has 812 features, and in ”for loop,” a total of 810 features were considered and the ”for loop” will run in reversed direction by reducing the features count to 10 after completing a single ”for loop” till the dataset size is reduced to 10 features. During the reduction of features after every ”for loop,” it is noticed that the accuracy of each model is kept on fluctuating at different feature counts. As per Table 3, the accuracy of the K-Nearest Neighbor is 96.68% when the machine learning model is considering 410 and 420 features and from accuracy improved to 97.05% when the model considered 430 features.

Table 3: Deep Analysis on K-Nearest Neighbor using ANOVA

K-Nearest Neighbor	
Feature Count	Accuracy
450	97.05%
440	97.05%
430	97.05%
420	96.68%
410	96.68%

6.3.2 Principal Component Analysis

In the case of principal component analysis, firstly same ”for loop” has been run in the reverse direction from 810 features by reducing the feature count to 10 after completing a single ”for loop” till only ten features are left in the dataset. During the reduction of features after every ”for loop,” it is noticed that the accuracy of each model is kept on improving so again the similar ”for loop” is rerun from 10 features by reducing the feature count by one, and it runs till one feature is left in the dataset. As per Table 4, the accuracy of the K-Nearest Neighbor improves until 12 features after which, the accuracy of the model is not stable.

Table 4: Deep Analysis on K-Nearest Neighbor using PCA

K-Nearest Neighbor	
Feature Count	Accuracy
13	97.05%
12	97.05%
11	96.86%
10	96.50%
9	96.68%

6.4 Discussion

There were some significant challenges while working on this project which is stated below:

1. Removing missing values from the dataset was one of the most significant decisions to make, (Robinson et al.; 2017) removed the columns which were having more than 25% of missing values. This research was tested by removing columns that were having more than 20% and 30% of missing values, and there was not much sign of the difference in terms of all the four evaluation metrics. To avoid the data loss, this research was tested by removing the columns which were having more than 20% of missing values.

2. The second major issue was to do binning on the dependent variable and to decide the range of energy consumption units in British thermal unit (Btu) for each of the four classes. Multiple times sampling of energy consumption range had been done to fit-in a higher number of rows in each category.

3. The issue of handling class imbalance was the last and essential part of the data pre-processing. This significantly reduced the size of the dataset. This might harm the interpretation and performance of the machine learning algorithms.

Two different approaches have been followed in this research project. One is feature selection method, i.e., analysis of variance (ANOVA) and another one is dimensionality reduction method called principal component analysis (PCA) to classify energy consumption in four categories namely, 'very-low,' 'low,' 'medium,' and 'high' for the commercial buildings in the United States. A comparative analysis has been done, and the output of the analysis has been visualized in Figure 4. As per the results, the K-Nearest Neighbor algorithm provided the best f1 score of 97% for both ANOVA and PCA. Using ANOVA, K-Nearest Neighbor achieved an f1 score of 97% by considering 430 features, and using PCA, K-Nearest Neighbor achieved a 97% f1 score by considering 12 features. It is followed by the Random Forest algorithm, which achieved an f1 score of 96% for ANOVA by considering 350 features and 97% of f1 score for PCA by considering seven features. The Gaussian Naive Bayes algorithm achieved an f1 score of 89% for ANOVA by considering 420 features and f1 score of 93.37% for PCA by considering two features. The Logistic Regression algorithm achieved an f1 score of 79% for ANOVA by considering 270 features and f1 score of 69% for PCA by considering 30 features. Based on this comparative analysis, it can be concluded that PCA performed better than ANOVA in predicting energy consumption in multiple classes as it provides more than 90% of the f1 score for three algorithms.

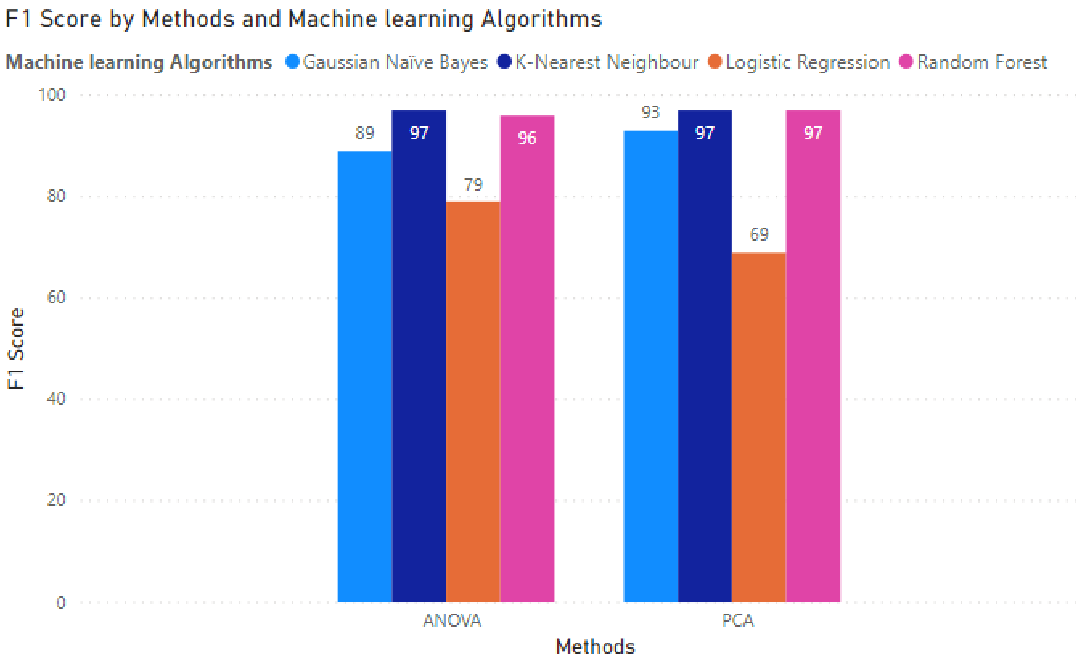

Figure 4: Comparative Analysis of ANOVA and PCA

7 Conclusion and Future Work

In this paper, the real-world dataset of Atlanta city (CBECS dataset) was downloaded and cleansed by removing columns which were having 20% of missing values. Both feature selection (ANOVA) and dimensionality reduction (PCA) methods were used after binning the dependent variable and handling the class imbalance. A total of four classification machine learning algorithms were applied to both approaches. As per the result, amongst all, K-Nearest Neighbor models performed better by achieving 97% f1 score for both ANOVA and PCA, followed by Random Forest, which provides a 97% f1 score for PCA and 96% f1 score for ANOVA. The dimensionality reduction method PCA tends to perform better than the feature selection method ANOVA as it provides more than 90% f1 score for the Random Forest, K-Nearest Neighbor, and Gaussian Naive Bayes.

In the future, all the four classification models developed in this research can be applied to all the commercial buildings in prime city areas to include as a synopsis of total city energy consumption. This research is carried out on the 2012 CBECS dataset, which contained data of approximately 5.6 million commercial buildings[12]. In the year 2020, the 2018 CBECS dataset will be available on the EIA website, and these four models can be merged with the climate projection dataset to analyze what kind of changes will come on the energy consumption environment of diverse cities under different climatic scenarios.

[12] https://www.eia.gov/consumption/commercial/reports/2012/energyusage/

Acknowledgement

I would like to appreciatively acknowledge my supervisor, Dr. Muhammad Iqbal, for supporting, encouraging, and providing feedback for consecutive 12 weeks on the research project. In each of the one to one meeting, my supervisor used to respond to the queries promptly and patiently. His continuous feedback helped me to present this research project in an insightful and sophisticated manner. This experience with the supervisor helped to explore more aspects of machine learning concepts and methods.

I am eternally gratefully to my family for the love, constant support, steadfast belief in me, and letting me pursue my dreams.

References

Ahmad, M. W., Mourshed, M. and Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption, *Energy and Buildings* **147**: 77–89.
URL: *http://dx.doi.org/10.1016/j.enbuild.2017.04.038*

Ahmad, T., Chen, H., Huang, R., Yabin, G., Wang, J., Shair, J., Azeem Akram, H. M., Hassnain Mohsan, S. A. and Kazim, M. (2018). Supervised based machine learning models for short, medium and long-term energy prediction in distinct building environment, *Energy* **158**: 17–32.

Amasyali, K. and El-Gohary, N. (2018). Deep Learning for Building Energy Consumption Prediction, *Leadership in Sustainable Infrastructure, CSCE* (February).

Bin, Z. (2009). "Solid oxide fuel cell (SOFC) technical challenges and solutions from nano-aspects", *International journal of energy research* **31**(August 2007): 135–147.

Casteleiro-Roca, J. L., Gómez-González, J. F., Calvo-Rolle, J. L., Jove, E., Quintián, H., Diaz, B. G. and Perez, J. A. M. (2019). Short-term energy demand forecast in hotels using hybrid intelligent modeling, *Sensors (Switzerland)* **19**(11): 1–18.

Deng, H., Fannon, D. and Eckelman, M. J. (2018). Predictive modeling for US commercial building energy use: A comparison of existing statistical and machine learning algorithms using CBECS microdata, *Energy and Buildings* **163**: 34–43.
URL: *http://dx.doi.org/10.1016/j.enbuild.2017.12.031*

Dong, J., Munk, J., Cui, B., Boudreaux, P. R. and Kuruganti, T. (2018). Machine-Learning Model of Electric Water Heater for Electricity Consumption Prediction, *5th International High Performance Buildings Conference* (July): 1–10.

García-Martín, E., Rodrigues, C. F., Riley, G. and Grahn, H. (2019). Estimation of energy consumption in machine learning, *Journal of Parallel and Distributed Computing* **134**: 75–88.
URL: *https://doi.org/10.1016/j.jpdc.2019.07.007*

Geyer, P. and Singaravel, S. (2018). Component-based machine learning for performance prediction in building design, *Applied Energy* **228**(October 2017): 1439–1453.
URL: *https://doi.org/10.1016/j.apenergy.2018.07.011*

Liu, D. and Chen, Q. (2013). Prediction of building lighting energy consumption based on support vector regression, *2013 9th Asian Control Conference, ASCC 2013* (3).

Naganathan, H., Chong, W. O. and Chen, X. (2016). Building energy modeling (BEM) using clustering algorithms and semi-supervised machine learning approaches, *Automation in Construction* **72**: 187–194.
URL: *http://dx.doi.org/10.1016/j.autcon.2016.08.002*

Nateghi, R. and Mukherjee, S. (2017). A multi-paradigm framework to assess the impacts of climate change on end-use energy demand, *PLoS ONE* **12**(11): 1–23.

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

Platon, R., Dehkordi, V. R. and Martel, J. (2015). Hourly prediction of a building's electricity consumption using case-based reasoning, artificial neural networks and principal component analysis, *Energy and Buildings* **92**: 10–18.
URL: *http://dx.doi.org/10.1016/j.enbuild.2015.01.047*

Rahman, A. and Smith, A. D. (2017). Predicting fuel consumption for commercial buildings with machine learning algorithms, *Energy and Buildings* **152**: 341–358.
URL: *http://dx.doi.org/10.1016/j.enbuild.2017.07.017*

Ribeiro, M., Grolinger, K., ElYamany, H. F., Higashino, W. A. and Capretz, M. A. (2018). Transfer learning with seasonal and trend adjustment for cross-building energy forecasting, *Energy and Buildings* **165**: 352–363.
URL: *https://doi.org/10.1016/j.enbuild.2018.01.034*

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

Salakij, S., Yu, N., Paolucci, S. and Antsaklis, P. (2016). Model-Based Predictive Control for building energy management. I: Energy modeling and optimal control, *Energy and Buildings* **133**: 345–358.

Seyedzadeh, S., Pour Rahimian, F., Rastogi, P. and Glesk, I. (2019). Tuning machine learning models for prediction of building energy loads, *Sustainable Cities and Society* **47**(March): 101484.
URL: *https://doi.org/10.1016/j.scs.2019.101484*

Shen, P., Braham, W., Yi, Y. and Eaton, E. (2019). Rapid multi-objective optimization with multi-year future weather condition and decision-making support for building retrofit, *Energy* **172**: 892–912.
URL: *https://doi.org/10.1016/j.energy.2019.01.164*

Smpokos, G., Elshatshat, M. A., Lioumpas, A. and Iliopoulos, I. (2018). On the Energy Consumption Forecasting of Data Centers Based on Weather Conditions: Remote Sensing and Machine Learning Approach, *2018 11th International Symposium on Communication Systems, Networks and Digital Signal Processing, CSNDSP 2018* pp. 1–6.

Touzani, S., Granderson, J. and Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings, *Energy and Buildings* **158**: 1533–1543.
URL: *http://dx.doi.org/10.1016/j.enbuild.2017.11.039*

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

Wahid, F. and Kim, D. H. (2016). A prediction approach for demand analysis of energy consumption using K-nearest neighbor in residential buildings, *International Journal of Smart Home* **10**(2): 97–108.

Xu, X., Wang, W., Hong, T. and Chen, J. (2019). Incorporating machine learning with building network analysis to predict multi-building energy use, *Energy and Buildings* **186**: 80–97.
URL: *https://doi.org/10.1016/j.enbuild.2019.01.002*

Zhong, H., Wang, J., Jia, H., Mu, Y. and Lv, S. (2019). Vector field-based support vector regression for building energy consumption prediction, *Applied Energy* **242**(March 2019): 403–414.
URL: *https://doi.org/10.1016/j.apenergy.2019.03.078*