8-2019

A Machine Learning Framework for Energy Consumption Prediction

Chakara Rajan Madhusudanan
Clemson University, chakararajan@gmail.com

A MACHINE LEARNING FRAMEWORK FOR ENERGY CONSUMPTION PREDICTION

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Industrial Engineering

---

by
Chakara Rajan Madhusudanan
August 2019

Accepted by:
Dr. Sandra Duni Eksioglu, Committee Chair
Dr. Michael Carbajales-Dale, Committee Member
Dr. Burak Eksioglu, Committee Member

ABSTRACT

Energy needs to be used very efficiently in today's world. With fast paced improvements in the industrial sector, demand is increasing, and energy efficiency programs become vital to reduce the energy wastage while also meeting the demand. The analysis of several scenarios used by policy makers suggest that for the global temperature to raise by less than 2° C by the end of this century, it is necessary to reduce industrial energy consumption increase by at least a half. To be on track with these scenarios and to achieve the desirable targets, it is important that we incorporate a dependable forecasting tool that can be used to predict the energy consumption based on several expected parameters. In this thesis, a survey is conducted on energy consumption forecasting algorithms to compare the advantages and disadvantages of each, explaining for what applications they would be the best fit. Also discussed in this thesis is a machine learning supported regression model that has a higher accuracy when compared to conventional regression models. The Industrial Assessment Center database contains data from all assessments conducted on manufacturing facilities that include plant area, production hours, number of employees, annual sales and the region the facility is from. These variables, along with average annual temperature are the independent variables and represent the various factors affecting energy consumption. The dependent variable is annual energy consumption. The suggested model incorporates random forest feature selection to identify the most important variables in the dataset. The dataset is first divided into 3 groups based on the value of the most important variable, production hours. Each of these groups is further divided into three groups based on the value of the second most

important variable, plant area. The algorithm then fits linear, polynomial and support vector regression models to each of these 9 groups for training. While testing, the model uses the respective regression plane based on the testing data's value of the most important two variables. This approach to regression gave 23% lesser percentage deviation than conventional regression modelling. Polynomial regression works better for the entire dataset whereas linear regression performs equally good in the subsets, suggesting that the linearity of data increases as the dataset is divided into homogenous subsets. Production hours and plant area have the highest impact on energy consumption. To reduce energy consumption, these two factors must be analyzed. This model can be used by various Industrial Assessment centers to find out future expected energy consumption of clients to give a more accurate figure of the payback periods of various recommendations.

ACKNOWLEDGEMENT

I would like to start by thanking the Almighty for giving the strength and determination to write my master's thesis. I am forever indebted to my committee members Dr. Sandra Duni Eksioglu, Dr. Michael Carbajales-Dale and Dr. Burak Eksioglu for their constant support and constructive criticism to put together this work of research. I would also like to mention Phillip Litherland, the assistant director of Clemson Industrial Assessment Center for being a wonderful mentor during my time there. No number of words would be enough to explain the contribution of my family: my parents Madhusudanan Parthasarathy & Jayasree Madhusdanan, my aunt Vaijayanthi Parthasarathy and my girlfriend Priyanka Pitchumani. They have been instrumental in all my efforts in life and continue to do so until this day.

I would like to extend my sincere thanks to my friends, Vishnu Narayan and Nitin Srinath for helping me during the progress of my thesis with their constant comments and suggestions for improvement of the research. I would also like to thank two other friends of mine, Devesh Kumar and Srinivasan Nagarajan for their guidance in compiling and formatting my master's thesis. Without their help, a report of this scale wouldn't have been possible. I am grateful for the help of my teammates at the center: Lakshana Nagaraj, Vikas Garg, Harish Lakshmi Srinivasan, Satwik Dhumal, Murgesh Awati for their significant help in data collection and management.

TABLE OF CONTENTS

TITLE PAGE i
ABSTRACT ii
ACKNOWLEDGEMENT iv
LIST OF TABLES iii
LIST OF FIGURES iv

1. INTRODUCTION 1

1.1. BACKGROUND 1

1.2. CURRENT SCENARIO AND CHALLENGES 3

1.3. OVERVIEW OF INDUSTRIAL ASSESSMENT CENTER PROGRAM 7

1.4. MOTIVATION: NEED FOR ACCURATE ENERGY PREDICTION 8

1.5. ORGANIZATION OF THESIS 10

2. LITERATURE REVIEW 11

2.1. INTRODUCTION 11

2.2. REGRESSION MODELS 11

2.3. MACHINE LEARNING INTRODUCTION 13

2.4. RANDOM FOREST MODELS 14

2.5. SUPPORT VECTOR MACHINE MODELS 15

2.6. NEURAL NETWORK MODELS 17

2.7. COMPARISON OF DIFFERENT APPROACHES 20

2.8. SEARCHES AND DATABASES 23

3. METHODS 25

3.1. REGRESSION 26

3.2. MACHINE LEARNING 28

3.3. SUPPORT VECTOR REGRESSION 30

3.4. RANDOM FORESTS 30

4. MODEL 34

4.1. VARIABLE DESCRIPTION 35

4.2. DATA COLLECTION & PRE-PROCESSING 37

4.3. CLASSIFICATION APPROACH 38

4.4. CONVENTIONAL REGRESSION MODEL 39

4.5. PROPOSED REGRESSION MODEL WITH RANDOM FORESTS 43

4.6. MODULES USED 48

4.7. MODEL VALIDATION 50

5. DISCUSSIONS AND FUTURE WORK 53

6. REFERENCES 56

LIST OF TABLES

Table 2-1: Advantages and disadvantages of each method 23

Table 4-1: Description of different variables used in the model 35

Table 4-2: List of variables with their respective upper and lower bounds 38

Table 4-3: Kurtosis and Skewness values for linear and polynomial regression 40

Table 4-4: Test statistics and statistical significance of each variable 41

Table 4-5: Test statistics and statistical significance of each variable 42

Table 4-6: Variable importance based on random forests 44

Table 4-7: Summary of error terms in conventional and proposed approaches 51

Table 4-8: Summary of model validation 51

LIST OF FIGURES

Figure 1-1 Energy need in developing countries 5

Figure 1-2: United States map showing all the current Industrial Assessment Centers 7

Figure 3-1: Regression model with two independent, one dependent variable $y = 2x_1 + x_2$ 27

Figure 3-2: Support Vector Regression underlying optimization model 30

Figure 3-3: A simple classification problem 31

Figure 3-4: Working of decision trees to classify the data points into region of interest . 32

Figure 4-1: Flow chart of common energy usage in a manufacturing plant 34

Figure 4-2: Regions of the USA as specified by Central Bureau 36

Figure 4-3: Subset division demonstration. 39

Figure 4-4: Flowchart of conventional regression modelling 40

Figure 4-5: Distribution of errors in linear and polynomial regression 41

Figure 4-6: Flowchart of proposed regression modelling 43

Figure 4-7: Flowchart of first level of division 45

Figure 4-8: Flowchart of the divisions after 2 levels of division 47

1. INTRODUCTION

1.1. BACKGROUND

In the year of 2014, the industrial sector consumed almost 154 exajoules of energy. This was approximately 36% of the global total final energy consumption (TFEC). It was also seen that industrial sector's TFEC increased by 1.3% [1]. It is possible to differentiate energy consumption based on the type of consumption namely buildings, travel, energy generation etc. Out of these, buildings are the highest single contributor to world energy consumption and greenhouse gas emissions [2]. Therefore, building energy consumption in the industrial sector is a major contributor to global energy consumption. The international energy agency (IEA) has laid out a 2DS pathway for global climate change mitigation. This pathway was laid out in consistence with IEA's effort of having at least a 50% chance of keeping global temperature increase lesser than 2 degree Celsius by 2100 [3]. To be on the 2DS pathway it is required that annual growth in energy consumption be limited to 1.2%. The current growth is 2.9%. So, even though production demands are increasing in major world markets, it is very essential that we reduce the current growth in energy consumption [4]. With the advent of energy efficiency studies, industrial sector is moving towards best energy and waste management practices. Process energy efficiency is highly concentrated upon. [1].

Following this rapid and unprecedented growth in the demand, supply and prices of energy services has always been volatile and unpredictable. This uncertain nature of the energy market would affect the industrial sector the most, because of the high consumption

[5]. There are several challenges in setting sustainability goals that will result in energy saving and environmental protection. It will require major changes in both demand and supply behavior of commercial and personal energy consumption [6]. Research in the field of energy efficiency should go beyond a mere strategic talk between stakeholders of public and private utilities. There needs to be an understanding from society to the scientific community and vice versa [6].

Several diagnoses in the environmental sustainability situation needs to be addressed by academic research. This will help in collecting important results for civil society and policy makes. To implement actions based on the results, any sort of seed funding that strengthens the collaboration between utility organizations, academic researchers and small-medium scale enterprises is highly helpful [6]. The development and implementation of energy efficient technologies offer a lot of environmental benefits like:

- Lesser harmful emissions
- Reduced Waste
- Economic benefits from reduced energy & other resource usage
- Improved recycling systems [7]

To be on track for one or more of the sustainability goals mentioned above, it requires a lot of planning for government, private utility organizations and policy makers. Energy prediction algorithms, both basic and more sophisticated have been used to get an appropriate figure of future global energy demand. With the advent of artificial intelligence, machine learning algorithms have been able to strengthen conventional energy prediction algorithms or replace them in the long run. Comparison studies have also

been made to check the relevance of each approach.

1.2. CURRENT SCENARIO AND CHALLENGES

Carbon emission tariffs have been constantly increasing in the industrial sector. To reduce the tariffs paid and to fight the rising energy costs, energy usage efficiency is very important [7]. Due to recent rapid climate change, several approaches are discussed worldwide with regards to sustainability target. There are several new policies that are being discussed to replace current policies and achieve the sustainability targets for the century in a more efficient way. Increasing interest in this area has led to developing more ambitious energy efficiency targets by the end of 2050 [8]. To make these targets possible, goals must be assessed accurately. Providing efficient and reliable tools are becoming challenging. The number of tools has increased, providing system and procedure specific goals for sustainability [9]. The following are the five main energy challenges [10]:

- **Increasing energy consumption:** After the industrial revolution in the 1800s and 1900s ignited the trend of rapid growth in global energy demand. Almost every future energy scenario predicts global energy consumption to continue rising, though every research has highly variable numbers regarding the same. Based on several literature it is safe to assume that the average increase in world energy would be by a factor of 3 by the end of 21st century. The range suggests that it might be anywhere from 2.5 to 5.5 [11]. There are several barriers to implement energy efficient technologies, even though they are found to have very low or even negative costs. From our experience with the Industrial center at Clemson University we have observed a few of those barriers. These barriers pose a lot of

threat to the decision-making leadership's positive intent on energy projects:

- High capital investment and longer payback periods
- Long term payback periods are bigger risks to SMEs.
- Inability to overhaul systems by stopping production
- Management's assumption that energy projects are secondary to the basic production targets and income of the firm.
- High labor rates (in case of companies with no in house electricians)

- **Lack of energy access:** Availability of energy is not constant across different parts of the world. It is seen that the lower 75% of the economic population uses only 10% of the total global energy consumption. The global number of people lacking proper energy access is 1.5 billion and number of people lacking access to modern energy appliances is 3 billion [12]. Most rural families in developing countries still use more conventional methods for their cooking, heating/cooling needs.
- **Climate change:** As discussed in the background section of this chapter, the annual increase in energy consumption needs to be reduced by a half. A major reason of climate change is greenhouse gas emissions. It is seen that energy production and consumption is the single largest contributor to greenhouse emissions, with its share increasing year after year. This is expected to grow faster by the end of 21st century if there are no strong policy changes [14]. This is in sync with the International Energy Agency's predictions discussed earlier in this chapter. The global mean surface temperature is expected to increase by 4 to 5° C compared to pre-industrial levels [15].

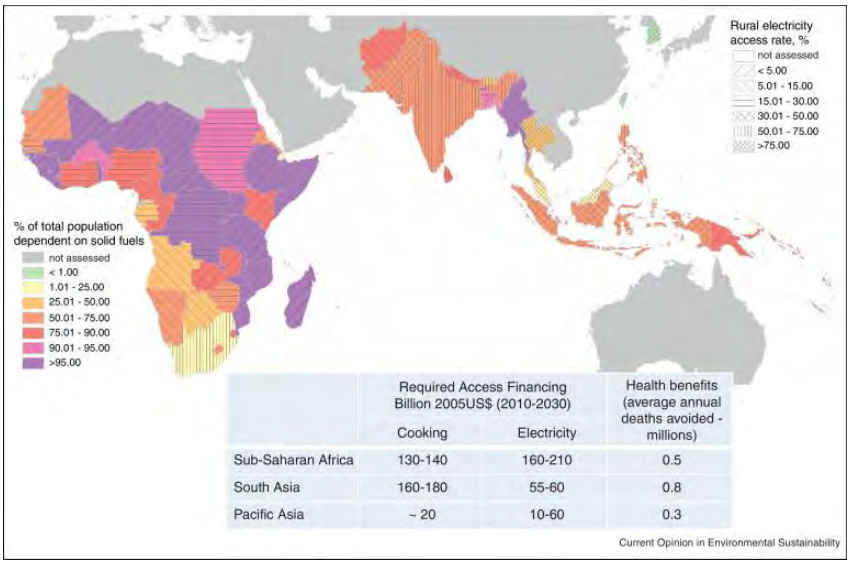

Figure 1-1 Energy need in developing countries [10, 13]

- **Land and water systems:** In addition to energy efficiency measures, there is a growing need for alternate systems of energy to mitigate the global energy loss in energy production. There can be impact on water and energy systems even with production of renewable or alternate energy. In case of bio energy, the production may result in scarcity of land for other forms of production like food and biodiversity preservation. Bio energy production, especially using special type of crops require large areas of land and high levels of clean water usage [16]. Several issues with bio-fuels indirectly contributing to greenhouse emissions have been reported in previous literature. Bio-fuel production uses a lot of nitrogen fertilizers and fuels. There is also an associated emission of CO_2 because of bio-fuel production associated land use pattern change [17]. Impacts arise from other alternate energy source as well, but they are mostly on the local side and balance out when comparing with current energy production scenarios.

- **Energy Security:** Energy security is the ability to provide all necessary forms of energy to countries. In a few cases it also means the availability of a certain percentage of necessary energy from renewable or alternate resources [18]. It is seen that in most emerging energy markets, energy security is not at the level that the government and other agencies would want it to be. There are several other vulnerabilities in these markets because of market uncertainties. Some issues are:
    - Insufficient and inefficient storage mechanisms
    - High energy intensity usage
    - Rapid growth in demand

All the literature mentioned above suggest the application of energy efficiency measures in addition to changing to alternative sources of energy and also changing energy usage patterns. Several governmental and non-governmental agencies have been involved in such efforts. The United States Department of Energy (DOE) is the governing body that deals with policies regarding energy management and nuclear fuel. For the past few decades, DOE has been working with research and academic research organizations to set up energy efficiency measures across sectors and all scales of companies. The DOE hosts a lot of program offices to deal with various issues in the energy sector. The program office of Energy Efficiency and Renewable Energy (EERE) is focused on the strengthening and accelerating the growth in various sectors employing energy efficiency measures in their facilities, offices and real estates. This office was found in 1981 and has hence been working on these measures. Another important effort by the DOE to implement energy efficiency measures is the Industrial Assessment Center (IAC) program.

1.3. OVERVIEW OF INDUSTRIAL ASSESSMENT CENTER PROGRAM

The Industrial Assessment Center (IAC) (formerly called the Energy Analysis and Diagnostic Centers) was established by the Department of commerce to fight the oil embargo and rising energy costs. In 1978, the program was moved to the DOE as an effort to incorporate energy efficiency measures in small and medium scale companies. The first center was started in 1976. Since then the DOE has been reviving, adding or removing centers from the program based on performance and necessity in the region. Currently there are 28 active centers that work within their stipulated area to implement energy audits. IACs have thus far conducted 18,792 energy audits and recommended 142,000 actions, resulting in 4.5 billion US dollars savings for the companies concerned (data as of 06-26-2019). The database of all assessments is available on the IAC website [19].

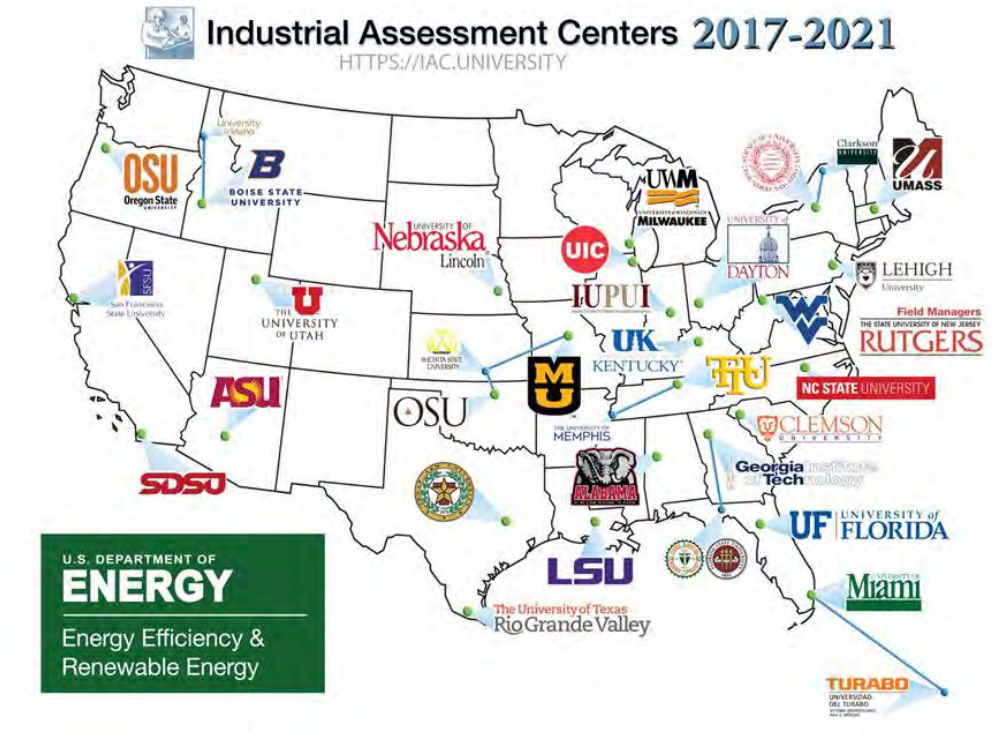

Figure 1-2: United States map showing all the current Industrial Assessment Centers

The Clemson Industrial Assessment Center was started in the year 2017 as a part of the program. Thus far, the center has conducted over 30 assessments in the Carolinas, resulting in an average of 10% energy costs reduction for clients. The center has dealt with a variety of manufacturers: automobile & allied manufacturing, other metal manufacturing, furniture manufacturing, waste water treatment plants, asphalts plants. One of the challenges faced by these companies in implementing energy efficiency measures is the uncertainty of future parameters based on which they should be making informed decisions. If there's a way to quantify future energy consumption of the future, it would be helpful for companies to take a decision based on future consumption patterns. This thesis will be focusing on one such case study from the Industrial Assessment Center database in addition to giving an extensive literature review on the algorithms that have been historically used for the same purpose.

1.4. MOTIVATION: NEED FOR ACCURATE ENERGY PREDICTION

As mentioned in the earlier part of this chapter, buildings account for a large percentage of global energy consumption and CO_2 emissions. Therefore, the prediction of building energy consumption is very important in improving their energy consumption. The physical process of energy consumption is quite complex because of high detail variation in building profiles and type of energy used for different purposes. It is very difficult to generalize the energy usage in any given building to consistently predict energy consumption since several factors like weather, occupancy & behavior, system independent energy using equipment such as lighting influence it to a very high level. Accurate energy consumption prediction of building is difficult due to the complexity of

the problem [20].

It is vital for the community to have required levels of dependable energy. With more and more countries developing economically the energy requirements will increase globally at a rapid pace. Utility organizations, mainly the large ones supplying to a large geographical area will be held responsible by consumers for proper maintenance and improvement of their service. To meet the expectation of customers, utility companies maintain databases of energy usage patterns and trends in industrial/domestic sector. Even after the advent of artificial intelligence, the most used methods today for forecasting using the databases are statistical regression approaches. That said, the rapid rise in popularity of data mining and other estimation approaches have encouraged the application of more sophisticated algorithms like decision trees and neural networks [21].

A majority of lifetime of people is spent inside residential or commercial buildings. On a global level, building energy performance is given a lot of importance by governmental, public and private utility companies. To implement building performance enhancements and to trace the improvements in terms of energy and cost savings, it is very important to have a robust energy prediction algorithm that can be applied to relevant available data. The future energy demand forecasting approach is currently of keen interest in the energy industry due to need of the hour: predicting peak energy demand which might in turn solve issues related to the electric distribution network. This approach can solve prime issues like discrepancies in planning for future energy requirements, amount and availability of sources in different points of time.

1.5. ORGANIZATION OF THESIS

This thesis is divided into six chapters and an introductory abstract. The abstract will briefly describe the reader the whole work that has gone into writing of this thesis. The first chapter discusses at length the background of the thesis. The chapter also speaks about the current challenges in energy industry, motivation behind energy prediction and a short overview of the IAC program. The second chapter discusses about different approaches taken to predict energy consumption in buildings. The chapter speaks about how each approach (regression, support vector machines, random forests, artificial neural networks) performed in predicting energy after giving a short introduction about each approach. The third chapter describes in detail the methods used in this thesis' case study from the IAC database. The chapter describes in depth the working of regression, support vector regression and random forests in prediction of parameters. The fourth chapter describes the dataset used for prediction, explaining each variable. The chapter also lists out the different equations used in the case study, while also explaining the application of methods from chapter 3 to this thesis' dataset of choice. The fifth chapter describes the results calculated out of the model built in chapter 4 and analyses the performance of the model. The chapter also discusses a few inferences from the result. The final chapter concludes the thesis, specifying the possible future work.

2. LITERATURE REVIEW

2.1. INTRODUCTION

The purpose of this chapter is to give a detailed analysis of the work that has been carried out in the energy prediction area. This chapter will consist of multiple topics, all of which are related to energy prediction algorithms and results that identify best algorithm. Best fit regression equations have been used for prediction and explanation historically [20]. This can be considered as a 'go to' method for deriving equations using past data and determining future variables of interest. Modern techniques of machine learning are replacing or working together with classical approaches like regression to better predict the future. Since this thesis aims at comparing these two approaches, this chapter will speak about past work on regression, machine learning algorithms like decision trees, support vector machines, artificial neural networks specifically in reference to energy prediction.

2.2. REGRESSION MODELS

The application of regression in predicting energy consumption was presented in [22, 23]. [22] states that the amount of energy utilized by buildings justifies the need for meticulous study and modelling in that area. The article further states that in developing energy models, statistical methods are a very convenient option when the user has access to only past data and not the several values that are required for engineering equations. Linear regression was found to be a relatively simple and reasonably accurate implementation when compared to other statistical methods. In this study hourly and daily data from a house dedicated to research was used. This difference in time duration gave the authors a chance to also study the relevance of the frequency of data collection in the

accuracy of the model. Independent variables were outdoor temperature and solar radiation. Energy consumption was the dependent variable. The following models were tested: simple & multiple linear regression, quadratic linear regression. This was done to see if the added complexity arising from the quadratic regression was justified by better quality of results.

It was found out that the time interval was a very important factor determining the quality of the model. The longer the time period, the better is the quality of the model. To explain this, the authors suggested that energy consumption anomalies show high discrepancies when data is collected over a shorter time period. When a longer duration of data was collected, the anomalies average out with each time period. The coefficient of determination was increased while using a multiple linear regression model. But in this approach, root mean square error (RMSE) got deteriorated. It was found out that multiple linear regression had the overall best quality of energy prediction when both these parameters were considered. Also, daily time intervals gave the best model parameters.

[23] presented a multiple regression model for the prediction of heating energy demand. The authors say that heating energy demand is a vital estimator. It is used in the design stage of building to estimate what amount of energy will be needed for space conditioning throughout its life. The independent variables considered were global energy loss coefficient of a building, south equivalent surface & temperature difference. [23] is a simpler model compared to [22] because it used only 3 variables and a single prediction model. [23] also mentions that in case of large datasets like the one used in this study, regression can be applied with good level of success, even though it is much simpler and

easier to build. This article used the principle of "black-box" to develop the model. This principle is used when input and output variables are known, and the user is required to fit in the best curve possible (also referred to as "black-box" due to the unknown nature) to establish a generalized relationship between independent and dependent variables. Though least squares estimate is the commonly used technique, it might sometimes result in errors not being normally distributed, which is a good validation of any curve fitting model [23]. Hence an iteratively reweighted least squared method was used. This method adjusts the weight of the coefficients in the regression equation to reduce the effect of residual outlier. This way of finding the best fit curve gives an improved least square estimate overall.

After training the model, it was tested on 17 blocks of flats. The model was found to be highly accurate with an R^2 value of 0.9744. It was also found out that 90% of the computed values had relative errors of 20% or under. [23] also compares the model to multiple dynamic solutions and states that the proposed model runs faster while giving comparable results. From [23], we learn that when generating models for datasets with smaller number of variables, regression is the best approach in terms of model quality and also speed. Since this thesis will also deal with smaller number of independent variables due to lack of complex data availability with SMCs, [23] was an important resource to establish how a regression model can be built.

2.3. MACHINE LEARNING INTRODUCTION

Machine learning can be defined as the use of computational methods trained by past data that help in making decisions related to a particular system. Generally, such tools are used to improve performances or making accurate predictions about the future.

Machine learning methods are currently being used in various fields as replacements or in association with classical regression/statistical models to predict future using past data [24]. For an extensive reading on the practical applications of machine learning, one might refer to [24]. A number of machine-learning algorithms, including but not limited to random forests [25], support vector machines [26, 27] and neural networks [28, 29, 30, 31, 32] are currently being used to predict energy consumption. There are several publications that suggest that machine learning algorithms, are at least on par with classical approaches. Such publications are discussed in detail in this chapter.

2.4. RANDOM FOREST MODELS

A comparison between the classical approach of linear regression and random forests was provided in [25]. Random forest is a classification and regression algorithm. It works by creating multiple decision trees based on the training of data [33]. That means, the output for a given set of input is based on the training. The algorithm can be used for both categorical and numerical output data. In case of numerical data (regression), the output from random forests is the mean of the values generated by several decision trees. In case of categorical data (classification), output is the mode of values generated by several decision trees. For e.g., if 7 out of 10 decision trees suggest that the output is 'Yes' and 3 suggest that the output is 'No', the output of the random forests would be 'Yes'. Decision trees are sought after machine learning algorithms due to their execution speeds. They also have straight forward training, which means that generalizing over a large dataset can be simple.

The effect of eight input variables on cooling and heating load of residential buildings was studied in [25]. The eight input variables were relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, glazing area distribution. The relation strength of each input variable to either of the output variables was statistically tested in order to determine the most related input. After that, based on the variables of interest, a classical linear approach was compared with random forest decision trees to identify the better prediction tool. 768 different residential buildings were chosen for running simulations on. Ecotect was used to determine the better of the two models. The parameter of interest was low mean absolute error deviation. For heating load the value was 0.51 and for cooling load it was 1.42. The results showed the possibility of using machine learning algorithms as a viable alternative to classical regression methods.

The problem with classical approaches is the necessity of data to be normalized [25]. Deviation from normality contributes to errors in addition to the errors arising from the model assumption. Whereas in machine learning algorithms, training using known data impacts how predictions are made. Normalization of the dataset doesn't become a default necessity as with classical approaches. In addition to iteratively training the model with multiple independent and dependent variable values, feedback is created after each iteration. This feedback modifies the model and increases the model's efficiency in making precise predictions [34].

2.5. SUPPORT VECTOR MACHINE MODELS

Another important machine learning algorithm, support vector machines (SVMs) was presented for performing energy usage prediction in [26, 27]. SVMs map the input

variables non-linearly to a multi-dimensional space. The output variable is constructed as a linear decision surface that extracts data from the space to predict new output variables based on the learning [35]. For e.g., in case of a model with 2 independent variables and 1 dependent variable, the independent variables are plotted against each other on a x-y graph. The dependent variable is a line that separates the data points into two parts. Based on which part the independent variables for the future belong to, the model predicts the dependent variable. The objective of [26] was to investigate the feasibility and applicability of SCMs in the field of building load forecasting. Random election was done on 4 buildings in the country of Singapore. In [26], 3 weather parameters, monthly average outdoor dry-bulb temperature, relative humidity and global solar radiation were taken as input variables. Data was collected from utility bills provided by landlords of 4 buildings in Singapore central business district. Average monthly bill amounts were considered to be the output feature.

SVMs were found to have coefficients of variance (CV) at about 3% and percentage error of about 4%. It is also mentioned in [26] that the increased accuracy of SVMs in load forecasting might be due to the fact that small data pools were used. Building energy data from 4 buildings on monthly basis was used for this research. This meant that not a lot of abnormalities were present in the data when compared to hourly or daily energy consumption. Also, SVMs work on the principle of structural risk minimization. While most machine learning algorithms try to reduce the training error in the models, the aforementioned principle works to minimize the upper bound of error [26]. It is seen that this might have resulted in the overall decreased error parameters. The authors also mention

that since SVMs have fewer parameters of interest that need to be optimized while building the model, accuracy might have been better.

Support vector regression was applied to model and predict Turkey's energy usage in [27]. The training set was relatively small, hence a ε-SVR model was used. This research is a bit difficult because building energy consumption was not modelled. The modelling was done on the overall energy consumption on Turkey and it was modelled as a function of socio-economic factors such as population, gross national product, imports and exports. Data used was the national energy consumption from 1975 to 2006. Prediction was done until 2026. This research, for the sake of simplicity, created multiple SVMs for every independent variable and later combined it into a wholistic SVM. Root mean squared error (RMSE) was the variable of interest in this research. The research proved that SVMs can be used for energy prediction with error rates around 3%. It is also seen from the [27] that RMSE values for training data is not included. Generally, RMSE values for training data is low compared to the testing data because the model is built on training data. Since the lower end of the values are left out of the average, the error calculated would be skewed towards the higher end. This means that even the upper end of the errors is acceptable, which gives an indication of the accuracy of SVMs. But based on the comprehensive review in [20], SVMs do not have ease of use and have low running speeds. Due to these reasons, applying SVMs to SMCs where data in hand is already low might not be a practical solution.

2.6. NEURAL NETWORK MODELS

Application of artificial neural networks (ANNs) in prediction of building energy

usage was discussed in [28, 29, 30, 31, 32]. [36] compared ANNs with simulation modelling. ANNs are machine learning algorithms inspired by the working of human body's central nervous system [37]. The human brain works in layers. A person's perception of a particular object or situation depends on how the several relevant experiences (layers) affected the person's thought process. Similarly, based on past data, it is possible to design an artificial neural network. This neural network will learn from different data points in the past as several layers. In addition to developing input-output relations, ANNs also learn from new outputs (similar to a human perception changing when relevant experiences change), making it easier to incorporate anomalies in energy consumption into the model [38]. [28] presented a backpropagation neural network to predict energy consumption based on three input variables. Backpropagation is the process of computing the errors in output and propagating it back to the network for altering the weights. Insulation, orientation angles and transparency ratios were taken as input variables to computed energy demand. Calculated values were compared with predicted values and it was found that ANNs give a prediction rate of about 94.8 – 98.5%. The deviation was also as low as 3.43%.

Adaptive ANNs, in which the neural networks can adapt to changes in the input data was presented in [29]. This research used both simulated and measured data to train and test the neural networks. This research presented two different approaches to adaptive ANNs: accumulated training & sliding window technique. In accumulated technique, when there is new data, it is added to the present dataset. The model retains all the data since the first data point. Sliding window technique is used for comparatively larger datasets where

accumulating window technique might take long durations to run. Older data is removed from the model as newer data is added. The indicator used was coefficient of variance. Both the approaches gave almost similar results when used on simulated data. But the sliding window technique gave better results with real measurements. This difference could be caused by the large amount of data with the former approach, making it obsolete with critical newer data.

[30] presented a neural network approach to predict building energy usage that was supported by statistical procedures. The statistical procedures were used to systematically treat both the relevance of input variables and number of free parameters whereas neural networks were used to predict the energy consumption based on data. It was found that neural networks supported by statistical tools improved the prediction accuracy. [31] used a multilayer perceptron model to predict Greek long-term energy prediction. There were 4 input variables used versus a single output desired. This ANNs had 5 hidden layers hence increasing the accuracy of the neural network. It was found out that ANNs performed very well in predicting energy consumption. The error rate was around 2% and the deviation from other prediction algorithms was around 1%. This literature is closely associated with the models discussed in the thesis because of the number of variables considered and the comparison between models. It is also seen from this research that ANNs are user friendly and changes to the model in case of anomalies is easier.

[32] presented the application of ANNs in energy prediction based on the consumption in a passive solar building. A fully insulated building and partly insulted building were checked in two seasons: winter and summer. The ANNs were trained using

simulated building data ranging from 15 cm to 60 cm thickness. It is seen that ANNs perform extraordinarily well when it comes to prediction energy consumption with building thickness as input. The coefficient of determination was close to 0.9991. This approach is a very good example of the superiority of ANNs in terms of model power and amount of data required.

[36] presented a comparison between ANNs and detailed model simulation based on physical principles. Energy plus was used to evaluate the influence of several input factors on energy consumption. It was found out that energy plus forecasts had an error range of about 13%. This error range was consistent with 80% of the database. In ANNs, the error range was about 10%, which is significantly lower than the former 13%. This shows that brute force calculations based on equations might not be as accurate and incorporation of historical data training using modern machine learning algorithms like ANNs improves the accuracy of prediction.

2.7. COMPARISON OF DIFFERENT APPROACHES

Although there is a lot of literature like the ones mentioned above, about different forms of building energy prediction, there is very limited literature that compare more than two or more algorithms to determine which one is best. Since this thesis will focus on comparing two approaches, we can benefit from such literature that establish a strong comparison principle. [20, 21] are examples of this type of study. [20] is a review paper that compares several approaches, broadly classifying them as elaborate or simple engineering methods, statistical methods and artificial intelligence methods. Elaborate engineering methods require large amounts of data that might not be readily available to

the users. These methods generally require data about the space conditioning, construction material about each and every room of a large building. This means that a lot of time and effort have to be invested in making this method a reality. Mostly, engineering methods are carried out by simulation modelling. But there is an area of concern in simulation modelling. Calibrating of inputs needs to be done carefully for matching the real time energy behavior of buildings as much as possible. After several steps of calibration, energy models show high accuracy. But calibration is a time-consuming process and hence can be tedious to the user. [39] reviewed two simplified forms of these engineering methods, degree day method and temperature frequency method. These methods could potentially save the time and effort.

Several literatures have been discussed previously in this chapter for demonstrating the application of regression energy modelling. Here we will discuss about the advantages and disadvantages as mentioned in [20, 21]. Regression works by training a model based on historical data to predict the future. Before training the model one needs to collect enough historical data. Since energy consumption patterns might always present anomalies that might not have happened in the past, the usage of regression in such cases is highly compromised. [21] suggests that due to the underlying principle of regression: establishing relationship, the user can never be sure of the underlying causal mechanism. But both [20, 21] support the simplicity of use in regression modelling.

Neural networks are suitable for solving problems of non-linear fashion and an effective approach to applications with complicated data structures. Neural networks can be most useful while trying to analyze or optimize sub-level components and related

behavior. They have been used in the past decade for analysis of HVAC systems. It was also seen that neural networks were very suitable in cases where the annual energy consumption had high fluctuation from the mean [20]. Neural networks can be used in applications where the mathematical relationship between several variables is unknown to the user. Lesser prior knowledge suggests the usage of neural networks. But the disadvantage with using neural networks is it does not describe the model as well as a regression model would. For eg, p-values of different variables from regression can be used to establish the relationship. But neural networks give only the expected output for the relevant input. Neural networks are hence used more as classifiers than as regressors [21].

SVMs are most useful in solving non-linear relationships that have a very small quantity of data to train the model with. It was proven that SVMs can perform well on monthly and yearly building energy. SVMs showed the best performance of prediction tools in on annual energy consumption of buildings and hence can be used to predict data of that sort. The disadvantages of SVMs match that of regression modelling where in establishing relationship is not enough to provide proof of causality [21].

Decision trees have been used in the past for segmentation of the data by applying a simple set of rules. The purpose of segmentation is to divide the dataset into smaller subsets and build more efficient prediction models for each subset. The most common methods used in decision trees are chi-squared automatic interaction detection (CHAID), classification and Classification and regression trees (CART). The major advantage of decision trees is that it [produces results based on easily understandable logic. So, it is possible to explain to plant personnel about how the prediction algorithm works. The main

disadvantages of decision trees is that they do not perform well with non-linear data. They are more susceptible to errors because of noise in the dataset. In general, random forests are more suitable for predicting categorical variables [21].

Table 2-1: Advantages and disadvantages of each method

MODEL COMPLEXITY	EASE OF USE	COMPUTATIONAL SPEED	AMOUNT OF DATA REQUIRED	ACCURACY
Simplified Engineering Methods	Regression Analysis	Artificial Neural Networks	Elaborate Engineering Methods	Artificial Neural Networks
Artificial Neural Networks	Simplified Engineering Methods	Simplified Engineering Methods	Support Vector Machines	Simplified Engineering Methods
Support Vector Machines	Elaborate Engineering Methods	Regression Analysis	Simplified Engineering Methods	Elaborate Engineering Methods
Elaborate Engineering Methods	Artificial Neural Networks	Support Vector Machines	Artificial Neural Networks	Support Vector Machines
Regression Analysis	Support Vector Machines	Elaborate Engineering Methods	Regression Analysis	Regression Analysis

2.8. SEARCHES AND DATABASES

Searches were conducted on databases on the application of various statistical, engineering and machine learning models on the prediction of energy consumption. Searches were also conducted on several databases maintained by the Department of Energy, International energy Agency about the societal, climatic and geographical changes because of rise in energy consumption. Machine learning models and the underlying principles were studied to be implemented on our dataset. To understand the relative efficiency of different models, searched were conducted to find literature that compare

different methods and list the best methods in several ways.

The databases that were searched include the following journals: Renewable and Sustainable Energy Reviews, Energy and Buildings, Energy Simulation in Building Design, Building and Environment, Machine Learning, Advances in Engineering Software.

3. METHODS

In this thesis we present a method combining both machine learning algorithms and statistical regression modelling for energy consumption forecasting. This thesis will specifically focus on data collected by several Industrial Assessment Centers throughout the country. The Industrial Assessment Center deals with industrial energy assessments in small and medium scale companies. Based on our interaction with plant personnel in several assessments, the challenge with reducing energy consumption in these companies is not being able to follow trends in usage. This is mainly due to the small amount of data available to the plant managers for analysis. [20] states that the energy consumption in buildings is based on two groups of parameters:

1. Physical Environmental parameters: Outside temperature, amount of solar radiation, humidity, wind speed etc.,
2. Artificial designing parameters: Transparency ratio, type of lighting used, number of units produced, orientation, material properties etc.,

It came to our notice that companies do not have the means or work force to measure such a wide variety of data. So, the necessity to identify a prediction tool with high accuracy while dealing with datasets containing independent variables that are measured over a longer period becomes vital. The IAC database has yearly data of several variables (explained in the model section) for public use. This research has tried to fit in a prediction model for the dataset to predict energy consumption of small and medium scale manufacturers. Below is a brief theoretical explanation of all the models used in this thesis. The equations used, and parameters selected will be discussed in the models section.

3.1. REGRESSION

Regression is a technique used for analyzing the impact of change in one or many variables on the change of another variable and it's used in variety of science and engineering disciplines for the same [21]. A simple linear regression is used to analyze the relationship between bivariate data i.e., the impact of one variable change on another variable change. Multiple regression is used for analyzing the impact of change in multiple variables on the change of one variable. The variable(s) which impact(s) another is called predictor variable (generally denoted by Xs) and the variable which is impacted is called response variable (generally denoted by Y) [40]. In this thesis, we will be dealing with multiple regression since our dataset involves multiple predictor variables and a single response variable.

One of the most common methods of establishing a correlation between variables is determining a "best fit" regression equation to go through the dataset consisting of multiple independent variables and a dependent variable. The best fitting equation to a set of data points would be the equation that is closest to all or most of the data points [40]. This can be achieved by determining the equation with least total vertical distance from the data points. This vertical distance can be defined as the "random error" generated due to the generalization of the whole dataset into an equation. Random error can be both positives and negatives, resulting in several values cancelling out or affecting other values. To avoid this discrepancy, the random errors are squared and summed. This resulting value is called "sum of squared errors" [40]. The goal of regression is to minimize this value "sum of squared errors."

Essentially, the process of determining the best fit equation is minimizing the sum of squared errors of the equation from the actual values. The equation with the minimum sum of squared errors shall be declared the regression equation for the dataset. By plugging in the values of expected independent variables in the future, it is possible to predict the value of dependent variable. This is one form of energy prediction that will be dealt in this thesis. A common example of a regression equation is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k$$

y is the independent variable, x are dependent variables, β_0 is the intercept and all other β are coefficients of different variables

Below are the estimates for least square functions [40]:

$$L = \sum_{i=1}^{n} (\in i)2$$

L is the least squares function. $\in_i$ is the distance of the fitted line from point "i"

$$L = \sum_{i=1}^{n} (yi - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij})^2$$

y_i is the original value of y at i.

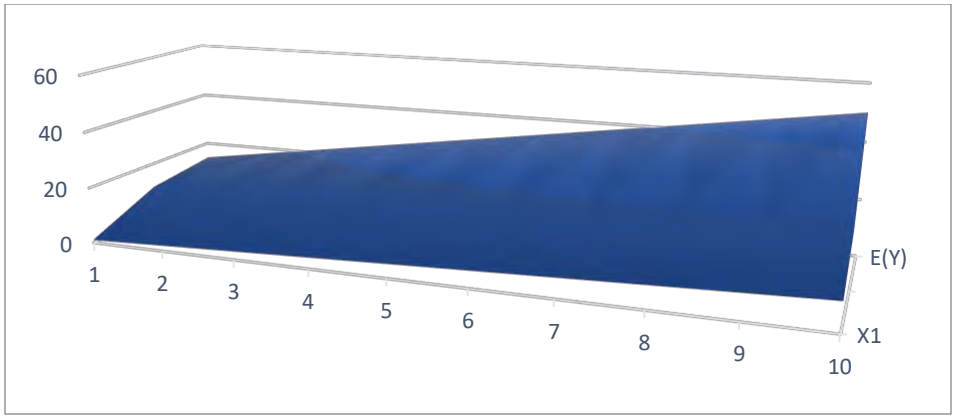

Figure 3-1: Regression model with two independent, one dependent variable $y = 2x_1 + x_2$

A machine learning algorithm, random forests, is another form of energy prediction that is of interest in thesis.

3.2. MACHINE LEARNING

Machine learning can be defined as computational methods used to improve performance metrics and predict future scenarios. Any machine learning algorithm uses data from the past to perform these activities. 'Data' includes any information available electronically to the user which one might use to train his algorithm [41]. The most important attributes of data used in machine learning are quality and size, both contributing to the level of learning and accuracy of the prediction. Machine learning has been used in several fields of study over the past few decades starting from weather prediction, stock market analysis. Machine learning can be used for one of the following standard tasks [41]:

- Classification: Classification tasks deal with assigning each element to a particular category. These tasks are used when there are multiple variables that decide what qualitative variable will be the output. For e.g., in electricity bills, demand charges are calculated only for manufacturing facilities. So, based on electricity bills from the past, a machine learning algorithm can classify the customer as a manufacturing facility if the customer is paying for demand charges. In classification problems, there is no real measure of how far an output is from the real-world scenario. There is only a pass or fail result of the machine learning algorithm [41].
- Regression: Regression in machine learning denotes any prediction in which a real value is the output. Machine learning does not use conventional regression formulas, regression here just denotes that a real value is predicted using past or

known data. For e.g., predicting a future electricity bill using past and present bills, consumption data. This thesis will focus on regression using machine learning algorithms. In regression problems, there is a measure for how far an output is from the real-world scenario. This can be useful in calculating a penalty proportional to the measure. Similar to how a regression equation will change when new data points are added, even a machine learning algorithm will keep learning from new data points. This makes accommodating new data easy [41].

- Ranking: Ranking tasks include arranging items according to the users' requirements, priority, deadlines etc. Common examples are web searches that arrange search results according to a keyword, calendar applications that arrange one's tasks according to time, natural language processing systems that rank words according to relevance etc. [41].
- Clustering: Clustering problems involve grouping of a set of items on a dataset together based on homogeneity. Clustering can be done based on area, gender, sex or any common factor. A lot of clustering problems are on social media: clustering profiles to help users identify other people within a particular organization or community. Like classification problems, there is no real measure of how wrong these algorithms might be. The output can only be classified as a pass or fail based on real world scenarios [41].
- Dimensionality reduction, also known as manifold learning: These problems involve transforming a preliminary representation of items into a lower order representation. These algorithms also preserve the properties of the preliminary

representation. One can explain it by comparing it to digital copies of older version of cameras [41].

3.3. SUPPORT VECTOR REGRESSION

Support vector regression is a form of machine learning regression. Support vector regression has been used extensively in the past for energy prediction. Support vector machines work on the principle of structural risk minimization [26]. As discussed earlier, statistical regression works on the principle of least squares estimate. On the other hand, support vector regression works on the principle of least maximum error threshold. This means that the algorithm works to keep the maximum error from a data point within a particular range or desired value [42]. Below is a simple illustration of a support vector regression based on the principles mentioned on [42].

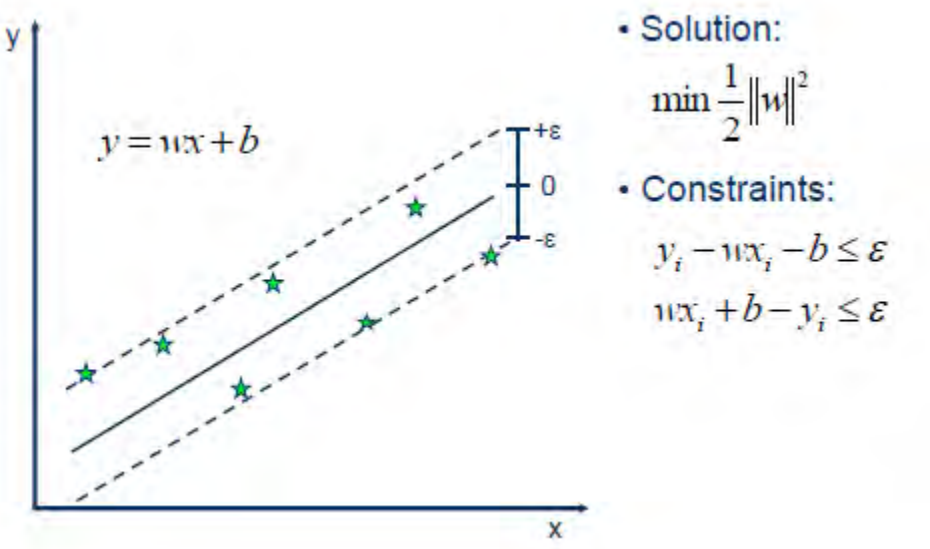

Figure 3-2: Support Vector Regression underlying optimization model

3.4. RANDOM FORESTS

Random forests fall into the category of classification machine learning algorithms.

Random forests have also been used in a few cases for quantitative output determination but will not be discussed in this thesis. The basic output behavior of random forests was discussed in the second chapter. This chapter will discuss the underlying working mechanism and explain why a particular feature called *wrapper aspect* would be useful in our case study. Random forests are a combination of decision trees, with each tree depending on the values of a random vector of the training dataset sampled independently. The forest then infers value from the several decision trees that are part of it. The efficiency of a random forests depends upon the strength of each tree that is part of the forest [43]. A simple illustration of a decision tree is shown below:

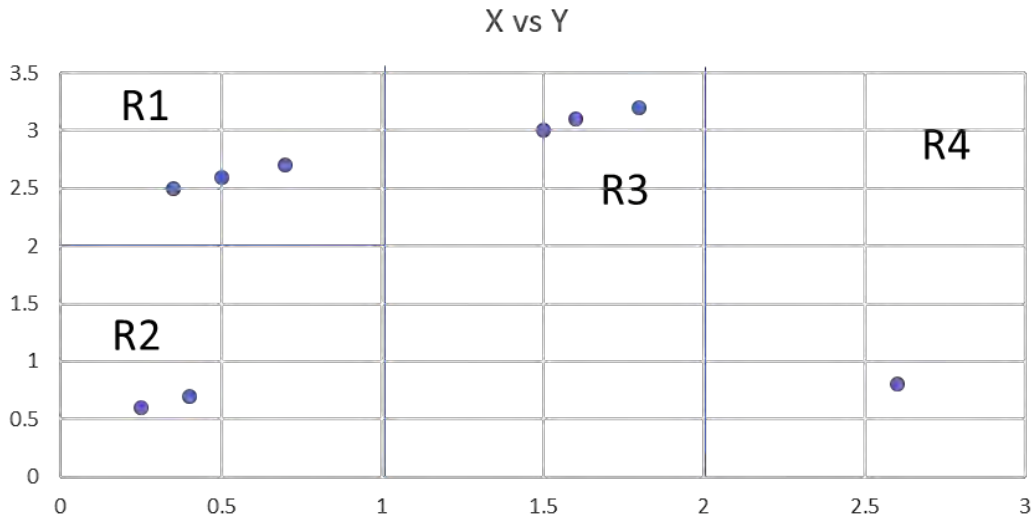

Figure 3-3: A simple classification problem

From the illustrated example below, it is seen that the decision tree is able to group the data points of interest by first selecting the variable that has the potential to make the biggest possible classification. In the case of the illustrated example, the condition "Is X > 2" has the potential to directly assign data points into R4. No other condition can directly

contribute a first order classification without further checking for another condition. In other words, the algorithm tries to create the shortest decision tree possible to complete the classification problem. This means that the major classifying feature is selected to be the first feature of importance, in this case, X. It is seen that out of the 3 decisions taken by decision tree, 2 decisions are based on the feature X, further adding value to the premise that X is the first feature of importance for this particular problem. While trying to fit a prediction model on dataset with several independent features that have a high range, it is important to group the data points into several subsets and then try to fit in a prediction model. This grouping will make sure that the skewness in the entire dataset will not affect the prediction model results. Random forest classification will give us an output that says what the important features are in a dataset, along with specifying the percentage importance. Based on the percentage importance, the user can divide the primary dataset into smaller localized datasets of a variable. Localized datasets are expected to give better fitting models due to lesser skewness.

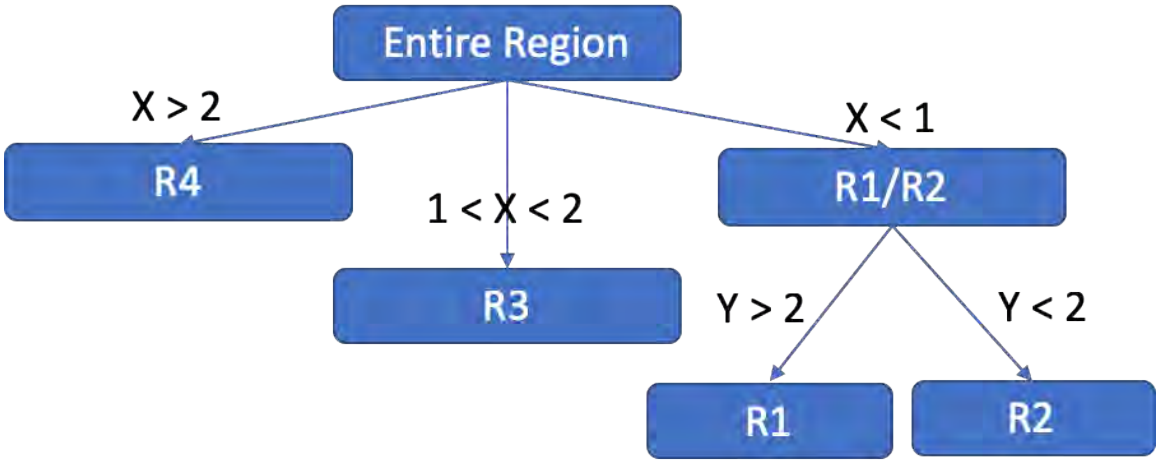

Figure 3-4: Working of decision trees to classify the data points into region of interest

It is to be noted that the random forest algorithm does not assess the variable importance individually as done by *f* or *p*-test in regression modelling. The variables are assessed jointly assessed for all the variables that are part of the dataset. Then the variables are ranked based on the strength on association with the output variables. This feature of random forests is called 'wrapper aspect' [25]. In this study we use this aspect of random forests to select the input variables to be used as the basis of classification for the regression models. The random forests have been used to find the features of importance ranked one by one so that grouping of the data points can be done based on those features. Each level of division has been done based on a particular variable. The top most classification has been done on the variable of most importance and so on. A detailed flowchart of how the division has been done is presented in the next chapter in this thesis.

4. MODEL

The first step in establishing a prediction model is determining how a model is going to predict and what a model is going to predict. As mentioned in the methodology section, regression makes use of single or multiple independent variables to predict the value of a dependent variable. So, the first step would be establishing variables for the regression models. From [25], it is seen that the three common electricity consuming equipment in a manufacturing plant irrespective of its size is:

1. Lighting equipment
2. Motor equipment
3. Heating/Cooling Equipment

But with our experience at the energy audits conducted by Clemson IAC, we have seen that motors are not the only equipment involved in production. To simplify the whole model, the major electricity consuming equipment mentioned in [25] are changed to suit our needs as follows:

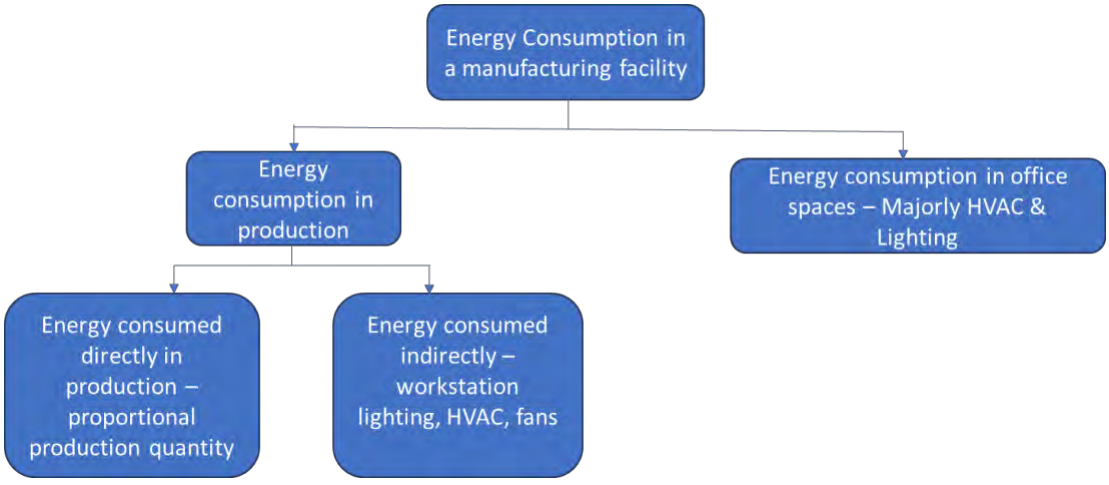

Figure 4-1: Flow chart of common energy usage in a manufacturing plant

4.1. VARIABLE DESCRIPTION

The flowchart above summarizes the various causes of energy consumption. The variables below capture those causes to an extent. Energy consumed in production is captured by the annual sales and production hours. The energy consumed in space conditioning is captured by the plant area, number of employees and outside temperature. To account for the deviation in energy consumption due to the region from which the data was taken, 4 binary variables have been taken. These binary variables will create one combination for each region and will convert the categorical variable region to numerical variables to be used in the model.

Table 4-1: Description of different variables used in the model

DESCRIPTION	VARIABLE
Region	x_1, x_2, x_3, x_4
Average Temperature (F)	x_5
Sales ($)	x_6
Number of Employees	x_7
Plant Area (Sq. Ft)	x_8
Annual Production Hours	x_9
Energy Usage (kWh)	y

The United States of America is divided into 4 major regions by Census Bureau. The designated regions are:

- Northeast: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, New Jersey, New York, Pennsylvania.
- Midwest: Illinois, Indiana, Michigan, Ohio, Wisconsin, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota.

- South: Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, District of Columbia, West Virginia, Alabama, Kentucky, Mississippi, Tennessee, Arkansas, Louisiana, Oklahoma, Texas.
- West: Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming, Alaska, California, Hawaii, Oregon, Washington.

Figure 4-2: Regions of the USA as specified by Central Bureau

x_1, x_2, x_3, x_4 are binary variables used to incorporate the categorical data with 4 levels, as mentioned above in numerical models. The average temperature x_5 in F is the average temperature measured throughout the year of assessment at the county of the participating university. The IAC database does not contain the exact location of clients. Since the companies should be within 150 miles radius of the participating university, the average temperature from the university's county is considered approximate. The sales x_6 in $ is the amount of gross sales done by the company throughout the year. This represents

the amount of product delivered by the company. y in kWh represents the total energy consumed by the plant throughout the year. Another feature of interest could be the demand consumption in kW which is the peak demand by the company every month. But peak demand will not be dealt with in this research.

4.2. DATA COLLECTION & PRE-PROCESSING

The Department of Energy (DOE) saves data from every assessment made by past and present Industrial Assessment Centers (IAC) throughout the country. It is available for public use. The database can be found at IAC's website https://iac.university/. The database contains data of over 18000 assessments by 57 past and present centers. The data range from 1981 to 2019. For the purpose of this research, data over a five-year period, 2014 – 2018 was used. For having an assessment conducted by an IAC, companies must fall within the following guidelines:

- Within Standard Industrial Codes (SIC) or The North American Industry Classification System (NAICS) codes.
- Located less than 150 miles of a participating university
- Gross annual sales below $100 million
- Fewer than 500 employees at the plant site
- Annual energy bills more than $100,000 and less than $2.5 million
- No professional in-house staff to perform the assessment

This research has also concentrated on companies following all the above criteria. The DOE does offer waiver on some of these regulations for IACs to be able to help customers in the university's locality. The data from such companies do not fit into our

model for creating a predictive tool to be used in small and medium scale manufacturing. So, a 3-standard deviation from the mean for each variable was considered the primary dataset for this research. All data points with blank entries were removed from the dataset. Data loss because of this process was about 2%. The variables from the IAC database was used to build this model. Data with NAICS codes 23, 31-33 (manufacturing firms) were only used. Total number of datapoints was 752.

Table 4-2: List of variables with their respective upper and lower bounds

VARIABLE	AVERAGE	STANDARD DEVIATION	UPPER BOUND	LOWER BOUND
x_5	57.02	8.41	82.25	31.80
x_6	22915458.48	21694248.12	87998202.84	0
x_7	89.04	70.13	299.43	0
x_8	103477.30	134365.32	506573.25	0
x_9	4197.29	1814.41	9640.53	0

The temperature data for each IAC county was derived from National Oceanic and Atmospheric Administration's National Centers for Environmental Information webpage.

4.3. CLASSIFICATION APPROACH

The model presents a combined approach of machine learning and regression modelling to build an energy prediction model for small and medium scale companies' data from the IAC database. It was seen from the dataset that all the variables had a very high range. It is difficult to fit a prediction model to fit into such a dataset without having one extreme of a particular variable skewing the other extreme. So, to make a substantial prediction model, the best way would be to split the variables in to bins based on one or few variables and then try to fit a prediction model into the subset based on the remaining

variables. To implement this approach this research has used the wrapper aspect of random forests to first identify the variables of importance. Then these variables shall be used one after the other to split the dataset. Once a prediction model with reasonable accuracy is achieved or not enough data points per group are available, the splitting will be stopped.

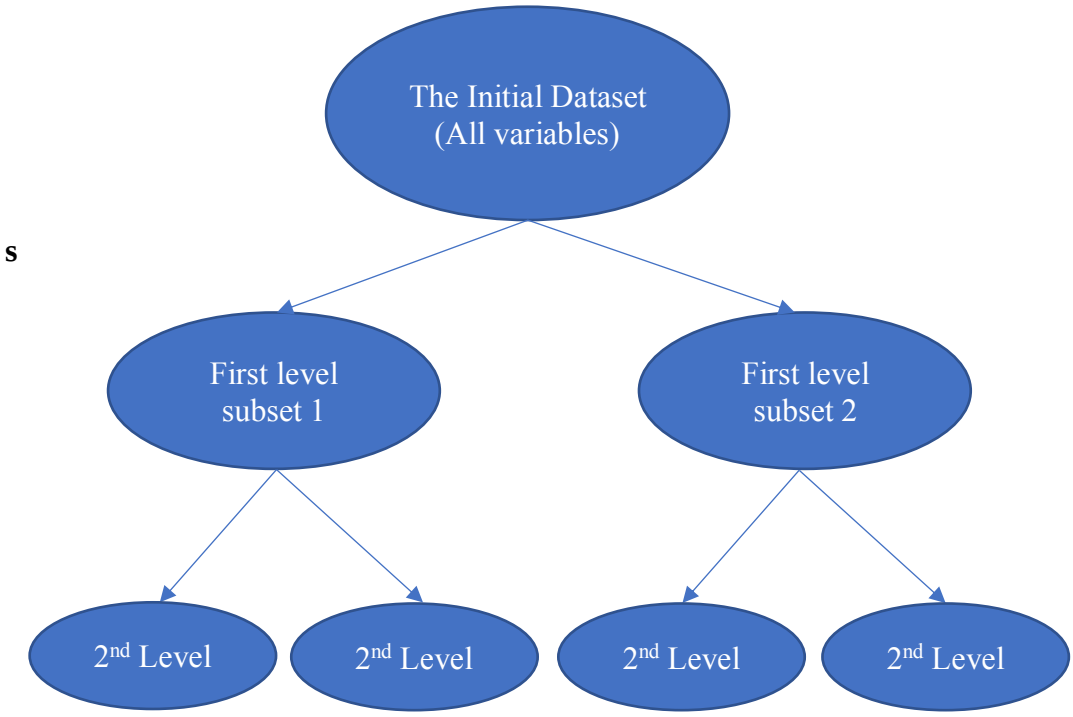

Figure 4-3: Subset division demonstration.

The basic idea behind this model is instead of fitting a large-scale prediction model on the entire dataset, divide the dataset into homogenous groups and then implement prediction algorithms, in this case linear regression, polynomial regression and support vector regression on those datasets.

4.4. CONVENTIONAL REGRESSION MODEL

Based on literature, it is seen that historically regression modelling involves the fitting of best fitting curves (in terms of least weighted or iteratively least weighted squares)

into the dataset and validating the model using R^2, RMSE or percentage deviation values.

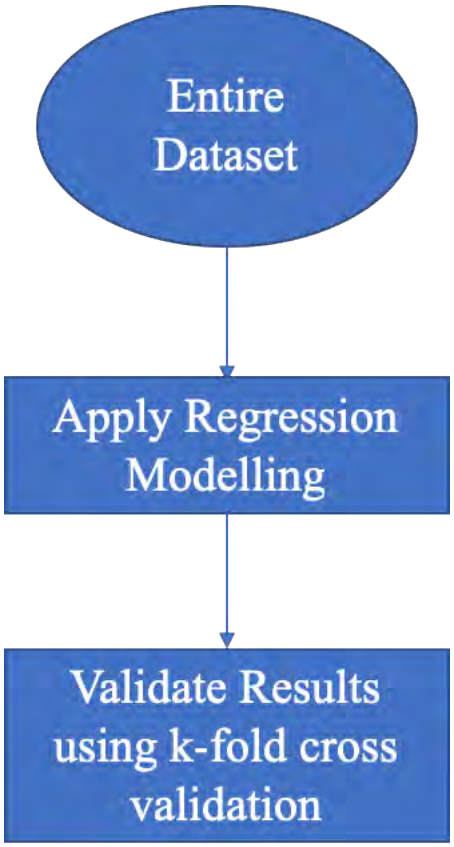

Figure 4-4: Flowchart of conventional regression modelling

The first step in regression modelling would be seeing if the dataset follows the assumption that residuals are normally distributed [45]. There are several told to check for the normality of a residuals. In this thesis, kurtosis and skewness of the residuals have been checked to assure normality of the residuals.

Table 4-3: Kurtosis and Skewness values for linear and polynomial regression

ITEM	SKEWNESS	KURTOSIS
Linear Regression	0.32	-0.31
Polynomial Regression	-0.375	-0.3

Kurtosis value between 2 & -2 acceptable. Skewness is between -0.5 & 0.5 the

distribution is approximately normal [46].

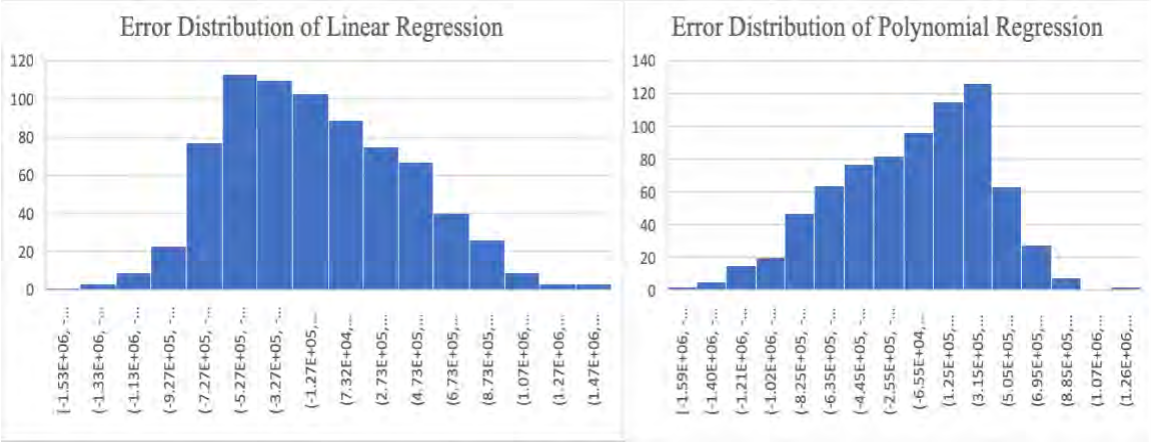

Figure 4-5: Distribution of errors in linear and polynomial regression

The next step in regression modelling would be to find out individual variable importance by performing a hypothesis test. For this purpose, a T test was conducted. The test statistic for this test is [46]:

$$T_0 = \frac{\beta_i - \beta_0}{\sqrt{Cjj\sigma 2}}$$

Null hypothesis: (H_0): $\beta_i = 0$, variable x_i is not significant, for every i = 1, 2….9

Alternate hypothesis: (H_1): $\beta_i \neq 0$, variable x_i is significant, for every i = 1, 2….9

For this test, considering a 5% significance, the rejection region would be $T_{0.025}$ (T test is a two tailed test). If the absolute value of T_0 is greater than $T_{0.025}$, the null hypothesis has to be rejected.

Table 4-4: Test statistics and statistical significance of each variable

VARIABLE	TEST STATISTIC	$T_{0.025}$	SIGNIFICANCE
1	-1	1.96	No
2	1.4	1.96	No
3	1.5	1.96	No
4	1.2	1.96	No
5	0.6	1.96	No

6	3.9	1.96	Yes
7	5.7	1.96	Yes
8	5.3	1.96	Yes
9	9.6	1.96	Yes

It is seen from the table that first five variables (binary variables denoting region and temperature variable) are insignificant in this model. But the presence of those variables might have affected the significance of other variables. Also, past research has shown that the ambient temperature is indeed an important factor in energy consumption. So, the binary variables were removed from the dataset and the remaining variables were checked for statistical significance. Following is the second hypothesis test.

Null hypothesis: (H_0): $\beta_i = 0$, variable x_i is not significant, for every i = 5,….9

Alternate hypothesis: (H_1): $\beta_i \neq 0$, variable x_i is significant, for every i = 5,….9

For this test, considering a 5% significance, the rejection region would be $T_{0.025}$ (T test is a two tailed test). If the absolute value of T_0 is greater than $T_{0.025}$, the null hypothesis has to be rejected.

Table 4-5: Test statistics and statistical significance of each variable

VARIABLE	TEST STATISTIC	$T_{0.025}$	SIGNIFICANCE
5	7.3	1.96	Yes
6	4.2	1.96	Yes
7	5.2	1.96	Yes
8	6.2	1.96	Yes
9	11.7	1.96	Yes

From the above hypothesis, it is seen that all the variables are significant in this case. The presence of additional binary variables that were insignificant affected the model's capacity to identify that temperature was significant. Now it is seen that the temperature is also a significant variable. Therefore, all the five variables from the above

table shall be used for running the conventional regression model. This gives the model improved accuracy as opposed to a linear regression model that tries to incorporate all the variables in a dataset.

4.5. PROPOSED REGRESSION MODEL WITH RANDOM FORESTS

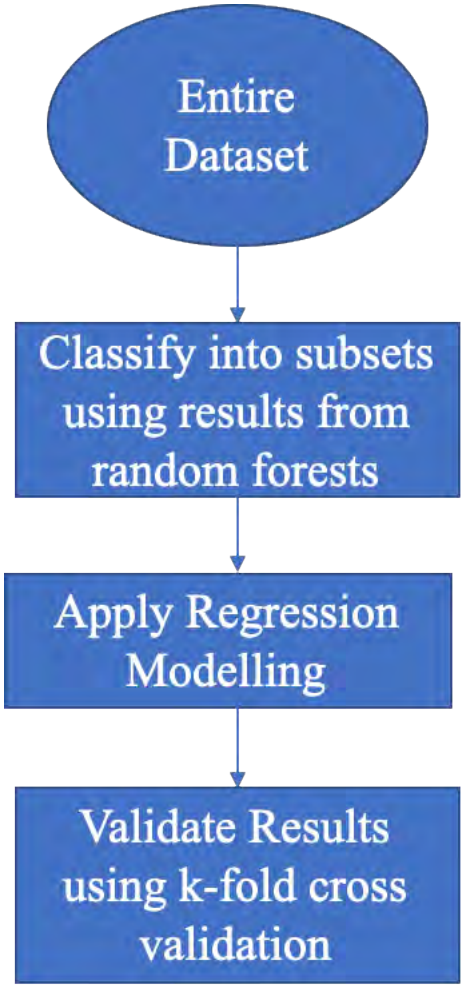

Figure 4-6: Flowchart of proposed regression modelling

First, the random forest algorithm was run on the complete dataset to check for the most relevant variables. These most relevant variables will later be used to split the dataset into homogenous groups. The results were as follows:

Table 4-6: Variable importance based on random forests

VARIABLE	% IMPORTANCE
x_1, x_2, x_3, x_4	0
x_5	14.4
x_6	19.2
x_7	18
x8	21.9
x9	26.5

It is seen that the total importance of the variables sums up to 100. It is also seen that the importance of region (denoted via the four binary variables having no importance). The algorithm gives the names of the top two important variables: plant area and production hours in this case. The least important variable is 14% important, which is the average temperature. In a residential setting, the major consumer of energy is HVAC. It accounts for up to 50%, but HVAC is only 20% of the total energy consumption in the USA [45]. This shows that in the industrial sector, which is another major energy consumer, HVAC systems account for lesser of the energy consumption than that of the residential sector. So, the dependency of energy consumption on the outside average temperature (which affects HVAC consumption) is reduced. That being said, 14% contribution to the energy consumption is still done by average temperature. So, the variable is still incorporated in the models. The first level of classification is done based on most significant variable, that is production hours. The levels of classification will be

- Companies with one shift a day (lesser than 2960 working hours per year)
- Companies with two shifts a day (more than 2960 hours but lesser than 5840 hours per year)
- Companies with three shifts a day (more than 5840 hours per year).

This division was made after conversations with plant personnel during our IAC visits and all plants work on an 8-hour shift, with one, two or three shifts a day.

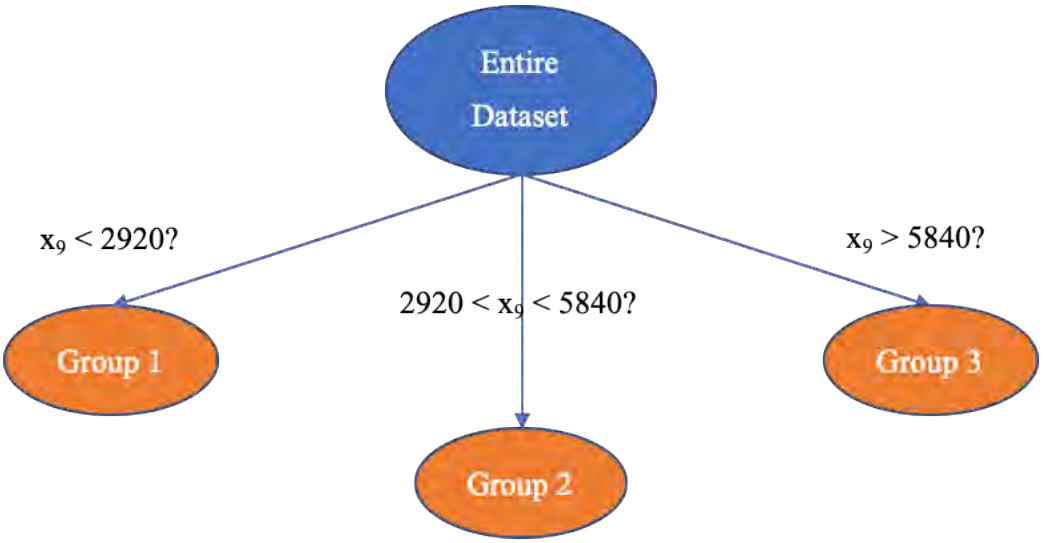

Figure 4-7: Flowchart of first level of division

For the first level of split, it is seen that there was a considerable percentage deviation for all the group. This suggests scope for improvement on the model, particularly using further divisions such as the second level of division using random forests. The second division shall be based on the second variable of importance. The second variable of importance is plant area. This division is done into the following categories.

- Plants with area lesser than 50,000 sq.ft
- Plants with area lesser than 100,000 sq.ft but greater than 50,000 sq.ft
- Plants with area greater than 100,000 sq.ft.

The division is represented via a flowchart in the next page. When the split in this level was made, first 2 categories were tried, and the deviation reduced substantially from the previous test with one level of classification. When the classification was further

extended to three categories, the improvement in deviation was very less, of the order 0.5 to 1%. Hence, further division might not necessarily mean a better deviation.

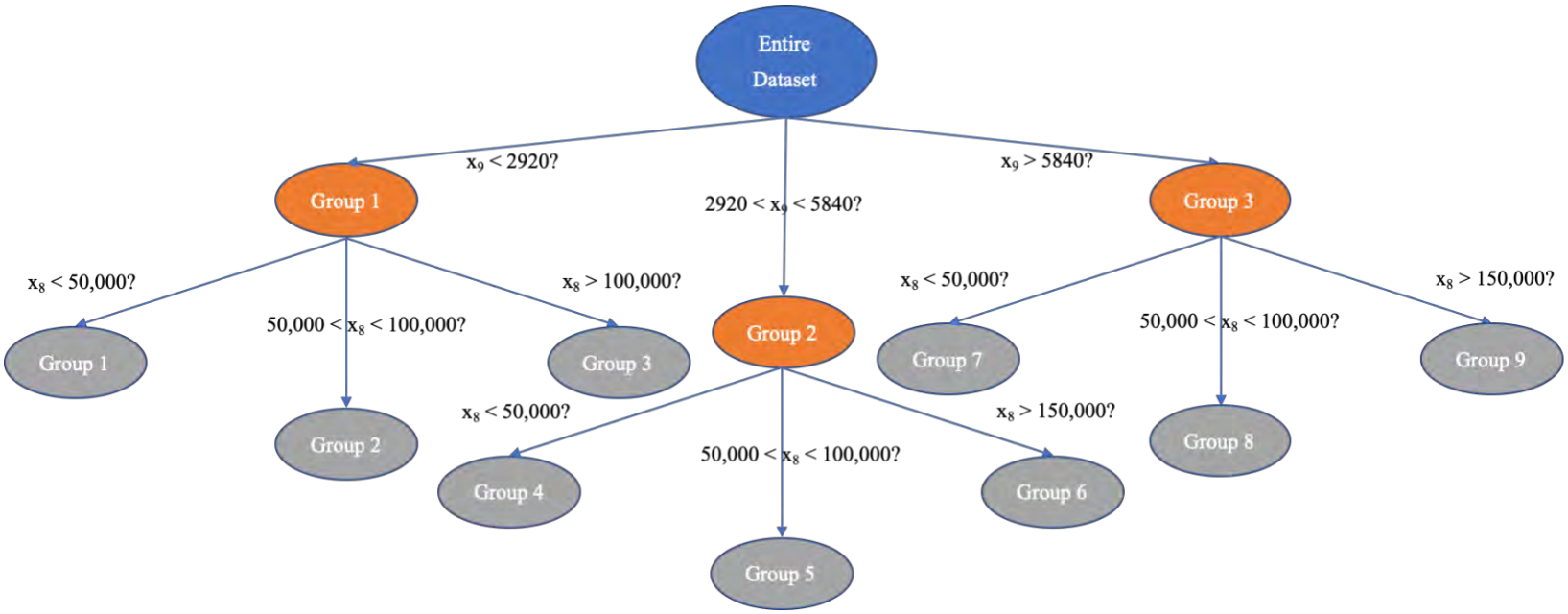

Figure 4-8: Flowchart of the divisions after two levels of division

4.6. MODULES USED

Python was used to code the algorithms and manipulate the data for implementing the algorithms. Below is a brief description of how each algorithm was implemented using python's libraries and modules.

- **Linear Regression:** A linear regression model was created using sklearn library's LinearRegression class. Pandas library was used to import the dataset. sklearn library's LabelEncoder and OneHotEncoder classes were used to generate binary values for the categorical variables. Both these classes were fit to independent matrix of features, only on the column of the categorical variable. This column was later transformed into an array of binary features. sklearn library's train_test_split module was used to split the data into training and testing datasets. An 87.5% test 12.5% train split was used. Based on previous literature it was seen that anywhere between 70-30 (high customizability and low precision) and 95-5 (low-customizability and high precision) train test splits were used. 87.5-12.5 split was used to balance customizability and precision. sklearn library's metrics class was used to calculate mean absolute, mean squared and root mean squared error.
- **Polynomial Regression:** A polynomial regression model was created using sklearn library's PolynomialFeatures class. Pandas library was used to import the dataset. sklearn library's LabelEncoder and OneHotEncoder classes were used to generate binary values for the categorical variables. Both these classes were fit to independent matrix of features, only on the column of the categorical variable. This column was later transformed into an array of binary features. sklearn library's

train_test_split module was used to split the data into training and testing datasets. An 87.5% test 12.5% train split was used. The PolynomialFeatures generated variables of the specified degree. For eg, for a specified degree 2, the class creates squares of every variable and also every second-degree interaction possible. By trial and error, it was found that second degree polynomial had the least deviation and hence degree 2 was specified for the polynomial features. sklearn library's metrics class was used to calculate mean absolute, mean squared and root mean squared error.

- **Support Vector Regression:** A support vector regression model was created using sklearn library's SVR class. Pandas library was used to import the dataset. sklearn library's LabelEncoder and OneHotEncoder classes were used to generate binary values for the categorical variables. Both these classes were fit to independent matrix of features, only on the column of the categorical variable. This column was later transformed into an array of binary features. sklearn library's train_test_split module was used to split the data into training and testing datasets. An 87.5% test 12.5% train split was used. sklearn library's metrics class was used to calculate mean absolute, mean squared and root mean squared error.
- **Random Forest Feature Selection:** A random forest feature selection model was created using sklearn library's RandomForestRegressor class. Pandas library was used to import the dataset. sklearn library's LabelEncoder and OneHotEncoder classes were used to generate binary values for the categorical variables. Both these classes were fit to independent matrix of features, only on the column of the

categorical variable. This column was later transformed into an array of binary features. The number of estimators used was 100. It means that 100 decision trees are created, and the most repeated values are given as the output.

4.7. MODEL VALIDATION

Any machine learning model needs to be validated for consistency throughout the dataset. A model that is predicting well on one part of the dataset might not be or one form of test train split might not be predicting to the same degree of accuracy on another train test split. Bias might arise from the way training and testing data has been divided. So, it is essential to try multiple combinations of train-test splits to judge whether the model performs to the same level. If the model gives consistent results, it is safe to assume that the model will give the same level of performance in further applications as well. This was found out after testing of the model on different train and test combinations. To validate this model to satisfy the above-mentioned criteria, k-fold cross validation is a very good fit. This method of validation has been used as per [46, 47] to validate the proposed model in this research. The validation procedure is as follows:

- Randomize the data points by shuffling.
- Divide dataset into k equal folds or groups.
- Train with k-1 folds and test on 1 fold.
- Follow step 4 for every possible combination.
- Find percentage deviation for every combination created.
- If the values of all folds are similar, model output is consistent.

For the purpose of this research, 8 folds have been selected. This gives an 87.5 – 12.5 train

test split. 10 folds are considered reasonably good for any estimate [48]. A simple illustration of this methodology is as follows:

The results from validating the model by this procedure are:

Table 4-7: Summary of error terms in conventional and proposed approaches

ITEM	Linear Regression % Deviation		Polynomial Regression % Deviation		Support Vector Regression % Deviation	
	w RF	w/o RF	w RF	w/o RF	w RF	w/o RF
FOLD 1 ERROR	57.63	33.31	52.80	31.64	59.83	35.71
FOLD 2 ERROR	55.62	33.99	52.53	31.94	61.23	35.62
FOLD 3 ERROR	56.56	34.76	53.00	32.90	59.28	36.06
FOLD 4 ERROR	55.30	33.05	52.48	32.05	58.90	34.54
FOLD 5 ERROR	55.43	33.30	51.33	32.63	59.31	37.26
FOLD 6 ERROR	57.58	33.60	53.9	32.55	61.90	35.84
FOLD 7 ERROR	57.19	33.10	52.35	32.53	58.00	36.48
FOLD 8 ERROR	53.40	33.30	52.94	32.96	56.56	36.63
AVERAGE ERROR	56.10	33.55	52.54	32.35	59.37	36.06

Table 4-8:Summary of model validation

TYPE OF REGRESSION	% DEVIATION WITHOUT RANDOM FORESTS	% DEVIATION WITH RANDOM FORESTS
LINEAR REGRESSION	56.1	33.55
POLYNOMIAL REGRESSION	52.54	32.35
SUPPORT VECTOR REGRESSION	59.37	36.06

It is seen from the table that the proposed approach performs consistently better than conventional approaches of several types of regression. It is also seen that there is 20-23% improved by using the proposed approach. The percentage deviation is calculated as:

$$\left(\frac{Absolute\ Value\ (Original\ Value - Predicted\ Value)}{Originial\ Value}\right) * 100$$

5. DISCUSSIONS AND FUTURE WORK

It was seen from the model section that the region was a statistically insignificant variable. This makes us question if energy consumption is constant throughout the country. The two main causes of energy consumption varying in various regions of the country would be: ambient temperature and working hours. But as explained in the variable description, both these factors have been represented as separate variables already. This might be a reason for region not being a significant variable in this research. It is also seen that there are almost equal number of datapoints from each region, making the variable even lesser significant. It was also seen that the presence of region variables affected the model by making the temperature variable insignificant. This is seen in the case of T-test conducted. But in the case of random forests, the presence of binary variables of region did not affect the significance of other variables. The algorithm was able to identify the statistical significance of temperature variable even with the presence of binary variables denoting region, which the model assigned as having zero significance. This establishes the superiority of machine learning tools. Statistical procedures like regression, rely on equations of test statistic and the corresponding tables to make decisions about variables.

But in the case of machine learning models, though insignificant variables might be part of the dataset, the model learns over time that the values of such variables do not alter the output. Since there is no strict necessity of a plane or curve to be fit, the learning is purely based on historical data and no assumptions. It is also seen that there were assumptions that a dataset needs to satisfy in order to be used for regression models. But in case of machine learning tools, such assumptions need not be met. This is because the

algorithm makes predicted decisions based on past naturally occurred decisions, giving an edge to machine learning models. It also seen that the outside temperature is the least significant variable. Assuming that the outside temperature has a considerable impact on energy consumption due to the presence of HVAC systems is common. The lesser significance of temperature can be attributed to the HVAC usage trends of the country. The International Energy Agency states that 50% of residential energy consumption is from HVAC whereas only 20% of the overall energy consumption from this country is because of HVAC. This explains why outside temperature is a not a significant variable in industrial energy consumption.

Validation of the model also shows that polynomial regression performed better than linear regression on the entire dataset. But on localized datasets after division, linear regression performed equally good. This shows that the linearity of data increases as the dataset is split into homogenous subgroups. Support vector regression performed the worst in predicting energy consumption (on both entire dataset and subsets). As mentioned earlier, support vector machines try to minimize the maximum error. In case of a dataset with a very high variability like the one used in this research, the algorithm tries to fit a plane that reduce the maximum error at several points, increasing the average error. This is the reason for better performance of statistical regression that least average error estimates.

This research can further be expanded to the following applications.

- Shorter time periods: This research was conducted based on energy consumption of industrial facilities from the IAC database. But the data is yearly. Literature has

shown that weekly and monthly energy consumption data give a better estimate. So, incorporating the model in such scenarios might give a better accuracy.

- Usage of data loggers: The dataset used has number of employees, plant area as variables. However, the values for these variables are constant. If data loggers can be used to monitor the number of employees in the plant at a particular time period and also the area being cooled/heated, that would be a more accurate representation of energy consumption.
- Application in residential energy consumption: This research has focused on industrial energy consumption. But this model can also be used to predict energy consumption of small-scale residential apartments.
- Use in IACs: In Industrial Assessment Center, we make recommendations for clients to make necessary changes for reduced energy usage. To give the clients an idea on return on investment, we speak to them about payback period. But payback periods would be more accurate figures, if we can incorporate the possible changes in the facility like number of employees, plant area while calculating. This model can be used for such predictions.
- Price of electricity: An important factor in electricity consumption is the price of electricity. Industries might use lesser amount of energy if there is high cost of electricity. Incorporating the cost of electricity of a particular variable might strengthen the model's efficiency of prediction.

6. REFERENCES

1) International Energy Agency. 2017. "Tracking progress: Industry", https://www.iea.org.
2) Allouhi, A., Y. El Fouih, T. Kousksou, A. Jamil, Y. Zeraouli, and Y. Mourad. 2015. "Energy Consumption and Efficiency in Buildings: Current Status and Future Trends." *Journal of Cleaner Production* 109 (December): 118–30.
3) International Energy Agency, 2017. "Tracking Clean Energy Progress 2017", *Informing energy sector transformations* (June).
4) International Energy Agency. 2017. "ETP 2017 data visualization", *Energy Technology perspectives.*
5) Hong, Tao, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J. Hyndman. 2016. "Probabilistic Energy Forecasting: Global Energy Forecasting Competition 2014 and beyond." *International Journal of Forecasting* 32 (3): 896–913.
6) Diedrich, Amy, Paul Upham, Les Levidow, and Sybille van den Hove. 2011. "Framing Environmental Sustainability Challenges for Research and Innovation in European Policy Agendas." *Environmental Science & Policy* 14 (8): 935–39.
7) Dovì, Vincenzo Giorgio, Ferenc Friedler, Donald Huisingh, and Jiří Jaromír Klemeš. 2009. "Cleaner Energy for Sustainable Future." *Journal of Cleaner Production* 17 (10): 889–95.
8) Santoyo-Castelazo, Edgar, and Adisa Azapagic. 2014. "Sustainability Assessment of Energy Systems: Integrating Environmental, Economic and Social Aspects." *Journal of Cleaner Production* 80 (October): 119–38.

9) Ness, Barry, Evelin Urbel-Piirsalu, Stefan Anderberg, and Lennart Olsson. 2007. "Categorising Tools for Sustainability Assessment." *Ecological Economics: The Journal of the International Society for Ecological Economics* 60 (3): 498–508.

10) Vuuren, D. P. van, N. Nakicenovic, K. Riahi, A. Brew-Hammond, D. Kammen, V. Modi, M. Nilsson, and K. R. Smith. 2012. "An Energy Vision: The Transformation towards Sustainability—interconnected Challenges and Solutions." *Current Opinion in Environmental Sustainability* 4 (1): 18–34.

11) Metz, B., O. R. Davidson, P. R. Bosch, R. Dave, and L. A. Meyer. 2007. "Issues Related to Mitigation in the Long-Term Context (Chapter 3)." In *Climate Change 2007: Mitigation. Contribution of WG III to the Fourth Assessment Report of the IPCC*, edited by B. Metz, O. R. Davidson, P. R. Bosch, R. Dave, and L. A. Meyer. Cambridge: Cambridge University Press.

12) Keywan Riahi, 2017. "Energy Pathways for Sustainable Development." *Global Energy Assessment* 6: 1208–1301.

13) Shonali Pachauri, Abeeku Brew-Hammond. 2019. "Energy Access for Development." *Global Energy Assessment* 8: 1403–1453.

14) Clarke, Leon, Jae Edmonds, Volker Krey, Richard Richels, Steven Rose, and Massimo Tavoni. 2009. "International Climate Policy Architectures: Overview of the EMF 22 International Scenarios." *Energy Economics* 31 (December): S64–81.

15) Van Vuuren, D. P., M. Meinshausen, G-K Plattner, F. Joos, K. M. Strassmann, S. J. Smith, T. M. L. Wigley, et al. 2008. "Temperature Increase of 21st Century Mitigation

Scenarios." *Proceedings of the National Academy of Sciences of the United States of America* 105 (40): 15258–62.

16) Bringezu, Stefan, Helmut Schütz, Meghan O'Brien, Lea Kauppi, Robert W. Howarth, and Jeff McNeely. 2009. "Assessing Biofuels: Towards Sustainable Production and Use of Resources."

17) Searchinger, Timothy, Ralph Heimlich, R. A. Houghton, Fengxia Dong, Amani Elobeid, Jacinto Fabiosa, Simla Tokgoz, Dermot Hayes, and Tun-Hsiang Yu. 2008. "Use of U.S. Croplands for Biofuels Increases Greenhouse Gases through Emissions from Land-Use Change." *Science* 319 (5867): 1238–40.

18) Kruyt, Bert, D. P. van Vuuren, H. J. M. de Vries, and H. Groenenberg. 2009. "Indicators for Energy Security." *Energy Policy* 37 (6): 2166–81.

19) "IAC: Industrial Energy Assessments." https://iac.university/

20) Zhao, Hai-Xiang, and Frédéric Magoulès. 2012. "A Review on the Prediction of Building Energy Consumption." *Renewable and Sustainable Energy Reviews* 16 (6): 3586–92.

21) Tso, Geoffrey K. F., and Kelvin K. W. Yau. 2007. "Predicting Electricity Energy Consumption: A Comparison of Regression Analysis, Decision Tree and Neural Networks." *Energy* 32 (9): 1761–68.

22) Fumo, Nelson, and M. A. Rafe Biswas. 2015. "Regression Analysis for Prediction of Residential Energy Consumption." *Renewable and Sustainable Energy Reviews* 47 (July): 332–43.

23) Catalina, Tiberiu, Vlad Iordache, and Bogdan Caracaleanu. 2013. "Multiple Regression Model for Fast Prediction of the Heating Energy Demand." *Energy and Buildings* 57 (February): 302–12.

24) Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. *Foundations of Machine Learning*. MIT Press.

25) Tsanas, Athanasios, and Angeliki Xifara. 2012. "Accurate Quantitative Estimation of Energy Performance of Residential Buildings Using Statistical Machine Learning Tools." *Energy and Buildings* 49 (June): 560–67.

26) Dong, Bing, Cheng Cao, and Siew Eang Lee. 2005. "Applying Support Vector Machines to Predict Building Energy Consumption in Tropical Region." *Energy and Buildings* 37 (5): 545–53.

27) Kavaklioglu, Kadir. 2011. "Modeling and Prediction of Turkey's Electricity Consumption Using Support Vector Regression." *Applied Energy* 88 (1): 368–75.

28) Ekici, Betul Bektas, and U. Teoman Aksoy. 2009. "Prediction of Building Energy Consumption by Using Artificial Neural Networks." *Advances in Engineering Software* 40 (5): 356–62.

29) Yang, Jin, Hugues Rivard, and Radu Zmeureanu. 2005. "On-Line Building Energy Prediction Using Adaptive Artificial Neural Networks." *Energy and Buildings* 37 (12): 1250–59.

30) Karatasou, S., M. Santamouris, and V. Geros. 2006. "Modeling and Predicting Building's Energy Use with Artificial Neural Networks: Methods and Results." *Energy and Buildings* 38 (8): 949–58.

31) Ekonomou, L. 2010. “Greek Long-Term Energy Consumption Prediction Using Artificial Neural Networks.” *Energy* 35 (2): 512–17.

32) Kalogirou, Soteris A., and Milorad Bojic. 2000. “Artificial Neural Networks for the Prediction of the Energy Consumption of a Passive Solar Building.” *Energy* 25 (5): 479–91.

33) Tin Kam Ho. 1995. “Random Decision Forests.” *ICDAR '95 Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1):* 278-282.

34) Zoph, Barret, and Quoc V. Le. 2016. “Neural Architecture Search with Reinforcement Learning.”

35) Cortes, Corinna, and Vladimir Vapnik. 1995. “Support-Vector Networks.” *Machine Learning* 20 (3): 273–97.

36) Neto, Alberto Hernandez, and Flávio Augusto Sanzovo Fiorelli. 2008. “Comparison between Detailed Model Simulation and Artificial Neural Network for Forecasting Building Energy Consumption.” *Energy and Buildings* 40 (12): 2169–76.

37) Mohiuddin, Mao, and Jain. 1996. “Artificial Neural Networks: A Tutorial” 29 (March): 31–44.

38) Kalogirou, Soteris A. 2000. “Applications of Artificial Neural-Networks for Energy Systems.” *Applied Energy* 67 (1): 17–35.

39) Al-Homoud, Mohammad Saad. 2001. “Computer-Aided Building Energy Analysis Techniques.” *Building and Environment* 36 (4): 421–33.

40) Iain Pardoe. 2012. “Applied Regression Modeling – A Business Approach.” *John Wiley & sons.*

41) Mohri, M., Rostamizadeh, A. and Talwalkar, A. 2018. “Foundations of machine learning.” *MIT press.*

42) Saed Sayed. 2019. “An Introduction to Data Science.” *Rutgers School of Arts and Science.*

43) Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.

44) Pérez-Lombard, Luis, José Ortiz, and Christine Pout. 2008. “A Review on Buildings Energy Consumption Information.” *Energy and Buildings* 40 (3): 394–98.

45) Applied Statistics and probability for engineers by Montgomery

46) Wong, Tzu-Tsung. 2015. “Performance Evaluation of Classification Algorithms by K-Fold and Leave-One-out Cross Validation.” *Pattern Recognition* 48 (9): 2839–46.

47) Meijer, Rosa J., and Jelle J. Goeman. 2013. “Efficient Approximate K-Fold and Leave-One-out Cross-Validation for Ridge Regression.” *Biometrical Journal. Biometrische Zeitschrift* 55 (2): 141–55.

48) Ron Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, International Joint Conference on Artificial Intelligence (IJCAI), 1995.