Forecasting Electricity Consumption in a Moroccan Educational Institution

HOUDA DAKI (houda.daki@gmail.com)

Laboratory of Information Technologies, National School of Applied Sciences University of ChoEuaib Doukkali https://orcid.org/0000-0002-1344-7034

Asmaa El Hannani

Chouaib Doukkali University Faculty of Sciences: Universite Chouaib Doukkali Faculte des Sciences

Hassane OUAHMANE

Chouaib Doukkali University Faculty of Sciences: Universite Chouaib Doukkali Faculte des Sciences

Research

Keywords: Big data, Machine learning, SMACK architecture, Spark, Smart grid, Electrical consumption forecasting

Posted Date: February 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-248534/v1

RESEARCH

Forecasting Electricity Consumption in a Moroccan Educational Institution

Houda Daki*, Asmaa El Hannani and Hassan Ouahmane

*Correspondence:
daki.h@ucd.ac.ma
Laboratory of Information Technologies, National School of Applied Sciences, University of Chouaib Doukkali , Route d'Azemmour, Nationale No 1, ElHaouzia, 24002, El Jadida, Morocco
Full list of author information is available at the end of the article

Abstract

Recently, predictive analytic contributes very well for reliable electric power supply. It provides advanced techniques to process, interpret and analyze big energy data and make it more valuable. In this paper, we have presented a benchmark of the most used forecasting models in predicting electrical energy consumption for educational institutions. This study is based on a real use case, implemented using Big Data eco-system based on SMACK architecture. The proposed system analyzes six years of data sets that highly impact National School of Applied Sciences of El Jadida-Morocco energy consumption including planning data (courses, activities, holiday etc) and meteorological data (temperature, pressure,humidity etc). The aim of this benchmark is to evaluate the prediction performance of each forecasting model in order to choose the accurate one to predict electricity consumption in educational institutions.

Keywords: Big data; Machine learning; SMACK architecture; Spark; Smart grid; Electrical consumption forecasting

Introduction

In the last years, electrical consumption has exponentially increased in many sectors due to the new behavior and policies adopted by energy users. Buildings sector which includes residential and commercial structures has known a peak demand that causes a great number of challenges related to energy management. In the same approach, Morocco electrical consumption has increased by 2.4% in the second quarter of 2019 compared to the previous year, and between 2002 to 2016 Morocco saw an annual average growth of 5.4%. Many sectors contribute at this high electricity consumption, but the building sector shows the highest growth. Universities are among domestic institutions with high electrical usage of the total amount of the entire institutions in building sector. Especially, in last year with the new practices, advanced experiences and researches, as well as the increase of the number of students and faculties. In order to support this extraordinary demands, energy supplies try to manage their resources, re-dimension the grid and control energy supply and demand. Recently, utility companies adopt a new strategy based on real time and predictive analytic to improve customer satisfaction, increase their system reliability and use efficiently renewable energy. This technique guarantees height efficiency and accuracy by taking precaution from extreme events based on critical measures that can affect energy uses [1]. In general, forecasting is used in several domains for many purpose, each use has his own horizon and granularity depending on the nature of data and the desired results. Forecasting use high degree of probability and statistical techniques to predict future and unknown events, so

it relays on modeling, machine learning and data mining that analyze current and historical facts [2].

The aim of machine learning concept is automate and mechanize the acquisition of knowledge from experience. Furthermore, machine learning improves performance of computational methods and algorithms. Thus, machine learning is hardly linked to the data-driven model which finds relationships between the state variables of the system without using physical behavior of the system. Machine learning uses some approaches to select and extract features from input data sets to train data-driven models including: supervised, unsupervised, reinforcement and transfer learning. Data-driven methods always seeks to identify the relationship between inputs and outputs by applying data mining techniques, statistical analyses and machine learning algorithms. This process goes through three stages: training to run training data sets and produce results, validation to run evaluation data sets that are different from the training data set to provide an impartial assessment of the implemented algorithm and finally the testing step to evaluate the forecasting model performances using accuracy metrics [3, 4]. In general, to apply both machine learning and data-driven models to a problem, certain steps must be followed, starting by defining objectives then collect, explore, process and visualize data and finally, run, evaluate and compare performance of the model.

In this paper, we present a real use case to predict electrical energy consumption for educational institutions. This study analyzes six years of NSASE (National School of Applied Sciences of El Jadida) data from 2014 to 2019, using five forecasting models to evaluate the prediction performance of each model and choose the accurate one. The solution is a Big Data eco-system based on SMACK(Spark, Mesos, Akka, Cassandra, Kafka) architecture and it collects data with high impact on NSASE energy consumption including planning data (courses, activities, holiday etc) and meteorological data (temperature, pressure,humidity etc). The rest of the paper is organized as follows: first section is reserved for background, the next section for proposed methodology, section for presents results and discussion and conclusion is the section five.

BACKGROUND

In general, machine learning system flow specific steps as Fig.1 describes. So, this section describes these important axes for an efficient and performing forecasting system.

Figure 1 Machine learning workflow

Data Selection and preparation

To improve prediction accuracy, the system must collect all relevant data with high impact on school building energy consumption. Electrical consumption monitoring in building can be less accurate, if it does not include all factors that hardly affect the energy use. Thus, focusing on the most relevant information and eliminate irrelevant ones is a central step in forecasting problems. For concreteness, we reviewed some studies worked on forecasting electrical consumption for education institutions to

highlight the more relevant data to collect in this context. Tab.1 summarizes for each study the data used to forecast electrical consumption. In general, the most of these studies use temperature, humidity and wind speed to represent meteorological data. But a few works use the occupancy data for there building electrical forecasting, and even those studies which use it they don't detailed enough.

Table 1 Identification of relevant data of reviewed research papers

	[5]	[6]	[7]	[8]	[9]	[10]	[11]
Temperature	×		×	×	×	×	×
Irradiation	×					×	
Humidity	×	×				×	
Wind speed	×				×	×	
Electricity usage	×	×	×	×	×	×	×
Building type	×						
Day of the week	×	×			×		
Day of the year		×					
Hour of the day		×					
Event day		×					
Event type		×					

As a result, many studies have highlighted the significance of occupancy and weather data in electricity consumption prediction. However, most of these works either do not integrate these two main data types or do not use detailed occupancy data. This work explores the use of both occupancy and meteorological data, the study gives a daily occupancy more detailed than others works, which will ensure the accuracy and the quality of predictions. The proposed solution use in addition to electrical consumption:

- **occupancy data :** presents time data, which are strongly linked to daily occupancy and schedule data based on holidays, school year, semester period and even weekends. In fact, the occupancy behaviors and activities affect hardly the energy consumption patterns. The NSASE has a relational database in which are stored all planning and schedule data for different semesters over the year. These data sources will enhance the quality of prediction results.
- **Meteorological data:** presents many factors, which must be taken into consideration because it affects strongly occupancy behavior and materials consumption. Temperature, humidity, wind speed and pressure are all main factors that should be selected as input data to implement our forecasting model. These data can be collected from several devices such as sensors, weather data and also meteorological station data.

Model Selection and training

In the last decade, forecasting techniques for electrical consumption become an active research area. These techniques contribute very well for reliable electric power supply, provides advanced techniques to process, interpret and analyze big energy data and make it more benefit and valuable. Furthermore, forecasting electrical consumption guarantees height efficiency and accuracy by taking precaution from extreme events based on critical measures that can affect energy uses [1]. A great number of research studies have discuss various aspects of electrical energy prediction to identify the suitable granularity and the accurate models. In general, energy forecasting models can be classified into three categories [3, 12, 13, 14, 15, 16]:

- **White-box models:** is a technique based on detailed physical information, it uses known and conventional knowledge like physics equations to describe cases.
- **Black-box models:** is a technique based on historical data, it uses data mining techniques, statistical analyses and machine learning algorithms to get the relation between the input and the future outputs values.
- **Grey-box models:** is a technique based on the combination of both white-box and black-box models, by improving single data-driven techniques with optimization methods, or the combining several machine learning algorithms.

According to Runge et al. [17] study, in energy consumption forecasting 84% of research use Artificial Neural Network (ANN) models applied with black-box-based models, followed by white-box with 12% and finally grey-box models with 4%. Thus, for this work, we will focus on black-box models, in order to describe this forecasting process, its concept as well as the benchmark of the models that offer.

The work of Bourdeau et al.[18] presents a review on data-driven building energy modeling techniques. They introduce the most prevalent techniques and to further provide an up-to-date overview of recent studies and advancements in building energy consumption modeling and forecast studies. According to Bourdeau et al. works, the number of reviewed research papers from 2007 to 2019 for supervised machine learning for building energy consumption modeling and forecast show that ANN is the most studied, then in second place the Support Vector Machine (SVM) and in the third place the regression models. Wei et al. [19] also review the prevailing data-driven approaches used in building energy analysis. Wei et al. review many methods for prediction building energy including artificial neural networks, support vector machines, statistical regression, Decision Tree (DT) and genetic algorithm. they conclude that ANN, SVM and regression techniques are the most used in these cases. In the same context, Amasyali et al. [20] make an overview for the most used algorithms for energy consumption in buildings. They conclude that an overall of 47% of the energy consumption prediction models utilized ANN as machine learning algorithms, while 25% used SVM, 4% DTs and 24% other statistical models.

Some studies implements their own systems and run many algorithms on them to find the more appropriate model in the case of electrical consumption forecasting in the buildings. Grolinger et al.[6] explore many prediction intervals for electrical consumption in the context of event-organizing venues including daily, hourly, and 15-min. Grolinger et al. compare forecasting results accuracy for two machine-learning approaches, ANN and SVM. They achieved high consumption prediction accuracy with daily data better than hourly or 15-min readings, and using the ANN model instead of the SVM model. The work of Amberet al. [21] also interested on daily forecasting electricity consumption of building. Amberet al. compare prediction capabilities of five different intelligent system techniques including Multiple Regression (MR), Genetic Programming (GP), ANN, Deep Neural Network (DNN) and SVM. Similarly to Grolinger et al., Amberet al. results demonstrate that ANN performs better than all other four techniques. Kim et al. [22] study compares building electric energy prediction approaches that use a traditional statistical method (linear regression) and ANN algorithms. Kim et al. results illustrate that the ANN modeling was more accurate and stable than the linear regression method. As a

result, the majority of studies find that ANN are the primary models employed to evaluate and predict energy consumption [6, 21, 23, 24, 25].

ANN offers great number of models, but according to Runge et al. [17] study most research use Multi Layer Perceptron (MLP). Runge et al. review shows that 61% of the ANN models works use MLP with a large portion using a two hidden layer. The result of the analysis showed that MLP is followed by radial basis neural networks, non-linear autoregression neural network, general regression neural network, and Nonlinear Autoregressive Network with Exogenous Inputs Neural Network with 2–7% each. MLP is regression artificial neural network that have widely used in the building sector for supervised learning using ANN. Chammas et al. [26] propose a system based on MLP to predict energy consumption of a building. They compare their system against four other algorithms, namely: Linear Regression (LR), SVM, Gradient Boosting Machine (GBM) and Random Forest (RF). Chammas et al. achieve that MLP is the best system configuration energy consumption forecasting in a building. Wahid et al. [27] find also that MLP is the best model to predict electrical consumption in building. Wahid et al. study compare MLP and RF, and their results show that MLP achieved 95.00% accurate result, whereas the accuracy observed by RF was only 90.83%.

All of these works listed above like this article compare the efficiency of ANN to other machine learning models in the case of building power consumption forecasting. However, our work compares MLP to a large number of models that are the most used in this field of research. Thus, this research will give more precision on the performance of the MLP.

Model Testing and Validation

Accuracy metrics help to validate forecasting performances of data-driven algorithms. In fact, validation step is used in order to verify the quality of the model, error metrics are used to measure the difference between values predicted and the values actually observed. Bourdeau et al.[18] propose a overview study of the most used metrics based on reviewing many studies. They find that the root mean square error (RMSE), the coefficient of variation of RMSE (CV-RMSE) and the mean average error (MAE) assessed in 47%, 38% and 36% respectively. On the other hand, the coefficient of determination (R2), the Mean Square Error (MSE), the Mean Relative Error (MRE), the Mean Bias Error (MBE) and the Normalized Mean Bias Error (NMBE) are used only for 27%, 16%, 9%, 2% and 4% respectively. Zhang et al.[28] study shows that CV-RMSE must be the first performance measure to be selected followed by other metrics. Zhang et al. consider CV-RMSE is more important metric followed by RMSE, but if these two values were unavailable, then MAPE was selected. If MAPE was unavailable, then R2 was selected. If R2 was unavailable, then the most relevant error method presented was selected and indicated as others. However, this order is not usually respected, Runge et al. [17] have done a review of the most used error metrics, and they find that MAPE is predominately (38%) used as the main performance measure within forecasting papers, with CV-RMSE and R2 accounting for 17–20% of the performance metrics applied. The use of many metrics is essential to know the models performance because applying one or two metrics may give satisfying results when evaluated against other metrics can give

poor results. Most of the times studies do not use enough metrics to evaluate their models, however it is crucial to measure the performance of the model using large error metrics. In our work, we will cover different types of evaluation metrics to evaluate deeply the model performance. Tab2 describes all the metrics that we will use in our study:

Table 2 error metrics description

Metric	Description
RMSE [29]	It presents the concentration of data around the line of the best fit. Root Mean Squared Error is usually used in regression analysis for numerical predictions because it's a good good measure of accuracy in the case of comparing prediction errors of different models or model configurations for a particular variable.
MAE [30]	It's usually used with regression models, this model evaluation metric describes the average of the absolute values of all differences between the forecast and the real values expressing the same phenomenon.
R2 [31]	It's also known as the coefficient of determination, it's a statistical measure that shows how close are the data to the fitted regression line. It indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures is a percentage value between $0 - 100\%$ scale. In general, the higher the R-squared, the better the model fits the data.
MSE [32]	It's usually used with regression models, this model evaluation metric presents the mean of the squared prediction errors over all instances in the test set. In fact, it describes the difference between the real and the predicted results for an instance. Mean Squared Error presents how the predicted results are close to to real set of points. In general, The smaller value of this metric is the more accurate.

PROPOSED METHODOLOGY

Use Case Description

The purpose of this use case is to explore data analytic technologies to predict electrical consumption of NSASE. The aim of NSASE is to become a green school using a private smart grid powered by green energy that will cover 40% of its electrical need using private smart grid that implement photo voltaic panels. This system has some challenges in term of surplus production management, because NSASE can't inject the surplus on the Moroccan electrical infrastructure or store it using storage devices which are expensive or limited. As a result, the unavoidable solution is maintain the balance between electrical production and consumption. The proposed solution implements Big Data SMACK architecture, it collects and analyses all relevant data with high impact on NSASE energy consumption such as planning and meteorological data, the historical data used in this case study presents monthly meteorological and planning data of NSASE from 2014 to 2019.

The proposed solution is implemented using SMACK (Spark, Mesos, Akka, Cassandra, Kafka) Big Data architecture, the use of SMACK is not done randomly but after a deep benchmark between the more used Big Data architectures [33]. SMACK architecture takes advantages of the most architectures(Lambda and AKKA), because it guarantees data quality and high performance by balancing throughput and latency [34]. SMACK architecture is composed by four layers each layer use a specific tool to handle data [35]. Our cluster is contains tree nodes, the master node for resource management, scheduling Spark applications and data ingestion. This node deploys Mesos master and Kafka broker. Furthermore, two additional nodes as slaves deploy Mesos slaves, Cassandra for data persistence and Spark executors. All these machines have the same configuration: Ubuntu 18.04.1 LTS as operating system, 8 cores in CPU, 8 Go in RAM and 1T of storage capacity.

Data sets Description

The used data set contains 10 different variables, meteorological information (temperature, humidity, pressure), electrical energy consumption, and academics Schedule data(Date,level,course type, course duration, staff, building type). It was collected from several sources, Tab.3 summarizes the data flows sources and destination with the description of the flow content. The data was recorded every month for 5 years from 2014 to 2019 and it was split into 80% for training and 20% for testing. So, for prediction we have an historical data of 5 years that contains the monthly consumed energy, occupancy and meteorological data.

Table 3 Data Flows Description

Data Type	Flow Content	Source	Direction
Weather Data	Timestamp Temperature Humidity Pressure	Sensors	Kafka Topic
Electrical consumption	Timestamp Electrical usage	Meters	Kafka Topic
Academic schedule	Timestamp Building type Course type Course duration Level Staff	MySQL Database	Kafka Topic

Prediction

The proposed solution is based on Spark ML lib and streaming module. Spark ML lib library offers the most known machine learning and statistical algorithms and make it easy to use [36]. Spark ML lib module provides both classification and regression models, in our case the problem nature is regression, so we typically select models looking at this type of problems. The solution runs five machine learning models including DT, RF, GBT , SVM and MLP. We run models, we calculate error metrics to specify the optimal model. We store the accurate model to apply it on the streaming data using Spark Streaming module. All these steps are developed using Scala, which is an object-oriented and functional programming language. Our experiment is divided to two main steps:

- **Features:** the system has three data sources, i) structured data including occupancy data, ii) the semi-structured data coming from weather data in JSON format and the iii) unstructured data for electrical consumption coming from sensors and smart meters installed in NSASE in text format. the selection of these data is based on the reviewed papers, but to be more accurate, models were tested on several scenarios by using all the features or by omitting some of them. The experiment defines three scenarios that all use electrical consumption data in addition to other features depending on the scenario: i) prediction using all data sets ii) prediction using occupancy data only and ii) prediction using meteorological data only.
- **Classifiers:** to solve our prediction problem, the use of the right category of machine learning models will impact hardly the results. In our case, we select regression models because we deal with a regression problem. According to many approaches existing in literature for electrical consumption prediction

Table 4 Models parameters

Parameter	Value
DT	
Strategy	vairance
FeatureSubsetStrategy	auto
NumTrees	20
Seed	12345
RF	
Strategy	vairance
FeatureSubsetStrategy	auto
NumTrees	20
Seed	12345
SVR	
Epsilon	0.1
Kernel	rbf
C	0.1
MLP	
Layers	6/3/1
Convergence tolerance	1E-5
Block size	128
Seed	12345L
Maximum number of iterations	200
GBT	
Num iteration	10

study, we choose the most used algorithms in the field to do our comparison: DT, RF, GBT, MLP and Support Vector Regression (SVR); SVR uses the same principle as SVM, but for regression problems.To run these machine learning algorithms, extracting and transforming features is also needed because Spark ML Lib is waiting for data in two columns: Features and Labels. Features is an array of data points of all the features to be used for prediction and Labels contain the output label for each data point. In our case features are: Date, temperature, humidity, pressure and schedule list that contains course type ,course duration,level, staff,building type. and there is one output label is the electrical consumption.

RESULTS AND DISCUSSION

Table 5 Electrical prediction results

	RMSE	MAE	R2	MSE
Prediction results with all features				
DT	2180	2684	0.38	4753747
RF	2151	1819	0.4	4626617
GBT	2502	1856	0.59	2120327
SVR	733	876	0.63	829997
MLP	538	436	0.94	289997
Prediction results without schedule data				
DT	2352	2078	0.28	5530597
RF	2322	1019	0.32	5030597
GBT	3502	2792	0.44	4527419
SVR	2234	2684	0.56	922397
MLP	2089	1819	0.79	789907
Prediction results without meteorological data				
DT	2301	1219	0.35	4830597
RF	2251	1619	0.49	4627817
GBT	4732	3559	0.52	2239962
SVR	2134	2584	0.52	956397
MLP	624	919	0.83	389907

This section describes the results of applying machine-learning algorithms in collected data, the analysis procedure that was used and the discussion of the results.

Before fitting the model on the training data, the first step is to set up the model parameters to determine how it will make predictions,Tab. describes the parameters used for each model, then we will fit the model on the training data.

Our experiment run these models : DT, RF , SVM and MLP on all features (electrical consumption history, weather and schedule) and we compare all models results to find the accurate one. In order to verify the quality of the models, Spark Mlib evaluation library will be used to measure the difference between the predicted values and the values actually observed. Fig. 2 shows that MLP presents better predictions with considerably better error rate as Tab. 5 describes.

In addition to that, we prepare three scenarios to validate features using MLP: i) prediction using all data sets (electrical consumption history, weather and schedule) ii) prediction using occupancy data and electrical consumption history and ii) prediction using meteorological data and electrical consumption history. The importance of this result is that we omitted the most important feature used by the system and run each type of features using MLP. Our best system configuration is using all data sets(electrical consumption history, weather and schedule data), in which we used three kind of data : i) structured data including occupancy data, ii) the semi-structured data coming from weather data in JSON format and the iii) unstructured data for electrical consumption coming from sensors and smart meters, for each model. We also tried other models without some features, and all models showed poor results with a very high difference between reel and predicted electrical consumption. 5 shows prediction results for each scenario using all models.As a result, we conclude that the use of schedule and academic data is crucial to have an accurate prediction in this kind of systems.

Figure 2 Prediction results of all models with all features

Figure 3 MLP prediction results with all features scenarios

According the error metrics results in Tab 5, we can conclude that the best model is MLP, it was able to predict with an accuracy R2 of 94,1% when using all the features, and it showed better error rate for other scenarios.

The results presented by DT are not good, it presents only a RMSE of 2180 and this makes sense due to many factors that make our data non-linear. First of all, the number of students is increased because the NSASE has allowed new courses of study, moreover, the NSASE had construction works in 2014 and 2017 which directly impacts the consumption of electricity due to the use of electric machines. In fact, DT model is optimal in regression problems if the target variable is inside the range of values that it has seen in the train data set. In our case, the range of values is not very constant due to many factors listed before. random forest (RF) presents a RMSE of 2151 despite of the fact that this model is an improvement of DT algorithm, in our case we have just improve with 1,651% in the R2 coefficient which not enough to make it an optimal model. SVR shows a RMSE of 733 which is better than previous models but it still not optimal for this case. Compared

to the previous models MLP shows good accuracy, according to a RMSE of 538, Regression Artificial Neural Network indicates a good match between the observed and predicted data.

MLP as a regression artificial neural network model is more accurate with large quantity of data and using the optimal parameters. At this experiment for all scenarios tested, there isn't high volume of data sets, but this quantity will increase exponentially and data quality will be high in next years, so the application behavior on larger data sets will be more accurate.

CONCLUSION

In this paper, we have presented a Big Data solution to predict electrical consumption in an engineering school. The system collects meteorological data including humidity, temperature and pressure, and occupancy data including course duration, course type, level, Staff and building type(educational, administrative). To be more accurate, we define three scenarios to choose the optimal features to use: i) prediction using all data sets(electrical consumption history, weather and Occupancy) ii) prediction using occupancy data and electrical consumption history and ii) prediction using meteorological data and electrical consumption history. The experiments results show that the system have best error rate using all features.

The experiments show that the use of all features (electrical consumption history, occupancy and weather) with MLP presents 94,1% of all the variability of the response data. SVR in the second place with 63% then RF model is the third in the ranking with 40% of data accuracy, then DT with 38,27%. To conclude, experimental results on different features using Spark applications demonstrated that MLP is the best model in our scenario, contrary to other models which are so far from expectations.

List of abbreviations

Table 6 List of abbreviation

Abbreviation	Value
SMACK	Spark Mesos AKKA Cassandra Kafka
NSASE	National School of Applied Sciences of El Jadida
ANN	Artificial Neural Network
SVM	Support Vector Machine
DT	Decision Tree
MR	Multiple Regression
GP	Genetic Programming
DNN	Deep Neural Network
MLP	Multi Layer Perceptron
LR	Linear Regression
GBM	Gradient Boosting Machine
RF	Random Forest
RMSE	Root Mean Square Error
CV-RMSE	Coefficient of Variation of RMSE
MAE	Mean Average Error
R2	Coefficient of Determination
MSE	Mean Square Error
MRE	Mean Relative Error
MBE	Mean Bias Error
NMBE	Normalized Mean Bias Error
SVR	Support Vector Regression

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Funding
Not applicable.

Authors' contributions
All authors read and approved the final manuscript.

Acknowledgements
Not applicable.

Author details
Laboratory of Information Technologies, National School of Applied Sciences, University of Chouaib Doukkali , Route d'Azemmour, Nationale No 1, ElHaouzia, 24002, El Jadida, Morocco.

References

1. Dagnely, P., Ruette, T., Tourwé, T., Tsiporkova, E., Verhelst, C.: Predicting hourly energy consumption. can you beat an autoregressive model. In: Proceeding of the 24th Annual Machine Learning Conference of Belgium and the Netherlands, Benelearn, Delft, The Netherlands, vol. 19 (2015)
2. Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., Fouilloy, A.: Machine learning methods for solar radiation forecasting: A review. Renewable Energy, 569–582 (2017)
3. Bourdeau, M., Zhai, X.-Q., Nefzaoui, E., Guo, X., Chatellier, P.: Modelling and forecasting building energy consumption: a review of data-driven techniques. Sustainable Cities and Society (2019)
4. Liu, H.: Smart Cities: Big Data Prediction Methods and Applications. Springer (2020)
5. Amber, K.P., Aslam, M.W., Mahmood, A., Kousar, A., Younis, M.Y., Akbar, B., Chaudhary, G.Q., Hussain, S.K.: Energy consumption forecasting for university sector buildings. Energies **10**(10), 1579 (2017)
6. Grolinger, K., Capretz, M.A., Seewald, L.: Energy consumption prediction with big data: Balancing prediction accuracy and computational resources. In: Big Data (BigData Congress), 2016 IEEE International Congress On, pp. 157–164 (2016). IEEE
7. Ruiz, L.G.B., Rueda, R., Cuéllar, M.P., Pegalajar, M.: Energy consumption forecasting based on elman neural networks with evolutive optimization. Expert Systems with Applications, 380–389 (2018)
8. Moon, J., Park, J., Hwang, E., Jun, S.: Forecasting power consumption for higher educational institutions based on machine learning. The Journal of Supercomputing **74**(8), 3778–3800 (2018)
9. Allab, Y., Pellegrino, M., Guo, X., Nefzaoui, E., Kindinis, A.: Energy and comfort assessment in educational building: Case study in a french university campus. Energy and Buildings **143**, 202–219 (2017)
10. Amber, K., Aslam, M., Hussain, S.: Electricity consumption forecasting models for administration buildings of the uk higher education sector. Energy and Buildings **90**, 127–136 (2015)
11. Hong, W.-C., Li, M.-W., Fan, G.-F.: Short-term load forecasting by artificial intelligent technologies (2019)
12. Foucquier, A., Robert, S., Suard, F., Stéphan, L., Jay, A.: State of the art in building modelling and energy performances prediction: A review. Renewable and Sustainable Energy Reviews **23**, 272–288 (2013)
13. Ordiano, J.Á.G., Bartschat, A., Ludwig, N., Braun, E., Waczowicz, S., Renkamp, N., Peter, N., Düpmeier, C., Mikut, R., Hagenmeyer, V.: Concept and benchmark results for big data energy forecasting based on apache spark. Journal of Big Data **5**(1), 11 (2018)
14. Tardioli, G., Kerrigan, R., Oates, M., James, O., Finn, D.: Data driven approaches for prediction of building energy consumption at urban level. Energy Procedia **78**, 3378–3383 (2015)
15. Yildiz, B., Bilbao, J.I., Sproul, A.B.: A review and analysis of regression and machine learning models on commercial building electricity load forecasting. Renewable and Sustainable Energy Reviews **73**, 1104–1122 (2017)
16. Daut, M.A.M., Hassan, M.Y., Abdullah, H., Rahman, H.A., Abdullah, M.P., Hussin, F.: Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods: A review. Renewable and Sustainable Energy Reviews **70**, 1108–1118 (2017)
17. Runge, J., Zmeureanu, R.: Forecasting energy use in buildings using artificial neural networks: a review. Energies **12**(17), 3254 (2019)
18. Bourdeau, M., qiang Zhai, X., Nefzaoui, E., Guo, X., Chatellier, P.: Modeling and forecasting building energy consumption: A review of data-driven techniques. Sustainable Cities and Society **48**, 101533 (2019)

19. Wei, Y., Zhang, X., Shi, Y., Xia, L., Pan, S., Wu, J., Han, M., Zhao, X.: A review of data-driven approaches for prediction and classification of building energy consumption. Renewable and Sustainable Energy Reviews **82**, 1027–1047 (2018)
20. Amasyali, K., El-Gohary, N.M.: A review of data-driven building energy consumption prediction studies. Renewable and Sustainable Energy Reviews **81**, 1192–1205 (2018)
21. Amber, K., Ahmad, R., Aslam, M., Kousar, A., Usman, M., Khan, M.: Intelligent techniques for forecasting electricity consumption of buildings. Energy **157**, 886–893 (2018)
22. Kim, M.K., Kim, Y.-S., Srebric, J.: Predictions of electricity consumption in a campus building using occupant rates and weather elements with sensitivity analysis: Artificial neural network vs. linear regression. Sustainable Cities and Society **62**, 102385 (2020)
23. Ahmad, T., Chen, H., Guo, Y., Wang, J.: A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review. Energy and Buildings **165**, 301–320 (2018)
24. Rahman, A., Srikumar, V., Smith, A.D.: Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. Applied energy **212**, 372–385 (2018)
25. Ai, S., Chakravorty, A., Rong, C.: Household power demand prediction using evolutionary ensemble neural network pool with multiple network structures. Sensors **19**(3), 721 (2019)
26. Chammas, M., Makhoul, A., Demerjian, J.: An efficient data model for energy prediction using wireless sensors. Computers & Electrical Engineering **76**, 249–257 (2019)
27. Wahid, F., Ghazali, R., Shah, A.S., Fayaz, M.: Prediction of energy consumption in the buildings using multi-layer perceptron and random forest. IJAST **101**, 13–22 (2017)
28. Zhang, Y., Yang, Q.: A survey on multi-task learning. arXiv preprint arXiv:1707.08114 (2017)
29. Chai, T., Draxler, R.R.: Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. Geoscientific model development **7**(3), 1247–1250 (2014)
30. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. Climate research **30**(1), 79–82 (2005)
31. Acharya, M.S., Armaan, A., Antony, A.S.: A comparison of regression models for prediction of graduate admissions. In: 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), pp. 1–5 (2019). IEEE
32. Botchkarev, A.: Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. arXiv preprint arXiv:1809.03006 (2018)
33. Horgas, T.: Benchmarking of big data architecture trade-offs
34. Ounacer, S., Talhaoui, M.A., Ardchir, S., Daif, A., Azouazi, M.: A new architecture for real time data stream processing. International Journal of Advanced Computer Science and Applications, 44–51 (2017)
35. Estrada, R.: Fast Data Processing Systems with SMACK Stack. Packt Publishing Ltd, ??? (2016)
36. Pentreath, N.: Machine Learning with Spark. Packt Publishing Ltd, ??? (2015)

Figures

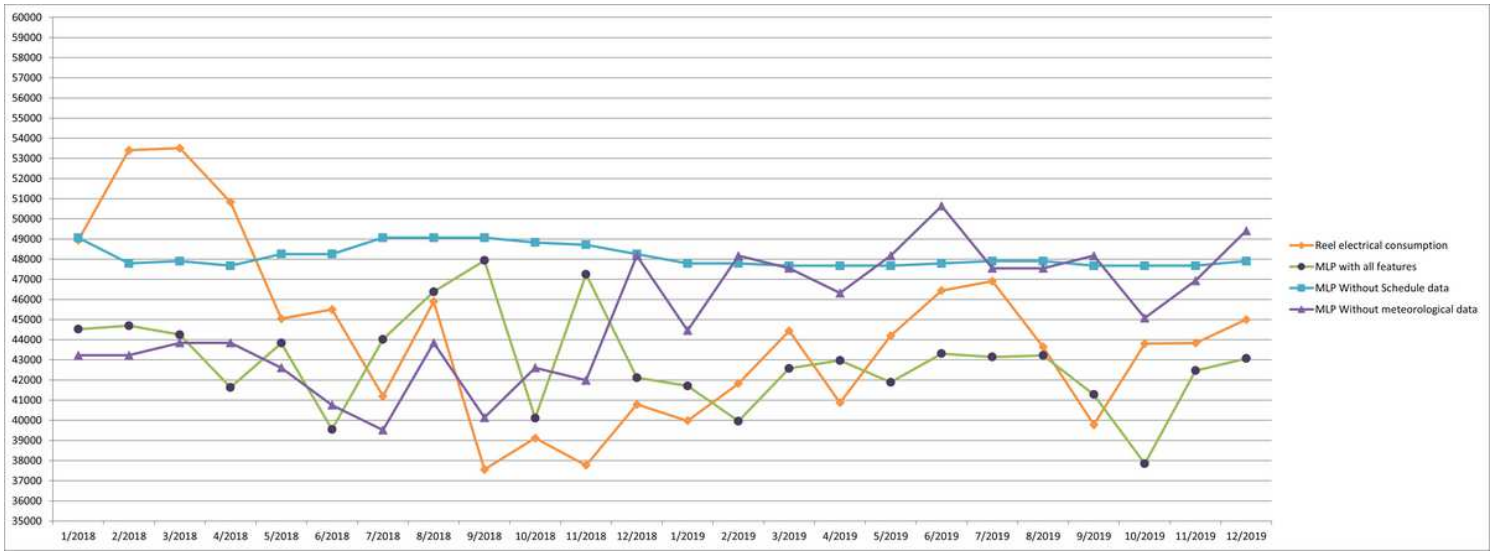

Figure 1

Machine learning workflow

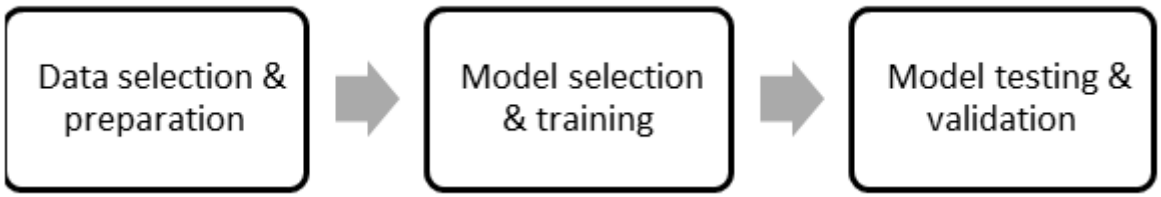

Figure 2

Prediction results of all models with all features

Figure 3

MLP prediction results with all features scenarios