National College of Ireland

Classification of Energy consumption for commercial buildings based on Energy usage.

MSc Research Project
Data Analytics

Pavan Kumar Sudhakar
Student ID: X17126738

School of Computing
National College of Ireland

Supervisor: Dr. Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Pavan Kumar Sudhakar
Student ID:	X17126738
Programme:	Data Analytics
Year:	2018-2019
Module:	MSc Research Project
Supervisor:	Dr. Christian Horn
Submission Due Date:	12/08/2019
Project Title:	Classification of Energy consumption for commercial buildings based on Energy usage
Word Count:	6155
Page Count:	20

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:	9th August 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):	

Contents

Classification of Energy consumption in commercial buildings by energy usage.

Pavan Kumar Sudhakar
X17126738

Abstract

The revolution of industries and thirst towards global progress has led to the construction of numerous buildings. These buildings were proved to be one of the significant sources of greenhouse gas emitters. Therefore, the need for the construction of energy-efficient buildings has become more demanding than it ever used to be. As a result of maintaining heating and cooling requirements, energy consumed by these buildings is of high level, which in turn emits a higher percentage of CO2 into the atmosphere. To control this, the energy consumed by the buildings needs to be predicted to make crucial decisions while designing the buildings. In recent times various machine learning models have been proposed to predict and classify the energy consumption of a building, but no research has developed a model which can be used to predict and classify the energy consumption of any given buildings based on the parameters related to the corresponding building. This paper involves addressing this research gap by proposing hybrid stacked machine learning model to classify the energy usage of 10 buildings based on the weather parameters related to respective buildings. The results obtained from this research concluded that the proposed model has outperformed all the other models with maximum accuracy of 88.12% in classifying the building energy.

Keywords: Energy Classification, Commercial Buildings, XG Boost, Ada Boost, Gradient Boost, Stacking.

1 Introduction

The emission of Co2 into the atmosphere has become a significant concern, as the global warming effect has already reached an alarming rate. One of the major contributors of greenhouse gases are the buildings, especially the commercial buildings, where the heating and cooling requirements are required to be maintained all the time. These buildings are emitting about 40% of the whole CO2 emitted into the atmosphere. One of the better approaches to reduce the amount of greenhouse gases being emitted is to control the amount of energy that is being consumed. This can be achieved by accurately predicting the energy consumed by the buildings and classifying them based on their consumption rate (Seyedzadeh et al.; 2018).
The idea for the current research comes from the thoughts behind Wang and El-Gohary (2019) effort in benchmarking the buildings, which is believed to be a proven methodology in understanding the energy efficiency of buildings. Also, the motivation to research in this area comes from Zhong et al. (2019), in which the article highlights the increase

of commercial buildings in China with an annual rate of 5.6% increase, which is almost 2.9 times greater than the overall world average. This high energy consumption is not beneficial for any country, especially the developing countries. Therefore, accurate energy prediction of buildings is given utter importance. For which, government, real estate firms and infrastructure maintenance companies have started to invest more, as it is considered to be the primary step in the reduction of energy consumption.

The proposed research work involves in extending the work of Wang and El-Gohary (2019) in benchmarking the buildings based on their energy consumption by involving additional weather parameters like wind direction, visibility, speed and time of the day. This study addresses the problem of how weather and time parameters can influence the energy consumption of buildings. Though previous researches were made in this area using weather parameters, there isnt single research that developed a model to predict and classify the energy consumption for more than one building. This is considered important because most of the research conducted in this area involves predicting the energy consumption of one particular building based on the values related to that building. This becomes a problem as the model developed would perform well only for the building under research and cannot be applied to the any other building. This research particularly addresses this problem gap, by proposing a data-driven machine learning model, which would predict and classify the energy consumption of any building by using the weather parameter related to the corresponding building. The building energy consumption is classified into three categories, namely High consumer, Medium consumer and Low consumer based on the cut off value set.

In this research, a benchmarking building dataset has been used to classify the buildings. This dataset is chosen because most of the existing research, involves the development of machine learning models that are suited to the building understudy, this limits the scope of the model only to that building and poorly performs when applied to a different building. This dataset is developed to attract all researcher to develop models using this dataset and compare their models performance (Miller and Meggers; 2017).

The following sections construct this paper: Related work section 2 categorised into subsections based on climatic variables impact and different techniques used in building energy prediction, Research methodology section 3 illustrating the flow of steps followed in order to address the research question, Implementation section 5 elaborating the technical aspects and then evaluation section 6 briefing the results and finally conclusion & future works 8.

[a]

Figure 1: Energy band used across Europe.

[a] https://ec.europa.eu/jrc/en/energy-efficiency/products/ecodesign

2 Related Work

Over the last decade, many researchers have performed many researches in understanding, predicting and classifying the energy consumption of commercial buildings which may include an office building, institutional building, hospital, shopping mall, etc. This section involves the discussing the different popular approaches that is carried out thus far in this research area, which helped in formulating the research methodology, implementation strategy, tuning and evaluation of models developed. This section involves following subsections:

2.1 Influence of weather parameters towards energy prediction.

Energy prediction research conducted by Wang and Srinivasan (2017) used the weather-related parameters like outdoor temperature, precipitation, solar radiation collected from the local weather station closer to the building under study. The data was collected on an hourly basis to predict energy consumption. The study was conducted using Multiple Linear Regression (MLR), Support Vector Regression (SVR), Artificial Neural Networks (ANN) and an ensemble model based on ARIMA, Random Forest, MARS, KNN, MLP, BT as base learners It could be seen that the ensembled approach has yielded better performance than the single model, but the single model has better reliability, easy implementation and high computational speed.

Wang, Wang, Zeng, Srinivasan and Ahrentzen (2018) has predicted the energy consumption of NV dormitory buildings in China for a period of 4 months. The author has used building outdoor air temperature and relative humidity. The data was collected only for a period of two hours (7pm-9pm), which is the buildings peak utilization time. For this research, along with the temperature parameters, the occupancy factor was also considered to predict energy consumption. Bootstrap aggregation technique was implemented, which classifies the buildings energy consumption by taking the votes from the bootstrapped samples. The bootstrapped model was compared with ANN, SVM and PMV model. The results show that the proposed bootstrapping model outperforms the other bench-marked models.

Amasyali and El-Gohary (2019) developed a data-driven model using simulated data and real-time data, which includes parameters namely occupant behaviour and usual weather-related parameters like temperature, humidity, dew point, atmospheric pressure, etc. In addition to this, temperature and humidity inside the building were also considered. These factors are assumed to influence the occupants behaviour. The collected data were normalised, and a model is developed using the ensembled approach. The feature reduction was performed using Principal Component Analysis (PCA), and the best variables were given to the model. Grid search CV is used to tune the base models, namely Gaussian process Regression, SVR, ANN and Linear Regression (LR). Data was divided into a 9:1 ratio for training and testing. The results of this research indicate that the hybrid model (ensemble) performed better than individual models.
Almalaq and Zhang (2018), used deep learning techniques to predict energy consumption, using the data collected from a primary school in Denver, Colorado, USA. The data is collected for 5 minutes interval for the whole 2012. The energy consumed is in the unit of kW per hour. It is observed that the building has high consumption during the

weekdays and low during the weekends which is no surprise, as the school doesnt operate on weekends and the need for electricity is very little during that period. This research is an extension of LSTM by wrapping the LSTM with Genetic Algorithm (GA) to make a hybrid model. The data is then split into 70:30 for training and testing, respectively. Out of 105408 steps of data, 73785 steps of data are used for training and the rest is used to test the model. The GA algorithm proved to be overcoming the non-deterministic polynomial (NP) problem of LSTM, and the model is compared with ARIMA, Multi-Layer Perceptron (MLP) and kNN. The results prove that the proposed GA-LSTM model has predicted energy consumption with the lowest RMSE and MSE value than other benchmarking models.

Zhong et al. (2019) collected data of a large building located in the city of Tianjin, China. This is a multi-purpose building which is used for entertainment, exhibition and office purposes. The thermal conductivity of outside and inside walls of the building are considered along with the usual weather parameters like dry-bulb temperature, relative humidity and same parameters are taken inside the building as well using sensors fitted at various locations. The history energy consumption data were collected at an hourly level, and the weather values are obtained from the nearest meteorological station. A novel approach of vector-based Support Vector Regression (SVR) is proposed in this research, which is developed using sigmoid and radial kernels. The proposed model is compared against the model developed using deep learning and Gradient Boosting Regression (GBR) for benchmarking purposes. The results show that the proposed model performed better than the benchmarking models with the lowest value of RMSE and MSE among all the models developed.

2.2 Techniques used in previous research

Numerous techniques and methodologies were proposed and implemented to predict the building energy consumption over the past few years. The following papers discuss the attempts in developing a sustainable model to predict and classify the energy consumption of buildings.
de Oliveira and Oliveira (2018) proposed a unique approach of combining bootstrap aggregation and decomposition to improve the performance of Uni-variate time series forecasting model to predict the energy consumption for short term (hourly) and for long term (weeks to months). Monthly energy data for the period of July 2016 to December 2016 was gathered from the International Energy Agency. In order to reduce the skewness, the data is first normalised. Then it is adjusted by employing the Box-Cox technique. This maintains the model to be on the positive side in the original scale. With the help of Seasonal Trend using Loess (STL) technique, the data is then separated into three major components, namely 1) Trend, 2) Seasonal, 3) Remainder. Data were collected for several countries in the Middle East region. Using moving block bootstrap method, the remainder for the time-series data was calculated, and at each bootstrap approach, 100 new series is generated, and every series is combined to create an inverted BC transformation. The proposed model performed better than the benchmarked models like Remainder Seive Bootstrap ARIMA and auto ARIMA with a lower value of MAPE and sMAPE.

Ribeiro et al. (2018) developed a novel model to predict the energy consumption of a school in Canada by using time series data, combined with weather parameters like temperature, dew point, etc., gathered inside and as well as outside the school building.

Overall, five different models were developed, two models using Hephaestus in which 1st model contains data for just one month, and the 2nd model has data for 12 months. Two models without Hephaestus algorithm with the same data split and the final model is just a plain ML model using 12-month data. The Hephaestus model works directly on top of feature and values without modifying the machine learning models, and it will work in both pre-processing and post-processing states. The predictions from another similar building can be used in the prediction of the building under study using Hephaestus model, and this makes it be used as a standard regression algorithm. This technique involves four stages 1) Time series adaption, 2) non-temporal domain adaption 3) standard ML and finally 4) Adjustment. The flow involves removal of time effect from the dataset, then aligning the non-temporal features so that it can be used together, then the processed dataset is feed into the Hephaestus algorithm and finally the labels are re-adjusted which were adjusted during the initial phase.

2.2.1 Implementation of Neural Networks in Energy prediction

Using Neural Networks (NN) to predict the energy consumption of buildings has been conducted by many researchers ever since the arrival of NN. It is an effective technique, which would model the non-linearity of the input vectors into target values. So many variants of NN has been tried and successfully implemented to predict the energy consumption of the buildings.

Recently, Ruiz et al. (2018) has developed a unique NN called Elman Neural Network (ENN) which predicts the energy consumption of building when used along with Genetic Algorithm (GA) to optimize the weights for the NN model. In Ruiz et al. (2018) ENN was implemented by coupling Non-Linear Auto Regressive (NAR) model and Non-Linear Auto-Regressive exogenous (NARX) model. This additional wrapping of NAR and NARX is added to improve the accuracy of the overall model. Energy consumption of a university is predicted in this research. The research showed the weakness of ENN, that, when the number of neurons is high, the model tends to overfit, so the neuron count is kept at a minimum value to carry out the research. Still, ENN performed better with 80% accuracy with the best-case scenario.

Then, Mohammadi et al. (2018) has proposed an improved ENN approach to perform energy prediction. The critical difference between the earlier ENN approach and the model proposed in this paper is, for the context layer, a self-feedback loop is added based on the co-efficient gain. Whereas in the earlier approach, a back-forward loop is structured for context and hidden layers of the NN. This addition has improved the memory effect and learning speed of the NN model while handling non-linear systems. This enables the NN model to be more sensitive to the history of the data owing to improved memory effect. The model is implemented for three commercial buildings in Iran. 50 days of historical energy data were gathered and transformed into hourly data. MAPE and RMSE evaluation factors were chosen to evaluate the model. The improved ENN model performed significantly better than the earlier ENN model with MAPE and RMSE of 3.02 and 2.07 when compared to 20.74 and 11.61 (ARIMA), 15.46 and 10.26 for SVR. Though the prediction error was reduced with the improved ENN model, still the long-term energy forecasting didnt perform as expected.

This identified problem area with ENN was addressed by Rahman et al. (2018) by proposing the Recurrent Neural Network (RNN) to predict the long-term energy need of the buildings. Here, the RNN model is used by combining Long Short-Term Memory (LSTM)

to predict the medium to long-term energy consumption. The model was applied to the dataset gathered for the period of 83 days from public safety building in Salt Lake City in the USA. The advantage of using this model is that it imputes the missing values by itself. Based on the actual data, the proposed RNN model gives weighted average predictions for before and after the missing data segment. Also, the results obtained from this research prove that the proposed RNN model has performed better than ENN models.

2.2.2 Ensemble approaches in Energy consumption prediction

A prediction model with just one model might be precarious and sometimes doesnt predict as expected. Trying different algorithms for the same problem might not work with all cases, because each algorithm has its own disadvantages and lags at a specific area while developing a model. This can be overcome by implementing an ensembled machine learning model, which would adjust or neutralise the errors or weakness of one machine learning model with another model. This becomes very important in cases like energy prediction because a minor improvement in the prediction accuracy could bring in a significant cost reduction in infrastructure maintenance (Wang, Wang, Zeng, Srinivasan and Ahrentzen; 2018).

Fan et al. (2014) has proposed a machine learning model based on ensemble ML approach to predict the next day energy consumption, also identifying peak power demand time. The tallest building in Hong Kong is chosen, and the historical energy data was collected. Data is cleaned to remove outliers. Then, by using Recursive feature elimination, dimensions are reduced. Finally using eight different base models, a hybrid ensemble algorithm is developed. The results of this research show that the ensemble model performed better than each of the individual base models with a prediction error of 2.32 (MAPE).

Wang, Wang and Srinivasan (2018) predicted the energy consumption using Ensemble Bagging Trees (EBT) using data obtained from the meteorological department closer to the building and the actual energy consumed by the structure under study. The proposed model performed well with better accuracy than single prediction models with prediction error range between 2.9% to 4% only. But, the model is too complicated and takes high computational time. Also, EBT involves additional data partitioning steps for training multiple base models.

Robinson et al. (2017) proposes an ensemble model to predict the building energy consumption using minimal features related to the dimensions of the building to train the model. Data was gathered from U.S Energy Information Administration (EIA), which contains just 6270 rows of data related to building under study. Gradient Boosting (GB) model was developed, and the results are validated using k-fold cross-validation technique. The proposed model is then compared with Linear regression, SVR, RF regressor, Ada Boost, etc. The proposed GB model performed better than other models with R^2 value of 0.81. This research also indicates that higher capacity models like GB can perform well even with limited input features, than using more features on general linear models. GB itself is an ensemble algorithm which grows a series of trees sequentially to improves the performance of the single tree-based model, where the performance of newly created tree is based on the performance of the previous tree. This algorithm works by the principle of assigning greater weights to the features in the new tree that performed poorly in the previous tree (Deng et al.; 2018).

For example, Multiple Linear Regression wont perform very well with non-linear prob-

lems, which can be neutralised by the application of ANN over the MLR model. But ANN has its own disadvantage of establishing a relationship with the buildings actual energy consumption and input physical parameters. Thus, each algorithm has its own pros and cons, depending on the problem, multiple algorithms shall be used to make a better prediction model (Wang and Srinivasan; 2017).

2.3 Conclusion

Though Neural Networks and Time-series approaches have performed considerably well than all the other models studied thus far, it still has a disadvantage of taking high computational time, complex model structure and requirement of large scale computational devices. Ensemble ML models have performed better than the other models without much complexity and computational support. But, the choice of algorithms to be used in ensemble model purely relies on the researcher expertise, and there is documentation or standard procedure to use ensembled ML model to address a specific problem (Wang and Srinivasan; 2017), (Wang, Wang, Zeng, Srinivasan and Ahrentzen; 2018). But based on the learning's and understandings from the literature study, it could be observed that there is a need to develop a machine learning model which would classify energy consumption of multiple building for every hour. From the knowledge consumed through the literature, it gives the idea of applying a hybrid machine learning model can yield better prediction accuracy and performance in addressing the research objective. Therefore, I have developed an ensemble machine learning model by using XG Boost, AdaBoost, Gradient Boost and SVM algorithms.

3 Methodology

3.1 Data Extraction

The dataset used in this research is a benchmarking dataset used for the purpose of evaluating the buildings energy prediction models with each other. Miller and Meggers (2017) has collected the data directly from the sites of about 507 buildings in the USA. This data provides the details of the buildings and their energy consumption in hourly basis for the year 2015 [1]. The buildings for which the data collected includes buildings related to Education, Government, Healthcare, Universities, Commercial property, Entertainment, etc,. Each building has 8760 records, which contains the hourly energy consumption details for whole year in kW/hour.
The weather parameters related to the buildings which includes, Temperature, Humidity, Dew Point, Wind speed, Wind direction, Wind degrees, Conditions, Visibility, Sea level pressure hPa and Gust speed, etc, were also gathered from the local meteorological station closer to the related buildings. Totally there are 31 different weather data collected for 507 buildings. The details of, which weather file corresponds to which building is maintained separately in a meta data file.

Note: All data-sets used in this project are publicly available and GDPR compliant, and there is no data related to gender, religion and other things that may disclose a person's identity.

[1] `https://github.com/buds-lab/the-building-data-genome-project/tree/master/data`

3.2 Data Pre-Processing

In this step, the dataset is cleansed and transformed, so that it fits to the model. The steps carried out to make the dataset to be ready for application of algorithm is explained in below sections.

3.2.1 Cleaning

The dataset was cleansed by following steps below:

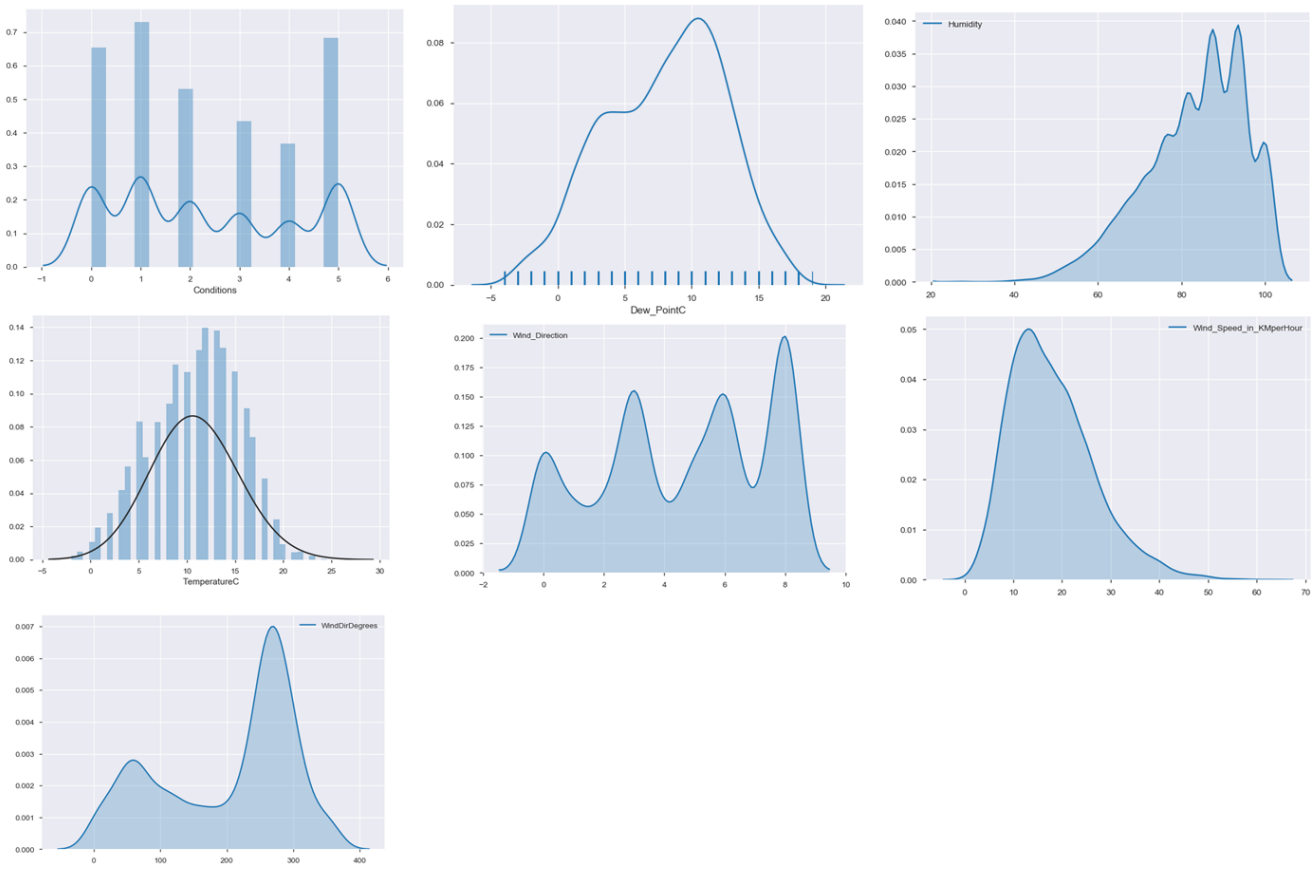

Figure 2: Variable Distribution plot

1. None of buildings in the energy dataset had data for all the years, few buildings had data only to 2014 and the remaining buildings had data for 2015 only. This was identified and the building data related to 2014 was removed. The reason to remove 2014 is because it had huge number of missing values. Since, the data gathered is very sensitive, imputation of missing values would have resulted in poor model prediction.

2. The missing values in other columns is also checked, few columns had the dummy value coded as -9999, this was also identified and removed from the dataset.

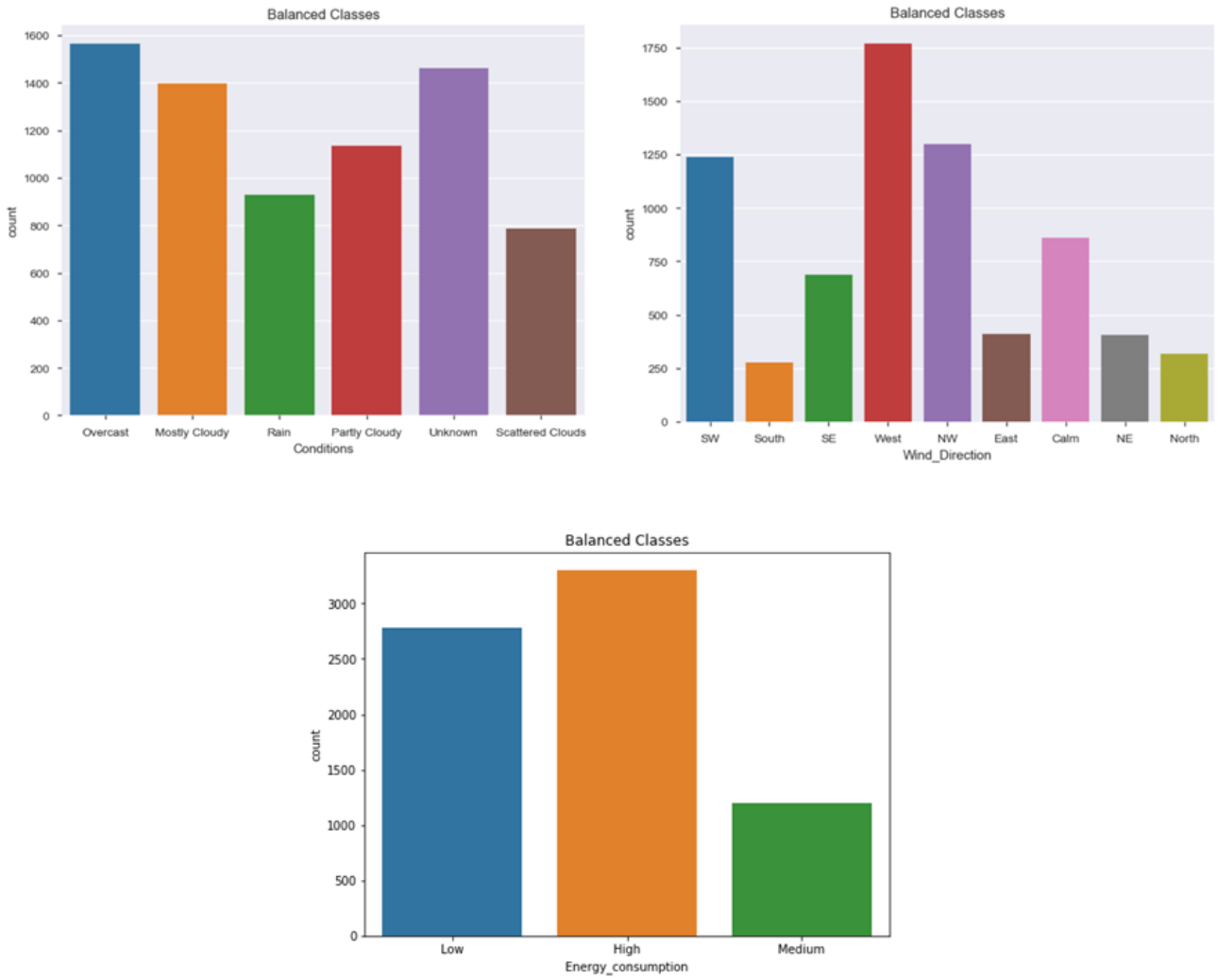

Figure 3: Class balance plot of categorical variables.

3. The distribution of all the independent features are analysed by plotting the variables. This is shown in figure 2.

4. The Outliers were identified with the help of box-plot technique and removed from the dataset.

5. The weather data had lot of redundant rows, which are identified manually with the help of Excel and removed from the dataset.

6. Variables like 'Events', 'Gust_Speed' and 'precipitation_in_mm' did not have data for almost 7500 rows out of 8016 rows. Those features were removed before feeding into the model.

7. Co-relation of all variables with the dependent variables were checked using Pearson correlation co-efficeint to identify the most influencing variables in the model. Multicollinearity is also checked among the independent variables by plotting variable heatmap co-relation plot from Seaborn package in Python.

8. The categorical variables are checked for the class imbalances by using countplot function in Seaborn. It is shown in the figure 3.

3.2.2 Variable Transformation

The below transformations were applied on the data to fit the model.

1. The weather dataset and building energy dataset was merged together in reference to date and hour.

2. The date column was split into hour and minutes columns and then minutes column was removed. All the column names was suitably renamed as per the metadata.

3. The wind speed column had values coded as 'calm' when there is absolutely no wind. This was replaced as '0' to make the whole column into float type.

4. The conditions column had around 25 classes, where most of them are redundant, with different names. This was identified and coded into 6 classes only.

5. The categorical variables are label encoded and the numerical variables are normalized between the range of 0 to 1, before feeding the data into the model.

3.3 Feature Engineering

After checking the co-relation among the dependent and independent variables using co-relation plot, a step wise feature selection was conducted using Chi-square analysis. Out of all available variables, the best number of features are selected using this test. The number of variables are adjusted using a for loop and the model is ran every time to identify the best number of best variables for the model.

3.4 Holdouts

Final stage of pre-processing is to split the data for training and test purposes. The train data will be used to train the models. The predicted values are compared against the test data set to measure the accuracy and error produced by the model. The data is split using train_test_split function in python's sci-kit learn library. The data is split in the ratio of 80:20 for training and testing the model.

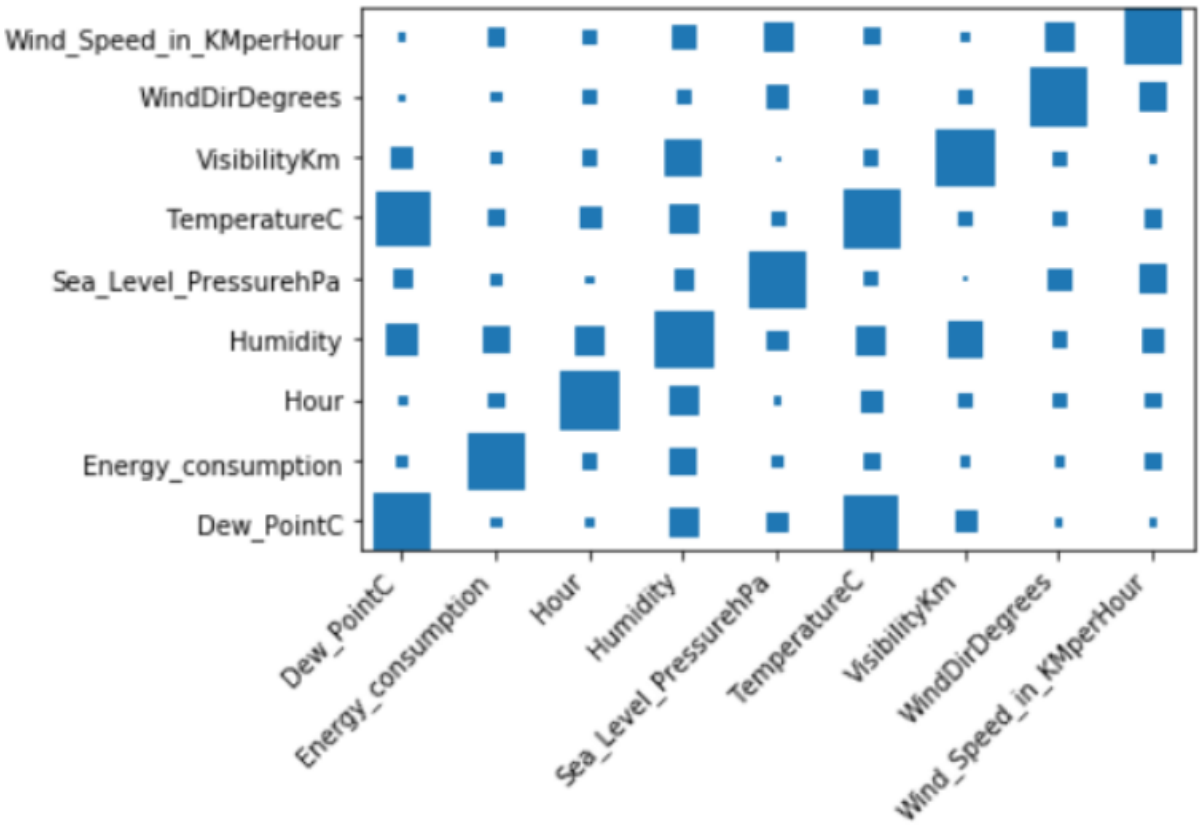

Figure 4: Co-Relation matrix

4 Design Specification

4.1 SVM

Support Vector Classifier has been chosen as one of the benchmarking models, as this model is considered to be a promising model in classification problem. It works by plotting the data points in n-dimensional space, where the value of each feature is used as co-ordinates. Classification of dependent variable happens by finding the right hyperplane across the plotted classes. Since, this research involves multiclass classification, selection of kernel is very essential. Out of all kernels, radial kernel is chosen as it performs very well for multiclass problems.

4.2 Boosting

In general, boosting is a technique of improving the model by making the weak learners in the model into strong learners. Ada Boost is an ensembled decision tree algorithm, which works by building the decision tree for all features, then calculation and assignment of weight at each depth. During each step, the errors made by the previous tree is considered while assigning the weights to the new tree. The weak learners are assigned higher weights while the strong learners are given low weights. This makes Ada Boost an efficient classification algorithm as it learns from the errors of previous trees and makes

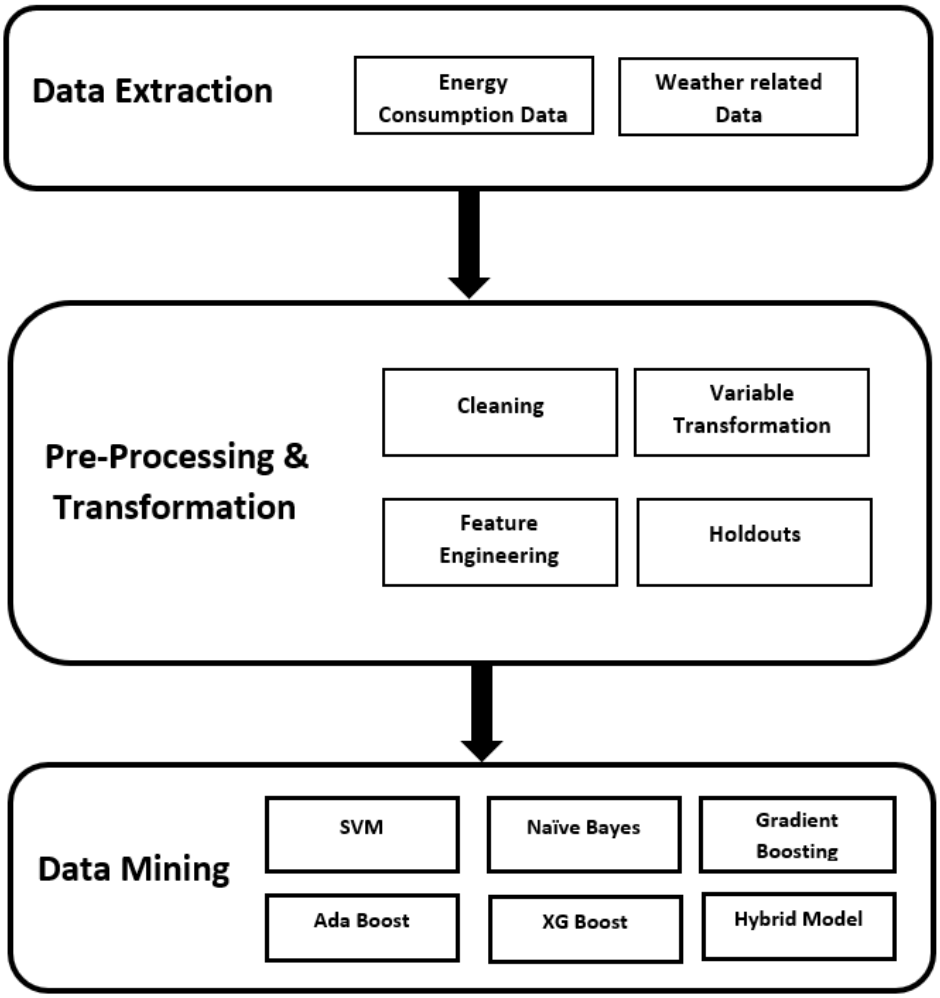

Figure 5: KDD Framework incorporated in Methodology.

full use of weak learners in the overall model. Gradient Boost is like Ada Boost, but instead of weights, GB model uses gradients at the loss function to improve the models performance.
XG Boost is simply eXtreme Gradient Boosting algorithm, where it improves the computational time of GB model by its ability to be highly scalable and fast computational power. XG Boost is far easier to execute than other boosting algorithms. Also, the plot_importance function in XG boost helps the researcher to identify the contribution of each variable in the model by giving the f score of each feature.

4.3 Hybrid Model

Combining multiple weak learners into a single learning algorithm improves the performance of the model. This is very useful in cases where a small improvement in prediction accuracy could bring in significant changes in the problem area. In this paper, 3 machine learning models namely 1) SVC, 2) Gradient Boosting, 3) Ada Boost were used as base leaners which would build new training dataset for the meta learner. Here, XG Boost is used as a meta learner which learns from the training set generated by the base learning

models. This potentially improves the predictive capacity of the model as the weakness of one model is compensated by the other.

5 Implementation

The overall research is implemented in Python (version 3) using Jupyter notebook in a 8GB RAM laptop. The machine is configured with i5 8th generation processor with a maximum clock speed of 2.20 GHz. The analysis was first conducted for a primary school Javier with the weather parameters collected from the nearest meteorological station. Firstly, data for XG Boost model is developed by using 80:20 holdouts for training and testing.The XG Boost package needs to be installed separately as it's not available in default python libraries. For this, Anaconda prompt window is run in administer mode to install and import the XG Boost library. Before running the model, the data should be converted into numeric values. This transformation step is mandatory, as XG Boost works only when the data is numeric. For this, the categorical variables are separated from the main dataset which is then encoded using Label Encoding technique. This would assign the classes a numeric value which can be applied to XG Boost model. The integer and float variables are separated and normalized, so that the values in different scale are in the range of 0 to 1. Finally, both the datasets were merged together which is now ready for algorithm application.

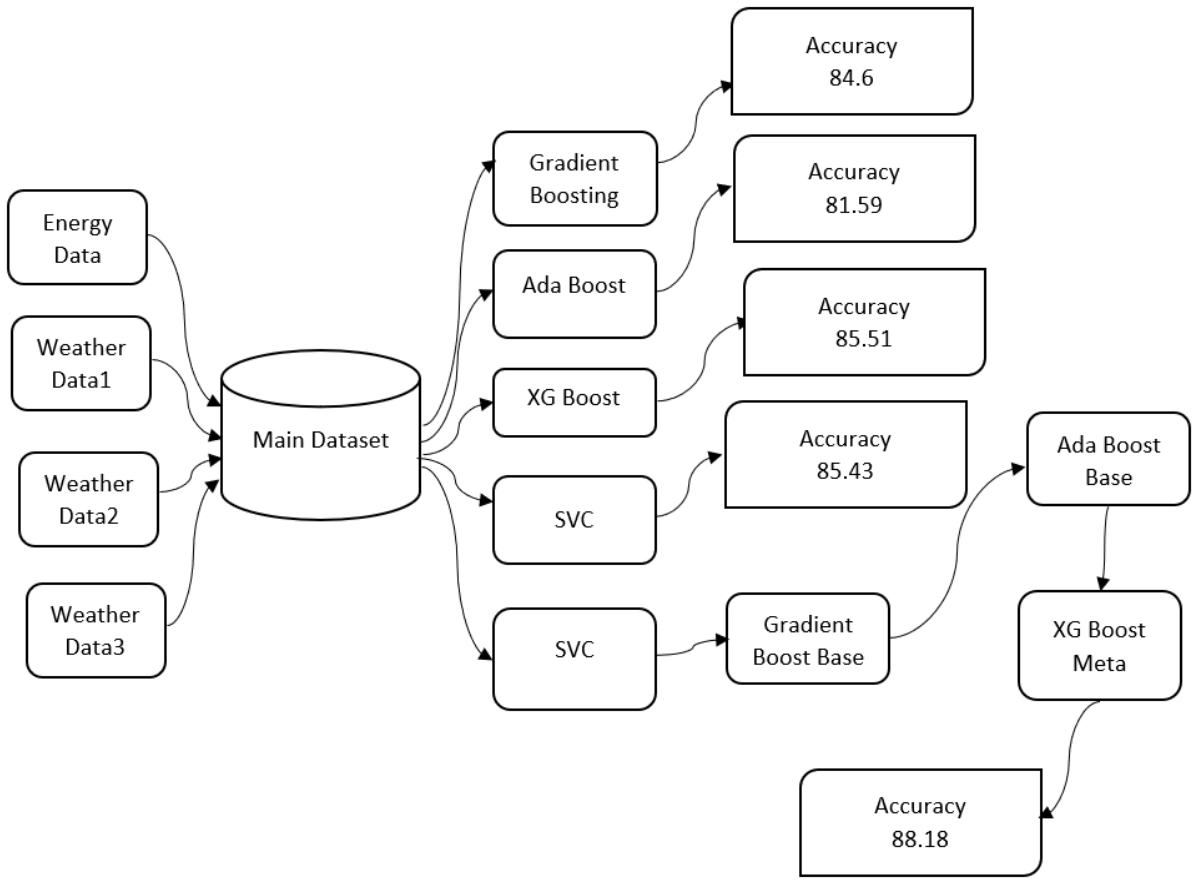

Figure 6: Architecture Implementation.

Initially, the model is with the default parameter settings with learning rate set as 0.1, max_depth = 5 and number of estimators = 10. This setting didn't produce the expected accuracy, therefore by using GridSearchCV the model is run for 100,000 times to identify the ideal parameters for the XG Boost model. The model performed best at the setting of learning rate = 0.43, max_depth = 26 and number of estimators = 100. The model with new hyper parameter setting has performed far better and the accuracy has improved to a significant level. The second and third model was developed using SVM and Gaussian Nave Bayes algorithm, these models were developed to evaluate the hybrid algorithms by keeping them as benchmarking algorithms. The performance of 2nd and 3rd model is not as much as the other models, which is anticipated. Gradient Boosting Classifier (GBC) model which is another hybrid machine learning algorithm, generally very good in handling classification problem, but it takes more time to finish the models computation. Even in this case, GBC model took the longest computational time than all the other models. Another hybrid model, Ada Boost was built using Decision Tree Classifier (DTC) as its base model. The DTC model is bagged and boosted with different tree depth setting to improve the model accuracy. AdaBoost is an eXtreme version of GB which mainly focusses in parallel operation and reducing the computation time taken by GBC. But the results obtained in this examination shows that, though GBC took more time, but still performed better than Ada Boost model.

Finally, the proposed stacked ensemble model is developed by using SVC, GBC and Ada Boost as base learning algorithms. The ensemble model is built using 'Vecstack' library in python. Like XG Boost, the vecstack library is also installed using anaconda prompt as it's not available in default python libraries. This library helped in stacking the three base models one after the other. The original training and testing data are passed into the 'stacking' function. This function will use the original holdouts for all base models and creates a new training data for the meta learner. On top of it, all base models are cross validated by 10-fold CV method, this happens internally while stacking the base algorithm. Here, XG boost algorithm is used as meta learner which will learn from the newly created training data, generated by the base learning models. By doing this, the performance of the overall model improves significantly as it has the combined effect of four different models.

S.No	**Building Name**	**Related weather data**
1	Primary school Jayla (Academic)	Weather2
2	Primary school Jaylin (Academic)	Weather3
3	Primary school Janiya (Academic)	Weather3
4	Primary school Janice (Academic)	Weather3
5	Office Jett (Office)	Weather2
6	Office Jerry (Office)	Weather3
7	Office Jackie (Office)	Weather2
8	Primary school Javier (Academic)	Weather1
9	Primary school Jeanette (Academic)	Weather1
10	Primary school Janis (Academic)	Weather1

Table 1: List of buildings and their weather details.

All the above models, except the hybrid model and naive bayes is tuned with the help of GridSearchCV to find the right hyperparameters to produce the maximum accuracy. A for loop is also used during development stage to tune the individual parameters, though this is not the ideal method to tune the parameters, it still performed very well. The problem with GridSearchCV is that it the model ran for 100,000 times to identify the best values for just 3 parameters (Learning rate, Max_depth and number of estimators). GridSearchCV has taken almost 2 days to complete, while the for-loop method was equally effective with just by developing 100 models for each parameter and took just 22 minutes. After tuning and running the models, it is validated by using k-fold cross validation method. This is performed to make sure the accuracy achieved is not because of biased sample. A 10-fold cross validation is performed to make sure all the data was used in training and testing purposes, and the reported metrics is the average metrics of all 10 folds.
To evaluate the research objective, 10 different building was chosen from 3 different locations so that the model is not a biased for just one building. The details of the buildings and their corresponding weather data is shown in below table:

6 Results and Evaluation

This is the crucial step in overall data-mining process. After the application of algorithms on data, it is essential to know how well the model has performed for the given set of data. In general, metrics like accuracy, precision, sensitivity, specificity, recall, Kappa score and Area Under Curve (AUC) curve are used to evaluate a classification model. The usage of these metrics depends on the type of problem and data used. As explained in 3.4, the model is evaluated using the test dataset.

6.1 Accuracy

Accuracy is the ratio of number of correct predictions to the total number of predictions made. This is one of the well know evaluation metric to analyse the performance of machine learning model. The accuracy achieved by each of the six developed model for the all the buildings were tabulated and shown in the below table.

Weather	**Office**	**XGBoost**	**SVC**	**Naive Bayes**	**Gradient Boosting**	**Ada Boosting**	**Stacked Ensemble Model (Proposed)**
W2	Jayla	80.72	67.9	66.4	78.7	80.7	78.9
W3	Jaylin	81.34	73	67.38	80.9	72.2	81.83
W3	Janiya	82.6	69.3	69.39	83	76.3	83.12
W3	Janice	79.59	68.1	61.1	78.3	70.8	78.96
W2	Jett	81.72	68.96	68.21	82.2	71.7	82.6
W3	Jerry	85.11	75.9	69.14	84.6	80.1	85.26
W2	Jackie	74.7	64.9	59.9	75	66.8	74.9
W1	Javier	86.26	85.43	82.41	87.5	87.5	88.18
W1	Jeanette	80.63	81.59	79.12	81.59	81.59	82
W1	Janis	76.23	75.41	69.64	78.29	78.29	78.84

Table 2: The accuracy achieved for all buildings.

6.2 Area Under Curve

The AUC curve is a widely used visualizing technique to analyse a classifier model. AUC is formed by plotting True positive Rate (Sensitivity) on the Y-axis and False Positive Rate (Specificity) on the X-axis. AUC is usually plotted for binary class problem. Since, in our case the number of classes is more than 2, micro averaging and macro averaging technique is used to plot the AUC graph. The macro averaging takes the average of all classes and computes the AUC metric, while the micro averaging technique will aggregate the contribution of all classes in order to calculate the AUC metric. Usually for a multiclass problem micro averaging technique is best suited, because the chances of having class imbalance is more in multiclass classification.

From the graph, it can be seen that model has performed exceptionally well with micro average AUC area is equal to 0.96 and macro average AUC area is equal to 0.83. The individual AUC class values are 0.84 for class0, 0.87 for class1 and 0.76 for class 2.

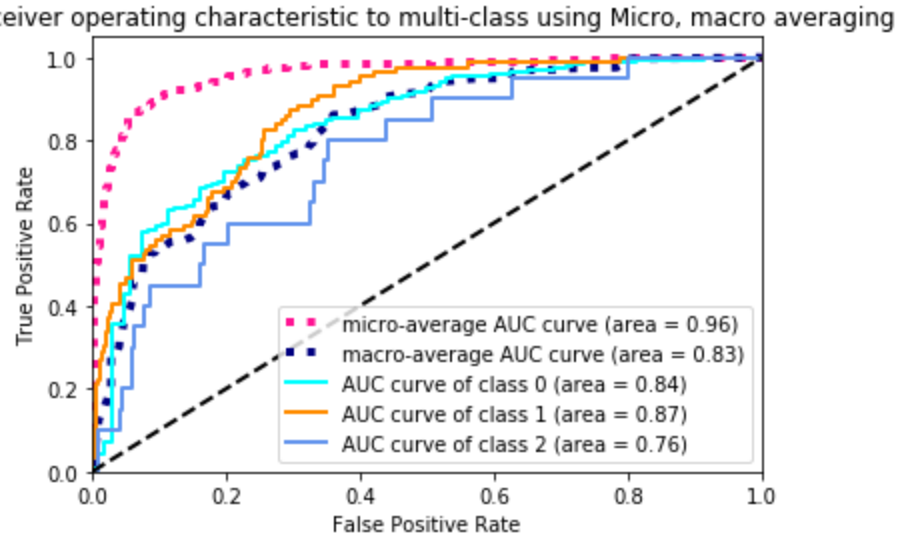

Figure 7: Area Under Curve

7 Discussions

From the above results, the performance of all six models shall be compared, and it is observed that the proposed stacking algorithm has provided a significant improvement in classifying the buildings energy consumption. Apart from the stacked model, Gradient boosting and XG Boost algorithms has also performed well with good accuracy. Other benchmarking models like SVC and Naive Bayes has performed moderately with very low accuracy between the range of 60-70. The tuning of hyper parameters has really helped in improving the performance of the model. Though GridSearchCV method is time consuming in identifying the right parameters for the right model, the parameters suggested by this function has boosted the models accuracy to almost 10%.

Statistical tests are also conducted , which concluded that there is a significant improvement in prediction while classifying the energy consumption using stacked ensemble model. The results of 10-folds were taken from all the models to conduct this experiment.

Data is first checked for normality using Shapiro-Wilk test which came out significant and then Mann-Whitney U-test was conducted with proposed ensemble model against all other single models. Table 3 shows the results of statistical tests and its significance.

```
# Checking for Normality
from scipy import stats
print(stats.shapiro(output))
(0.9317049384117126, 0.0023475424386560917)
#p-value of 0.002 shows that, data is not normally distributed
```

Shapiro-Wilk test for Normality

```
# Checking for Statistical Significance
from scipy import stats
stats.mannwhitneyu(output['Stack'],output['gnb'],use_continuity=True,
    alternative=None)
stats.mannwhitneyu(output['Stack'],output['svc'],use_continuity=True,
    alternative=None)
stats.mannwhitneyu(output['Stack'],output['adb'],use_continuity=True,
    alternative=None)
stats.mannwhitneyu(output['Stack'],output['xgb'],use_continuity=True,
    alternative=None)
stats.mannwhitneyu(output['Stack'],output['gb'],use_continuity=True,
    alternative=None)
```

Mann-Whitney U-test

p-Value	**Algorithm**	**Statistical Significance**
0.0000729	SVC	Yes
0.0000903	Naive Bayes	Yes
0.0069377	Adaboost	Yes
0.0924414	XGBoost	No
0.4547421	Gradient Boost	No

Table 3: Statistical Significance - Independent sample T-test

The table shows that proposed algorithm has a statistically significant improvement than Naive Bayes, SVC and Ada boost model. Though the test between proposed algorithm with GB and XGBoost is not significant, still there is noticeable improvement in accuracy among them.

8 Conclusion and Future Work

The overall study definitely has an impact in the research area, an unique model is developed in this research, which can be used to classify any building's energy consumption by using their energy consumption details and weather parameters from the nearby meteorological station or by the sensors fitted at the building. The energy consumption is classified as 'High', 'medium' and 'Low' consumers as per their energy usage. In summary, we can conclude that, multiple classification algorithms were developed and tested to classify the building's energy to identify the low to heavy energy consumers. The

proposed stacked hybrid model that has Ada Boost, XGB and SVC as base learner and XGB as meta learner has outperformed all the other models and this has been proved by a statistical test as well. This research used energy consumption data of 10 buildings and three different weather data to carry out the analysis. The main objective of this research is to build a model that would classify energy consumption of any building with same or closer level of accuracy. This research has satisfied that objective by classifying energy consumption of multiple buildings based on the weather parameters related to the respective buildings.
The impact of building size and number of occupants present in each floor in the building would really improve the model. At this moment, those data are not available. But this can be considered as an improvement and I would highly suggest to extend this research by considering the building size specification and number of occupants in the building.

9 Acknowledgement

I would like to thank my supervisor Dr.Christian Horn for his continuous support and guidance throughout this research. His advises were very supportive while taking crucial decisions in every step, without which the on-time completion of project would have not been possible.

References

Almalaq, A. and Zhang, J. J. (2018). Evolutionary deep learning-based energy consumption prediction for buildings, *IEEE Access* **7**: 1520–1531.

Amasyali, K. and El-Gohary, N. (2019). Predicting energy consumption of office buildings: a hybrid machine learning-based approach, *Advances in Informatics and Computing in Civil and Construction Engineering*, Springer, pp. 695–700.

de Oliveira, E. M. and Oliveira, F. L. C. (2018). Forecasting mid-long term electric energy consumption through bagging arima and exponential smoothing methods, *Energy* **144**: 776–788.

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.

Fan, C., Xiao, F. and Wang, S. (2014). Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques, *Applied Energy* **127**: 1–10.

Miller, C. and Meggers, F. (2017). The building data genome project: An open, public data set from non-residential building electrical meters, *Energy Procedia* **122**: 439–444.

Mohammadi, M., Talebpour, F., Safaee, E., Ghadimi, N. and Abedinia, O. (2018). Small-scale building load forecast based on hybrid forecast engine, *Neural Processing Letters* **48**(1): 329–351.

Rahman, A., Srikumar, V. and Smith, A. D. (2018). Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks, *Applied energy* **212**: 372–385.

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.

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**: 889–904.

Ruiz, L. G. B., Rueda, R., Cuéllar, M. P. and Pegalajar, M. (2018). Energy consumption forecasting based on elman neural networks with evolutive optimization, *Expert Systems with Applications* **92**: 380–389.

Seyedzadeh, S., Rahimian, F. P., Glesk, I. and Roper, M. (2018). Machine learning for estimation of building energy consumption and performance: a review, *Visualization in Engineering* **6**(1): 5.

Wang, L. and El-Gohary, N. M. (2019). Machine-learning-based model for supporting energy performance benchmarking for office buildings, *Advances in Informatics and Computing in Civil and Construction Engineering*, Springer, pp. 757–764.

Wang, Z. and Srinivasan, R. S. (2017). A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models, *Renewable and Sustainable Energy Reviews* **75**: 796–808.

Wang, Z., Wang, Y. and Srinivasan, R. S. (2018). A novel ensemble learning approach to support building energy use prediction, *Energy and Buildings* **159**: 109–122.

Wang, Z., Wang, Y., Zeng, R., Srinivasan, R. S. and Ahrentzen, S. (2018). Random forest based hourly building energy prediction, *Energy and Buildings* **171**: 11–25.

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**: 403–414.

List of Tables

List of Figures