Towards Prediction of Energy Consumption of HVAC Plants Using Machine Learning

Monika Goyal(✉) and Mrinal Pandey

Manav Rachna University, Faridabad, India
monikagoyal.er@gmail.com, mrinal@mru.edu.in

Abstract. Today energy optimization has become a great challenge as energy is being consumed at a fast rate in almost every sector including buildings, transport and industries. However, Buildings are the largest consumer of energy followed by Transport and Industry throughout the world. In buildings, most of the energy consumption depends upon the usages of air conditioning plants (Heating, Ventilation and Air Conditioning). Therefore, with the necessity to determine the energy consumption due to HVAC plant in building, this research focuses on Cooling Tower data of HVAC plant of a building as Cooling Tower is an important component of HVAC and carries a major responsibility of maintaining the ambient temperature within a building. In this paper, three popular Machine Learning techniques namely Multiple Linear Regression, Random Forests and Gradient Boosting Machines were experimented for predicting the energy consumption due to HVAC plant within a building. The findings of the experiments reveal that Random Forest outperforms in terms of error measures.

Keywords: Machine Learning · Energy optimization · Heating · Ventilation and Air Conditioning · Random Forests · Multiple Linear Regression · Gradient Boosting Machines

1 Introduction

Technology is spreading in each and every sector in the present world and it has put forward the challenge of quick consumption of energy. The demand of energy is increasing for the smooth functioning of various types of machines in homes, hotels, offices, industries and transport. Although advancements in technology have brought ease and comfort in human lives, but on the other hand, they have put burden on the natural non-renewable resources, because energy is conventionally produced by the burning of fossil fuels like coal, oil and natural gas, which take millions of years to form. If the precious energy continues to be consumed at such high pace, soon the non-renewable energy sources will deplete. Apart from that, burning of fossil fuels releases carbon, more the fossil fuels burnt more carbon is emitted which consequently leads to air and water pollution [1]. To conserve energy on a bigger scale, high energy consuming areas should be identified and targeted. Studies show that globally buildings account for approximately 40% of energy usage throughout the world [2–5] and it is greater than the energy consumed by other two areas, industry and transport which as per reports are 32% and 28% respectively.

U. Batra et al. (Eds.): REDSET 2019, CCIS 1229, pp. 254–265, 2020.
https://doi.org/10.1007/978-981-15-5827-6_22

These data motivated to target the biggest energy consumer i.e. buildings. The analysis of energy consumption pattern in a building, revels that the biggest portion of energy is consumed by HVAC (Heating, Ventilation and Air conditioning) system. An HVAC unit consists of different components like Chillers, Cooling towers, Primary Pumps, Secondary Pumps. Each of these components performs their designated task and consumes power to operate. It consumes around 40%–50% of the total energy consumed in a building [6, 7]. This further motivated to analyze and counter the energy consumption profile of HVAC. Therefore, in this research the energy consumption due to HVAC plant, particularly Cooling Tower data was analyzed using regression (Multiple Linear Regression) and two Ensembles based Machine Learning techniques namely Random Forest and Gradient Boosting Technique.

The organization of this paper is as follows: Sect. 1 introduces the problem and gives a brief overview of the Machine Learning techniques. Section 2 describes the Machine Learning techniques. Literature Survey is presented in Sect. 3. Methodology of the work done is described in Sect. 4. Section 5 explains the experiments performed and results obtained. Section 6 gives conclusion of this research.

2 Machine Learning

Machine Learning is the concept in which a machine learns and behaves in a certain manner when a particular type of data is fed as input. Machine Learning can be classified as Unsupervised Learning and Supervised Learning. A common technique in unsupervised learning is clustering in which the given dataset is grouped into a given number of clusters depending upon the similarity index. In Supervised Learning the input data is mapped to the desired output using a labelled set of training data. Two common techniques in Supervised Learning are Classification and Regression. Following section presents the Machine Learning Techniques used in this research.

Regression can be viewed as a statistical methodology generally used for numeric prediction. Regression can be classified as a) Linear Regression which involves finding the best line to fit two variables, such that one variable is independent called Predictor and can be used to predict the other variable which is dependent called Response. b) Non – Linear Regression which involves more complex calculations and finds the best curve instead of best line. A common example is Polynomial Regression. Ensemble based Regression techniques are becoming popular for last few years. In Ensemble techniques, regression is performed by integrating the results of several individual models with the objective of improving the accuracy and robustness of prediction in learning problems having a numerical response variable. Two most popular ensemble methods are bagging and boosting [8, 9].

2.1 Multiple Linear Regression

The concept of Multiple Linear Regression is similar to Linear Regression with a difference that the model consists of one Response variable B which is dependent upon multiple Predictor variables A_1, A_2, A_3 ..... A_n. Interpretation of results in MLR is more complex due to the correlation among different predictor variables [10].

2.2 Random Forests

This technique of Machine Learning lies under Ensemble technique category of model development. It is a tree based technique applicable for both Classification and Regression. Various features which enhance the appeal of Random Forests are: Prediction speed, suitability for high dimensional problems, handling of missing values and outliers etc. [11].

2.3 Gradient Boosting Machines

This technique is also a tree based ensemble learning technique, where additive regression models are built by iteratively fitting a simple base to recently updated pseudo residuals by the application of least squares at every consecutive iteration [12].

3 Literature Survey

Machine Learning techniques have been proved highly effective approach to address the issue of energy consumption in buildings [13]. Apart from applying the ML algorithms individually, Ensemble methods can be created by specifically combining different models, which improve the effectiveness of the model in terms of accuracy and performance. Following section presented the application of Machine Learning techniques for energy consumption particularly in buildings.

Authors in [14] proposed two frameworks for anomaly detection in HVAC power consumption. One was a pattern based anomaly classifier called CCAD-SW (Collective contextual anomaly detection using sliding window) which created overlapping sliding windows so that anomalies can be pointed out as soon as possible. This framework made use of bagging for improved accuracy.

Another was prediction based anomaly classifier called EAD (Ensemble Anomaly Detection) which used Support Vector Regression and Random Forests. Experiments were performed on HVAC power consumption data collected from a school in Canada and results show that EAD performed better than CCAD-SW in terms of sensitivity and reducing False Positive rate.

Decision Tree Analysis was performed in [15] to predict the cost estimations of HVAC while designing buildings. The HVAC sub systems are CP (Central Plant) system, WD (Water Side Distribution) system and AC (Air Conditioning) system. Different combinations of these sub systems result in different costs of HVAC plants. The study was carried out in office buildings in Korea. The study showed that AC component of HVAC has maximum impact on the cost followed by CP and then WD has minimum impact.

Authors in [16] applied six Regression techniques: Linear Regression, Lasso Regression, Support Vector Machine, Random Forest, Gradient Boosting and Artificial Neural Network on for estimating Energy Use Intensity in Office buildings and energy usage by HVAC, plug load and lighting based on CBECS 2012, micro data. Out of them, Random Forest and Support Vector Machine were found comparatively robust.

A sensor-based model was proposed by the authors of [2] for forecasting the energy consumed by a multi-family residential building in New York City. The model was built using Support vector regression. Authors analysed the prediction performance through the perspective of time and space and found that most optimal prediction was hourly prediction at by floor levels.

In another paper [3] the authors developed a framework in which they used clustering algorithm and semi-supervised learning techniques to identify electricity losses during transmission i.e. between source and destination. The technique also helps in optimizing the losses. Deep learning is used for semi-supervised machine learning because of its ability to learn both labelled and unlabelled data. The electricity consumption, heating, cooling and outside temperature data was obtained from a research university campus in Arizona.

The work done in [17] used various supervised classifiers- DT (Decision Trees), DA (Discriminant Analysis), SVM (Support Vector Machines) and KNN (K- Nearest Neighbors) to disaggregate the data of power consumption by multiple HVAC units into that consumed by individual HVAC, while the data was retrieved collectively from single meter to reduce cost and complexity. Power consumption information of individual appliances is necessary for accurate energy consumption monitoring. The experiment was performed by collecting data from a commercial building in Alexandria. The results show that K- nearest neighbour was most efficient in power disaggregation.

A component-based Machine Learning Modelling approach was proposed [18] to counter the limitations of Building Energy Model for energy demand prediction in buildings. Random Forest was selected and applied on the climate data collected from Amsterdam, Brussels and Paris. MLMs excel over BEM as they generalize well under diverse design situations.

The work done in [19] witnessed the collection of energy consumption data of a house in Belgium and outside weather data and application of four Machine Learning algorithms namely Multiple Linear Regression, Support Vector Machine with Radial Kernel, Random Forest and Gradient Boosting Machines to predict the energy consumption and to rank the parameters according to their importance in prediction. They proposed that GBM was best at prediction.

The author in [20] proposed an ensemble technique which is a linear combiner of five different predictor models: ARIMA, RBFNN, MLP, SVM and FLANN. The combiner model was applied on stock exchange data for predicting the closing price of stock markets and it proved to be better in terms of accuracy as compared to individual models.

In another paper [21] the authors applied several Supervised Machine Learning techniques including Classification, Regression and Ensemble techniques to estimate the air quality of Faridabad by predicting the Air Quality Index. The algorithms applied include Decision Tree, SVM, Naïve Bayes, Random Forest, Voting Ensemble and Stacking Ensemble. They concluded that Decision Tree, SVR and Stacking Ensemble outperform other methods in their respective categories.

A framework was proposed by the authors of [22] in which they selected 8 different characteristics of a residential building as input parameters and depicted their effect on the 2 output parameters- Heating load and cooling load. Linear Regression and Random Forests were applied and results showed that Random Forests were better at predicting Heating and Cooling load in terms of accuracy.

4 Methodology

The Methodology for this research follows the classical approach and starts with the Data Collection followed by Data Pre-processing which includes Data Cleaning, Integration and Transformation, after pre-processing Model construction and evaluation phases completes the process of traditional Machine Learning approach. Figure 1 shows steps of the methodology adopted.

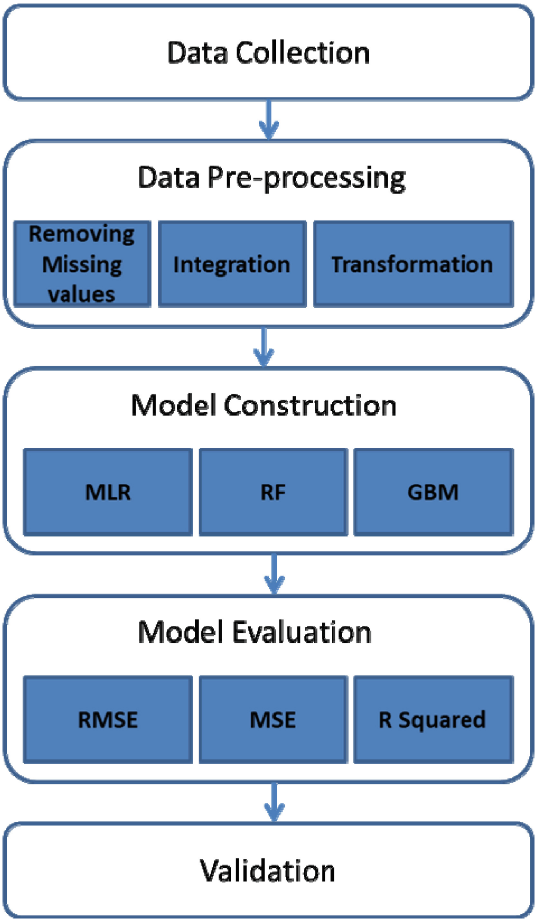

Fig. 1. Methodology

4.1 Data Collection

The dataset of HVAC plants was collected from a hotel building in New Delhi, India for this research. It consists of HVAC data from sensor recordings at every 5-min intervals for one year from Oct 2017 to September 2018. The data were categorized into two categories as Humidity (May–Nov) and Non-Humidity (Dec–April) depending upon two weather conditions. The data contains Inlet temperature, Outlet temperature and energy consumed by Cooling Tower, Dry Bulb Temperature and Relative Humidity. Energy consumed by Cooling Tower depends on Wet Bulb Temperature, so WBT has been calculated using DBT and RH. The experiments were performed on the Cooling Tower data. Table 1 describes the parameters of the data used along with the units in which each parameter is measured.

Table 1. Dataset parameter description

Parameter	Description	Unit
DBT(Dry Bulb Temperature)	Ambient temperature	°Celcius
RH(Relative Humidity)	Amount of water vapour in air relative to its temperature	%age
WBT(Wet Bulb Temperature)	Temperature brought down by water evaporation	°Celcius
CT_INLET	Temperature of water entering into Cooling Tower	°Celcius
CT_OUTLET	Temperature of water exiting from Cooling Tower	°Celcius
CT_POWER	Power consumed by Cooling Tower	Kilo Watts

In this research the data of April and August were considered for analysis & prediction. This is so because April represents non-humid weather and August represents humid weather condition particularly in NCR region of India. It must be mentioned that, HVAC consumes more energy in humid conditions to counter humidity rather than summer or winters.

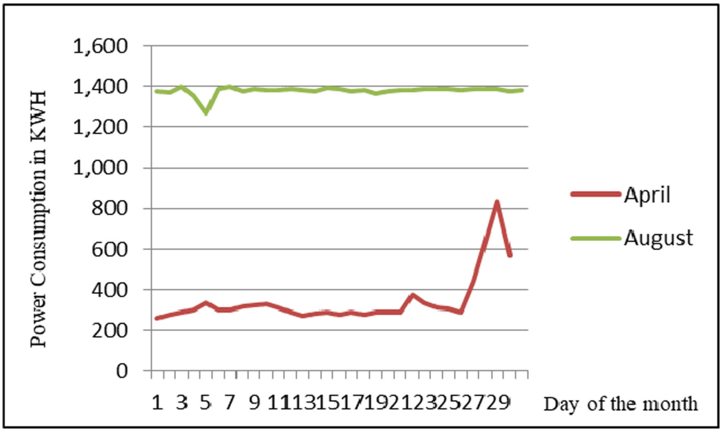

Fig. 2. Energy consumption in April and August

Figure 2 represents the day wise pattern of energy consumed by Cooling Tower in the months of April and August. Here x-axis denotes the days of month and y-axis denotes the power consumed in Kilo Watt Hour.

4.2 Data Pre-processing

Data pre-processing consists of filling the missing values, removing any outliers, transforming it into a form suitable for algorithm application, normalization, feature selection etc. [23]. The dataset used for this research was pre-processed in the following manner:

Although the data has been recorded by sensors at every 5-min intervals but at some instances the readings were mis-recorded at 1 or 2 min intervals. Such records were not complete and most of their fields were blank. In order to avoid any kind of flaws in result calculation, few extra records were removed while few records were filled with the mean of the values of that particular attribute.

As the data was recorded by different sensors installed at different places, there was a problem of integration of files for the values for different parameters into a single file. In order to ensure symmetry of data, extra records were removed from some files so that each record should represent the values of all parameters at same time instance. Data Transformation was done to convert the data into a form suitable for applying Machine Learning algorithms.

The original dataset consists of Dry Bulb Temperature and Relative Humidity as features, but the energy consumption of Cooling Tower also depends on Wet Bulb Temperature, therefore one more feature namely Wet Bulb Temperature is derived from Dry Bulb Temperature and Relative Humidity for the experiments.

4.3 Model Construction

For building the model, three Machine Learning algorithms namely Multiple Linear Regression, Random Forests and Gradient Boosting Machines were applied on the datasets of April and August months.

Data Partitioning. Datasets were partitioned according to 70%–30% rule into Training and Testing data. For partitioning, Random Sampling without Replacement was applied which resulted in 70% Training data and 30% Test data.

Parameter Tuning. In MLR default algorithm was applied on the data. In RF model, there are several parameters like ntree and mtry. Different values of ntree were tried where 500 is the default value. Best results were obtained by keeping ntree at 200. In GBM model also, certain variations were done in the values of some parameters. Best results were obtained by keeping shrinkage at 0.1 and interaction.depth at 6.

Libraries and Functions. All algorithms were implemented using RStudio. MLR was implemented using lm function and another function regr.eval was used to evaluate error metrics- RMSE, MSE, R Squared and MAE. The libraries included for implementing RF were- randomForest, miscTools and caret. GBM implementation used gbm library and function.

4.4 Evaluation

The results of the models were evaluated using three well known performance measures namely RMSE, MSE and R Squared.

Root Mean Square Error can be viewed as the standard deviation of residuals, where residuals indicate the distance of data points from the line of best fit i.e. these are the difference between actual and predicted values. Lower the value of RMSE better is the prediction. Formula for RMSE:

$$RMSE = \sqrt[2]{\sum\nolimits_{i=1}^{N} \frac{\left(\hat{\mathrm{Y}}_i - Y_i\right)^2}{N}}$$

Where

Yi : is the observed value for the ith observation
Ŷi : is the predicted value
N : is sample size

Mean Square Error can be defined as the average of the error squared. It is used as the loss function in least squares regression. MSE is the sum of the square of difference between predicted and actual target variables, spanning over all the data points, divided by the total number of data points. Formula for MSE:

$$MSE = \sum\nolimits_{i=1}^{N} \frac{\left(\hat{\mathrm{Y}}_i - Y_i\right)^2}{N}$$

R Squared is also known as Coefficient of Determination. It is used to statistically measure the closeness of data points to the fitted regression line. R squared can be defined by the following formula:

$$R^2 = 1 - \frac{\sum\left(Y_i - \hat{\mathrm{Y}}_i\right)^2}{\sum(Y_i - \bar{\mathrm{y}})^2}$$

Where Ŷi: is the predicted value of Y
Ȳ: is the mean value of Y

4.5 Model Validation

In order to validate the generality of the model, the above mentioned Machine Learning algorithms were also applied on another dataset collected from the UCI repository [24]. The dataset was used by researchers [22] to perform regression. The dataset consisted of eight independent variables describing various building parameters and two dependent variables: Heating Load (Y_1) and Cooling Load (Y_2). The comparative result summary is given in Sect. 5.

5 Experiments and Results

The R studio was used for performing various experiments of the research. Three Machine Learning algorithms namely Multiple Linear Regression, Random Forests and Gradient Boosting Machines were experimented on cooling tower data sets both for the month of April and August. The results of the experiments are described in Table 2 and Table 3.

5.1 Results for the Month of April

Table 2 summarizes the results obtained after applying all the three above mentioned algorithms on the dataset of April 2018 in terms of aforementioned performance measures.

Table 2. Results for the month of April 2018

Performance metric	ML algorithm		
	MLR	RF	GBM
RMSE	6.14	5.08	5.32
MSE	37.78	25.88	28.3
R Squared	0.23	0.5	0.65

The values of RMSE are 6.14, 5.08 and 5.32 for MLR, RF and GBM respectively. Corresponding MSE values are 37.78, 25.88 and 28.3 and R Squared values are 0.23, 0.5 and 0.65. It is clear from the Table 2 that RMSE and MSE error measures are minimum for Random Forest for the month of April.

5.2 Results for the Month of August

Table 3 summarizes the results obtained for August 2018. As per the results RMSE is 3.44 for MLR, 3.09 for RF and 3.72 for GBM. MSE values are 11.89, 9.57 and 13.83 for MLR, RF and GBM respectively. R Squared result values are 0.05, 0.43 and 0.57. It can be observed from the Table 3 that in terms of RMSE and MSE Random Forest outperforms among three algorithms.

Table 3. Results for the month of August 2018

Performance metric	ML algorithm		
	MLR	RF	GBM
RMSE	3.44	3.09	3.72
MSE	11.89	9.57	13.83
R Squared	0.05	0.43	0.57

5.3 Comparative Analysis of the Results for Hotel Building Dataset and UCI Dataset Using Random Forest

Random Forests has been proved as the best algorithm for this research in terms of RMSE and MSE. Therefore, same was applied on another dataset obtained from UCI repository. The recorded results are given in Table 4.

Table 4. Comparative Analysis of the results for Hotel Building dataset and UCI Dataset

Performance metric	Dataset			
	Hotel dataset	Hotel dataset	UCI dataset	UCI dataset
	April 2018	August 2018	Predictor Y1	Predictor Y2
RMSE	5.08	3.09	1.19	1.71
MSE	25.88	9.57	1.42	2.95
R Squared	0.5	0.43	0.98	0.96

As per Table 4, when RF was applied on the hotel building dataset, the obtained value of RMSE is 5.08 and 3.09 for the months April and August respectively while UCI dataset RMSE values are 1.19 and 1.71 for Heating Load and Cooling Load respectively. Similarly, the values MSE are 25.88 and 9.57 for April and August for the Hotel dataset while 1.42 and 2.95 for Heating Load and Cooling Load for UCI dataset. The values of R Squared are 0.5 and 0.43 for April and August months for Hotel dataset and 0.98 and 0.96 for Heating Load and Cooling Load for the dataset obtained from the UCI repository.

Figure 3 shows the results comparison of experiments performed on the hotel dataset and dataset collected from UCI repository.

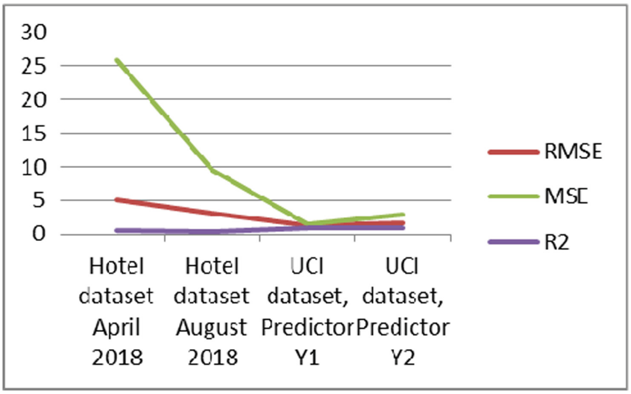

Fig. 3. Comparative analysis of both datasets

6 Conclusion and Future Scope

Energy being a precious resource needs to be utilized in the most efficient manner so that it is conserved while the comfort of consumers is also not compromised. Buildings are the largest consumer of energy globally and within a building HVAC accounts for the maximum energy consumption. This paper targeted cooling tower dataset of HVAC plant for analyzing the energy consumption in buildings. Three well known Regression algorithms namely Multiple Linear Regression, Random Forests and Gradient Boosting Machines were applied to predict the energy consumption due to HVAC plants in buildings. The results were compared on well-known performance measures namely Root Mean Square Error, Mean Square Error and R Square. The analysis of results proved Random Forests as the most suitable algorithm for this study. Further, Random Forest algorithm was also employed on another energy consumption dataset obtained from UCI repository for validation purpose.

The work in this area can further be carried on by experimenting with other powerful algorithms like Support Vector Machines and Extreme Gradient Boosting. Also optimization techniques can be explored and applied in order to work towards energy optimization.

References

1. Goyal, M., Pandey, M.: Energy optimization in buildings using machine learning techniques: a survey. Int. J. Inf. Syst. Manag. Sci. **1**(2) (2018)
2. Jain, R.K., Smith, K.M., Culligan, P.J., Taylor, J.E.: Forecasting energy consumption of multi-family residential buildings using support vector regression: investigating the impact of temporal and spatial monitoring granularity on performance accuracy. Appl. Energy **123**, 168–178 (2014)
3. Naganathan, H., Chong, W.O., Chen, X.: Building energy modeling (BEM) using clustering algorithms and semi-supervised machine learning approaches. Autom. Constr. **72**, 187–194 (2016)
4. Ahmad, M.W., Mourshed, M., Rezgui, Y.: Trees vs neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy Build. **147**, 77–89 (2017)
5. Chou, J.S., Bui, D.K.: Modeling heating and cooling loads by artificial intelligence for energy-efficient building design. Energy Build. **82**, 437–446 (2014)
6. Carreira, P., Costa, A.A., Mansu, V., Arsénio, A.: Can HVAC really learn from users? A simulation-based study on the effectiveness of voting for comfort and energy use optimisation. Sustain. Cities Soc. **41**, 275–285 (2018)
7. Drgoňa, J., Picard, D., Kvasnica, M., Helsen, L.: Approximate model predictive building control via machine learning. Appl. Energy **218**, 199–216 (2018)
8. Mendes-Moreira, J., Soares, C., Jorge, A.M., Sousa, J.F.D.: Ensemble approaches for regression: a survey. ACM Comput. Surv. **45**(1), 10 (2012)
9. Prasad, A.M., Iverson, L.R., Liaw, A.: Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems **9**(2), 181–199 (2006). https://doi.org/10.1007/s10021-005-0054-1
10. Krzywinski, M., Altman, N.: Multiple linear regression: when multiple variables are associated with a response, the interpretation of a prediction equation is seldom simple. Nat. Methods **12**(12), 1103–1105 (2015)

11. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
12. Friedman, J.H.: Stochastic gradient boosting. Comput. Stat. Data Anal. **38**(4), 367–378 (2002)
13. Banihashemi, S., Ding, G., Wang, J.: Developing a hybrid model of prediction and classification algorithms for building energy consumption. Energy Procedia **110**, 371–376 (2017)
14. Araya, D.B., Grolinger, K., ElYamany, H.F., Capretz, M.A., Bitsuamlak, G.: An ensemble learning framework for anomaly detection in building energy consumption. Energy Build. **144**, 191–206 (2017)
15. Cho, J., Kim, Y., Koo, J., Park, W.: Energy-cost analysis of HVAC system for office buildings: development of a multiple prediction methodology for HVAC system cost estimation. Energy Build. **173**, 562–576 (2018)
16. Deng, H., Fannon, D., Eckelman, M.J.: Predictive modeling for US commercial building energy use: a comparison of existing statistical and machine learning algorithms using CBECS microdata. Energy Build. **163**, 34–43 (2018)
17. Rahman, I., Kuzlu, M., Rahman, S.: Power disaggregation of combined HVAC loads using supervised machine learning algorithms. Energy Build. **172**, 57–66 (2018)
18. Singaravel, S., Geyer, P., Suykens, J.: Component-based machine learning modelling approach for design stage building energy prediction: weather conditions and size. In Proceedings of the 15th IBPSA conference, pp. 2617–2626 (2017)
19. Candanedo, L.M., Feldheim, V., Deramaix, D.: Data driven prediction models of energy use of appliances in a low-energy house. Energy Build. **140**, 81–97 (2017)
20. Nayak, S.C.: Escalation of forecasting accuracy through linear combiners of predictive models. EAI Scalable Inf. Syst. **6**(22) (2019)
21. Sethi, J.S., Mittal, M.: Ambient air quality estimation using supervised learning techniques. Scalable Inf. Syst. **6**(22) (2019)
22. Tsanas, A., Xifara, A.: Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy Build. **49**, 560–567 (2012)
23. Fan, C., Xiao, F., Wang, S.: Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Appl. Energy **127**, 1–10 (2014)
24. https://archive.ics.uci.edu/ml/datasets/Energy+efficiency