2019

# A Prototype for predicting energy consumption in buildings: a case of commercial office buildings.

Wachira, Paul Manasse Macharia

*Strathmore School of Computing and Engineering Sciences*

*Strathmore University*

# A Prototype for Predicting Energy Consumption in Buildings: A Case of Commercial Office Buildings

**Paul Manasse Macharia Wachira**

**Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Technology at Strathmore University**

*Faculty of Information Technology*
**Strathmore University**
**Nairobi, Kenya**

**June, 2019**

# DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.



Paul Manasse Macharia Wachira

23rd June 2019

**Approval**

The thesis of Paul Manasse Macharia Wachira was reviewed and a approved by the following:

Dr. Vincent Omwenga,

Senior Lecturer, Faculty of Information Technology,

Strathmore University.

Dr. Joseph Orero,

Dean, Faculty of Information Technology,

Strathmore University.

Prof. Ruth Kiraka,

Dean, School of Graduate Studies,

Strathmore University.

# ABSTRACT

Energy consumption remains the highest cost areas for businesses together with facilities, people and equipment but unfortunately, it is the only one that is not carefully monitored. For businesses to be able to manage energy consumption they must first be able to predict future energy consumption so as to aid in budgeting and planning for cost reduction strategies. This study proposed an energy consumption prediction prototype to help predict future consumption of energy in commercial office buildings thus aiding proper budgeting and cost reduction. To develop the prediction model the study used the 2012 CBECS (Commercial Buildings Energy Consumption Survey) dataset hosted by the Energy Information Administration (EIA) of the United States of America. After cleaning and reviewing the data set, 26 Features were selected for Features Engineering. Features Engineering enabled the research to choose the best 4 Features which were used for training and testing different Regression Based Machine Learning Algorithms. Using $R^2$ (R Squared), MAE (Mean Absolute Error) and RMSE (Root Mean Square Error) to determine performance, the study selected Gradient Boost Machines as the best algorithm for the prototype with an Accuracy of 97%. Python packages Pandas, NumPy, Matplotlib, Seaborn and Scikit-learn were used in data cleaning, descriptive statistics, features engineering, data visualization and training and testing the machine learning algorithms for the energy consumption prediction model. The prototype was developed using Flask (a Python micro web framework) to enable the Building owners provide the prototype with data via web browser related to the 4 features selected for energy consumption prediction. Usability Test was done with 48.1% of the users strongly agreeing and 44.2% agreeing to use the prototype in future for prediction of electricity consumption in their buildings.



Keywords: *energy consumption; prediction; commercial office building; features engineering; machine learning algorithms;*

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS/ACCRONYMS

**ABC** - Artificial Bee Colony

**ANN** - Artificial Neural Network

**ARIMA** - Autoregressive Integrated Moving Average

**ASHRAE** - The American Society of Heating, Refrigerating and Air-Conditioning Engineers

**BPNN** - Back Propagation Neural Network

**CBECS** - Commercial Buildings Energy Consumption Survey

**CDD** - Cooling Degree Day

**CIESOL** - The Centre for Solar Energy Research

**CMI** - Conditional Mutual Information

**DCs** - Data Centers

**DECC** – Department of Energy & Climate Change

**DFD** - Data Flow Diagram

**EIA** - Energy Information Administration

**FWPT** - Flexible Wavelet Packet Transform

**GDP** - Gross Domestic Product

**GWh** - Gigawatt Hours

**HDD** - Heating degree day

**HVAC** - Heating, Ventilation, and Air-Condition

**HWS** - Hot Water Supply

**KNBS** - Kenya National Bureau of Statistics

**KSh** - Kenyan shilling

**kWh** - Kilowatt hour

**LS-SVM** - Least Square Support Vector Machines

**LS-SVR** - Least Squares Support Vector Regression

**MAE** - Mean Absolute Error

**MAPE**- Mean Absolute Percentage Error

**MLR** - Multiple Linear Regression

**MSE** - Mean Squared Error

**NLSSVM** - Non-Linear Least Square Support Vector Machine

**OLS** - Ordinary Least Squares

**PBA** - Principal Building Activity

**RAD** - Rapid Application Development

**RMSE** - Root Mean Square Error

**SEA** - Sustainable Energy Africa

**SVM** - Support Vector Machines

**SVR** - Support Vector Regression

**UNIDO** - United Nations Industrial Development Organization

# LIST OF EQUATIONS

# DEFINITION OF TERMS

**Energy** – The power or capacity to do work e.g. moving an object by application of force. Energy can exist in a variety of forms such as chemical, mechanical, thermal, nuclear and electrical (Energy, 2019). **In this research when energy is mentioned it refers to electrical energy.**

**Commercial Building** – it is a building used for commercial use and can include office buildings, retail buildings or warehouses. (Commercial Building, 2019). **This research focuses on office buildings.**

**Office Building** – a structure that holds single or multiple firms whose primary focus of business relates to administration, consulting, client services and clerical services not related to retail sales (Office Building, 2019).

## ACKNOWLEDGEMENTS

First, I want to thank God for enabling me to work on my thesis. I also extend gratitude to Dr. Vincent Omwenga my supervisor, whose guidance, patience and suggestions have been crucial to the development of this thesis. Special thanks to Engineer Cyprian Njuki Njururi for the advice he extended to me on Energy Matters. Finally, I would like to thank my family, friends and colleagues for their support and prayers during this period. I will always be indebted to you all.

## DEDICATION

This research work is dedicated to Bilha Githara my loving wife, my three children - Tara Chelsea, David Asher & James Christian and my extended family especially my mother Joyce Wangari who inspired me to pursue this degree.

# CHAPTER 1: INTRODUCTION

## 1.1 Background of the Study

Energy consumption in commercial and residential buildings accounts for 40% of the total energy consumption of any country and is expected to increase to 50% by 2030 (Hassan, Zin, Abd Majid, Balubaid, & Hainin, 2014). Furthermore, approximately 90 percent of our time is either spent at home or in the office (United Nations Industrial Development Organization [UNIDO], 2009). Energy consumption by households accounts for approximately 30% of global energy (Gowda, 2016). Commercial buildings in the US consumed 18% of the total energy used (U.S. Energy Information Administration [EIA], 2018a). Energy consumption and demand patterns play a huge role in policy development and planning of power supply systems by governments. Energy consumption by buildings represented 40% of all energy used in Europe (Rousselot & Pollier, 2018). Commercial buildings, office buildings, and universities are among the buildings which consume the most energy (Department of Energy & Climate Change [DECC], 2013; Rhee & Chung, 2014). Improving the understanding of energy consumption in commercial buildings remains an important topic for those who seek to reduce energy consumption (Martin, 2013).

According to Kenya National Bureau of Statistics (KNBS) survey report of 2014, Kenyan households consumed around 200-kilowatt hours (kWh) of electricity and paid an average of Ksh. 3400 as by March 2017, compared to Ksh. 3042 in February 2013. In addition, the commercial and industrial consumption for the year 2012 was 3419GWh (Irungu, 2016). Kenya has got three types of energy consumers based on their monthly consumptions. There are Lifeline consumers who use less than 50 kWh per month, residential consumers who utilise at least 200 kWh which are the majority and often affected by an increase in prices and finally the Industrial or commercial consumers who exceed 1,500 kWh.

There are three main areas that can lead to greater energy efficiency, as well as decreases in carbon emissions. These are technological innovations for energy efficiency,

fluctuating the energy supply mix and promoting structural changes in the economy (Roula, 2016).

Accurate power development planning (generation, transmission, and distribution) in the country is critical for the country's development. The energy demand of the household is one of the main issues in planning and monitoring studies (Verdejo, Awerkin, Becker, & Olguin, 2017). Measuring grid growth and understanding the consumer behaviour for residential customers is vital. There are two majors' factors which influence building energy consumption categorized as internal factors and external factors. Internal factors such as inherent attributes, management attributes and technical attributes of a building and external factors such as meteorological parameters have an influence on the energy consumption of a building (Woo & Cho, 2018).

Energy consumption patterns and demands are critical inputs into an energy future consumption prediction model (Reade & Zewotir, 2016). The future consumptions can then help in planning and budgeting by consumers and by extension power suppliers and governments. The predicting model's accuracy depends on the accuracy of the data and assumptions underlying it. It is therefore paramount that these be as accurate as is reasonably possible.

## 1.2 Problem Statement

The demand for energy has over the years been on the rise because of the emergence of digital devices and modern property development (Mehar, Gill, & Matawie, 2018). Energy, along with human labour, product costs, facilities and equipment, is one of the largest cost areas for organizations, yet it is the only one not carefully monitored and managed (Winston, Favaloro, & Healy, 2017). Since it is not carefully monitored and managed it adds an extra financial burden on organizations as they have limited and calculated budget. (Sustainable Energy Africa [SEA], 2017; Singh, 2017). Winston et.al, (2017), also states that a company's approach to energy and carbon emissions now directly impacts its cost structure, risk profile, and resilience. They further highlight that organizations need to establish a clear strategy and goals to manage energy in a coordinated way.

Energy is the most volatile operating expense for a building (Energywatch, n.d.) and many energy cost drivers can be unpredictable (Schneider Electric, n.d.). In the old days budgeting for energy was a matter of looking at the current costs, adding two percent and hoping for the best (EnergyPrint Inc, n.d.). EnergyWatch Inc (n.d.) further states that it is not enough to simply take a three-year average of rate and consumption and project forward into the future. Schneider Electric (n.d.) also notes that some energy buyers forecast their energy budget needs by simply applying an arbitrary percentage increase over the previous year.

Monitoring and analysing the use of energy can reveal operational problems affecting costs, performance and quality. Machine learning models can predict the energy burdens thus leading to a better understanding of the energy consumption landscape. Therefore, effective energy consumption prediction is important in determining the demand and subsequent cost allocation (Amber et al., 2017).

## 1.3 General Objectives

The aim of the research was to develop an Energy Consumption Prediction Prototype for Commercial Buildings using Machine Learning on existing data to aid building owners get more accurate predictions of future energy consumption which is essential for better energy budgeting and planning.

### 1.3.1 Specific Objectives

i. To analyze factors that influence energy consumption in commercial office buildings.
ii. To review methods and techniques used in predicting energy consumption in buildings.
iii. To design and develop an energy consumption prediction prototype based on machine learning algorithms for commercial office buildings.
iv. To test the energy consumption prediction prototype.

## 1.4 Research Questions

i. What are the factors influence energy consumption in commercial office buildings?
ii. What are the methods and techniques used in predicting energy consumption in buildings?
iii. How can an energy consumption prediction prototype be designed and developed using existing machine learning algorithms?
iv. How can the energy consumption prediction prototype be tested?

## 1.5 Justification

Electricity is one of the basics for civilizations as it affects every aspect of human life (Kavaklioglu, 2018). Electricity is used in residential, offices, transportation and industrial activities nowadays. As it is contributing to bettering our life, electricity causes some issues in global warming depending on the source of energy used. Office buildings are amongst the heavy consumers of all the energy produced and that is becoming a problem not only to governments but to organizations as well. Organizations are spending more than what was expected in terms of energy needs.

Even though there has been a gradual installation of smart meters in residential buildings, various issues have been raised in regard to these devices. There are some concerns that smart meters can and do collect unnecessary information about hourly electricity use, thus violating users' privacy. According to Power Technology (2013), despite installing smart meters, energy providers encourage their customers to pay annually based on a yearly estimate. Meaning that you pay based on an "estimated" use, even if you have a smart meter (Power Technology, 2013). Experiments have shown that some smart meters over record the power used by up to 600% and some under record by 50% depending on the power measurement technique used. Lab tests conducted by Dutch scientists in the Netherlands have shown that some of today's ' smart ' electrical meters can give false readings that in some cases may be 582 percent higher than actual energy consumption (Cimpanu, 2017).

Therefore, there is a need for an accurate prediction of future power demand and electrical energy consumption patterns by consumers to better plan the energy use in prevailing economic condition. A prediction model that the consumers have control and can be able to factor in in their cost management strategies. One approach to address this problem is predicting energy consumption at the consumer level (office buildings) for better energy budgeting, demand management and effective utilization of the limited available financial resources. Thus, this study will contribute in the effort of closing the gap between what is budgeted for energy and what is consumed.

## 1.6 Scope of the study

There are various ways in which predicting of energy consumption could potentially help in cost reduction and adherence to budget in organization. In this thesis, the main focus is developing an energy consumption prototype using the 2012 CBECS Survey dataset released by the Energy Information Administration (EIA) of the United States of America in June 2015 and revised in August 2016. The dataset contains information for 6720 commercial buildings which was collected through professional interviews and other computerized survey tools. The data was collected between April to November 2013 using Building owners and managers (U.S. Energy Information Administration [EIA], 2015). The CBECS dataset contains several features such as energy consumption, climate of the building's region, building type, envelope materials, heating degree days (HDDs), number of laptops, cooling degree days (CDDs), number of computers, type of HVAC system, construction year, and so on. The aim of this dataset was to obtain factual data for commercial buildings in the U.S. on energy consumption. This proposed prototype can be applied by commercial office buildings in Kenya with aim of managing energy costs.

# CHAPTER 2: LITERATURE REVIEW

## 2.1 Introduction

This chapter describes the current literature on energy consumption and predicting in order to provide an in-depth understanding and insights into existing studies significant to this research. This review forms the foundation of this study and begins by reviewing existing literature on factors that influence energy consumption in commercial office buildings. Then current methods and techniques used in energy consumption prediction are reviewed and a conceptual framework with steps the researcher followed in developing the prototype's prediction model.

## 2.2 Factors affecting energy consumption in Commercial Office Buildings

The factors influencing energy consumption can be categorized into two:

i) The external factors - such as weather and economy which can influence the energy consumption of commercial office buildings (Delzendeh, Wu, Lee, & Zhou, 2017).

ii) The internal factors - such as air conditioning, heating and ventilation which proved to be the greatest factor affecting the high energy consumption of commercial buildings (Vakiloroaya, Samali, Fakhar & Pishghadam, 2014).

In China, lighting and other applications, namely office equipment, lifts (elevators) have influenced high energy consumption in commercial buildings (Fridley, Zheng & Zhou, 2008) while in the USA, HVAC applications, and space cooling/heating were the applications influencing the high energy consumption in commercial buildings (Department of Energy, 2012).

Energy consumption is related to population growth and economic development (Lombard, Ortiz & Pout, 2007). Furthermore, increased purchasing power, promoted developed nations' lifestyle, raised energy needs are factors making the need for energy to increase. Nababan (2015) puts forward the fact that analysing the energy consumption of households and commercial buildings, the availability of infrastructure, economic factors, lifestyle factors and mindset factors should also be considered.

According to U.S. Energy Information Administration [EIA] (2018b), the major factors that contribute to the largest use of electricity in commercial buildings include lighting, refrigeration, ventilation, office equipment (computers and related equipment), water heating and cooking. Furthermore, Gul and Patidar (2015) posit that lighting remains the major user of electricity in office buildings and managing lighting well can lead to a reduction in energy costs.

The importance of accurate prediction of future power demand and electrical energy consumption patterns need not be overemphasized. Inaccurate prediction will lead to an inadequate power system expansion resulting in an inadequate capacity to meet the demand (Ouedraogo, 2017). This would result in load shedding or the use of the costly peaking power plants for long periods because it takes a long time and huge capital resources to develop power plants. On the other hand, a high forecast will lead to the development of a huge power system that would be under-utilized leading to high electricity tariffs. Both low and high demand forecasts have adverse effects on the consumers whether residential or commercial. It is therefore important that the demand forecast be as accurate as is reasonably possible. Factors affecting energy consumption are further discussed in the below sub-sections.

### 2.2.1 Climatic Conditions

Many climate parameters influence the energy consumption of buildings (Kalamees et al., 2012). The prevailing local climate and the overall climate change in the coming years will substantially affect the total consumption of a building (Wan, Li, Liu & Lam, 2011). Wan et al. (2011) further states that some of the major weather parameters that influence building consumption are moisture content of the air, temperature, wind direction/speed and solar radiation. Furthermore, Jim and Peng (2012) identified weather as a huge influential factor in Hong Kong building energy consumption. Kalamees et al. (2012) also notes that Heating Ventilation and Air Conditioning (HVAC) systems are mostly used during the summer when temperatures are high and Hot Water Supply (HWS) systems are mostly used during winter when temperatures are low. Both these systems use a lot of electricity within buildings.

### 2.2.2 Building Related Characteristics

De Silva and Sandanayake (2012) identified the following building related characteristics as having influence on the energy consumption in buildings: building size, building age, building type, building class, building's worker density, building's usage hours, building's geographical location, building's orientation, building's envelop, building's construction quality, building's design, building's illumination, no. of lifts in the building, building's area covered by air-conditioning and building's indoor thermal quality. These characteristics have very high impact on building's energy consumption (De Silva & Sandanayake 2012).

In proportion to their volume, buildings with a higher shape factor have a larger surface area, resulting in greater energy consumption in cold climates. On the other hand Buildings with lower shape factor, require low energy demand. A slight difference in some building - related parameters would result in remarkable fluctuations in building energy consumption, according to Yun and Stemmers (2011). Furthermore, floor area / size as well as some layout parameters like building size, surface to heated volume ratio and some structural metrics e.g. ventilation of walls, windows and roofs influences energy consumption (Papadopoulos, Theodosiou, & Karatzas, 2002).

### 2.2.3 Occupant Related Characteristics

Ren et al. (2013) posits that occupants of building characteristics also is a factor in energy consumption. Patterns of household energy consumption are affected by occupancy patterns and behaviours such as when and how long household appliances, lighting, space cooling, hot water and space heating are operated (Ren et al., 2013) . Two main factors that could affect occupancy behaviours are: (a) the number of household residents and (b) the way of living of the inhabitant's. For example, energy used can differ from the moment the first individual wakes up in the morning and the last individual sleeps and the time the building is unoccupied throughout the day.

Often, lighting, cooling system, most appliances and heating systems are not used whenever the house is unoccupied. In addition, the building consumes less energy during

the sleeping hours in a regular appliance load profile. Buildings often use comparatively high energy in the evening for cooking, dishwashing, TV and computers (Ren et al., 2013).

Some recent studies have shown that tenant conduct in different households plays a significant role in the fluctuation in electricity consumption. Office buildings don't use energy, but individuals do. (De Silva & Sandanayake, 2012). As the tenancy level increases, energy consumption by lighting, elevators, plug loads and HVAC also increase. Janda (2011) argued that the role of occupants in a building is crucial but rarely discussed and sometimes ignored. People's role in the use of energy is more influential (Janda, 2011). A survey of Swedish households showed that only 17% of respondents put off lights regularly when they are leaving the building (Linden, Carlsson-Kanyama, & Eriksson, 2006).

Zero Carbon Hub (2015) conducted a study on construction design, evaluation and overheating of post-occupancy of a building. He found out that there was a substantial difference in electricity consumption in the very same building block between two apartments. This was as a result of different occupant characteristics, including: different home presence, occupancy levels and variations in the occupants' thermal preferences (Zero Carbon Hub, 2015).

Automated computer equipment lighting, and power saving features will help, but the only answer is often wholesale changes in company culture (Zou, Weidong, & Tang, 2017). Therefore, Minor changes in the behaviour of employees can have a significant effect on the amount of energy a company uses (Zou et al, 2017).

### 2.2.4 Building Systems and Services Related Characteristics

De Silva and Sandanayake (2012) identified the following building systems and services related characteristics as having influence on the energy consumption in buildings: building's services & systems load, building's services & systems specification, building's operation & maintenance, age of building's services & systems, efficiency of building's services & systems, appliance ownership and sub facilities offered.

### 2.2.5 Socio-Economic and Legal Characteristics

De Silva and Sandanayake (2012) identified several authors that conducted studies that showed level of education and income affect energy use. Mahapatra and Gustavsson, (2008) identified that Homeowner's age influences their energy usage behaviour with older homeowners less likely to adopt new energy reduction investment measures.

Saidur (2009) discussed the importance a countries' energy policy plays in influencing how energy is used in buildings and regulations such as energy efficiency labels, building codes and energy efficiency standards help reduce energy usage.

Yu, Fung, Haghighata, Yoshinoc and Morofskyd (2011) found out that electricity rates and primary heating/cooling sources help reduce energy usage in buildings by influencing behaviour of occupants since they know irresponsible usage of energy would lead to higher bills.

In response to the increase in energy consumption in buildings, public authorities and decision makers worldwide came up with new strategies which includes policies and measures to try to reduce energy consumption (Bull, Chang, & Fleming, 2012). These policies can be categorized into three (Annunziata, Frey, & Rizzi, 2013):

i. Regulatory measures such as building regulations, which have mandatory aspects and include minimum requirements.
ii. Soft instruments which consist mainly of voluntary standards such as certifications that go beyond the regulatory requirements.
iii. Economic incentive to motivate building owners and occupants to begin renovations or refurbishment works to improve their buildings' energy efficiency. Examples include energy savings performance contracting, tax exemptions and capital subsidies.

The implementation of these policies requires technical knowledge. Therefore, from a political point of view, it is difficult to monitor and evaluate the implementation of these policies and measures.

## 2.3 Methods and Techniques used in energy consumption prediction

Energy producers are constantly under pressure to meet the growing demand for energy from the commercial and domestic sectors. For all stakeholders, effective energy consumption predicting techniques are crucial ( Fayaz & Kim, 2018). Numerous energy - consuming elements such as lighting, HVAC systems, electrical devices, etc. must be considered in order to predict the power demand in a building. The element with the highest energy consumption in The Center for Solar Energy Research (CIESOL) building was the solar cooling system. (Hamzaçebi, 2016). In the below sub-section several methods and techniquees used in energy consumption prediction are discussed.

### 2.3.1 Sensor based energy prediction in commercial buildings

The predicting of energy consumption for both commercial and residential based on sensors has caught the interest of researchers in the past decade. However, the scarcity of residential data has led researchers to focus on commercial buildings (Jain, Smith, Culligan, & Taylor, 2014). After the Great Energy Predictor Shootout competition hosted by The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE), sensor-based energy prediction gained momentum (Jain et al., 2014; Kreider & Haberl, 1994). The competition consisted of building a predictive model that predicted the whole building electrical consumption. The winner of the competition developed a sensor-based energy predictive model using machine learning algorithms (Jain et al., 2014; MacKay, 1994). The outcomes of the competition attracted the interest of many researchers to explore the sensor-based energy prediction (Jain et al., 2014).

Doblander, Strohbach, Ziekew and Jacobsoen (n.d.) used sensors to predict energy loads of individual households as part of the PeerEnergyCloud Project which was funded by the German Federal Ministry of Economic Affairs and Energy to address application of technologies to improve load balancing on local energy grids. Figure 2.1. shows the analytical components of their system and Figure 2.2. shows the sensor devices used for data collection. Several regression methods and classification methods were used with varied results.

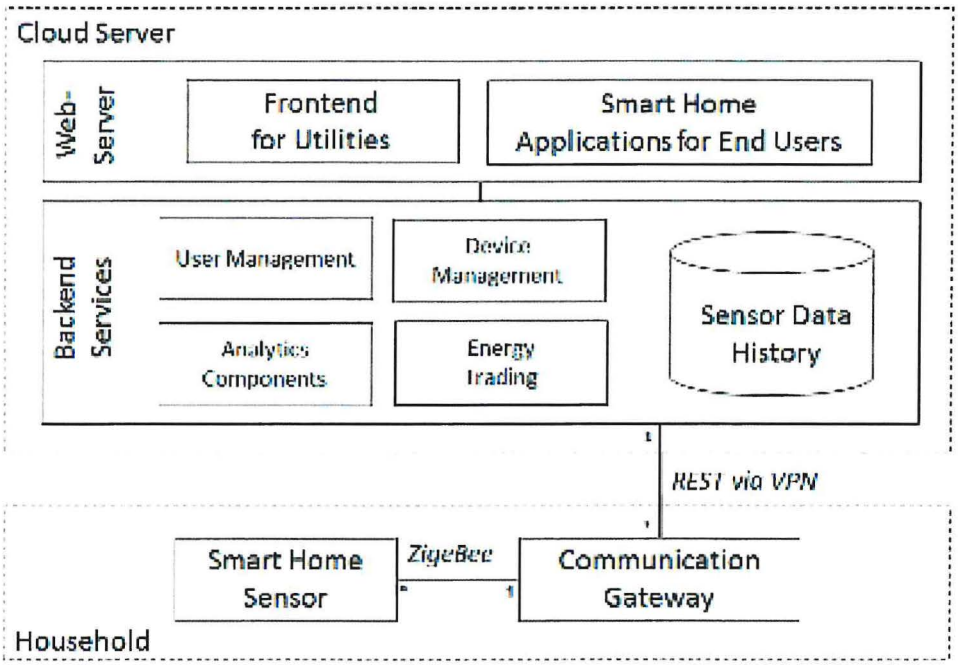


*Figure 2.1. High level architecture of the system deployment in the PeerEnergyCloud project. (Doblander, Strohbach, Ziekew, & Jacobsoen, n.d.)*

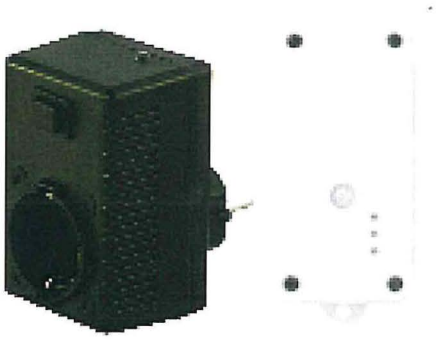

*Figure 2.2. Sensor Devices used in data collection. (Doblander, Strohbach, Ziekew, & Jacobsoen, n.d.)*

Smpokos, Elshatshat, Lioumpas and Illiopoulos (2018) forecast energy consumption of Data Centers (DCs) using machine learning based on remote sensor data of weather conditions. Smpokos et al. (2018) noted that cooling of the IT infrastructure of DCs used the most energy and thus outdoor weather conditions were important in developing the forecasting algorithm. The data collected by the sensors were the Independent Variables (Features) used by the Machine Learning Algorithms applied – Linear Regression as illustrated by Figure 2.3.

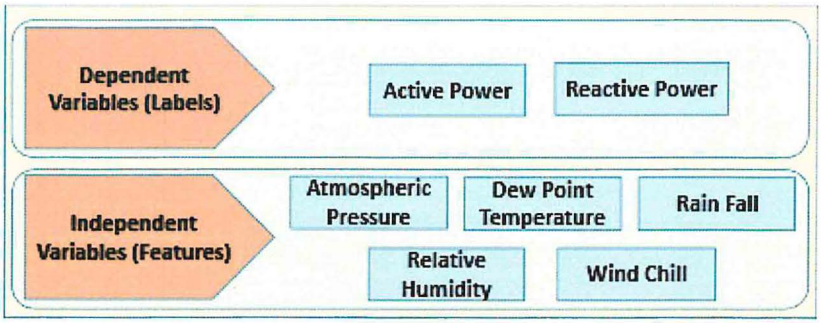


*Figure 2.3. Dependent and Independent variables (Smpokos, Elshatshat, Lioumpas, & Illiopoulos, 2018).*

### 2.3.2 Smart grid and smart metering forecast

To accurately forecast the price and demand in smart grids system remains an important challenge. The strong correlation between the price and demand in smart grid systems makes separate predicting to be ineffective. Thus, Ghasemia, Shayeghi, Moradzadeh and Nooshyar (2016) proposed a novel hybrid algorithm which forecast simultaneously the price and demand. The algorithm was classified into three parts. The first part made use of new Flexible Wavelet Packet Transform (FWPT) and a new feature selection method which uses conditional mutual information (CMI). The second part made use of a novel Multi-Input Multi-Output model based on Non-Linear Least Square Support Vector Machine (NLSSVM) and Autoregressive Integrated Moving Average (ARIMA) to model the price (linear) and load (nonlinear) correlation. The third part made use of a modified version of Artificial Bee Colony (ABC) algorithm based on time-varying coefficients and stumble generation. Jurado, Nebot, Mugica and Avellana (2015) proposed a hybrid

methodology that used feature selection based on entropies with soft computing and machine learning techniques such as Fuzzy Inductive reasoning, neural network and random forest to forecast the hourly energy predicting in buildings. The approaches used can be embedded in a second generation of smart meters.

Zhang, Grolinger and Capretz (2018) in their conference paper titled 'Forecasting Residential Energy Consumption using Support Vector Regression' used data collected from 15 households by London Hydro from 2014 to 2016 using Smart Meters.

Zhang et al. (2018) further developed a framework showing how their forecasting was to be done as shown by Figure 2.4.

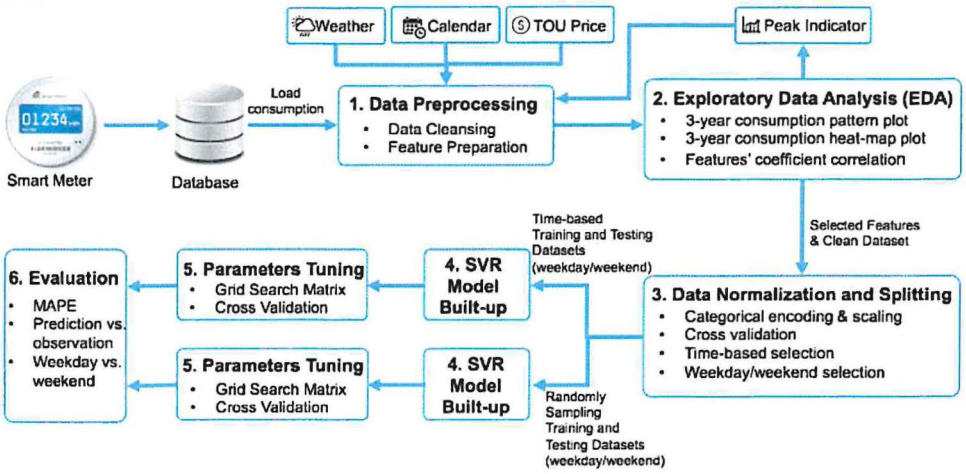


*Figure 2.4. Framework of individual household electricity forecasting (Zhang, Grolinger, & Capretz, 2018).*

Zhang et al. (2018) concluded that the use of smart metering technologies improved data analytics in energy management and that the prediction results would improve if occupancy of the house could be detected (so as to be used as a feature – user behaviour characteristics).

Ma and Ma (2018) in their journal article titled 'A review of forecasting algorithms and energy management strategies for microgrids' showed how various forecasting techniques

can be applied to short term load forecasting on microgrids. Figure 2.5. shows what Ma and Ma (2008) proposed.

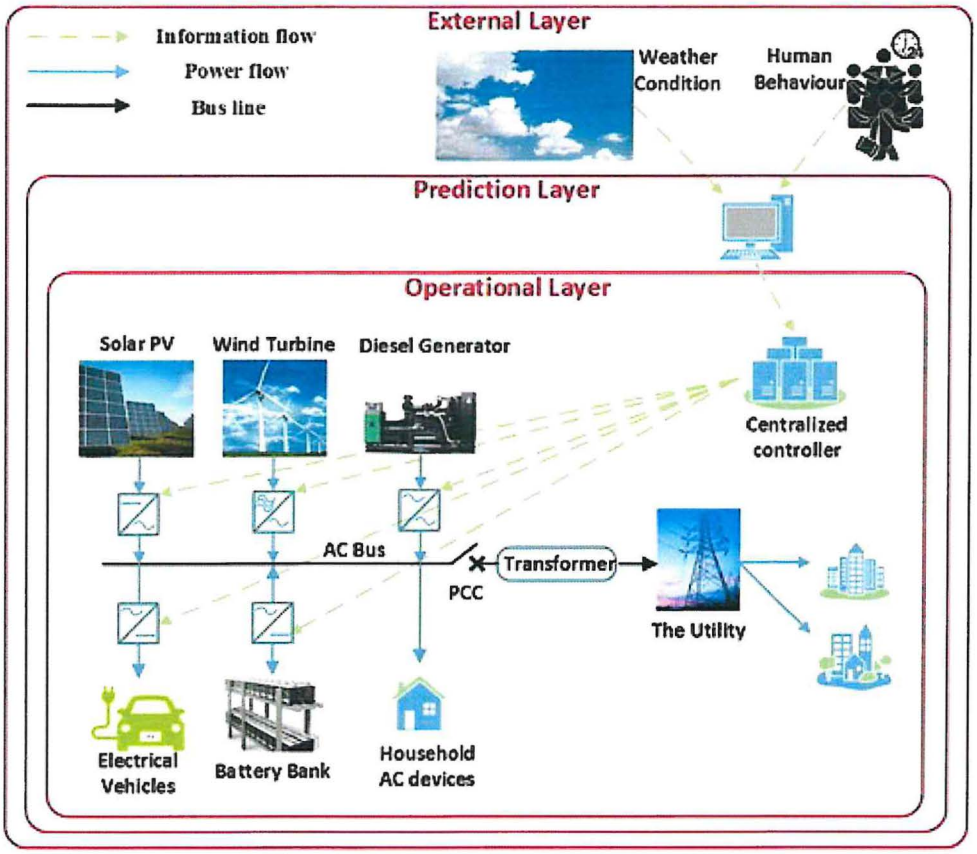


*Figure 2.5. A hierarchical microgrid structure (Ma and Ma, 2018).*

The structure Ma and Ma (2018) proposed had three layers with each having its function as follows:

- External Layer – dedicated to data collection.
- Prediction Layer – running the many advanced forecasting algorithms.
- Operation Layer – runs the different components of the microgrid and allows application of different energy management strategies.

A summary of the forecasting techniques proposed by Ma and Ma (2018) are summarised by Figure 2.6.

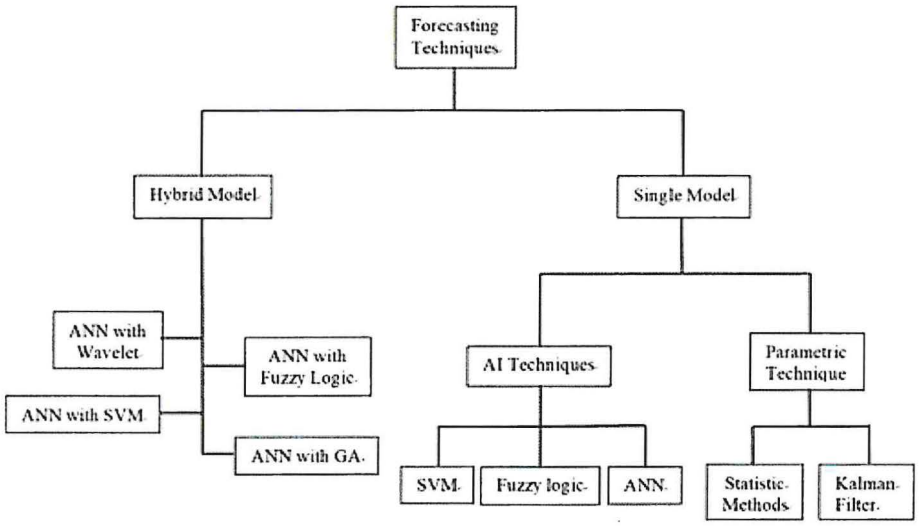


*Figure 2.6. Overview of the forecasting techniques (Ma, & Ma, 2018).*

### 2.3.3 Machine learning algorithms used in Prediction

Machine learning algorithms are techniques based on interdisciplinary concepts incorporating computer science power with statistical inference, including probability and optimization. Figure 2.7. shows a simple learning model's schematic flowchart with common phases known as:

1. Pre-processing stage.
2. Learning stage.
3. Performance Evaluation stage.

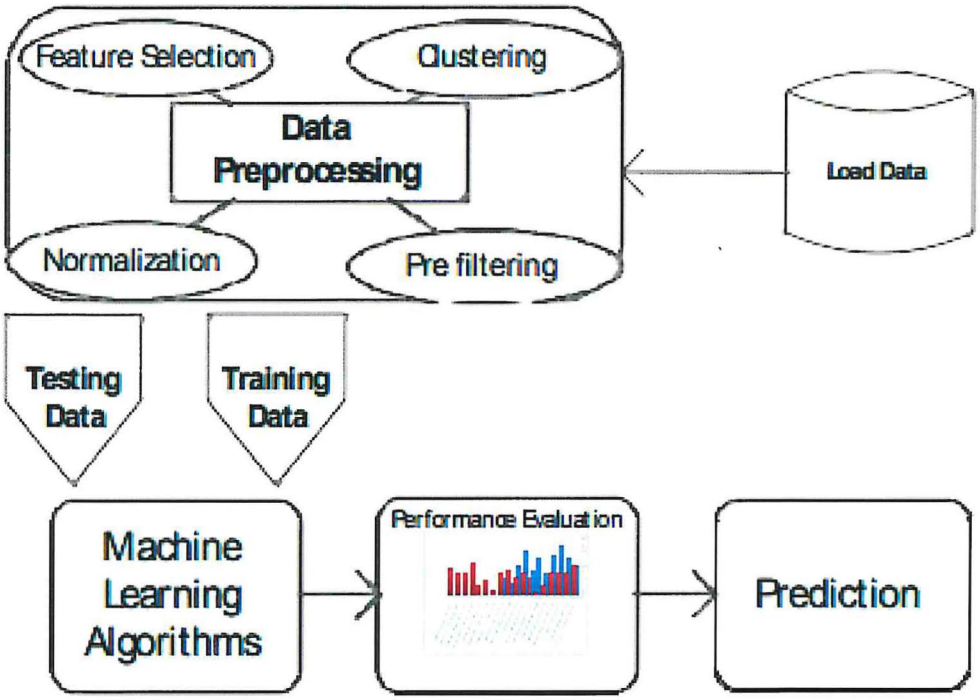


*Figure 2.7. Schematic flowchart of a simple machine learning model (Fallah, Deo, Shojafar, Conti, & Shamshirband, 2018)*

Previous studies in energy consumption prediction have used Support Vector Machines (SVM), Artificial Neural Network (ANN), Decision Trees and other algorithms. According to Amasyali and El-Gohary (2017), approximately 47% and 25% of researchers have applied ANN and SVM to build models respectively. Only 4% of the studies made use of Decision Trees. At the same time, 24% of the studies made use of other statistical algorithms such as Multiple Linear Regression (MLR), Ordinary Least Squares (OLS), and ARIMA. Figure 2.8. shows the distribution of the studies by type of machine learning algorithm.

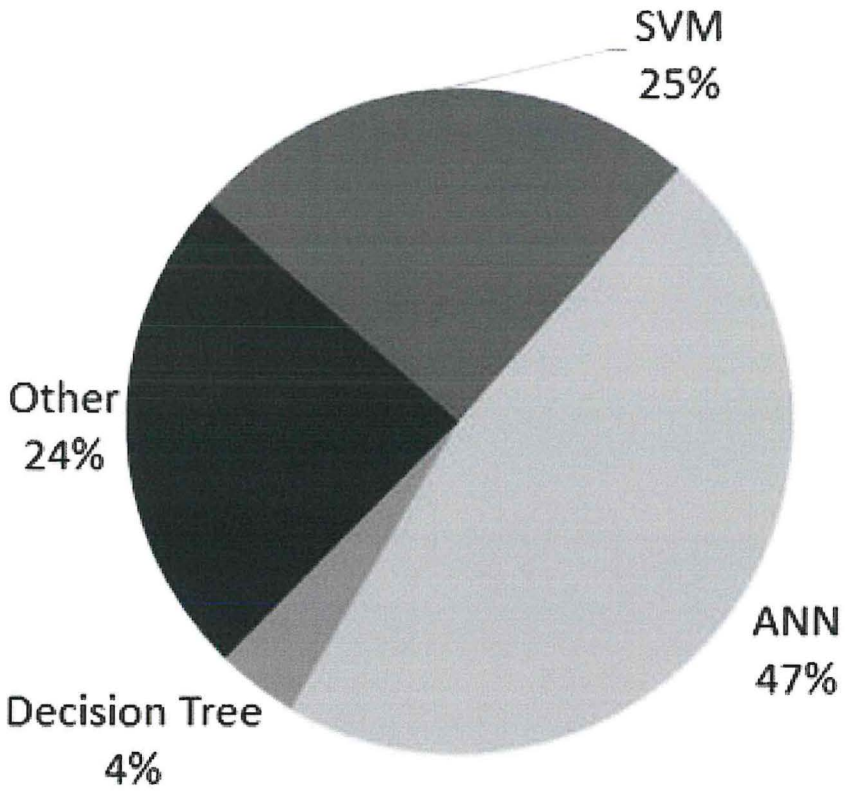


*Figure 2.8. Distribution of the studies by type of machine learning algorithm (Amasyali and El-Gohary, 2017)*

Due to the complexity in building electricity of building energy system, the ability of artificial neural networks to perform non- linear analysis proved to be an advantage in predicting electricity predicting (Ahmad et al., 2014). In a comparative study, Aydinalp, Ugursal and Fung (2004) acknowledged that neural network model was performing better than the engineering model. The disadvantages in using neural networks are that artificial neural network requires training to operate, and for large neural network, high processing time is required (Ahmad *et al.*, 2014).

In a study conducted by Raza and Khosravi (2015), artificial neural network model accuracy depended on certain parameters namely input combination, forecast model architecture, activation functions and training algorithm of the network.

Kaytez, Taplamacioglu, Çam, and Hardalac (2015) stated that several techniques are being used to forecast energy consumption needs and the new techniques being adopted for electricity consumption prediction are Least Squares Support Vector Machines (LS-SVM) and Support Vector Machines (SVM). The Least Square Support Vector Machines (LS-SVM) proved to be a quick and accurate prediction model.

### 2.4 Related work on energy prediction

Various studies have researched various aspect of electrical energy predicting namely nation's annual electricity consumption (Kavaklioglu, 2018), the annual energy consumption of the residential sector (Amiri, Mottahedi, & Asadi, 2015), the annual energy consumption of an industry sector (Kialashaki, & Reisel, 2014) and hourly energy needs using smart metering technology (Jain et al., 2014; Masuda, & Claridge, 2014).

Most of these studies have focused on predicting total energy consumption, cooling and heating of buildings. Tsai et al., (2016) developed a model for predicting monthly electricity consumption based on SVM for the prediction of overall energy based on mean outdoor dry-bulb temperature, relative humidity and global solar radiation. The model used three years data to train the model of four office buildings in Singapore. The results demonstrated that SVM can be employed for building energy consumption prediction.

Kialashaki and Reisel (2014) stated that "the industrial sector is the driving engine of economic development in the USA, and energy consumption in this sector might be considered as the fuel for this engine". Therefore, it was crucial to building an energy consumption predictive model for the industrial sector energy needs. Two models namely artificial neural network model and a multiple linear regression model were developed. Independent variables such as Gross Domestic Product (GDP) and the price of energy carriers was tested in the building of these models. Jain et al. (2014) proposed a model to enhance smart metering devices in the US using an integrated sensor-based predicting model. The result of the study showed that the most operative models are built with hourly consumption.

Li, Peng, and Meng (2010) compared SVM and back propagation neural network (BPNN) on predicting hourly cooling load of an office building. The SVM and BPNN-based models were trained using a month's data. They tested the effectiveness of a set of input parameters and selected the following features: mean temperature of the current hour, mean temperature of the previous hour, and mean temperature of the two hours ago (Amasyali & El-Gohary, 2017). The SVM-based model performed better than the BPNN-based model in predicting the hourly cooling load of the remaining four months

Edwards, New, and Parker (2012) did a study on hourly forecast of energy use for three research houses in Tennessee. The houses were vacant and equipped with automatically regulated equipment for simulated occupancy. They then used artificial neural networks (ANN) and least squares support vector regression (LS-SVR). LS-SVR perform better and achieved the best overall mean absolute percentage error (MAPE) ranged between 16% and 21% for the three households ( Edwards, New, & Parker, 2012).

Rodrigues, Cardeira, and Calado (2014) used ANN to forecast daily and hourly energy consumption. They presented results for two household for a small test period of three days where the MAPE were within 23.5% for the hourly loads (Rodrigues, Cardeira, & Calado, 2014). In a related study Ghofrani, Hassanzadeh, Etezadi-Amoli, & Fadali, (2011) undertook a forecast analysis for various short-term energy horizons i.e. 15, 30 and 60 min-ahead, however the results were reported for only a single test day and the overall MAPE was 12.9%, 18.3% and 30.4% for the three horizons respectively. Gajowniczek and Ząbkowski (2014) predicted a day ahead hourly loads of sample households by using ANN and Support Vector Regression (SVR) models. The results were reported for a single household where the MSE was around 0.1 kWh. It is important to note that example load profiles shown in the study had small volatility with load values less than 0.5 kWh (Gajowniczek & Ząbkowski, 2014).

Machine learning techniques are required to train energy consumption prediction models. Previous studies in predicting building energy consumption have utilized SVM, ANN, decision trees, and/or other statistical algorithms (Amasyali & El-Gohary, 2017). Artificial Intelligence and Machine Learning were the most used methods in predicting energy consumption (Ahmad et al., 2014; Ardjmand, Ghalehkhondabi, Weckman, &

Young, 2016; Demirkoparan, Kaynar, & Özekicioğlu, 2017; Suganthi & Samuel, 2012). Bee colony approach as a swarm-based algorithm was used by Gürbüz, Öztürk and Pardalos (2013) to build an energy consumption predictive model whereas Kavaklioglu (2018) used ANN to forecast electricity consumption in Turkey. Tso and Yau (2007) argued that decision tree and neural network models are suitable in predicting electricity consumption. Ahmad et al. (2014) also used support vector machine and neural networks to forecast electricity consumption. Furthermore, Ahmad et al. (2014), emphasized that Artificial Neural Network is the most used method in predicting electricity energy consumption. They then utilize several machine learning based approaches such as the Support Vector Machines (SVM) and neural networks to carry out accurate forecasting on energy usage.

Accurate energy consumption prediction influences the planning processes undertaken by organisations. In this study, unlike ordinary perspectives in the current research literature, the research has proposed a new prototype for effective building energy use analysis and prediction. In this work, Random Forest, Gradient Boosting, Linear Regression, K-Nearest Neighbours and Support Vector Machine based models are presented to predict energy consumption. Achieving higher accuracy in prediction requires inclusion of all factors that affect the overall electricity consumption. This is accomplished by initially analysing the dataset and identifying features that are contribute to predicting the electricity consumption accurately.

### 2.5 Proposed Model framework

The CBECS 2012 dataset used in developing the model was acquired from the U.S Energy Information Administration website and contains data for 6720 buildings with 1119 features for each building. The features can broadly be classified into Environmental, Physical or Building activities related.

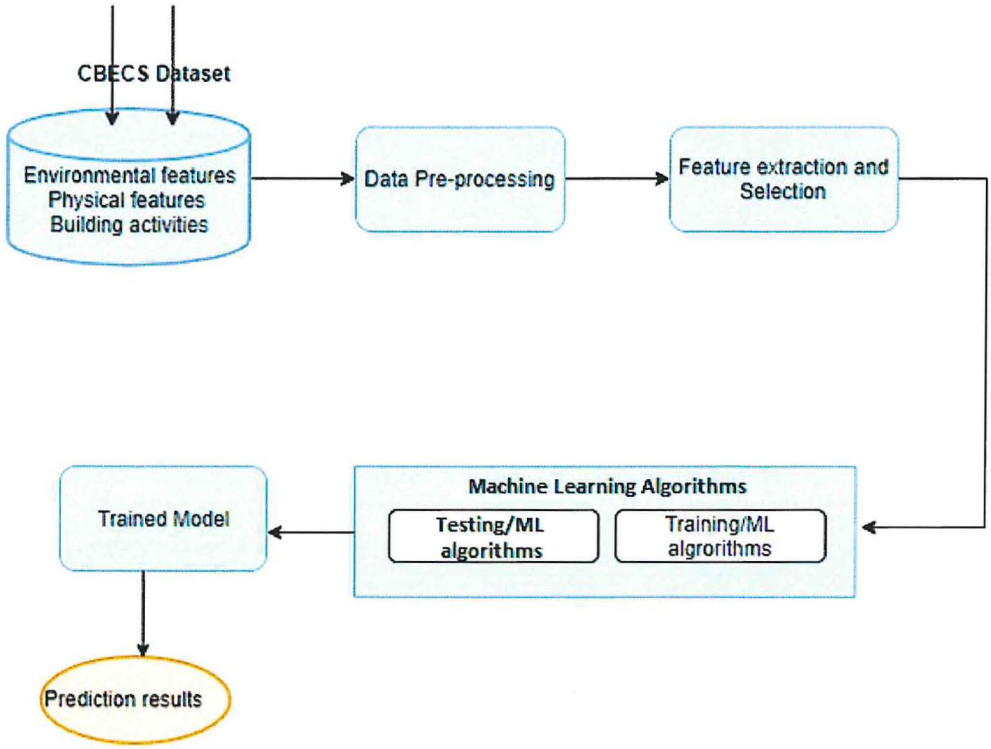


*Figure 2.9. Conceptual Design of the Model*

The first step involves various pre-processing and cleaning steps to check and deal with anomalies, missing data and outliers. For this descriptive statistics and data visualization is key to completing this step. The next step is Features extraction and Selection (also known as Features Engineering). In this step we use statistical tests to determine out of all the available features which are the best for predicting the target variable.

Using the best Features, we train and test various machine learning algorithms in the next step to determine which would perform the best on the data set. The mean absolute error (MAE), root mean square error (RMSE) and $R^2$ (R Squared) of each machine learning algorithm were compared to determine the best i.e. lower MAE/RSME means better performance by the machine learning algorithm and R squared nearer to 1 means better prediction accuracy. Finally, the best machine learning algorithm selected is used for predicting electricity consumption based on new data provided for the best features selected.

# CHAPTER 3: RESEARCH DESIGN AND METHODOLOGY

## 3.1 Introduction

All research is based on certain underlying philosophical assumptions as to what defines' valid' research and which research methods are best suited for knowledge development in a given study. Therefore, it is important to know the steps used to collect, analyse and report research findings. This chapter discusses methodological choice and the research design process and also the design strategies that underpinned this research study.

## 3.2 Methodology

According to Babbie and Mouton (2008), research methodology consists of all the techniques, approaches, and methods selected to conduct a research study. Creswell (2009), further explains that research includes three main components namely developing research question(s), data collection, and analysis which includes attempting to answer the research questions. This section discusses the research approach, data collection, analysis, sample, and ethical consideration which were employed in this study.

## 3.3 Research design

Research design is the general plan of how various aspects of a study will be structured together to ensure that the research problem is addressed adequately, by establishing how the data will be collected, measured and analysed (Thomas, 2010). According to de Vaus (2001), the aim of research design is to endure the evidence obtained enables a researcher to answer research questions. This research focuses on an existing problem with an aim of providing one of the ways of solving it. Therefore, the research design suitable for this is applied research. According to Dudovskiy (2018), applied research design is concerned with "finding a solution for an immediate problem facing a society, or an industrial/business organisation".

### 3.4 Data Analysis and Model Development

In the research some of the data analysis done included descriptive statistics of the features, dealing with missing values and data visualization to look for outliers. For Model Development the approach employed in this research is the use of regression techniques. Regression is the process of predicting a continuous outcome variable (y) based on the value of one or multiple predictor variables (x) (Abdelhamid, 2015). The goal of regression model is to build a mathematical equation that defines y as a function of the x variables. Next, this equation can be used to predict the outcome (y) on the basis of new values of the predictor variables (x).

The study made use of the following Python libraries; Pandas for general data analysis, NumPy for array handling, Matplotlib & Seaborn for chart generation (data visualization) and Scikit-learn for the Machine Learning algorithms. The Scikit-learn library is a library with implementations of various Machine Learning algorithms and error metrics for determining the quality of the predictions. The data analysis and model development steps are summarized by Figure 3.1.

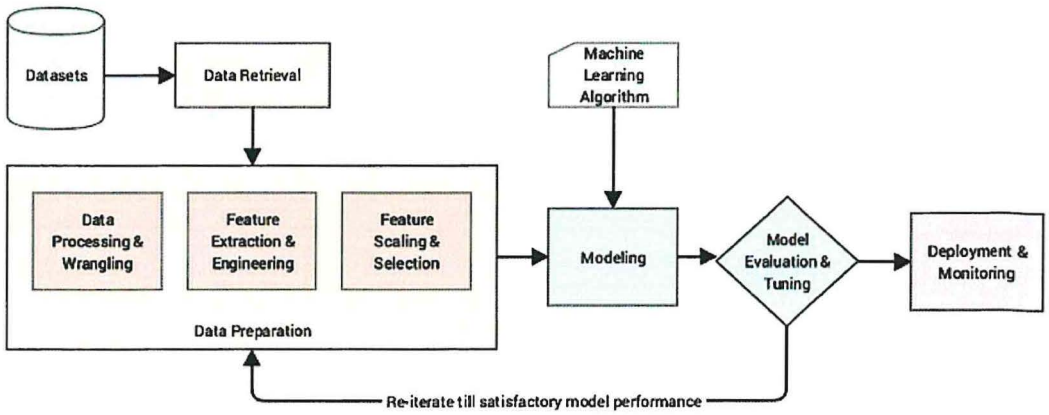


*Figure 3.1. A standard machine learning pipeline (Sarkar, Bali, & Sharma, 2017)*

### 3.4.1. Data acquisition

Data collection in energy consumption consists of collecting past/historical data to be used in developing the model (Amasyali & El-Gohary, 2017). The dataset used in this research was the 2012 CBECS data set released every 5 years by the U.S. Energy Information Administration:
(https://www.eia.gov/consumption/commercial/data/2012/index.php?view=microdata).
The latest dataset from 2012 to 2016 contained 6720 rows of data. Each row contained the features of an individual representative building through the CBECS 'Building Survey' questionnaire. These features included information such as the number of employees in the building, the square footage of the building, the principal building activity (PBA), heating degree day (HDD), cooling degree day (CDD) to mention a few (U.S. Energy Information Administration [EIA], n.d.).

### 3.4.2. Data Pre-processing

Data pre-processing comprise of data cleaning, data reduction and data transformation. Prior to data analysis, we removed 1586 records (5134 remaining), which included missing values for a precise analysis. Then the data was filtered to remain with 871 buildings whose principle building activity was Office. Due to the volume of the datasets as well as the potential existence of the correlations among the features, it was essential to reduce and select the most essential features.

### 3.4.3. Feature extraction and selection

The original dataset contained 1119 features and the research had to determine which features best predicted the target annual electricity consumption. Using domain knowledge, the Features were reduced to 26 as shown by Table 3.1. Then following features selection techniques were applied namely Univariate Selection, Feature Importance, Correlation Matrix with Heat map and Pearson's Correlation Coefficient which resulted in a final 4 Features used for modelling.

Table 3.1 *Selected Features using Domain Knowledge from the Dataset*

| Variable name | Label | Variable type |
|---|---|---|
| PBA | Principal building activity | Char |
| PBAPLUS | More specific building activity | Char |
| MAINCL | Main cooling equipment | Char |
| WTHTEQ | Water heating equipment | Char |
| SQFT | Square footage | Num |
| MONUSE | Months in use | Num |
| NWKER | Number of employees | Num |
| ELHT1 | Electricity used for main heating | Char |
| ELHT2 | Electricity used for secondary heating | Char |
| ELCOOL | Electricity used for cooling | Char |
| ELWATR | Electricity used for water heating | Char |
| ELCOOK | Electricity used for cooking | Char |
| RFGEQP | Refrigeration | Char |
| PCTERMN | Number of computers | Num |
| LAPTPN | Number of laptops | Num |
| PRNTRN | Number of printers | Num |
| SERVERN | Number of servers | Num |
| COPIERN | Number of photocopiers | Num |
| TVVIDEON | Number of TV or video displays | Num |
| FLUOR | Fluorescent bulbs | Char |
| BULB | Incandescent bulbs | Char |
| HALO | Halogen bulbs | Char |
| LED | Light-emitting diode (LED) bulbs | Char |
| HDD65 | Heating degree days (base 65) | Num |
| CDD65 | Cooling degree days (base 65) | Num |
| **Target name** | **Label** | **Variable type** |
| ELCNS | Annual electricity consumption (kWh) | Num |

### 3.4.4. Data Splitting

The dataset was split into training and the test dataset using the holdout method. In this method, a given dataset is split into two, the test set - 30% and training set - 70%. The training dataset is used to train the model and the unseen test dataset is used to test its

predictive performance (Zheng, 2015). The main reason for the separation between test and training is to ensure that the model performs well with data on which it has not been trained.

Data splitting was accomplished as follows:

Data set size was 871 records of Commercial Office buildings

Data set split = 70% : 30%

Training data set = 70% * Data set size = 610 records (buildings)

Testing data set = 30% * Data set size = 261 records (buildings)

### 3.4.5. Development of the Model

The model was developed using Python's Scikit-learn library which has various machine learning algorithms such as Gradient Boosting, Linear Regression, Random Forest, K-Nearest Neighbours and Support Vector Machines. Training data was used for fitting the model which was 70% of the data set and for testing the model the remaining 30% of the data set was used to determine how good the model performed. The process of developing the model is depicted by Figure 3.2.

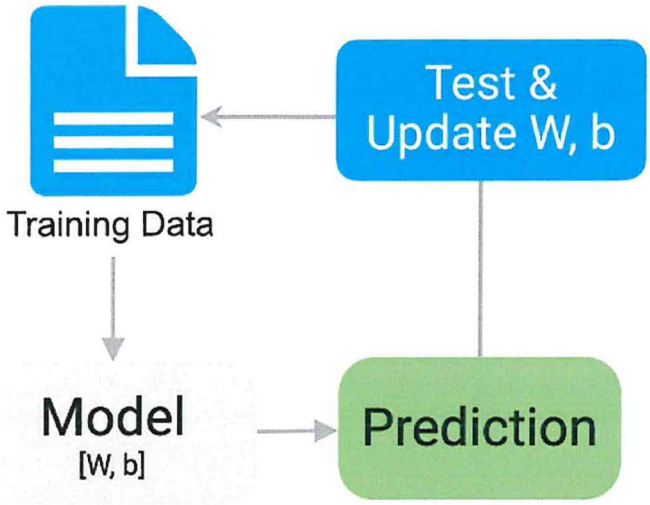


*Figure 3.2. Machine learning model process (Yufeng, 2017)*

### 3.4.6. Performance evaluation

The two main measures of performance of the model used in the research were Mean Absolute Error (MAE) - to compare performance of the different regression-based machine learning algorithms and R Squared ($R^2$) – to determine accuracy of the different regression based machine learning algorithms. Root Mean Square Error (RMSE) was also used - to compare performance of the different regression-based machine learning algorithms.

In statistics Mean Absolute Error (MAE) is the result of measuring the difference between two continuous variables and in machine leaning it summarizes the quality of a model (Minka, 2018). Equation 1 depicts MAE.

$$mae = \frac{\sum_{i=1}^{n} abs(y_i - \lambda(x_i))}{n} \quad (1)$$

$R^2$ is calculated as the square of correlation between the observed y values and the predicted $\hat{y}$ values. R-squared is also defined as the 'how much of the total variation is described by the regression line' or 'how much of the total variation in outcome target variable(y) is described by the independent predictors(x) (The Minitab Blog, 2013). Equation 2 depicts $R^2$.

$$\hat{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \quad (2)$$

The RMSE was used in various areas to evaluate prediction accuracy by considering the fundamental differences between the values forecasted by a model and the true values observed. These differences are referred to as residuals, and RMSE is often used to collate residuals into a single predictive power measure. An estimator's RMSE is defined with respect to an estimated parameter. Equation 3 depicts RMSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \quad (3)$$

RMSE is used to measure of the differences between values (sample and population values) predicted by a model and the values actually observed. A 'better' model is then chosen based on the one with lesser RMSE. RMSE is a relative measure. There is no "good", or "bad" RMSE, no "small" or "big" RMSE, it depends on the data (StackExchange, 2017).

### 3.5 System design and methodology

The prototype was developed using the Rapid Application Development (RAD) system development methodology which focus more on development rather than planning tasks (Despa, 2014). The RAD methodology phases are requirements gathering, user design, development of the prototype, evaluation and delivery of partially implemented software awaiting for customer feedback. This system methodology approach puts customer satisfaction at the core and also faster development time (Sharma, Sarkar, & Gupta, 2012). RAD is depicted by Figure 3.3.

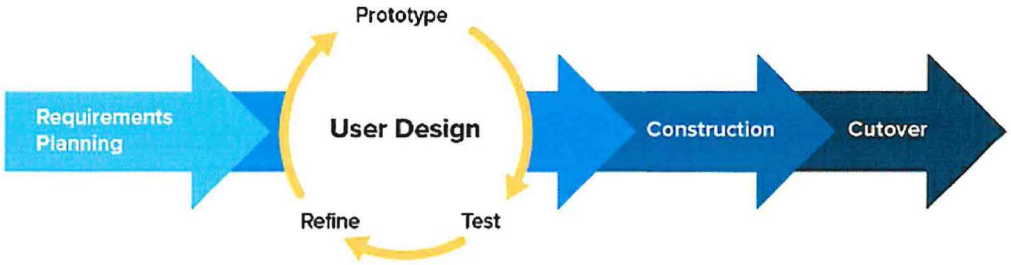


*Figure 3.3. Rapid application development cycle (Kissflow, 2018)*

### 3.5.1 Requirements Gathering

This was the first and critical step of the research's software development process. It involved understanding of the current problem, defining the requirements and finalizing the requirements of every stakeholder.

### 3.5.2 User design stage

Once all the requirements and objectives had been gathered and established, the next step was to start development by building out the user design through various prototype iterations.

### 3.5.3 Construction Phase

The next phase after designed of the proposed system was construction phase. This phase comprised of: Preparation for fast construction, Program and Application development, Coding, Integration, and System Testing. The prototype for this research was implemented using various Python libraries for developing the energy prediction model and Flask Web Framework for developing the user interface and run the prediction model in the background.

### 3.5.4 Cut over stage

This was the implementation phase where developed prototype was deployed. This step involved user training and evaluation of the prototype's performance.

## 3.6 Ethical Considerations

This study utilizes publicly available online data made available for researchers which is already anonymized. Therefore, the confidentiality and anonymity of the users is already adhered to.

# CHAPTER 4: SYSTEM DESIGN AND ARCHITECTURE

## 4.1 Introduction

Computer architecture is a detailed set of methods and rules that describes the functionality, organization and implementation of a computer systems (Dean, Patterson, & Young, 2018; Woods, 2016). The purpose of this study was to build a predictive model and to develop a prototype to predict the energy consumption. This chapter consists of discussing the system analysis and design used in the development of the prototype as well as the System Architecture.

## 4.2 System Requirements Analysis

Requirements analysis in software or systems engineering consists of gathering the conditions or needs of the new system or software to be developed (Chapman, Bahill, & Wymore, 2018). The system requirement analysis is categorized into two namely functional requirements and non-functional requirements.

### 4.2.1 Non-Functional Requirements

Non-Functional Requirements consists of the quality attributes of the developed system i.e. how the system should behave (Eriksson, 2012). The Non-Functional Requirements of the prototype were:

i. Since the prototype is hosted by a webserver it should be available for use at all times online for Building Owners and other interested individuals.
ii. The prototype should be scalable for future adjustment in the prediction model as more data is acquired and new key features emerge.
iii. The Response time of the prototype's prediction should be fast (less than 5 seconds).
iv. The Prototype should be easily maintainable i.e. code should be clear with comments explaining what sections of code do.

### 4.2.2 Functional Requirements

Functional requirements explain what has to be done by identifying the necessary task, action or activity that must be accomplished (Oinas-Kukkonen & Harjumaa, 2018). The Functional Requirements of the prototype were:

i. Link to website with Heating Degree Days for all locations in the world should be clickable and active.
ii. Prototype should allow the owner of the building to enter data on building's features – Surface Area of Building, Number of Employees in the Building, Number of Computers used in the Building and Heating Degree days of the Building's location.
iii. Prototype should only accept numeric data for the building's features.
iv. Clicking the 'Predict Electricity Consumption' button should only work if the building's features numeric data is present/provided.
v. Prototype should predict Electricity consumption using the data entered by the building owner.
vi. Prototype Maintenance should be available for the System Administrator via the Webserver.

## 4.3 System Design and Architecture

In this section the system is designed using different kinds of graphical representations of System components and processes. A brief system Architecture is also discussed.

### 4.3.1 System Architecture

The architecture of a system presents its interactions between the input, processes and outputs anticipated. The architecture of the prototype consists of a front-end interface, where building features will be uploaded, and the resulting predicted electricity consumption viewed. The goal of the system architecture diagram is to represent the general flow of information from the user to the systems on the web server as depicted by Figure 4.1.

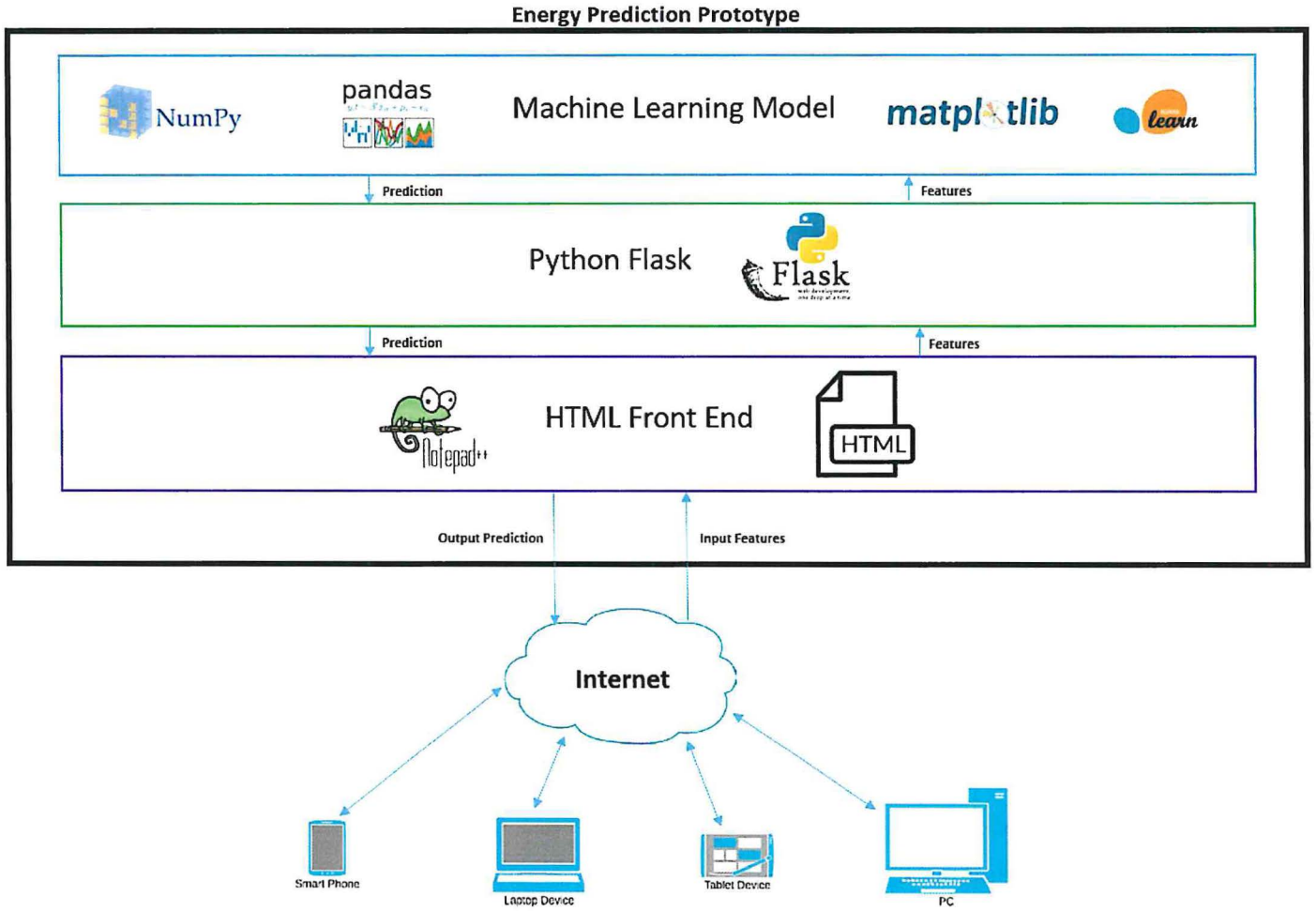


*Figure 4.1. System Architecture*

### 4.3.2 Use Case Representation

A use case diagram is graphic description of the interactions among the elements of a system. It is used to identify, categorize, explain and organize system requirements (Khurana, Chhillar, & Chhillar, 2016). The use case diagram indicates how the two core users, the Building Owner and the System Administrator will interact with the system. The System Administrator will monitor the performance of the prototype and ensure it is working and available online for Building Owners to access. The Building Owner will be able to view the result of electricity prediction. Prior to that, the Building Owner will have to provide the surface area, provide the number of employees, provide the number of computers and heat degree days of the building's location as depicted by Figure 4.2.

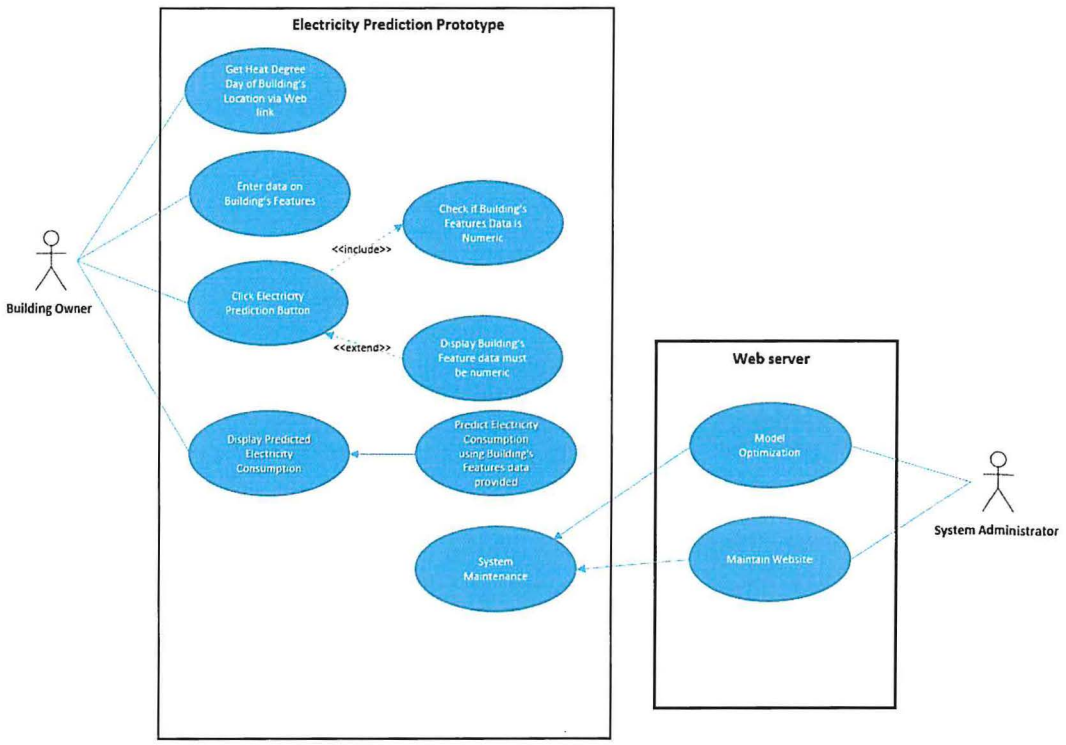


*Figure 4.2. Use Case Diagram*

### 4.3.3 Data Flow Diagram (DFD)

A data flow consists of representing the processes, external entities and flow of data in a system. It also represents the connecting data flows (Ibrahim & Yen, 2010).

#### 4.3.3.1 Context Diagram

It is a basic overview of the entire system being analysed and shows the system as a single high-level process with its relationship to external entities (Lucid Software, 2019). The context diagram of the prototype in this research is depicted by Figure 4.3.

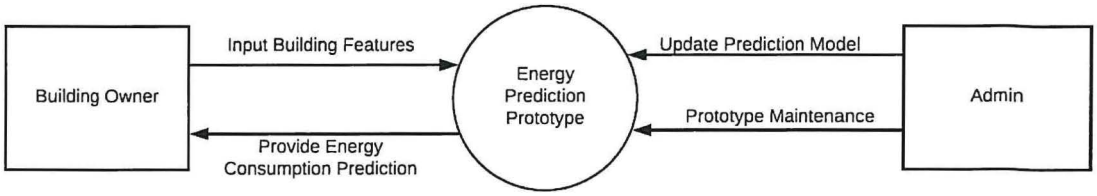


*Figure 4.3. Context Diagram*

#### 4.3.3.2 Level 1 Data Flow Diagram

This provides more details as compared to Context Diagram and highlights the main functions done by the system (Lucid Software, 2019). The context diagram of the prototype in this research is depicted by Figure 4.4.

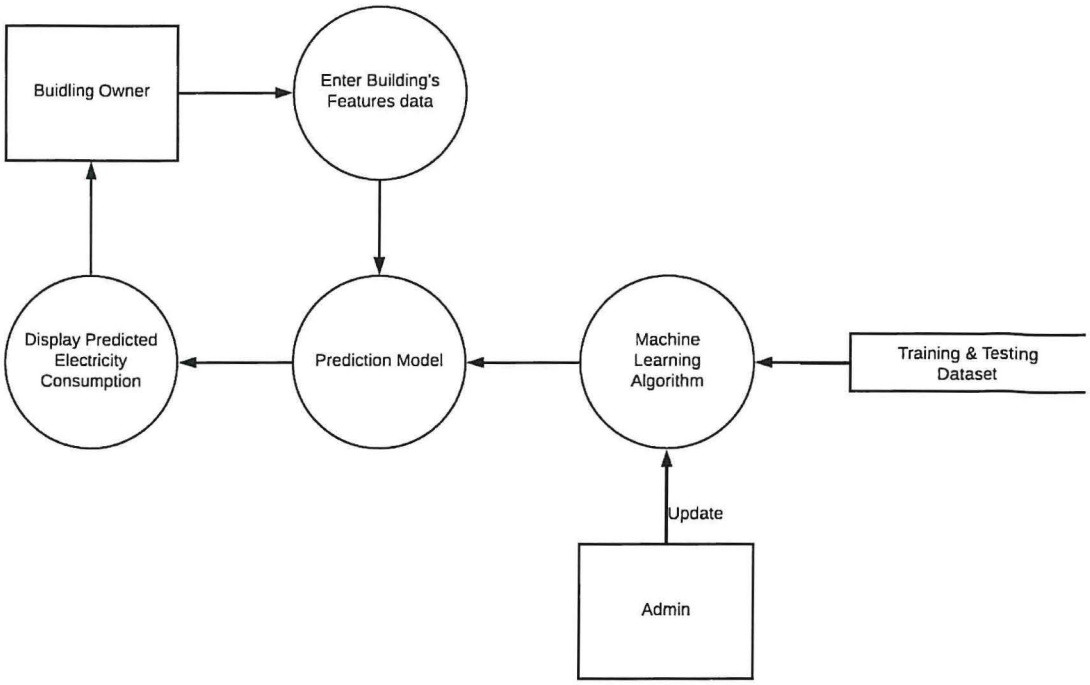


*Figure 4.4. Data Flow Diagram Level 1*

### 4.3.4 Sequence Diagram

A sequence diagram is an interaction diagram that shows how processes interact with each other and in what order. For this research the Sequence Diagram is depicted by Figure 4.5.

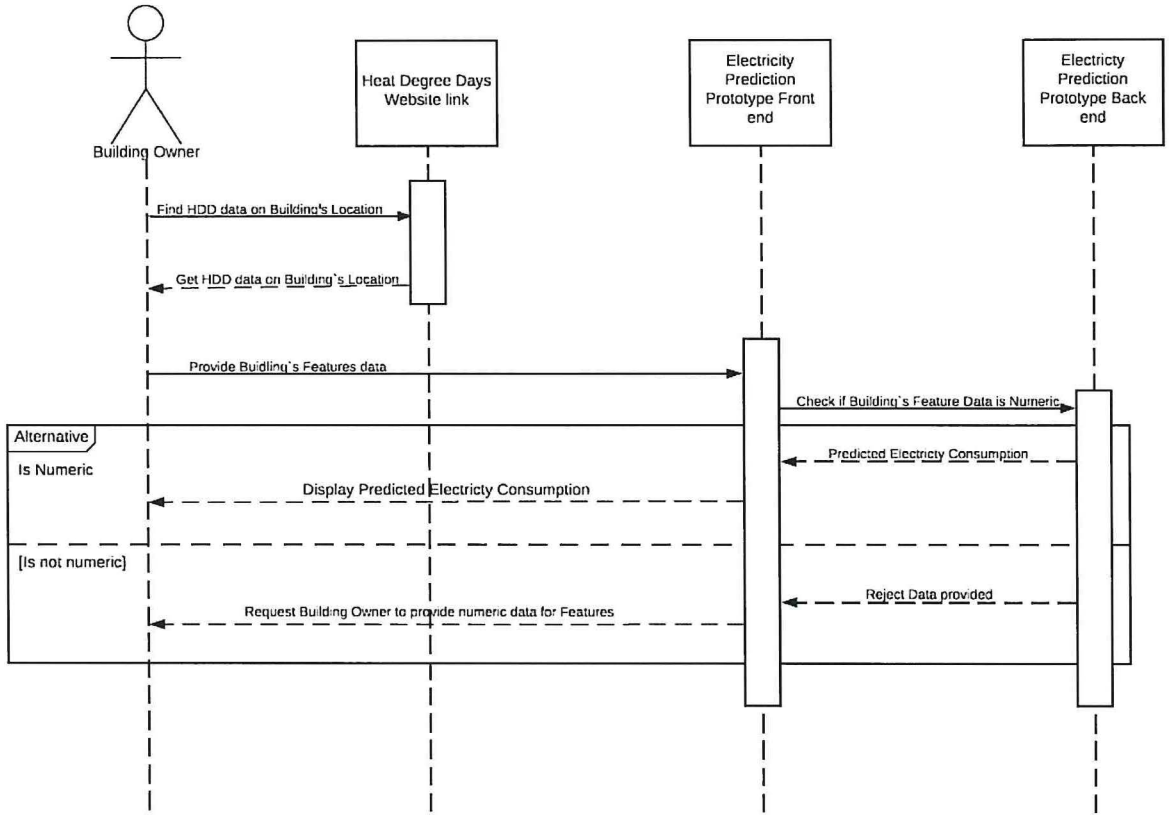


*Figure 4.5. Sequence Diagram*

### 4.3.5 Activity Diagram

An activity diagram is essentially a flow diagram showing a system's activities (Lucid Software, 2019). Figure 4.6. illustrates the activity diagram of the proposed prototype.

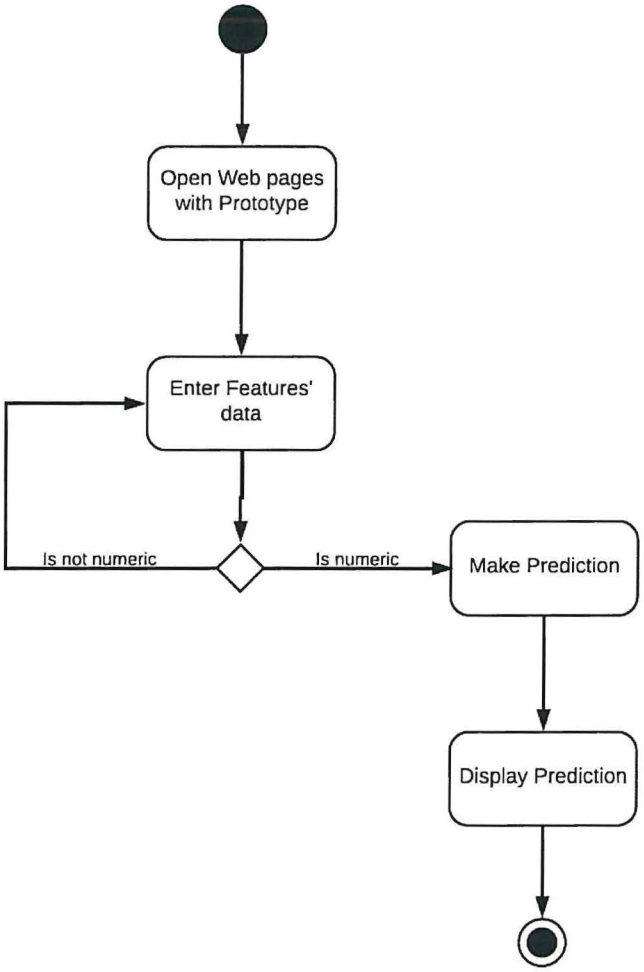


*Figure 4.6. Activity diagram*

# CHAPTER 5: SYSTEM IMPLEMENTATION AND TESTING

## 5.1 Introduction

This chapter discusses how the system was implemented and tested which involved highlighting specific software and hardware configuration settings, features engineering (features extraction and selection), training & testing of different Machine Learning Algorithms, evaluation & selection of the Machine Learning Algorithms for the model, development of the prototype using Flask and finally testing of the prototype.

## 5.2 Hardware and Software Environment

The implementation was done on a Windows 10 machine with the programming language of choice being Python version 3.7.3. The development environment used was Jupyter Notebook version 5.7.4. Notebook++ was used for developing the HTML front end of the prototype. Python packages used and their versions are summarized by Table 5.1 and the minimum system requirements by Table 5.2

Table 5.1 *Python Packages and Version used in developing the Prototype.*

| Package | Version |
|---|---|
| Flask | 1.0.2 |
| matplotlib | 3.0.3 |
| numpy | 1.16.2 |
| pandas | 0.24.2 |
| pipenv | 2018.11.26 |
| scikit-learn | 0.20.3 |
| seaborn | 0.9.0 |
| virtualenv | 16.4.3 |

Table 5.2 *Minimum Systems Requirements.*

| Requirement | Minimum |
|---|---|
| Processors | Intel Atom® processor or Intel® Core™ i3 processor |
| Disk space | 1 GB |
| Operating systems | Windows* 7 or later, macOS, and Linux |

### 5.3 Features Engineering

Features Engineering is the process of using domain knowledge and statistical tests to select the best features that would make machine learning algorithms work (DisplayR, n.d.; Shekhar, 2018). In this research domain knowledge was used to reduce the features from 1119 to 26. Then the following were applied - Univariate Selection, Feature Importance, Correlation Matrix with Heat map and Pearson's Correlation Coefficient which resulted in a final 4 Features used for modeling.

#### 5.3.1 Univariate Selection

Statistical test chi-squared ($chi^2$) was used to select 10 of the best non-negative features that had the strongest relationship with the output variable. The results are on Figure 5.1.

| | Specs | Score |
|---|---|---|
| 4 | SQFT | 5.165058e+08 |
| 16 | SERVERN | 7.452009e+06 |
| 6 | NWKER | 1.568935e+06 |
| 13 | PCTERMN | 1.411788e+06 |
| 23 | HDD65 | 8.706712e+05 |
| 14 | LAPTPN | 6.845760e+05 |
| 15 | PRNTRN | 3.895415e+05 |
| 18 | TVVIDEON | 3.623493e+05 |
| 17 | COPIERN | 8.365097e+04 |
| 1 | PBAPLUS | 7.100296e+02 |

*Figure 5.1. Top 10 best features using SelectKBest class*

### 5.3.2 Feature Importance

An Inbuilt Tree Based Classifiers was used to calculate a score for each of the feature in the data and the higher the score the more important/ relevant the feature was towards the variable. The results are on Figure 5.2.

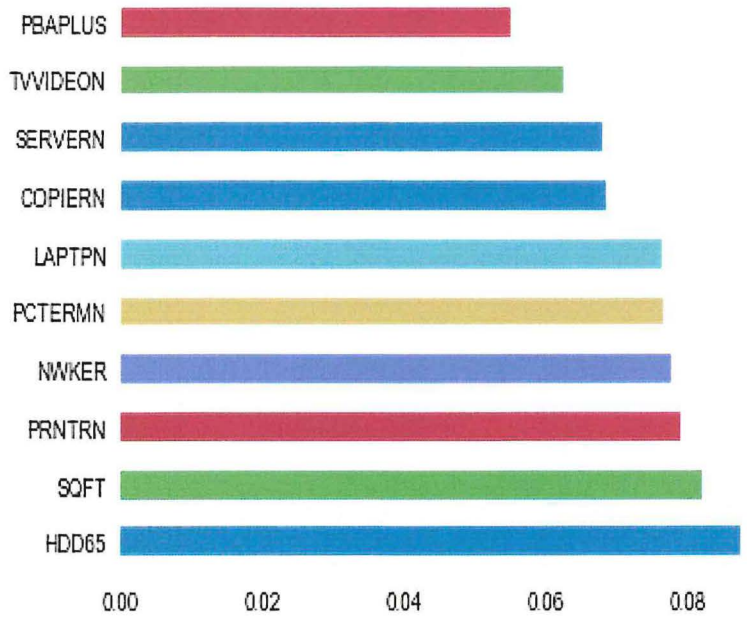


*Figure 5.2. Feature importance*

### 5.3.3 Correlation Matrix with Heatmap

A Heatmap which showed Correlation which is how the features are related to each other and the target variable was generated. The results are on Figure 5.3.

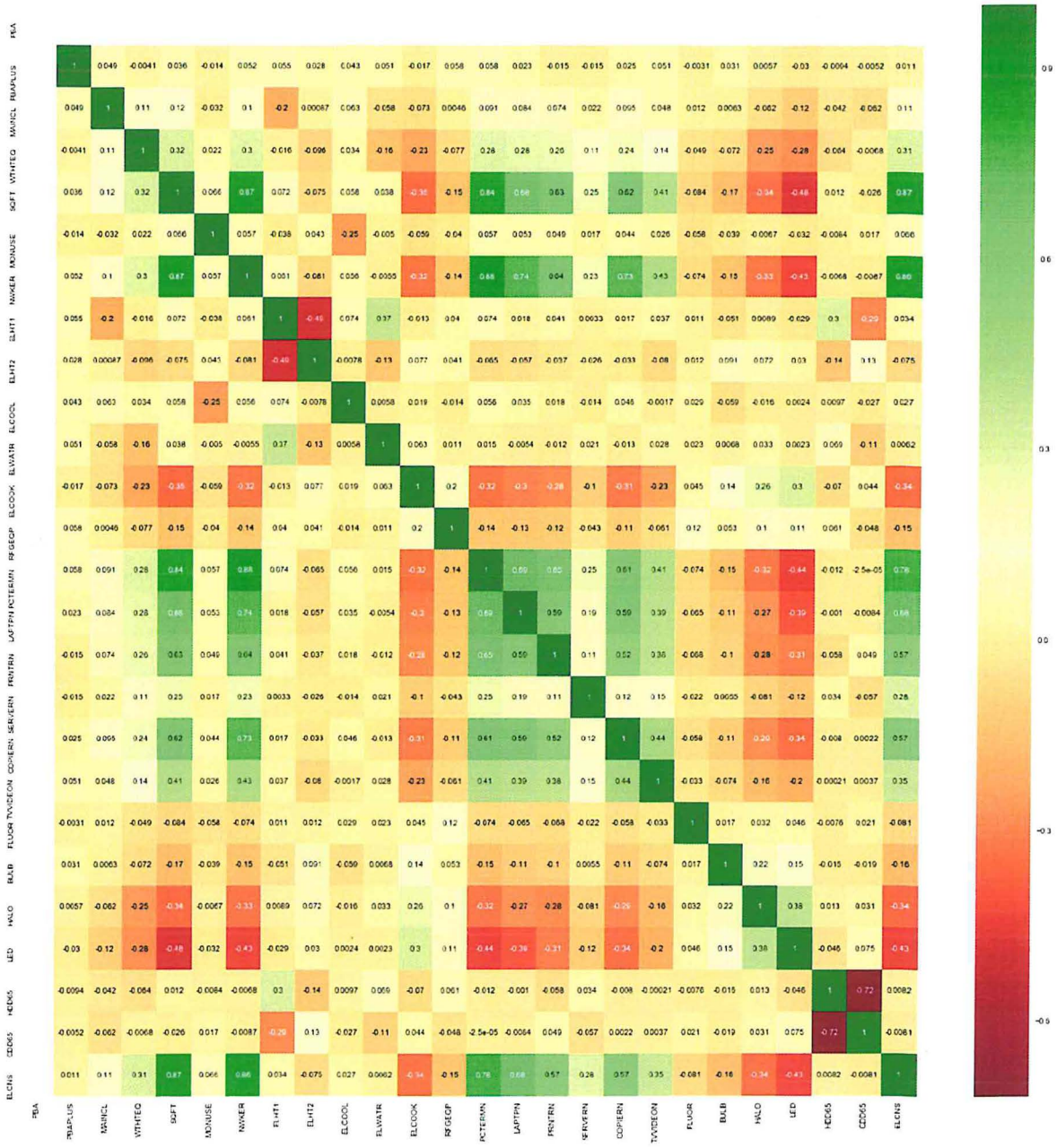


*Figure 5.3. Correlation matrix with Heatmap*

### 5.3.4 Pearson's Correlation Coefficient

The most negative and positive correlations were calculated using this method and the results are on Figure 5.4.

```
LED       -0.431164
HALO      -0.343388
ELCOOK    -0.341065
BULB      -0.162104
RFGEQP    -0.148399
FLUOR     -0.080782
ELHT2     -0.075321
CDD65     -0.008106
ELWATR     0.006187
HDD65      0.008196
Name: ELCNS, dtype: float64

WTHTEQ        0.313579
TVVIDEON      0.345949
COPIERN       0.568628
PRNTRN        0.574814
LAPTPN        0.678329
PCTERMN       0.780344
NWKER         0.858616
SQFT          0.868655
ELCNS         1.000000
PBA                NaN
Name: ELCNS, dtype: float64
```

*Figure 5.4. Pearson's Correlation Coefficient*

### 5.4 Training and Testing of different Regression-based Machine Learning Algorithms

This involved the following steps done on Jupyter Notebook:

i. Splitting the data into training and testing sets as depicted by Figure 5.5.

ii. Normalization of the Features so that different units do not affect the algorithms as depicted by Figure 5.6.

iii. Training and testing of different Regression-based Machine Learning algorithms as depicted by Figure 5.7. The Mean Absolute Error as depicted by Figure 5.8. was calculated to see which algorithm was best suited based on the test data set (the lower the MAE the better suited the ML algorithm is to the data).

```
In [48]: #Splitting Data into training and testing sets

         # Extract the buildings with no ELCNS and the buildings with a ELCNS
         no_score = finaldata1[finaldata1['ELCNS'].isna()]
         score = finaldata1[finaldata1['ELCNS'].notnull()]

         print(no_score.shape)
         print(score.shape)

         (0, 5)
         (871, 5)

In [49]: # Separate out the features and targets
         features = finaldata1.drop(columns='ELCNS')
         targets = pd.DataFrame(finaldata1['ELCNS'])

In [50]: from sklearn.model_selection import train_test_split

         # Split into 70% training and 30% testing set
         X, X_test, y, y_test = train_test_split(features, targets, test_size = 0.3, random_state = 42)

         print(X.shape)
         print(X_test.shape)
         print(y.shape)
         print(y_test.shape)

         (609, 4)
         (262, 4)
         (609, 1)
         (262, 1)
```

*Figure 5.5. Splitting data into Training and Testing set*

```
In [53]: train_features = X
         test_features = X_test
         train_labels = y
         test_labels = y_test

In [54]: #Imputing missing values and scaling values
         from sklearn.preprocessing import Imputer, MinMaxScaler

         #ML
         from sklearn.linear_model import LinearRegression
         from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
         from sklearn.svm import SVR
         from sklearn.neighbors import KNeighborsRegressor

         #Hyperparameter tuning
         from sklearn.model_selection import RandomizedSearchCV
         from sklearn.model_selection import GridSearchCV

In [55]: #Lets normalize the features so that the different units do not affect the algorithms.

         #Creating a scaler object with a range of 0 - 1
         scaler = MinMaxScaler(feature_range = (0, 1))

         #Fit on the training data
         scaler.fit(X)

         #Transform both the training and testing data
         X = scaler.transform(X)
         X_test = scaler.transform(X_test)

         #Converting y to one-dimensional array
         y = np.array(train_labels).reshape((-1,))
         y_test = np.array(test_labels).reshape((-1,))
```

*Figure 5.6. Normalization of the Features*

```
In [57]:

         #Linear Regression
         lr = LinearRegression()

         lr_mae = train_test_evaluate(lr)

         print ('Linear Regression Mean Absolute Error: %0.4f' %lr_mae, '\n')

         #Support Vector Machines
         svm = SVR(C = 1000, gamma = 0.1)
         svm_mae = train_test_evaluate(svm)

         print ('SVM Mean Absolute Error: %0.4f' %svm_mae, '\n')

         #Random Forest
         random_forest = RandomForestRegressor(random_state = 60)
         random_forest_mae = train_test_evaluate(random_forest)

         print ('Random Forest Mean Absolute Error: %0.4f' %random_forest_mae, '\n')

         #Gradient Boosted Machines
         gradient_boosted = GradientBoostingRegressor(random_state = 60)
         gradient_boosted_mae = train_test_evaluate(gradient_boosted)

         print ('Gradient Boosted Regression Mean Absolute Error: %0.4f' %gradient_boosted_mae, '\n')

         #K-Nearest Neighbours
         knn = KNeighborsRegressor(n_neighbors = 10)
         knn_mae = train_test_evaluate(knn)

         print ('K Nearest Neighbors Mean Absolute Error: %0.4f' %knn_mae)
```

*Figure 5.7. Training of different Regression-based Machine Learning algorithms*

```
In [56]: #Calculating the Mean Absolute Error using a Funciton
         def mae(y_true, y_pred):
             return np.mean(abs(y_true - y_pred))

         #Training, testing and evaluating a model
         def train_test_evaluate(model):

             #Train
             model.fit(X,y)

             #Test
             model_pred = model.predict(X_test)

             #Evaluate
             model_mae = mae(y_test, model_pred)

             #Return performance metric
             return model_mae
```

*Figure 5.8. The Mean Absolute Error Function*

## 5.5 Evaluation of the different Regression-based Machine Learning Algorithms

The MAE was visualized using Seaborn a Python Data Visualization Package on Jupyter Notebook as depicted by Figure 5.9. The visualization showed that Random Forest and Gradient Boosted Machines as depicted by Figure 5.10. had the lowest MAE and thus were better suited for the final Model.

```
In [59]: plt.style.use('seaborn')

         #A dataframe to hold the results
         model_comparison = pd.DataFrame({'model': ['Linear Regression',
                                                    'Support Vector Machine',
                                                    'Random Forest',
                                                    'Gradient Boosted Machines', 'K-Nearest Neighbours'],
                                          'mae': [lr_mae, svm_mae, random_forest_mae, gradient_boosted_mae, knn_mae]})

         #Horizontal bar chart of MAE
         model_comparison.sort_values('mae', ascending = True).plot(x = 'model',
                                                                    y = 'mae',
                                                                    kind = 'barh',
                                                                    color = 'black',
                                                                    edgecolor = 'black')

         #Plot formatting
         plt.ylabel('Models', size = 14)
         plt.yticks(size = 12)
         plt.xlabel('Mean Absolute Error')
         plt.xticks(size = 12)
         plt.title('Model Comparison on Test MAE', size = 16)

         #We can see that there is a use for ML because all the models significantly outperform the baseline.
```

*Figure 5.9. Using Seaborn to visualize the MAE of the different ML algorithms*

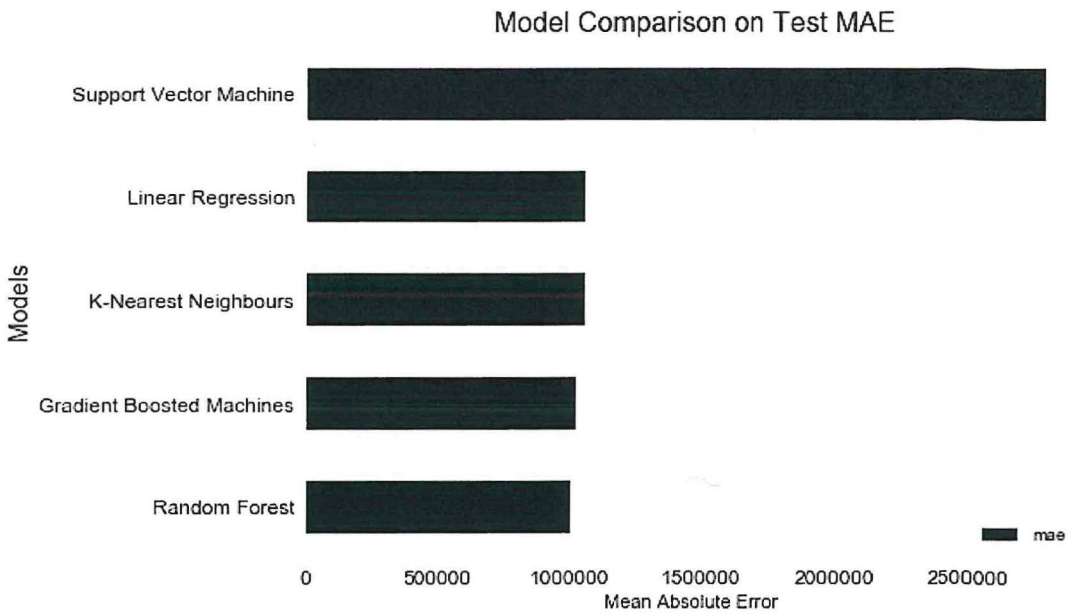


*Figure 5.10. Visualization the MAE of the different ML algorithms*

### 5.6 Selection of the best Machine Learning Algorithm for the Prototype

To select the best Machine Learning Algorithm the research used statistical test R Squared which was used to determine the accuracy of the Machine Learning Algorithm. Since the difference between the MAE values for the Linear Regression, K-Nearest Neighbour, Gradient Boosted Machines and Random Forests algorithms was not significant the R Squared and Accuracy was calculated on Jupyter Notebook for each, to determine which algorithm was best for the prototype as summarized the below sub-sections. Gradient Boosting Machines was selected with an R Squared value of 0.97 and Accuracy of 97%.

### 5.6.1 Linear Regression

The R Squared result for Linear Regression was 0.783 as depicted by Figure 5.11. The Accuracy was 78% as calculated using Equation 4 below:

$$Accuracy = r^2 * 100 \quad (4)$$

```
In [67]: import numpy as np
         import statsmodels.api as sm
         import statsmodels.formula.api as smf
         import pandas as pd
         from sklearn.model_selection import train_test_split

         Electconsumption = finaldata1['ELCNS']
         sqfoot = finaldata1['SQFT']
         noworkers = finaldata1['NWKER']
         noofpc = finaldata1['PCTERMN']
         degreehd = finaldata1['HDD65']

         y = Electconsumption
         x = np.column_stack((sqfoot,noworkers,degreehd,noofpc))
         x = sm.add_constant(x, prepend=True)

         x_train,x_test,y_train,y_test=train_test_split(x,y,test_size = 0.3, random_state = 42)

         results = smf.OLS(y_train,x_train).fit()
         print(results.summary())
```

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | ELCNS | R-squared: | 0.783 |
| Model: | OLS | Adj. R-squared: | 0.782 |
| Method: | Least Squares | F-statistic: | 545.6 |
| Date: | Tue, 16 Apr 2019 | Prob (F-statistic): | 6.86e-199 |
| Time: | 15:30:36 | Log-Likelihood: | -9774.7 |
| No. Observations: | 609 | AIC: | 1.956e+04 |
| Df Residuals: | 604 | BIC: | 1.958e+04 |
| Df Model: | 4 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 8.679e+04 | 2.08e+05 | 0.418 | 0.676 | -3.21e+05 | 4.95e+05 |
| x1 | 8.6679 | 0.701 | 12.365 | 0.000 | 7.291 | 10.045 |
| x2 | 3583.6930 | 328.623 | 10.905 | 0.000 | 2938.310 | 4229.076 |
| x3 | 36.7744 | 48.995 | 0.751 | 0.453 | -59.446 | 132.995 |
| x4 | -1024.4472 | 315.972 | -3.242 | 0.001 | -1644.984 | -403.910 |

| | | | |
|---|---|---|---|
| Omnibus: | 794.333 | Durbin-Watson: | 1.862 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 153594.389 |
| Skew: | 6.453 | Prob(JB): | 0.00 |
| Kurtosis: | 79.723 | Cond. No. | 7.13e+05 |

*Figure 5.11. R Squared Calculation for Linear Regression Algorithm.*

### 5.6.2 K-Nearest Neighbour

The R Squared result for K-Nearest Neighbour was 0.81 which was an Accuracy of 81% as depicted by Figure 5.12.

```
In [75]: from sklearn.model_selection import GridSearchCV
         params = {'n_neighbors':[2,3,4,5,6,7,8,9]}

         knn = neighbors.KNeighborsRegressor()

         model = GridSearchCV(knn, params, cv=5)
         model.fit(x_train,y_train)
         model.best_params_

Out[75]: {'n_neighbors': 9}

In [76]: # loading library
         from sklearn.neighbors import KNeighborsRegressor
         from sklearn.metrics import accuracy_score

         # instantiate learning model (k = 9)
         knn = KNeighborsRegressor(n_neighbors=9)

         # fitting the model
         knn.fit(x_train, y_train)

         # predict the response
         pred = knn.predict(x_test)

         # evaluate accuracy
         from sklearn.metrics import r2_score

         print(r2_score(y_test, pred))

         0.8100266896857792

In [77]: # evaluating accuracy
         accuracy = r2_score(y_test, pred) * 100
         print('\nThe accuracy of OUR regression is %d%%' % accuracy)

         The accuracy of OUR regression is 81%
```

*Figure 5.12. R Squared and Accuracy Calculation for K-nearest Neighbour Algorithm.*

### 5.6.3 Random Forest

The R Squared result for Random Forest was 0.962 which was an Accuracy of 96% as depicted by Figure 5.13.

```
In [78]: from sklearn.ensemble import RandomForestRegressor

         regressor = RandomForestRegressor(n_estimators=20, random_state=42)
         regressor.fit(x_train, y_train)
         y_pred = regressor.predict(x_test)

In [84]: from sklearn import metrics

         print(r2_score(y_test, y_pred))

         0.9622098664595038

In [85]: # evaluating accuracy
         accuracy = r2_score(y_test, y_pred) * 100
         print('\nThe accuracy of OUR regression is %d%%' % accuracy)


         The accuracy of OUR regression is 96%
```

*Figure 5.13. R Squared and Accuracy Calculation for Random Forest Algorithm.*

### 5.6.4 Gradient Boosting Machines

The R Squared result for Gradient Boosting Machines was 0.97 which was an Accuracy of 97% as depicted by Figure 5.10.

```
In [81]: import numpy as np
         import pandas as pd
         %matplotlib inline
         import matplotlib.pyplot as plt
         from sklearn.model_selection import train_test_split
         from sklearn import datasets
         from sklearn.metrics import mean_squared_error

         from sklearn import ensemble

In [82]: # Fit regression model
         params = {'n_estimators': 500, 'max_depth': 4, 'min_samples_split': 2,
                   'learning_rate': 0.01, 'loss': 'ls'}
         model = ensemble.GradientBoostingRegressor(**params)

         model.fit(x_train, y_train)

Out[82]: GradientBoostingRegressor(alpha=0.9, criterion='friedman_mse', init=None,
                     learning_rate=0.01, loss='ls', max_depth=4, max_features=None,
                     max_leaf_nodes=None, min_impurity_decrease=0.0,
                     min_impurity_split=None, min_samples_leaf=1,
                     min_samples_split=2, min_weight_fraction_leaf=0.0,
                     n_estimators=500, n_iter_no_change=None, presort='auto',
                     random_state=None, subsample=1.0, tol=0.0001,
                     validation_fraction=0.1, verbose=0, warm_start=False)

In [86]: from sklearn.metrics import mean_squared_error, r2_score
         model_score = model.score(x_train,y_train)
         # Have a look at R sq to give an idea of the fit ,
         # Explained variance score: 1 is perfect prediction
         print('R2 sq: ',model_score)
         y_predicted = model.predict(x_test)

         # evaluating accuracy
         accuracy = r2_score(y_test, y_predicted) * 100
         print('\nThe accuracy of OUR regression is %d%%' % accuracy)

         R2 sq:  0.9708783592044348

         The accuracy of OUR regression is 97%
```

*Figure 5.14. R Squared and Accuracy Calculation for Random Forest Algorithm.*

### 5.7 Implementation of the Prototype using Flask

The implementation of the Prototype using Flask was done on Windows 10 using the command console and consisted of the following steps:

i) The first step involved creating a virtual environment for running the Prototype as depicted by Figure 5.15.

ii) Installed all dependencies needed for the Prototype to run i.e. NumPy, Pandas, Matplotlib, Flask and Scikit-learn as depicted by Figure 5.16.

iii) Ran the Prototype as depicted by Figure 5.17. The front end of the Prototype could be accessed by typing the local web address http://127.0.0.1:5000 on any web browser.

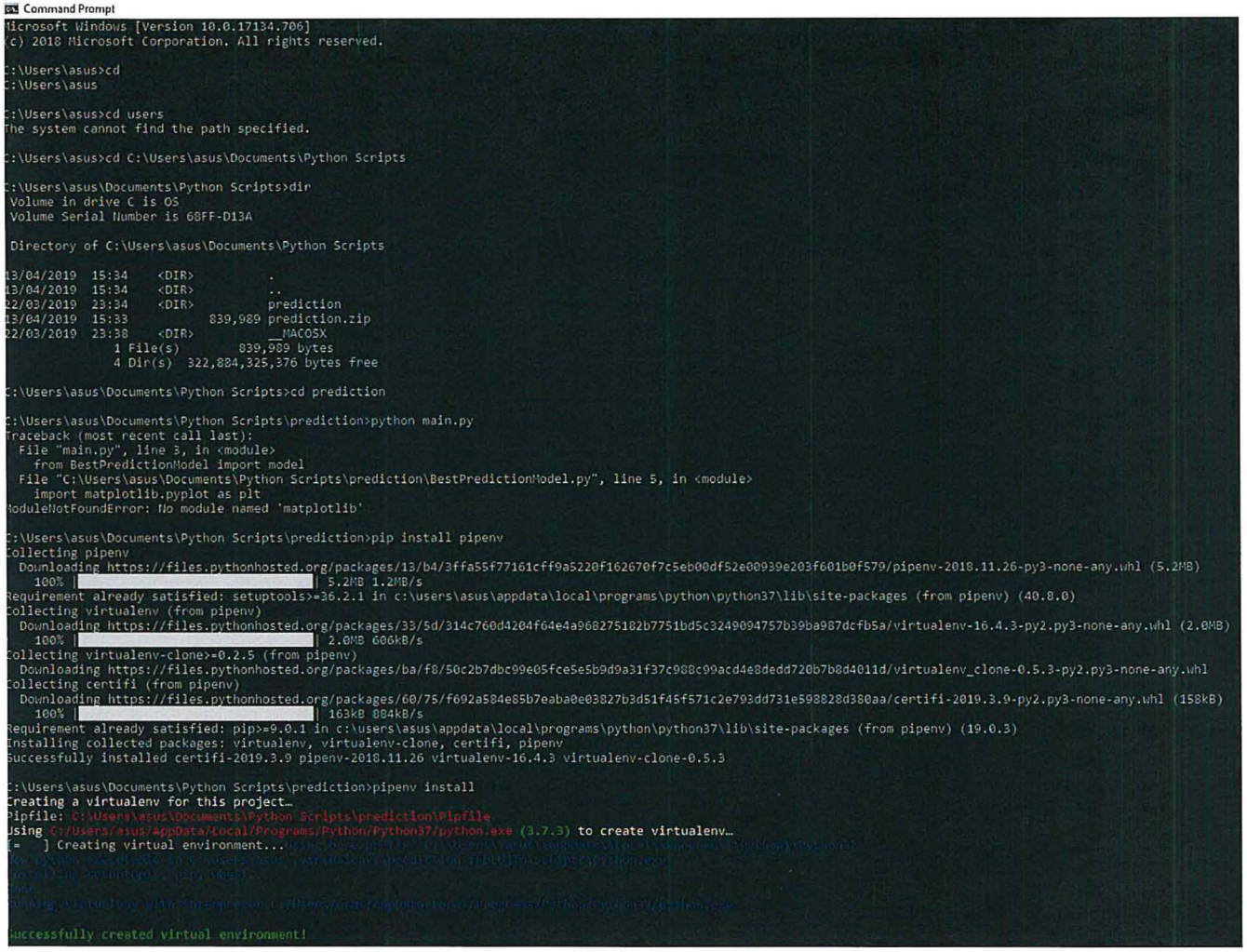


*Figure 5.15. Creation of a virtual environment for running the Prototype.*

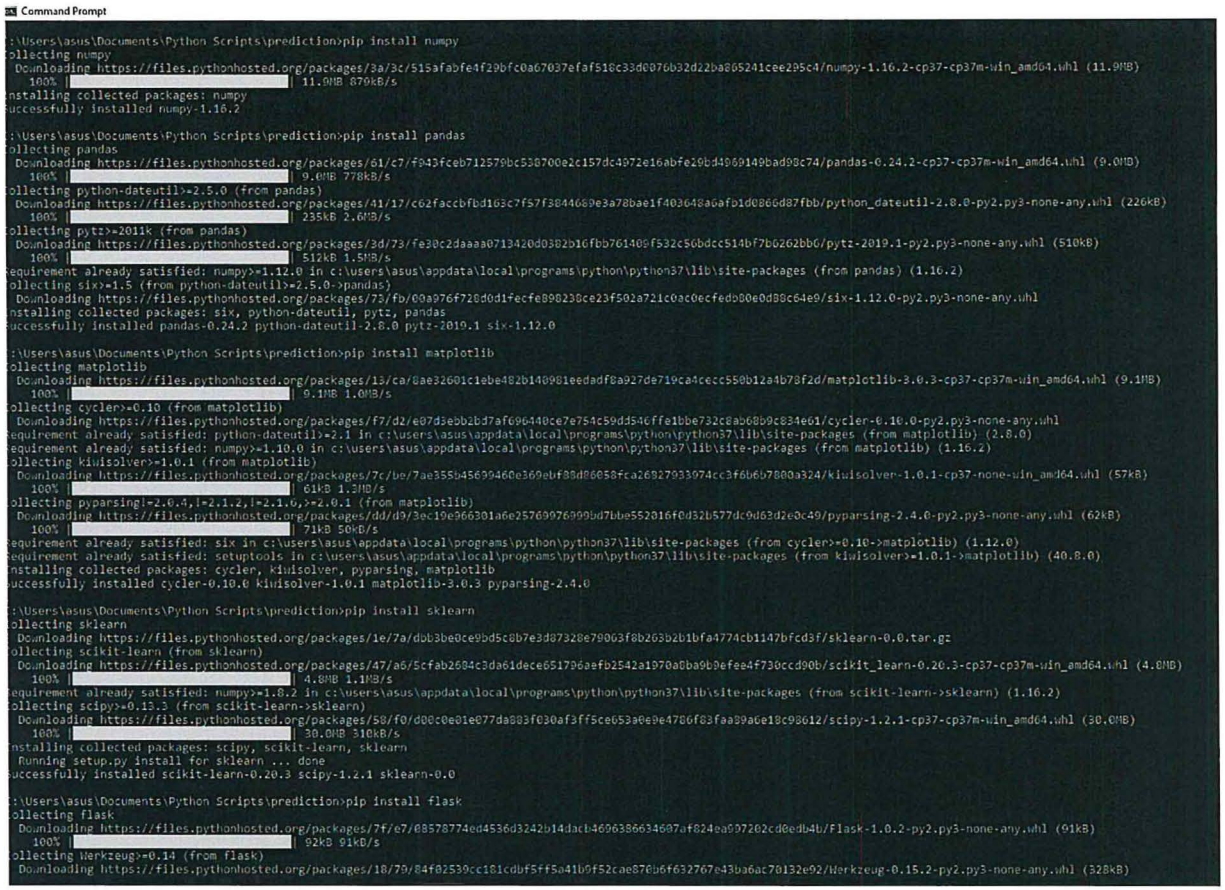


*Figure 5.16. Installation of Dependencies*

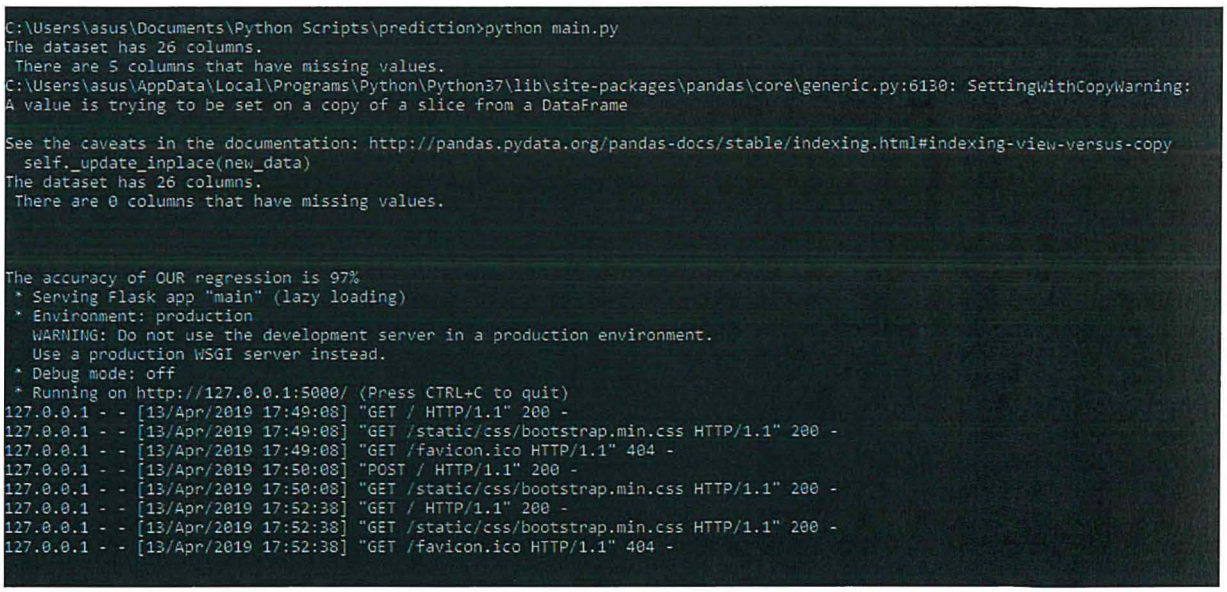


*Figure 5.17. Running the Prototype*

### 5.7.1 Components of the Flask Web Application

The Flask web application comprised of:

a) A Python file with the electricity consumption prediction model exported from Jupyter Notebook.

b) A HTML file to capture user input and display the results.

c) A Python file to render the HTML file with the data computed by the prediction model.

The Flask package contains a built-in development server which allows integration of the 3 parts above. Figure 5.18. shows the API data flow between the Front end, Python Flask and ML model. The code that runs the above three parts of the Flask web application is depicted by Figure 5.19.

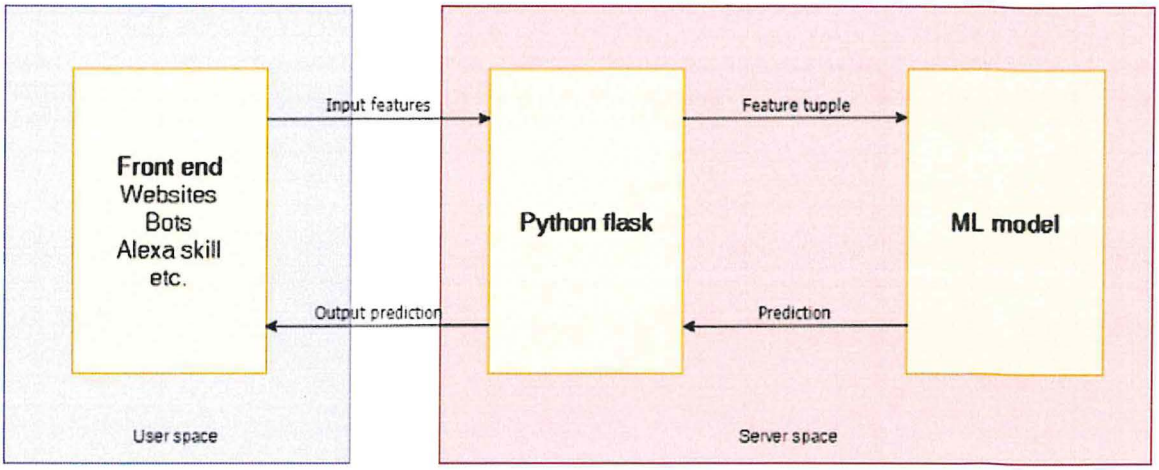


*Figure 5.18. API data flow (Kumar, 2018)*

```python
from BestPredictionModel import model
from flask import Flask, render_template, request

app = Flask(__name__)

@app.route("/", methods=["POST", "GET"])
def entry():
    data = {"errors": [], "analysis": None}
    if request.method == "POST":
        # capture values from the form
        sqft = request.form.get("sqft")
        data["sqft"] = sqft

        nwker = request.form.get("nwker")
        data["nwker"] = nwker

        pctermn = request.form.get("pctermn")
        data["pctermn"] = pctermn

        hdd65 = request.form.get("hdd65")
        data["hdd65"] = hdd65
        # validate all fields are required
        if all([sqft, nwker, pctermn, hdd65]):
            try:
                # Ensure values submitted are numbers
                sqft = float(sqft)
                nwker = float(nwker)
                pctermn = float(pctermn)
                hdd65 = float(hdd65)
            except Exception:
                data["errors"].append("All values must be numbers")
            else:
                # Calculate the prediction
                prediction = model.predict([[sqft, nwker, pctermn, hdd65]])
                data["analysis"] = prediction[0]
        else:
            data["errors"].append("All fields are required")

    return render_template("main.html", data=data)


if __name__ == "__main__":
    app.run()
```

*Figure 5.19. Main code to combine the 3 components of the Flask web application*

### 5.7.2 Front end of the Application

The front-end application was developed using HTML using Notepad++ and the HTML code snippet is depicted by Figure 5.22. The screenshot of the front-end application is depicted by Figure 5.20. The user is required to enter the required data for prediction. Figure 5.21. shows the prediction result after the user has entered the required data.

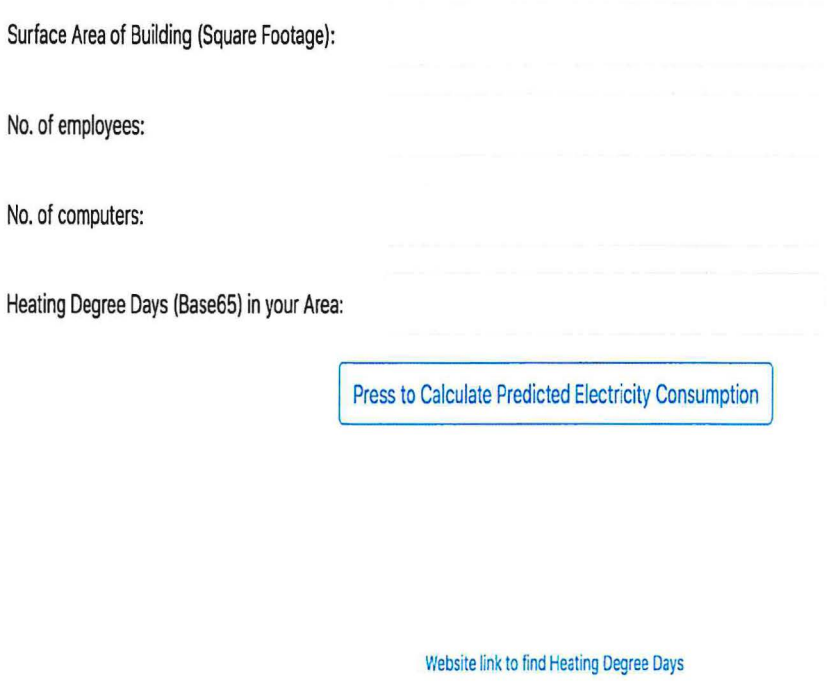


*Figure 5.20. Front End Application*

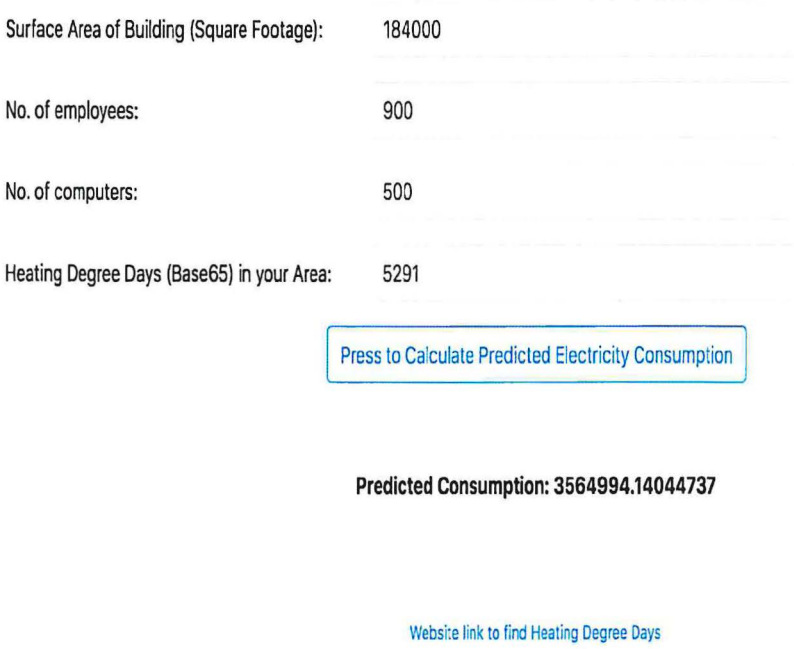


*Figure 5.21. Front End Application – After Prediction is done*

```
<div class="justify-content-md-center">
  <div class="card col col-md-12">
    <div class="card-body">
      {% if data.errors %}
      <p class="text-center">Please correct the following errors:</p>
      <div class="text-danger text-center">
        {% for error in data.errors%}
        <p>
          {{ error }}
        </p>
        {% endfor %}
      </div>
      {% endif %}
      <form action="" method="POST">
        <div class="form-group row">
          <label for="sqft" class="col-sm-4 col-form-label">Surface Area of Building (Square Footage):</label>
          <div class="col-sm-5">
            <input type="number" step="any" required class="form-control" id="sqft" name="sqft" value="{{data.sqft}}">
          </div>
        </div>
        <div class="form-group row">
          <label for="nwker" class="col-sm-4 col-form-label">No. of employees:</label>
          <div class="col-sm-5">
            <input type="number" step="any" required class="form-control" id="nwker" name="nwker" value="{{data.nwker}}">
          </div>
        </div>
        <div class="form-group row">
          <label for="pctermn" class="col-sm-4 col-form-label">No. of computers:</label>
          <div class="col-sm-5">
            <input type="number" step="any" required class="form-control" id="pctermn" name="pctermn" value="{{data.pctermn}}">
          </div>
        </div>
        <div class="form-group row">
          <label for="hdd65" class="col-sm-4 col-form-label">Heating Degree Days (Base65) in your Area:</label>
          <div class="col-sm-5">
            <input type="number" step="any" required class="form-control" id="hdd65" name="hdd65" value="{{data.hdd65}}">
          </div>
        </div>
        <div class="form-group row justify-content-center">
          <button type="submit" class="btn btn-outline-primary">Press to Calculate Predicted Electricity Consumption</button>
        </div>
      </form>
    </div>
    <div class="card-body">
      <div class="text-center">
        <strong>
          {% if data.analysis %}
            Predicted Consumption: {{ data.analysis }}
          {% endif %}
        </strong>
      </div>
    </div>
```

*Figure 5.22. HTML code snippet*

## 5.8 Compatibility Testing

The proposed prototype was tested for compatibility using Microsoft Edge (42.17134.1.0) as depicted by Figure 5.23. and Firefox Quantum (66.0.3) as depicted by Figure 5.24. The prototype worked on both and the same data was input on both giving the same prediction.

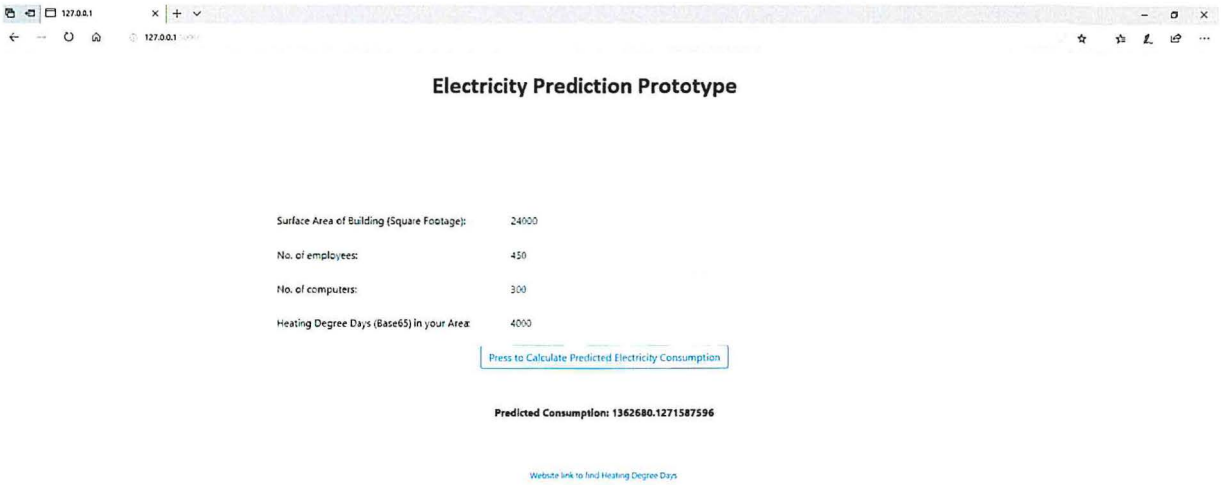


*Figure 5.23. Electricity Prediction Prototype testing on Microsoft Edge*

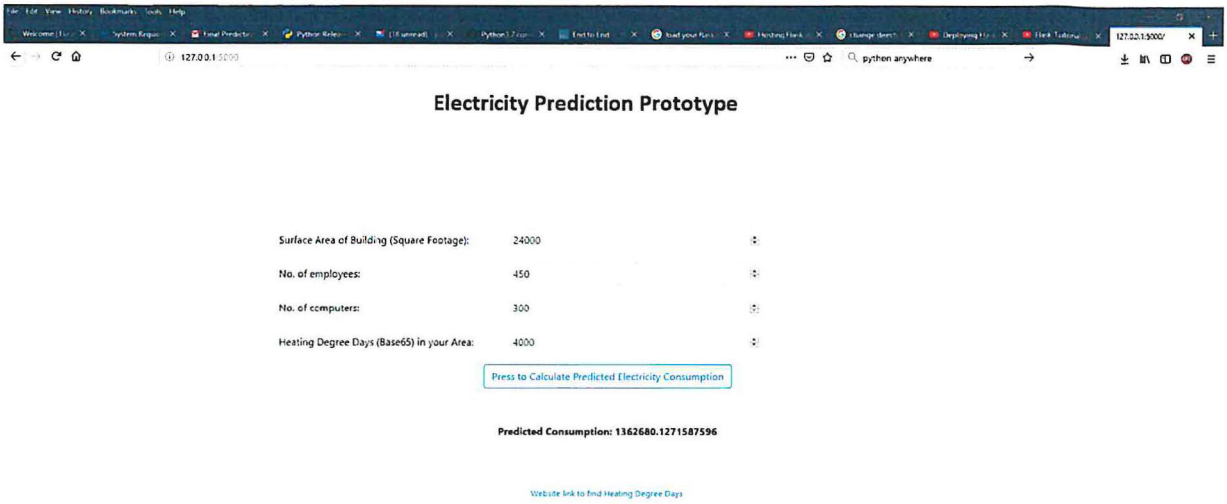


*Figure 5.24. Electricity Prediction Prototype testing on Firefox Quantum*

## 5.9 Prototype Testing

Prototype testing involves studying the software project under development in details in the early stages of work to make corrections in accordance with the set goals (XBSoftware, 2019). Prototype Testing involves collection of quantitative, qualitative and behavioural data while evaluating the user experience (Tutorialspoint, 2019). To ensure that the prototype met the user requirements, the following tests were done as summarized by Table 5.3

Table 5.3 *Test Cases*

| ID | Case | Expected Outcomes | Comments |
|---|---|---|---|
| 1 | Clicking link to Heating Degree Days Website | User directed to https://www.degreedays.net/ | Pass |
| 2 | Entering non-numeric data for the different features | Error Dialog Box | Pass |
| 3. | Clicking predict Electricity consumption button | Display of Predicted Electricity Consumption | Pass |

### 5.10 Prototype Usability Testing

Usability testing involves testing a prototype or system using real users to see how easy it is to use it (ExperienceUX, 2019). Users are asked to complete tasks on the prototype/system to see if they encounter problems or experience confusion.

For this research users were provided access to the prototype and a questionnaire was administered afterwards to get feedback on their experience. The outcomes of the Questionnaire are as follows:

#### 5.10.1 Education Level of the Users

This demographics question was meant to help the research determine what factors may influence a respondent's answers and opinions. It enabled the research to cross-tabulate and compare subgroups to see how responses varied between these groups.

As shown by Figure 5.25. 55.8% of the Users were Degree holders, 21.2% of the Users were Masters holders, 15.4% of the Users were Diploma holders and 7.7% of the users were Certificate holders.

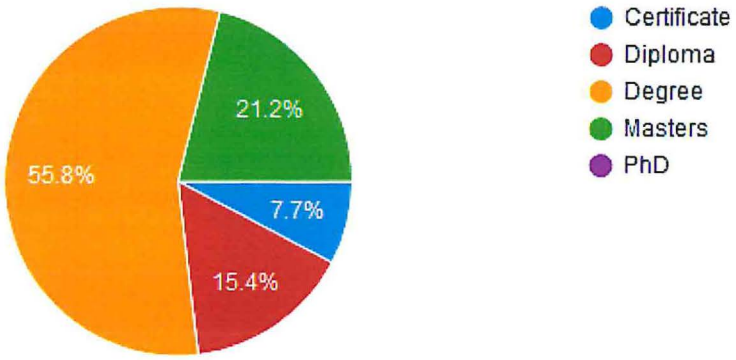


*Figure 5.25. Response on Education Level*

### 5.10.2 Labelling of Menu Items on Prototype

This question was asked to determine if the labelling and arrangement of the menu items on the user interface were good i.e. easy to understand.

As shown by Figure 5.26. 63.5% of the Users Strongly Agreed and 36.5% of the Users Agreed that the menu items were well labelled and arranged on the prototype. Based on Education Level as indicated by Table 5.4, the higher the Education Level of the User the higher the percentage of the users that Strong Agree within that group.

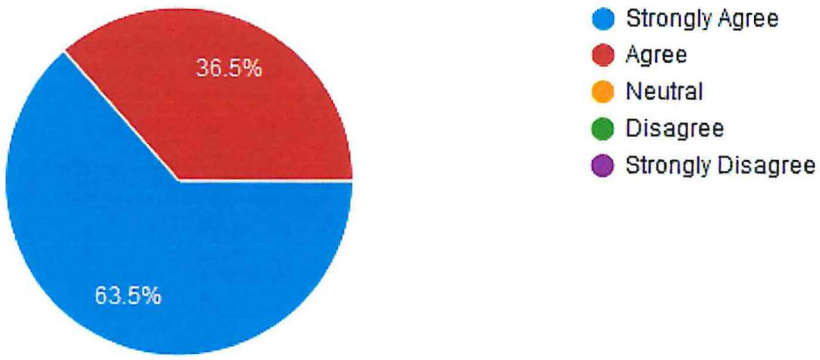


*Figure 5.26. Response on labelling and arrangement of Menu Items*

Table 5.4 *Response on labelling and arrangement of Menu Items per Education Level*

| Education Level | Strongly Agree | Agree |
|---|---|---|
| Certificate | 25% | 75% |
| Diploma | 50% | 50% |
| Degree | 69% | 31% |
| Masters | 73% | 27% |

### 5.10.3 Prototype Ease of Use

This question was asked to determine if the Prototype was easy to use. If the Prototype was hard to use it makes user experience bad for the users and they may not use the Prototype again in future.

As shown by Figure 5.27. 67.3% of the Users Strongly Agreed and 32.7% of the Users Agreed that the Prototype was easy to use. Based on Education Level as indicated by Table 5.5, the higher the Education Level of the User the higher the percentage of the users that Strong Agree within that group.

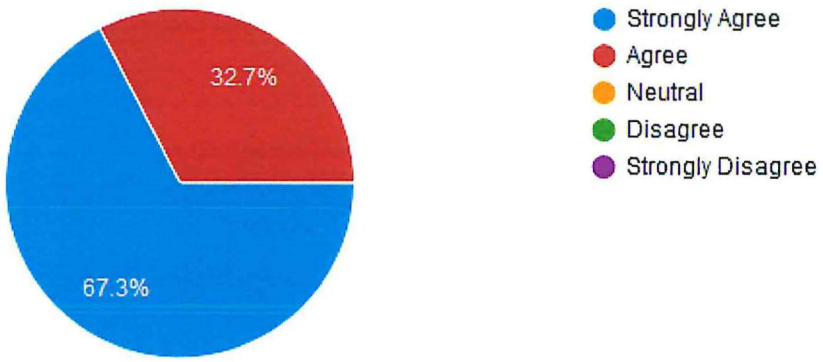


*Figure 5.27. Response on Ease of Use*

Table 5.5 *Response on Ease of use per Education Level*

| Education Level | Strongly Agree | Agree |
|---|---|---|
| Certificate | 25% | 75% |
| Diploma | 50% | 50% |
| Degree | 76% | 24% |
| Masters | 73% | 27% |

### 5.10.4 Future use of Prototype

This question was asked to determine if the users would use the Prototype in future to predict Electricity consumption in their buildings i.e. was the prototype useful enough for them to continue using it.

As shown by Figure 5.28. 48.1% of the Users Strongly Agreed, 44.2% of the Users Agreed and 7.7% of the Users were Neutral on the Future Usage of the Prototype. Table 5.6 shows the response of users based on their education level.

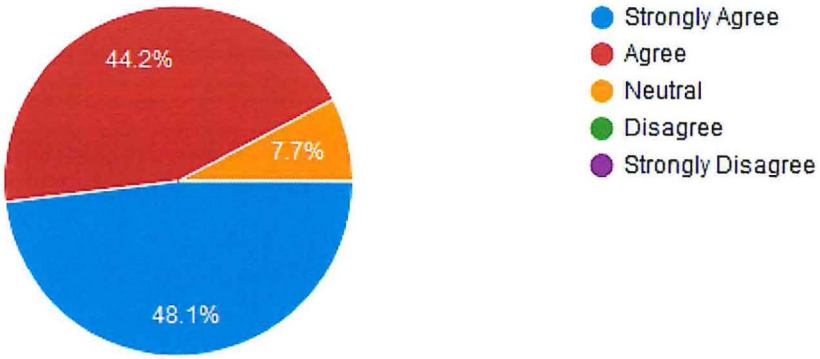


*Figure 5.28. Response on Future Usage of Prototype*

Table 5.6 *Response on Future Usage of Prototype per Education Level*

| Education Level | Strongly Agree | Agree | Neutral |
|---|---|---|---|
| Certificate | 25% | 50% | 25% |
| Diploma | 50% | 50% | 0% |
| Degree | 55% | 38% | 7% |
| Masters | 36% | 55% | 9% |

### 5.10.5 Recommending Prototype to other users

This question was asked to determine if the users thought the Prototype is good enough to recommend to other users i.e. can other users benefit from the use of the prototype.

As shown by Figure 5.25. 53.8% of the Users are Very Likely, 34.6% of the Users are Likely and 11.5% of the Users were Neutral on recommending the Prototype to other users. Table 5.5 shows the response of users based on their education level.

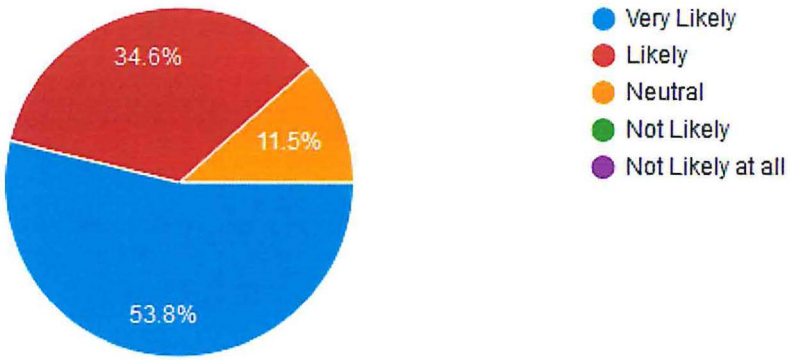


*Figure 5.29. Response on Recommending Prototype to other users*

Table 5.7 *Response on Recommending Prototype to other users*

| Education Level | Very Likely | Likely | Neutral |
|---|---|---|---|
| Certificate | 25% | 0% | 75% |
| Diploma | 50% | 50% | 0% |
| Degree | 62% | 31% | 7% |
| Masters | 45% | 45% | 10% |

# CHAPTER 6: DISCUSSION

## 6.1 Introduction

This chapter discusses the results of the study in relation to the objectives. It relates the results to the literature review. It outlines how this was accomplished through literature review discussions, through the processes of design and implementation. It also discusses the metrics used to check the model final results. There is also a summary of the results and suggestions for future research.

Energy management is the key to saving energy in an organization. This stems from the global need to save energy as this affects energy prices, emissions targets, and legislation. The research's aim was to come up with a solution to energy usage prediction to enable organisation plan on how control and reduce their energy consumption (reduce their operating costs). Energy expenses are often one of the major expenses in business operating costs in Kenya. An electricity prediction prototype was developed to enable commercial office buildings predict future electricity consumption, with the assumption that other factors remain constant.

From the reviewed literature, it was established that managing energy consumption effectively is a process and plays a critical role in reduction of operational costs. It was also found that Energy savings that come from behavioural changes (e.g. getting people to switch off their computers before going home) need ongoing attention to ensure that they remain effective and achieve their maximum potential.

## 6.2 Dataset description

The original dataset had 6720 rows (buildings) and 1119 features. Each building falls into one of 20 different classes according to the buildings principal building activity, or PBA as show by Figure 6.1.

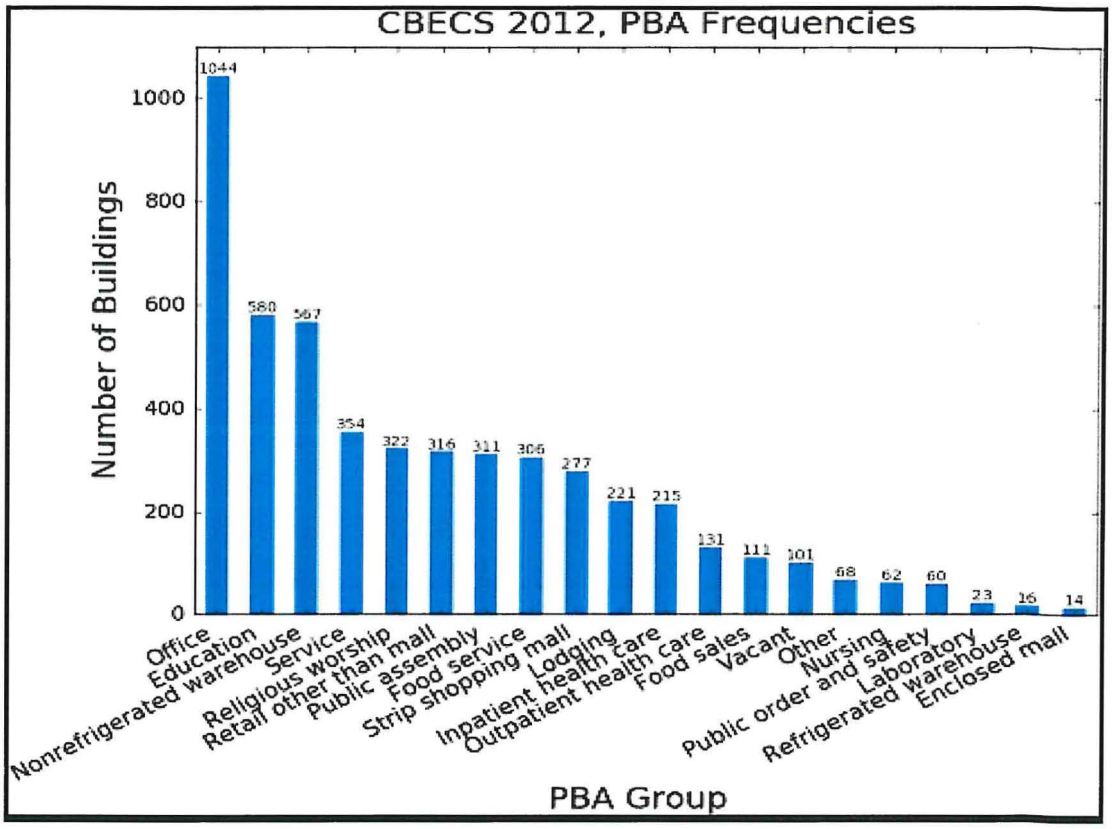


*Figure 6.1. PBA Frequencies*

### 6.3 Model Validation

To validate the Model the data was first split into training and test datasets using a ratio of 70%: 30% respectively. The training dataset was then trained using several regression-based machine learning algorithms and the Mean Absolute Error (MAE) was compared across the different algorithms as shown by Figure 6.2. For the dataset used in this research the best algorithms based on MAE was Gradient Boosted Machines and Random Forest.

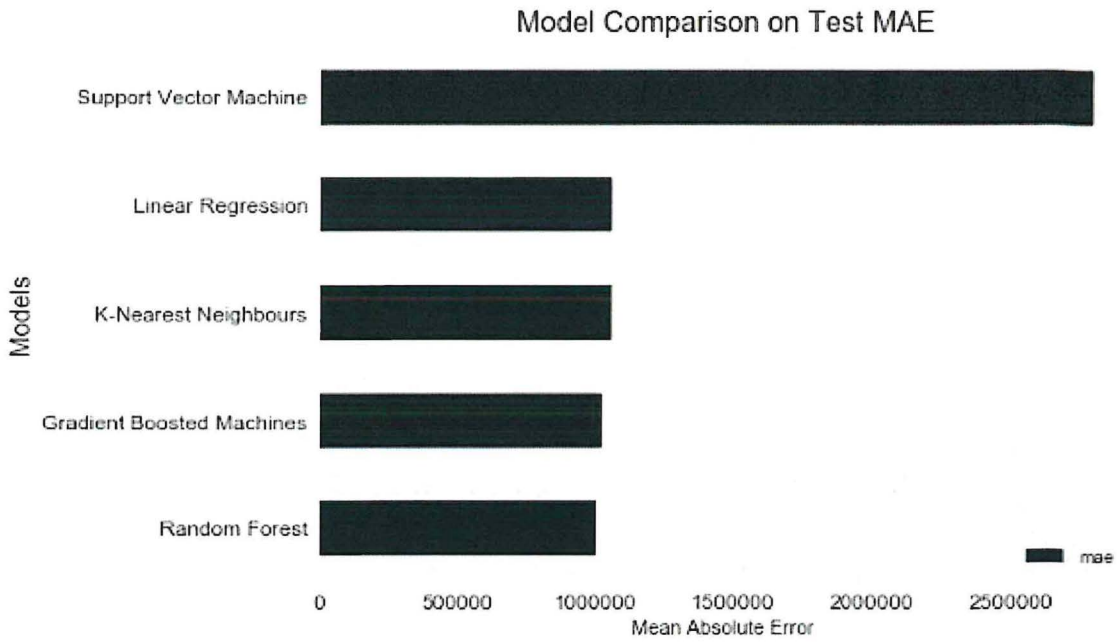


*Figure 6.2. Model Comparison*

### 6.4 Design Process Implementation

The process of accessing the prototype once operational was through a front end interface, where data would be input by the building owner and the data would be used in predicting electricity consumption. To aid the process, the flow of information was documented through the use of UML design diagrams. This included use case diagram, data flow diagram, sequence diagram and activity diagram.

### 6.5 System Functionality

Functionality of the prototype was tested using Compatibility Testing on different web browsers, Prototype Testing to see if the prototype behaves as expected when interacting with the user (building owner) and finally a usability test was conducted with a short questionnaire to get user feedback on their experience.

### 6.6 Accuracy of Outputted Results

To measure the accuracy of the predicted results was based on the R2 score and the score across the different machine learning algorithms tested was as summarized by Table 6.1.

Table 6.1 *Algorithm performance comparison*

| Model | R2 Score | Accuracy |
|---|---|---|
| Linear Regression | 0.783 | 78% |
| K-Nearest Neighbours | 0.810 | 81% |
| Random Forest | 0.962 | 96% |
| Gradient Boost Machines | 0.970 | 97% |

### 6.7 Research Contribution

The prototype was focused at predicting future electricity consumption in commercial office buildings and the predicted consumption would improve the decision-making process with regards to cost reduction and monitoring. With accurate prediction of electricity consumption, it makes it easy for owners of commercial office buildings to invest in equipment that can reduce or improve energy consumption in their buildings.

### 6.8 Limitations

The limitations of the prototype and energy predicting in general are summarised as follows:

i) Due to the existence of Commercial Office Buildings that may be defined as 'Outliers' the electricity prediction prototype is recommended for Buildings with more than 1000 square footage and less than 100,000 square footage.

ii) Due to the size of the training set of 1044 Commercial Office Buildings the prototype's prediction model would perform better if the size of the data set is increased and more features are discovered as a result.

# CHAPTER 7: CONCLUSIONS AND RECOMMENDATIONS

## 7.1 Conclusions

This research proposed an energy prediction prototype for commercial office buildings. To build the prediction models, the research identified and selected the features and building characteristics from the dataset. To evaluate the performance of the models, diverse experiments using four machine learning algorithms of KNN, SVM, Gradient boosting and linear regression were performed. The main objective of the research was to come up with a prototype to predict energy consumption and facilitate managing and monitoring energy consumption in commercial office buildings. This involved the use of already existing datasets in energy consumption in past years, as a basis for the prediction.

The research discussed the steps taken to mitigate the problem in the past, and the call for more institutions and individuals to come up with even more solutions. Other objectives of the research included performing literature reviews on the data used during previous researches, challenges faced, and implementations done in the past in an attempt towards providing solutions. The research focused on predicting electricity consumption as a means to help manage and monitor costs. This is because Energy remains the highest cost areas for businesses.

## 7.2 Recommendations

The following recommendations were made with regards to the research. Improvement of the prototype to include more features and a larger data set to be used for training. Also any new emerging regression-based machine learning algorithm can be tested to see if it can give a better prediction using more features.

## 7.3 Suggestions for future Research

This research can be extended to consider the possibility of predicting electricity tariffs (electricity price forecasting) or developing web-based/mobile-based prototypes for predicting energy consumption in residential buildings.

# REFERENCES


Abdelhamid, N. (2015). Multi-label rules for phishing classification. *Applied Computing and Informatics*, *11*(1) 29-46.

Ahmad, A., Hassan, M., Abdullah, M., Rahman, H., Hussin, F., Abdullah, H., & Saidur, R. (2014). A review on applications of ANN and SVM for building electrical energy consumption forecasting. *Renewable and Sustainable Energy Reviews, 33*(1), 102-109.

Amasyali, K., & El-Gohary, N. (2017). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, *81*(1), 1192-1205.

Amber, K. P., Aslam, M. W., Mahmood, A., Kousar, A., Younis, M. Y., Akbar, B., . . . Hussain, S. K. (2017). Energy Consumption Predicting for University Sector Buildings. *Energies 2017, 10*(10), 1579.

Amiri, S. S., Mottahedi, M., & Asadi, S. (2015). Development of Multi-Linear Regression Model to Predict Energy Consumption in the Early Stages of Building Design. *AEI 2015,* 54-65.

Annunziata, E., Frey, M., & Rizzi, F. (2013). Towards nearly zero-energy buildings: The state-of-art of national regulations in Europe. *Energy, 57*(C), 125-133.

Ardjmand, E., Ghalehkhondabi, I., Weckman, G. R., & Young, W. A. (2016). Application of decision support systems in scheduling/planning of manufacturing/service systems: A critical review. *International Journal of Management and Decision Making*. Retrieved March 10, 2019, from https://www.researchgate.net/publication/311446389_Application_of_decision_support_systems_in_schedulingplanning_of_manufacturingservice_systems_A_critical_review

Aydinalp, M., Ugursal, V. I., & Fung, A. S. (2004). Modeling of the space and domestic hot-water heating energy-consumption in the residential sector using neural networks. *Applied Energy*, *79*(2), 159-178.

Babbie, E., & Mouton, J. (2008). *The practice of social research, South African edition.* Cape Town: Oxford University Press Southern Africa.

Bull, R., Chang, N., & Fleming, P. (2012). The use of building energy certificates to reduce energy consumption in European public buildings. *Energy and buildings, 50*, 103-110.

Chapman, W. L., Bahill, T., & Wymore, W. (2018). *Engineering modeling and design.*

Cimpanu, C. (2017, March 17). Millions of Smart Meters May Over-Inflate Readings by up to 600%. Retrieved March 15, 2019, from https://www.bleepingcomputer.com/news/hardware/millions-of-smart-meters-may-over-inflate-readings-by-up-to-600-percent/

Commercial Building. (2019). *Definitions.com*. Retrieved March 14, 2019 from https://www.definitions.net/definition/Commercial+Building

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. London: Sage Publications.

De Silva, M. N., & Sandanayake , Y. G. (2012). Building Energy Consumption Factors: A Literature Review And Future Research Agenda. *World Construction Conference 2012 – Global Challenges in Construction Industry*, Colombo, Sri Lanka, pp. 90-99.

De Vaus, D. A. (2001). *Research design in social research.* London, England: SAGE Publications..

Dean, J., Patterson, D., & Young, C. (2018). A new golden age in computer architecture: Empowering the machine-learning revolution. *IEEE Micro*, *38*(2), 21-29.

Delzendeh, E., Wu, S., Lee, A., & Zhou, Y. (2017). The impact of occupants' behaviours on building energy analysis: A research review. *Renewable and Sustainable Energy Reviews*, *80*, 1061-1071.

Demirkoparan, F., Kaynar, O., & Özekicioğlu, H. (2017). Forecasting of Turkey's Electricity Consumption with Support Vector Regression and Chaotic Particle Swarm Algorithm. *Journal of Administrative Sciences*. Retrieved from https://www.researchgate.net/profile/Halil_Ozekicioglu2/publication/328496067_Forecasting_of_Turkey's_Electricity_Consumption_with_Support_Vector_Regression_and_Chaotic_Particle_Swarm_Algorithm/links/5bd172e345851537f59907bc/Forecasting-of-Turkeys-Electri

Department of Energy & Climate Change [DECC]. (2013). *An investigation of the effect of EPC ratings on house prices.* Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/207196/20130613_-_Hedonic_Pricing_study_-_DECC_template__2_.pdf

Despa, M. L. (2014). Comparative study on software development methodologies. *Database Systems Journal, 5*(3), 37-56.

DisplayR. (n.d.). What is Features Engineering?. Retrieved March 28, 2019, from https://www.displayr.com/what-is-feature-engineering/

Doblander, C., Strohbach, M., Ziekow, H., & Jacobson, H. (n.d.). Real-time Load Prediction with High Velocity Smart Home Data Stream. Retrieved from https://arxiv.org/ftp/arxiv/papers/1708/1708.04613.pdf

Dudovskiy, J. (2018). *The Ultimate Guide to Writing a Dissertation in Business Studies: A Step-by-Step Assistance* (January 2018 ed.).

Edwards, R. E., New, J. R., & Parker, L. (2012). Predicting Future Hourly Residential Electrical Consumption: A Machine Learning Case Study. *Energy and Buildings, 49*, 591-603.

Energy. (2019). *Dictionary.com*. Retrieved March 14, 2019, from https://www.dictionary.com/browse/energy

EnergyPrint Inc. (n.d.). Measuring to Manage: How to Budget for Building Energy Costs. Retrieved June 6, 2019, from https://energyprint.com/resources/budget-energy-costs-2019/

EnergyWatch. (n.d). Energy Budgets & Accruals. Retrieved June 6, 2019, from https://energywatch-inc.com/what-we-do/utility-expense-data-management/budgets-accruals/

Eriksson, U. (2012, April 5). Functional vs Non Functional Requirements. Retrieved May 2, 2019 from https://reqtest.com/requirements-blog/functional-vs-non-functional-requirements/

ExperienceUX. (2019). What is usability testing? Retrieved April 14, 2019, from https://www.experienceux.co.uk/faqs/what-is-usability-testing/

Fallah, S. N., Deo, R. C., Shojafar, M., Conti, M., & Shamshirband, S. (2018). Computational Intelligence Approaches for Energy Load Predicting in Smart Energy Management Grids:State of the Art, Future Challenges,and Research Directions. *Energies 2018, 11*(3), 596.

Fayaz, M., & Kim, D. (2018). A Prediction Methodology of Energy Consumption Based on Deep Extreme Learning Machine and Comparative Analysis in Residential Buildings. *Electronics 2018, 7*(10), 222.

Fridley, D. G., Zheng, N., & Zhou , N. (2008). Estimating Total Energy Consumption and Emissions of China's Commercial and Office Buildings. Retrieved from https://china.lbl.gov/sites/all/files/lbl-248e-commercial-buildingmarch-2008.pdf

Gajowniczek, K., & Ząbkowski, T. (2014). Short Term Electricity Predicting Using Individual Smart Meter Data. *Procedia Computer Science, 35*, 589-597.

Ghasemia, A., Shayeghi, H., Moradzadeh, M., & Nooshyar, M. (2016). A novel hybrid algorithm for electricity price and load forecasting in smart grids with demand-side management. *Applied Energy 2016, 177*, 40-59.

Ghofrani, M., Hassanzadeh, M., Etezadi-Amoli, M., & Fadali, M. S. (2011, August). *Smart meter based short-term load predicting for residential customers.* Paper presented at 2011 North American Power Symposium, Boston, MA, USA.

Gowda, A. P. (2016, May). *Factors influencing energy consumption among moderately low income residents in multifamily rental apartments.* (Unpublished master's thesis). Georgia Institute of Technology, Atlanta, Georgia.

Gul, M. S., & Patidar, S. (2015). Understanding the energy consumption and occupancy of a multi-purpose academic building. *Energy and Buildings, 87*, 155-165.

Gürbüz, F., Öztürk, C., & Pardalos, P. (2013). Prediction of electricity energy consumption of Turkey via artificial bee colony: a case study. *Energy systems, 4*(3), 289-300.

Hamzaçebi, C. (2016). Primary energy sources planning based on demand forecasting: The case of Turkey. *J Energy South Africa, 27*(1), 2-10.

Hassan, J. S., Zin, R. M., Abd Majid, M. Z., Balubaid, S., & Hainin, M. R. (2014). Building Energy Consumption in Malaysia: An Overview. *Jurnal Teknologi, 70*(7), 2180-3722.

Ibrahim, R., & Yen, S. Y. (2010). Formalization of the data flow diagram rules for consistency check. *International Journal of Software Engineering & Applications (IJSEA), 1*(4), 95-111.

Irungu, D W. (2016). *Simulation of the future electricity demand and supply in kenya using the long range energy alternative planning system.* (Unpublished master's thesis). University of Nairobi, Nairobi, Kenya.

Jain, R. K., Smith, K. M., Culligan, P. J., & Taylor, J. E. (2014). Predicting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. *Applied Energy, 123*, 168-178.

Janda, K. B. (2011). Buildings don't use energy: people do. *Architectural Science Review, 54*(1), 15-22.

Jim, C. Y., & Peng, L. L. (2012). Weather effect on thermal and energy performance of an extensive tropical green roof. *Urban Forestry & Urban Greening, 11*, 73-85.

Jurado, S., Nebot, A., Mugica, F., & Avellana, N. (2015). Hybrid methodologies for electricity load forecasting: Entropy-based feature selection with machine learning and soft computing techniques. *Energy, 84*, 276-291.

Kalamees, T., Jylhä, K., Tietäväinen, J., Jokisalo, J., Hyvönen, R., & Saku, S. (2012). Development of weighting factors for climate variables for selecting the energy reference year according to the EN ISO 15927-4 standard. *Energy and Buildings, 47*, 53-60.

Kavaklioglu, K. (2018). Principal components based robust vector autoregression prediction of Turkey's electricity consumption. *Energy Systems*, 1-22.

Kaytez, F., Taplamacioglu, M. C., Çam, E., & Hardalac, F. (2015). Forecasting electricity consumption: A comparison of regression analysis, neural networks and least squares support vector machines. *International Journal of Electrical Power & Energy Systems, 67*, 431-438.

Khurana, N., Chhillar, R. S., & Chhillar, U. (2016). A novel technique for generation and optimization of test cases using use case, sequence, activity diagram and genetic algorithm. *Journal of Software, 11*(3), 242-250.

Kialashaki, A., & Reisel, J. R. (2014). Development and validation of artificial neural network models of the energy demand in the industrial sector of the United States. *Energy, 76*, 749-760.

Kissflow. (2018). Rapid Application Development: Changing How Developers Work. Retrieved December 14, 2018, from https://kissflow.com/rad/rapid-application-development/

Kreider, J. F., & Haberl, J. S. (1994). Predicting hourly building energy use: The great energy predictor shootout--Overview and discussion of results. *1994 American Society of Heating, Refrigerating, and Air Conditioning Engineers (ASHRAE) annual meeting.* Orlando, FL, United States.

Kumar, V. (2018). Deploy Machine Learning Models for Free. Retrieved March 14, 2019, from https://medium.com/analytics-vidhya/how-to-deploy-simple-machine-learning-models-for-free-56cdccc62b8d.

Li, Q., Peng, R., & Meng, Q. (2010). Prediction model of annual energy consumption of residential buildings. *2010 International Conference on Advances in Energy Engineering.* Beijing, China.

Linden, A. L., Carlsson-Kanyama, A., & Eriksson, B. (2006). Efficient and inefficient aspects of residential energy behaviour: what are the policy instruments for change?. *Energy Policy,.34*(14), 1918-1927.

Lombard, L. P., Ortiz, J., & Pout, C. (2007). A review of buildings energy consumption information. *Energy and Buildings, 40,* 394-398.

Lucid Software Inc. (2019). What is a Data Flow Diagram?. Retrieved March 10, 2019, from https://www.lucidchart.com/pages/data-flow-diagram.

Ma, J., & Ma, X. (2018) A review of forecasting algorithms and energy management strategies for microgrids. *Systems Science & Control Engineering, 6*(1), 237-248.

MacKay, D. J. (1994). Bayesian nonlinear modeling for the prediction competition. *ASHRAE transactions, 100*(2), 1053-1062.

Mahapatra, K. & Gustavsson, L. (2008). An adopter-centric approach to analyze the diffusion patterns of innovative residential heating systems in Sweden. *Energy Policy, 36,* 577–590.

Martin, C. (2013). Generating low-cost national energy benchmarks: A case study in commercial buildings in Cape Town, South Africa. *Energy and Buildings, 64,* 26-31.

Masuda, H., & Claridge, D. E. (2014). Statistical modeling of the building energy balance variable for screening of metered energy use in large commercial buildings. *Energy and buildings, 77,* 292-303.

Mehar , A. M., Gill, A. Q., & Matawie, K. (2018). Analytical Model for Residential Predicting Energy Consumption. *2018 IEEE 20th Conference on Business Informatics.* Vienna, Austria.

Minka, E. (2018). Mean Absolute Error ~ MAE [Machine Learning(ML)]. Retrieved January 19, 2019, from https://medium.com/@ewuramaminka/mean-absolute-error-mae-machine-learning-ml-b9b4afc63077.

Nababan, T. S. (2015). The Factors Affecting the Household Energy Consumption, Energy Elasticity and Energy Intensity in Indonesia. *International Conference on Entrepreneurship, Business, and Social Sciences (ICEBSS).* Semarang and State University of Jakarta, Yogyakarta, Indonesia.

Office Building. (2019). *InvestorWords.com.* Retrieved March 26, 2019 from http://www.investorwords.com/14602/office_building.html

Oinas-Kukkonen, H., & Harjumaa, M. (2018). Persuasive Systems Design: Key Issues, Process Model and System Features. *Communications of the Association for Information Systems, 24*(28), 485-500.

Ouedraogo, N. S. (2017). Africa energy future: Alternative scenarios and their implications for sustainable development strategies. *Energy Policy, 106*, 457-471.

Papadopoulos, A.M., Theodosiou, T.G., & Karatzas, K.D. (2002). Feasibility of energy saving renovation measures in urban buildings the impact of energy prices and the acceptable pay back time criterion. *Energy and Buildings, 34*(5), 455-466

Power Technology. (2013, September 16). The case against smart meters – are people right to be suspicious?. Retrieved March 4, 2019, from https://www.power-technology.com/features/feature-case-against-smart-meters-people-right-to-be-suspicious/

Raza, M. Q., & Khosravi, A. (2015). A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. *Renewable and Sustainable Energy Reviews, 50*, 1352-1372.

Reade, S., & Zewotir, T. (2016). Modelling household electricity consumption in eThekwini municipality. *Journal of Energy in Southern Africa, 27*(2), 38-49.

Rhee, K. E., & Chung, H. M. (2014). Potential opportunities for energy conservation in existing buildings on university campus: a field survey in Korea. *Energy and Buildings*, *78*, 176-182.

Rodrigues, F., Cardeira, C., & Calado, J. M. (2014). The Daily and Hourly Energy Consumption and Load Forecasting Using Artificial Neural Network Method: A Case Study Using a Set of 93 Households in Portugal. *Energy Procedia, 62*, 220-229.

Rousselot, M., & Pollier, K. (2018). Energy efficiency trends in buildings. Retrieved November 12, 2018, from http://www.odyssee-mure.eu/publications/policy-brief/buildings-energy-efficiency-trends.html

Saidur, R. (2009). Energy consumption, energy savings, and emission analysis in Malaysian office buildings. *Energy Policy, 37*(10), 4104–4113.

Sarkar, D., Bali, R., & Sharma, T. (2017). *Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems.* Berkeley, CA: Apress.

Schneider Electric. (n.d.). Energy Budget Development. Retrieve June 6, 2019, from https://www.schneider-electric.com/en/work/services/energy-and-sustainability/energy-procurement/energy-budget-development.jsp

Sharma, S., Sarkar, D., & Gupta, D. (2012). Agile Processes and Methodologies: A Conceptual Study. *International Journal on Computer Science and Engineering (IJCSE), 4*(5), 892-898.

Shekhar, A,. (2018, February 14). What is Features Engineering for Machine Learning?. Retrieved from https://medium.com/mindorks/what-is-feature-engineering-for-machine-learning-d8ba3158d97a

Singh, K. (2017). Financing for Whom by Whom? Complexities of Advancing Energy Access in India. Retrieved from https://www.cgdev.org/sites/default/files/financing-whom-whom-complexities-advancing-energy-access-india.pdf

Smpokos, G., Elshatshat, M. A., Lioumpas, A., & Illiopoulos, I. (2018). On Energy Consumption Forecasting of Data Centers Based on Weather Conditions: Remote Sensing and Machine Learning Approach. Retrieved from https://arxiv.org/pdf/1804.01754.pdf

StackExchange. (2017). Good and Bad RMSE Values?. Retrieved October 11, 2018, from https://stats.stackexchange.com/questions/295785/good-and-bad-rmse-values

Suganthi, L., & Samuel, A. A. (2012). Energy models for demand forecasting—A review. *Renewable and Sustainable Snergy Reviews*, *16*(2), 1223-1240.

Sustainable Energy Africa [SEA]. (2017). Household energy access. Retrieved March 2, 2019, from http://www.sustainable.org.za/userfiles/household.pdf

The Minitab Blog. (2013, May 30). Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?. Retrieved from https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

Thomas, P. Y. (2010). *Towards developing a web-based blended learning environment at the University of Botswana.* (Doctoral dissertation). Retrieved from http://uir.unisa.ac.za/handle/10500/4245

Tsai, S. B., Xue, Y., Zhang, J., Chen, Q., Liu, Y., Zhou, J., & Dong, W. (2016). Models for predicting growth trends in renewable energy. *Renewable and Sustainable Energy Reviews, 77*, 1169-1178.

Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy,* 32(9), 1761-1768.

Tutorialspoint. (2019). Prototype Testing. Retrieved from https://www.tutorialspoint.com/software_testing_dictionary/prototype_testing.htm

United Nations Industrial Development Organization (UNIDO). (2009). Sustainable energy regulation and policymaking for Africa. Retrieved September 26, 2018, from https://www.unido.org/sites/default/files/2009-02/Module18_0.pdf

U.S. Department of Energy. (2012). 2011 Buildings Energy Data Book. Retrieved from http://large.stanford.edu/courses/2015/ph240/davidson1/docs/2011_BEDB.pdf

U.S. Energy Information Administration (EIA). (2015). How was the 2012 CBECS Buidlings Survey conducted?. Retrieved September 21, 2018, from https://www.eia.gov/consumption/commercial/reports/2012/methodology/conducted.php

U.S. Energy Information Administration (EIA). (2018a). Energy use in Commercial Buildings. Retrieved September 18, 2018, from https://www.eia.gov/energyexplained/index.php?page=us_energy_commercial#tab2

U.S. Energy Information Administration (EIA). (2018b). How the United States uses energy. Retrieved September 13, 2018, from https://www.eia.gov/energyexplained/index.php?page=us_energy_use

U.S. Energy Information Administration (EIA). (n.d.). 2012 CBECS Survey Data. Retrieved September 25, 2018, from https://www.eia.gov/consumption/commercial/data/2012/

Vakiloroaya, V., Samali, B., Fakhar, A., & Pishghadam, K. (2014). A review of different strategies for HVAC energy saving. *Energy Conversion And Management, 77*, 738-754.

Verdejo, H., Awerkin, A., Becker, C., & Olguin, G. (2017). Statistic linear parametric techniques for residential electric energy demand forecasting. A review and an implementation to Chile. *Renewable and Sustainable Energy Reviews, 74*, 512-521.

Wan, K.K.W., Li, D.H.W., Liu., D. & Lam, J.C. (2011). Future trends of building heating and cooling loads and energy consumption in different climates. *Building and Environment, 46*(1), 223-234.

Winston, A., Favaloro, G., & Healy, T. (2017). Energy Strategy for the C-Suite. Retrieved from https://hbr.org/2017/01/energy-strategy-for-the-c-suite.

Woo, Y. E., & Cho, G. H. (2018). Impact of the Surrounding Built Environment on Energy Consumption in Mixed-Use Building. *Sustainabilty, 10*(3), 832.

Woods, E. (2016). Software architecture in a changing world. *IEEE Software*, *33*(6), 94-97.

XBSoftware. (2019). Prototype Testing. Retrieved from https://xbsoftware.com/qa-software-testing/full-qa-cycle/prototype-testing/

Yu, Z., Fung, B.C.M., Haghighata, F., Yoshinoc, H., & Morofskyd, E. (2011). A systematic procedure to study the influence of occupant behaviour on building energy consumption. *Energy and Buildings, 43*(6), 1409–1417.

Yufeng, G. (2017). The 7 Steps of Machine Learning. Retrieved November 20, 2018, from https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e

Yun, G. Y., & Steemers, K. (2011). Behavioural, physical and socio-economic factors in household cooling energy. *Applied Energy,, 88*(6), 2191-2200.

Zero Carbon Hub. (2015). Post-occupancy Evaluation - Rowner Research Project - Phase Two. Retrieved from http://www.zerocarbonhub.org/sites/default/files/resources/reports/ZCH-RownerResearch-Phase-II.pdf.

Zhang, X. M., Grolinger, K., & Capretz, M. A. M. (2018). Forecasting Residential Energy Consumption Using Support Vector Regressions. Retrieved from https://www.researchgate.net/publication/330474705_Forecasting_Residential_Energy_Consumption_Single_Household_Perspective.

Zheng, A. (2015, Ocober 16). Evaluating Machine Learning Models. Retrieved December 10, 2018, from Oreilly: https://www.oreilly.com/ideas/evaluating-machine-learning-models/page/4/offline-evaluation-mechanisms-hold-out-validation-cross-validation-and-bootstrapping.

Zou, J., Weidong, L., & Tang, Z. (2017). Analysis of Factors Contributing to Changes in Energy Consumption in Tangshan City between 2007 and 2012. *Sustainability, 9*(3), 452.

# APPENDICES

**Appendix: Questionnaire**

**Electricity Prediction Prototype Usability Test Questionnaire**

**Question 1: What is your highest Level of Education?**

- Certificate
- Diploma
- Degree
- Masters
- PhD

**Question 2: Are the menu items well labelled and arranged on the prototype?**

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

**Question 3: Is the prototype easy to use?**

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

**Question 4: Would you use the porotype in future to predict electricity consumption for your Building?**

- Strongly Agree
- Agree
- Neutral
- Disagree

- Strongly Disagree

**Question 5: How likely are you to recommend this prototype to other users?**

- Very Likely
- Likely
- Neutral
- Not Likely
- Not Likely at all