**Can data analysis be used on electricity usage data to effectively predict the occupancy of the household to identify circumstances in which elderly or   
disabled residents may require aid**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

A Dissertation and Proposal

Presented to

Faculty of Engineering, Environment and Computing

Coventry University

Coventry, UK

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In partial fulfilment

of the Requirements for the Degree

BSc in Computer Science

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

By

Shahzeb Dawood - 705511

2018

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Supervisor

Dr. David Croft

Keywords: Electricity Usage Data; Data mining; Predictive algorithms; Performance; Household

**MEMBER PROFILES**

**Primary Assessor:** David Croft, Ph.D.Lecturer

Faculty of Engineering, Environment and computing,

Coventry University

**Secondary Assessor:** Matthew England. Ph.D.  
Senior Lecturer

Faculty of Engineering, Environment and Computing,

Coventry University

**Chair Assessor:** Peter Every

Principal Lecturer and Course Leader for the Creative Computing Pathway course,

Associate Head of Department for Computing,

Coventry University

**DECLARATION OF ORIGINALITY**

**300COM / 303COM Declaration of originality**  
I Declare that This project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.

**Statement of copyright**

I acknowledge that the copyright of this project report, and any product developed as part of the project, belong to Coventry University. Support, including funding, is available to commercialise products and services developed by staff and students.  Any revenue that is generated is split with the inventor/s of the product or service.  For further information please see www.coventry.ac.uk/ipr or contact ipr@coventry.ac.uk.

**Statement of ethical engagement**  
  
I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking)

   
**Signed:** **Date:** 29-04-18

Please complete all fields.

|  |  |
| --- | --- |
| First Name: | SHAHZEB |
| Last Name: | DAWOOD |
| Student ID number | 7015511 |
| Ethics Application Number | P68075 |
| 1st Supervisor Name | DAVID CROFT |
| 2nd Supervisor Name | MATTHEW ENGLAND |

This form must be completed and included with your project submission to Turnitin. Failure to append these declarations may result in your project being rejected for marking.



**Certificate of Ethical Approval**

**Applicant:**

Shahzeb Dawood

**Project Title:**

The use of data analytics to identify circumstances in which elderly or disabled residents may  
require aid through changes in electricity usage patterns.

This is to certify that the above-named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Medium Risk

**Date of approval:**

23 March 2018

**Project Reference Number:**

P68075

**ACKNOWLEDGMENTS**

*A very heartfelt thanks to all my teachers who, throughout the course of this degree, taught me the required relevant skills and equipped me enough to put forward this dissertation.*

*A dissertation that wouldn’t have been possible without the help and support of my family, friends. colleagues and my supervisor Dr. David Croft for his dataset, guidance and mentoring.*

**TABLE OF CONTENTS**

Member Profilesi

Declaration of originalityii

AcknowledgmentsIv

**ABSTRACT5**

**CHAPTER 1 - INTRODUCTION6**

* 1. Overview6
  2. Problem description and motivation6
  3. Overview7
  4. Problem description and motivation7

CHAPTER 2 - Literature Review8

CHAPTER 3 - METHODOLOGY4

3.1 The Dataset5

3.1.1 Contender6

3.1.2 Compendium6

3.1.3 Collection Methodology6

3.2 Environment5

3.2.1 Tools6

3.2.2 Libraries6

3.3 Algorithms6

3.3.1 Regression & Classification6

3.3.2 k-NN Algorithm6

3.3.3 Linear Regression6

3.3.4 Neural Network Classification6

3.4 Statistical Analysis6

3.4.1 Test Statistics6

3.4.2 Calculation of Test Statistics6

3.4.3 Definition and formulae of Statistical computations6

CHAPTER 4 - ANALYSIS4

4.1 Visual Aids5

4.1.1 k-NN Regressor6

4.1.2 k-NN Classifier6

4.1.3 Linear Regressor6

4.1.4 Neural Network Classifier6

4.1.5 Notation6

4.2 Statistical Analysis Report6

4.3 Comparison & explanation of the results6

CHAPTER 5 – PROJECT MANAGEMENT4

5.1 Project Plan5

5.1.1 Initial and final Gantt Chart6

5.1.2 Sprint Backlog6

5.2 Social, legal and ethical considerations5

5.3 Risk Management5

5.4 Quality Assurance5

CHAPTER 67 – APPENDIX A (Code and related material)4

CHAPTER 67 – APPENDIX B 4

CHAPTER 67 – APPENDIX C – Project Proposal 4

**ABSTRACT**

The worldwide widespread of smart meters means more understandable & workable data will be produced for every specific household in the UK, given the smart meter bill was passed on the 28th of November 2017, its enforcement will lead to raw electricity consumption data being produced which will then automate and efficiently collect and measure electricity usage (Gov.uk, 2013). This data can be mined and analysed for more than just producing efficient bills and smarter readings. This project intends to provide an achievable, affordable and a minimal resource required solution for independent senior citizens and the disabled by analysing their electricity data and learning specific trends and electricity consumption footprints. Every person has a different routine and lifestyles which impact and reflect a different electricity usage trend. Learning those trends can be useful and can enable us to make a judgement if something is unusual. This is partially implemented and made possible in this research by using various data mining techniques both supervised and unsupervised. Taking as inputs power consumption and active occupancy to data analysis algorithms, such as KNN regression and classification, Linear Regression, k-Means clustering and a Neural Network classifier, to predict the occupancy of the house on a minute to minute basis based on just power consumption. The dataset used for this model was recent and short term, recorded over a period of 2 months in the year of 2018 from a home in the city of Coventry. The said algorithms are then evaluated on their predictive performance using rigorous scrutinising statistical analysis, test cases and primarily F-measure to deduce which data mining algorithm is best suited to the given data set. The furtherment of this research work anticipates laying forth minimalist, resourceful and an economical alternative solution to assisted living with the aspiration of helping independent living senior and disabled citizens.

**CHAPTER 1**

**INTRODUCTION**

* 1. **Overview**

This dissertation talks about and compares different data mining algorithms (k-NN regression, k-NN classification, Neural Network classification and Linear regression) based on their ability to correctly predict the number of occupants of a household or simply just the presence of people using the homes electricity usage data. A rigorous statistical analysis of the algorithms was performed to compare their accuracy and precision and concludes which specific algorithm best translates itself within the given dataset and type of computational environment. The findings are furthered by providing an application area where seniors or disabled residents can benefit by the implementation of such a system and continue living independently and abdicate assisted living or other equivalent unreasonably priced solutions to the challenges that come with age, disabilities and living by one’s self.

* 1. **Problem definition and Motivation**

In a research survey carried out by the NHS it was revealed that around one in three adults over the age of 65 are susceptible to fall over at least once a year (NHS, 2015). Around half of whom are likely to fall more frequently. Another similar extended research concluded that two thirds of those who have fallen once are more susceptible to fall again in the next 6 months (Senior Health & Wellness Blog, 2018). In the UK alone, the cause of death amongst seniors aged 75 and over is the most common and unfortunate occurrence (Age UK, 2017).

The government and private institutions are trying their best to get a hold on the situation and minimise these forlorn happenings. Care homes, government allotted nurses, private carers and volunteer organisations are all viable solutions, but extortionate and according to a survey, unfavourable to the elderly as they fear losing their independence more than they fear illness and death. (Hough 2010)

An article published by the Washington state university briefly outlined the average cost of assisted living being somewhere around 70,000 dollars per annum which equates to more than £49,000 per year. (Staff 2018)

There needs to be an affordable solution for seniors and citizens who, not only, want to keep their independence, but also need to be able to afford it, on their subordinate pension and income.

The application and motivation of this research project is aimed at the elderly and disabled people who reside on their own. As mentioned above, problems like getting around the house and doing daily chores etc can become a struggle at their fragile age and accidents often happen.

Many propositions and steps have already been taken to prevent such events from happening but most of the times they require additional equipment like infrared or heat sensors in homes to monitor movement of the residents which, yet again by being expensive, violates the conditions of a well-suited solution for people who want to be able to afford independence.

Electricity usage data is readily available for, if not all then, most households and the required data is being recorded but going to waste. The imminent implementation of smart meters will, more sooner than later, ensure the recording and storage of the consumption data for every household in the United Kingdom. Leaving us with workable and copious amounts of data that can mined for trends and patterns which in turn can be a less expensive, reasonable and economical alternative solution to the said problem.

Every human being on the planet has a contribution, a footprint. A footprint is described as the physical impression left sand or floor by a foot or shoe. There is another, deeper surrogate meaning of the word “Footprint”. It doesn’t relate to the physical aspects of impressions, but the ecological footprint of a human being. It is elementally defined as the demand of humans on nature. This specific footprint, much like fingerprints, varies from human to human. Every one of us leads a different lifestyle, works different jobs, has different hobbies and uses electricity differently. It’s only fair to have a just and customised judgement for a varying variable as such.

The Electricity consumption data of a household collected over a specific period can be analysed to learn trends. These trends, or “electricity footprint” if you may, once learned, can be used to perform additional actions such as prediction, plotting and forecasting using data science and data mining methodologies. This electricity footprint and trends in the data can, in the scope of this research, be compared with everyday usage to then, based on the electricity usage of the minute, predict the availability of a person at home.

Abnormalities in the data can range from between anything like an electricity spike not being shown roughly at around 7pm when the inhabitant uses the kettle for their evening tea. Or the system predicting the person not being home when the resident is in fact home.

When abnormalities as such are visible, the system can choose to notify a user of the curious occurrence and inconsistency in the predicted data.

* 1. **Application**

The application of this research is to propose and partially implement a system that can be furthered to offer an affordable solution to independent living amongst the elderly and disabled.

This project aspires to implement different data mining algorithms that are capable of noticing abnormalities and deviations from normal trends, not directly and in a conventional straight-forward method but, by the uncomplicated comparison amongst the predicted data of occupancy against actual occupancy in electricity consumption entries. This will be made possible through the rigorous prospecting of data and uncompromising statistical tests on the algorithms used followed by a conclusion affirming which algorithm comparatively does the job with admirable accuracy.

try and overcome issues such as mobility and getting around the house that come with old age and disabilities and provide a secondary method of alerting emergency services if a problem arises. The only access to the problem being the persons electricity usage data.

* 1. **Implications**

The last sections assisted in grasping a reasonable understanding of the problem, motivation and need for such a system. This section very briefly outlines the intentions and methodology of this dissertation.

This research implements a diverse collection of data mining algorithm to predict the availability of people at home using supervised learning techniques and manipulating data. The success of the algorithm not only assures the partial implementation of a system that can be furthered to help in a real-world scenario, by making use of real-time power readings, but also counsels which machine learning algorithm works best from a handful of chosen and implemented ones.

The incessant need of the project to be implemented with limited resources and yet produce a capable solution dictated many paths in deciding aspects that were imperative to the project. For instance, the implementation of the algorithms. The algorithms were drafted in the programming language Python using Spyder as the Integrated Development Environment (IDE). The machine learning library “Scikit-learn” was used in the implementation of the algorithms, namely, k-NN regression and classification, Linear regression, Linear classification and Neural networks.

**CHAPTER 2**

**LITERATURE REVIEW**

Machine learning is a vital field in, not only computers, but every day professions such as retail, finance, medical sciences, education and so on (Alpaydin 2014). In layman terms, it’s the ability to make a computer system learn and work intelligently using statistical techniques that are performed on data. It is believed that it has originated from the evolution of pattern recognition, computational learning theory and artificial intelligence. A book extract dedicated to machine learning defines it as the technological modification and automation of a previously unreachable problem (Alpaydin 2014). Calculations and trends, patterns and behavior of data that was once innumerable and computable by the naked eye, was made possible by astute algorithms that learn by reaching into depths and relations between seemingly one-dimensional data to extract meaningful results and outputs that prove to be extremely helpful in various application areas.

Supervised learning is a machine learning approach which involves two separate attribute sets of data, an input set and a derivate output set. The job of this technique is to reckon the mapping between the input and output pairs such, that once it acquires this “mapping function”, with a degree of admirable success, the system can then generate a probable output for a given input instance from the same data set (Brownlee 2016). Examples of supervised learning methods include regression and classification algorithms.

Unsupervised learning, by then name alone, displays a contradictory nature to its dissimilar sibling, the supervised learning technique. This approach considers input data that has no reciprocal or corresponding output data and is “unlabelled”. The goal of this technique is to deduce and extract features from the input unlabelled data. Its structure and elemental patterns and trends are learnt. This methodology is often referred to as “exploratory data analysis” (Brownlee 2016). Due to the unavailability of a reference or labelled desired output, there is no way of telling if the results and findings from an unsupervised learning algorithm are, in fact, correct (Goodfellow, Bengio and Courville, 2016) which is a downside to such algorithms. Its use is limited to areas where patterns, grouping and visualisation is required based on latent similarities or trends within the data. Basic procedures of unsupervised learning include Clustering (e.g. k-Means), group assignation, anomaly detection and discretized representation (Self Organising Maps).

There is plethora of research work that contributes to the area of computer science that deals with data mining and analysis, but only a handful of work that directly associates with the research work carried out in this dissertation. The chosen literature’s, much like this research work, adapt different methods of data analysis to interpret and analyse the fitting of their work in a specific application. They also exhibit similarities ranging from the data collection methodologies to statistical analysis of algorithms.

The unpopularity, lack of full fledge implementation and availability of smart meters has forced many to take it upon themselves to create a homemade smart-meter. A paper published by the International Research Journal of Engineering and Technology (IRJET) implemented such a system using a microcontroller based design and implementation of an energy meter primarily to replace human interference in meter readings which makes the collections susceptible to error hence having an adverse effect on the economy. In addition to that, services such as monitoring energy consumption from a web page through device IP address and electricity theft detection by monitoring meter tampering were also made available. Due to these applications, a homemade smart meter was built based on the Internet of Things (IOT) technology in collaboration with existing electricity meters. The physical aspects of the system included an Arduino Nano Board, ESP 8266 Wi-Fi module (Node MCU), 16\*2 LCD display, buzzer, power supply and The Allegro® ACS712 for current sensing. The software aspects included a webpage to display load energy usage reading in terms of Watts and present device information in a more detailed analysis in both description and visualization. (IOT based smart energy monitoring, 2018). The hardware used for the data collection in said journal article has a lot in common with the data collection methodology used for the electricity data that was collected for this research as mentioned in detail in Chapter 3.1.3 – Data Collection Methodology.

Coming towards algorithm based related work, a research extract from a book, that consisted of 29 contributions from researches all thoroughly reviewed and validated, analysed electricity data of a household to extract its specific trends which were then, along with the number of inhabitants of the household, used as input variables to a fuzzy model, to forecast how probable it was for a house apparatus to be used or started within the next 60 seconds (Zhu, 2016). The electricity data set collection for this research was long term and collected over a period of 12 months. A case-study on fuzzy tools along with neural network based algorithm for efficient energy management was presented. That was written to predict the electricity usage 24 hours ahead of its time. Lastly, a cost analysis was presented to signify the importance of energy management and its implications. The findings of this specific research in this book, heavily relates to the “Electricity data analysis” area of this dissertation. Their success with the neural network algorithm, although much more complex and broken down, was the reason behind it being considered in this research. Initially only algorithms that followed a similar framework such as k-Means and k-NN were to be implemented.

Subsequently, a research specifically relating to electricity and anomaly detection with extensive results in the form of figures, diagrams and tables carried out data analysis on electricity consumption data using 2 different approaches: Statistical approach and Clustering (Jakkula and Cook, 2010). The two methods were used with the aim of mining the electricity data and counting the outliers using both techniques to compare them against the known outliers to see the efficiency of either methodologies. The dataset for this research was power consumption collected using a smart CASAS (Centre for Advanced Studies in Adaptive Systems) environment. It concluded that the 2 outlier detection mechanisms presented wrongly identified an outlier with a ratio of 1:43 in Clustering to Statistical approach. In layman terms, this means the possibility of wrongly identifying an outlier in clustering as opposed to statistical analysis was 1 in 43. The results and finding of this research paper showed that clustering would be inaccurate in its given task 2.3% of the times whereas Statistical approach would wrongly identify an outlier 97% of the time. The application of this research was, like in the previous related work researches, to ultimately conserve energy by identifying abnormal power consumption and hence taking a step towards energy efficiency and conservation, predict abnormalities earlier on, like in the case of too much power being consumed in a smart home the possibility to warn residents on the failure of appliances etc.

Another similar research worked towards forecasting electricity demand in the UK. This research made use of existing deterministic calendar information of weekdays, weekends, bank and national holidays to create a data set of binary data, to distinguish between weekdays and weekends, In collaboration with a KNN- regression model for load forecast and prediction for the upcoming 24 hours. A k-Nearest-Neighbour algorithm is an efficient way to estimate the value of an unknown function in a given point using its values in training points. It is defined as the mean function value of the nearest neighbours. A softer version of which takes a weighted average, where the weight of each neighbour is proportional to its proximity to the input testing instance (Navot et al., 2014). The existing calendar data was used as a base to compare input data against to sort what kind of day the load was being forecasted for (weekday, bank holiday, or weekends). The dataset used for this research dated from November 2001 to November 2008. Almost 7 years with values of power on an hourly basis making that up to be more than 66,500 instances. The results of this research were compared to existing benchmark models (Not created in the research) to the written Multivariate k-Nearest Neighbour Regression model. It concluded that amongst other benchmarks models. The method of choosing best algorithm being mean absolute percentage deviation value. The multivariate system, created within the research took the crown amongst other contenders including univariate k-NN regression model and other statistical benchmark models such as MOVAV (7)24, MOVAV (7)168, RW24 and RW168(Al-Qahtani, 2013). This research coupled with the Clustering and statistical approach comparison research paper have made it credible that a benchmark methodology or linear statistical approach is not a strong method to go by when it comes to electricity data. The extensive, exhaustive and variating nature of the electricity data makes it rather impossible to fuse well and it takes complex much intelligent algorithms to learn the ways and motifs.

k-NN algorithms have been immensely popular in data analysis (Suguna, 2010) and specifically prediction applications, for the most part. Despite classification being the conventional way to go, the k-NN algorithm is often implemented using regression as well. The testing instance in either case is determined using the mean of the nearest neighbours.

There is a controversy revolving ‘k’ algorithms (k-NN, K-Means etc) when it comes to choosing an optimum value for the number of nearest neighbours (k), since it is user defined. In a research carried out by the Mu’tah University – Jordan tried out various values of n ranging from 1 to the square root of the sum of samples in the training dataset and proposed a solution to the selection of the value of ‘k’ depending on ensemble learning (Basheer, Ali Abbadi and Awad, 2014).

An archaic yet substantial research paper proposed that “large values of K smoothen the results of classification” (Bishop, 1995). With insufficient evidence as to why, but it is also considered that the square root of the sum of samples in the training dataset is an optimal assigned value for ‘k’.

Neural Networks primarily consist of three layers. An input, hidden and output layer made of (variables). Connection between layers are based on relevance of the assigned value (weight) of that input unit. The more relevance, the more the weight and higher the weight, higher its importance (Palaniappan and Awang, 2008). Because of the highly non-linear relation between the input and output parameters, sigmoid and linear transfer-function are sometimes used to generate their outputs (Yu et al., 2007). They are suitable for training large amounts of data with few inputs and is often used when other techniques are unsatisfactory.

They also happen to be an immensely popular choice when it specifically comes to price prediction in electricity data. The complex multiple layer model and nature of the artificial neural networks makes it possible to work with much convoluted scenarios with more than one variable and attributes. A comparable research that made use of a 3 layered ANN model forecasting price, based on the electricity usage data, for the next 168 hours. This research, however, used the Levenberg-Marquardt algorithm for training purposes. The accuracy and efficiency of the systems predictions were computed using a real-world case study established on the electricity market (Catalão et al., 2007).

A comprehensive research carried out by Haque and Kashtiban, in the field of Neural Networks, read and compared published papers dating from years 1990-1996 and 2000-2005 listing 5 areas having attracted most attention amidst which was Load forecasting. A reasonable and in-demand application of electricity data seems to be load forecasting and outlier detection in order to battle energy inefficiencies. Even load forecasting can further be broken down into three sub categories of prediction areas based on periods of forecasts: Long Term (LTLF), Mid Term, and Short-Term Load forecasting. The forecasts lasting anywhere 5 – 20 years for long term, a month to 5 years and short term being defined as anywhere between an hourly to weekly forecasting system. All three of which have different applications and scenarios that call for different forecasting periods.

The conclusion of this research is conflicting yet informative. Despite the many benefits of using Neural networks for power based training and predictions like its ability to deal with stochastic pattern of variating increasing data and fast classification and implied filtering of results, it also stated how and when it would be discouraged to use ANN model. Problems such as training time, training vector and “noisy” data (data containing errors such as missing elements due to faults in recording etc) can cause complications in the algorithm. It also stated that such an algorithm should be used as an “additional tool” not as a “replacement” to other AI techniques (Haque and Kashtiban, 2005).

Anomaly detection is another area that hugely relates to this project. The application of this dissertation is to detect, effectively, occurrences and instances of data that can be considered as anomalous and swiftly acting upon those said occurrences. A survey-based journal article spoke in depth about the different kinds of outliers in data analysis and the challenges faced in isolating a given anomaly from normal regions in a traditional 2-dimensional dataset (Chandola, Banerjee and Kumar, 2009). This survey particularly proved helpful in narrowing down concepts and techniques most effective in successful outlier isolation and detection. In addition to that, it also spoke about techniques for Anomaly detection such as Neural networks, rule based, Bayesian Network, Support vector, Clustering and Nearest neighbour based techniques etc. Most of the mentioned techniques in this survey have been used for this dissertation for not outlier detection, but prediction based on a training dataset. Another research relating to anomaly detection proved to be more practical than survey based and aimed to detect abnormalities between an aged and a new transformer in the UK based on data and variables from the transformers such as temperature, vibration, moisture, load current etc in the 2 transformers (Catterson, McArthur and Moss, 2010). The data readings were calculated using sensors. This paper concluded that the Conditional Anomaly Detection (CAD) technique can be used in the online monitoring of the transformers.

An unpopular machine learning model in the field of electricity data is Linear regression. Not many existing research works has primarily used linear regression for prediction. One of the reasons of which can be the availability of other much more refined and successful models which rendered this model unnecessary. The linear regressor works in such a way that it computes the change and effect on the mean of an output Y by the variance of a single regressor X. An upside to this model is the correlation between errors. Since errors often happen to be uncorrelated, in a linear model, so are the responses (Montgomery, Peck and Vining, 2007).

Whilst other machine learning models were implemented in this research because of their reasonable success and popularity in the field of machine learning, the linear regressor model was implement for the verification of its unpopularity and its inability to deliver.

The final deliverable of this research, in addition to implementation of assorted predictive machine learning algorithms, covers a section for statistical analysis and comparison of the enforced algorithms based on their predictive performance and success using various arithmetic formulary. However, it primarily relied on F Score to calculate, prove and thereupon crown one algorithm victor. One such research study, published by the Australasian Joint Conference on Artificial Intelligence, discussed evaluation measures for different characteristics of machine learning algorithms. It speaks of commonly-accepted performance measuring techniques and touches base with and proposes new unconventional approaches for measures derived from “fault avoidance” such as Youden’s index γ, the likelihood values ρ−, ρ+, and Discriminant Power DP (Sokolova, Japkowicz and Szpakowicz, 2006). It is argued in this paper that performance measures other than accuracy, F-score, precision, recall or ROC do apply and can be beneficial.

Although the work Ms. Sokolova was aimed towards providing an unprecedented approach to algorithm measure, the areas of her research that resonated with this research work was the discussion of conventional measures such as precision, recall, sensitivity, specificity, accuracy and F measure. There is substantial evidenced work done using these statistical methodologies and a “fault avoidance” approach was unbefitting.

**The research rudimentarily defined the measures as follows (**For extensive and apprehensible definitions with corresponding formulae refer to the methodology section**):**

* **Receiver Operating Characteristic (ROC):**

Shows a relation between the sensitivity and the specificity of the algorithm

* **Accuracy:**

Approximation of the effectiveness the algorithm displays by showing the probability of the true value of the class label; in other words, it assesses the overall effectiveness of the algorithm;

* **Specificity:**

Specificity approximates the probability of the negative label being true; in other words, it assesses the effectiveness of the algorithm on a single class;

* **Sensitivity:**

Sensitivity approximates the probability of the positive label being true; in other words, it assesses the effectiveness of the algorithm on a single class;

* **F score:**

A composite measure which benefits algorithms with higher sensitivity and challenges algorithms with higher specificity.

For said measures, quality of classification is measured and built from a confusion matrix which records correctly and incorrectly recognized examples for each class. These classes are categorised into test cases based on the nature of the prediction if the prediction correctly identifies a positive condition, it is known as a true positive prediction and so on for the remaining test cases namely true negative, False positive, False negative.

Most of the existing work done in relation to this project takes things as far as analysing electricity usage data to extract trends, test different data analytical techniques and anomaly detection. The major application of those finding being the need to make electricity consumption more efficient and conserve energy where possible. The Anomaly Survey research discussed entirely different application areas of anomaly detection but none of them came close to the intended work and application of this project which, along with other related work, makes it evident that the deliverable is new and hence will contribute to a novel application of occupancy and outlier detection. The Multivariant k-NN regression model research showed sweeping results in the success of a nearest neighbour regression model in the forecasting of load data which stands to show that a k-NN model might be a viable technique for triumphant results while working with electricity consumption data.

This project intends to take things in a different direction, it aims to predict the occupancy of a household in not just a Boolean value, but in number of occupants as well. The intended application of this project is to identify an anomaly in data and alert emergency services to get the independent senior or disabled resident the help they need. The abnormality detection will root as an extension to the proposed occupancy predicative system. If the system produces a false negative, that is, no inhabitant is predicted against the known data telling us that there is in fact someone home, it’s clear that something is wrong. The said application of this project, with hope, stands out and is derived from personal motivation and empathy towards humanity.

Despite there being many such algorithms capable of successful prediction of occupancy based on electricity data, like load forecasting, not a lot of research is done in this domain of predicting occupancy based on electricity consumption let alone a statistical comparison of such algorithms and the intended application of this project.

The existing work seems to be aimed at crowning technology victor against the existing linear benchmark and statistical methods. It is a given that an expert learnable algorithm will perform well than a system with a linear standard point of reference. To some extent it’s safe to say there is no comparison there. But even amongst the various learnable algorithms there must be one that performs better than the other. In addition to the many ways listed above, this specific reason of a rigorous statistical analysis and comparison between different algorithms, also, makes this dissertation project novel and different from its related existing work. LIMITATIONS OF EXISTING WORK – COMPARISON OF RESEARCHERS

CHAPTER 3

METHODOLOGY

This section of the research focuses on the methods applied and implemented to set the objectives into motion. The dataset, algorithms, testing and statistical analysis methodologies implemented.

**3.1 The Dataset**

**3.1.1 Contender**

The search for a dataset was a crucial part of the research as it would lay the foundations on which the primary research would be based. Not many datasets were even considered but one specific dataset that was taken into close consideration was an open-source Electricity consumption dataset of 19 households in London over the course of 2 years. The data set consisted of separate csv files for household information, half-hourly electricity consumption, daily electricity consumption, acorn details, half hourly weather and daily weather. These separate 6 csv files contained information such as tariff, what block the house is in, its acorn group, a smart meter measurement of power on a half hourly and daily basis, power measures, minimum, maximum, mean, median, sum and std. Details of the acorn group, their profile and the people in the group. In addition to that it had 2 other files containing weather readings, again, on a half hourly and a daily basis.

Despite the data being publicly available, clean, workable and with the added benefit of having data from 19 different households and exhaustive readings, it was ruled out as the project aimed to display no threat or necessity of peoples personal and sensitive information. It was stressed and an unescapable part of research to make the system starved to maintain economic implications yet yield comparable results.

Working with this dataset would have been easy and not challenging because of its resourceful definition and extensive variables, but people would have been led to believe that minimum requirement for the application and implementation of a system as such requires innumerable readings and a convoluted combination of attributes. Ultimately this dataset was dropped as it was aimed to keep the project unambiguous for people yet competent in its goal.

**3.1.2 Compendium**

The dataset that was ultimately chosen for the project was not a publicly available one. It was a personal household electricity dataset that was recorded and belonged to my supervisor.

The dataset was a singular csv file containing three attributes, namely, Date and time, Power and “at home”. The power was calculated and recorded for every minute of each over the period of 2 months (January and February 2018) adding up to 37270 instances.

Since there is a reading for every minute of each day, there are roughly 1448 occurrences of the attributes for a single day. This makes the data highly exhaustive and broken down hence very specific and precise.

|  |  |
| --- | --- |
| Attribute | Description |
| datetime | The date and time |
| power | The number of watts consumed in kw/h |
| unit | An additional field added to serialise the datetime variable to represent minutes. |
| athome | A variable denoting the presence and availability of people inside the household with 0, 1 and 2 (indicating number of people) |

Table 2. Dataset Variables

The benefits of this data set were the additional variable “athome” which can be used to and linked to patterns and behaviours in the electricity usage and how the power goes up or down based on occupancy, which will make the system more learnable and proficient.

**3.1.3 Collection Methodology**

Given that the use of smart meters is not popular in every household thus far, this specific dataset was collected using manual techniques which will, soon enough in the future, become frivolous once smart meters installation is enforced by law.

A homemade smart meter was created for the effective collection of the electricity data which included the following components:

* **Node MCU:**   
  A Lua (scripting language) based ESP8266-E12 Wi-Fi SOC (System on Chip) module used commonly for the implementation of internet of things. It has 16 GPIO (General Purpose Input Output) and works on 3.3v. The term "NodeMCU" by default refers to the firmware rather than the dev kits. The firmware uses the Lua scripting language (Vaideek and Vasundhara, 2011)
* **Raspberry Pi:**“An inexpensive credit card-sized computer that plugs into a computer monitor or TV, uses a standard keyboard and mouse. It is a capable little device that enables people of all ages to explore computing, and to learn how to program in languages like Scratch and Python. It’s capable of doing everything you’d expect a desktop computer to do, from browsing the internet and playing high-definition video, to making spreadsheets, word-processing, and playing games” (Raspberry Pi, 2018)
* **SQL Lite Database:**For storage of the recorded electricity data.
* **Light Sensor:**For registering light flashes and hence computing the time between consecutive flashes.

Electricity meters that were installed over the past few years have been electronic and come equipped with an LED light. The job of this LED is to flash, proportionally to the amount of power passing through the meter. A light detector was used to calculate the power consumption. The sensor would register the time between consecutive light flashes on the electricity meter. The relationship between the meter LED to the number of kWh being passed is known, so it is just a matter of computing the rate of flashes to determine the power being exported per minute.

Given the rate of flashing of the LED, the power being consumed is simply calculated by the formula:

The process of calculating the seconds between the light flashes can be done manually but, as the flashing is dependent on power consumption which is seldom constant hence inconsistent, inaccurate results can be registered. To limit any chance of error this process was automated using microcontroller called NodeMCU and Raspberry Pi. This flashing rate can be stored or used to make immediate calculations of power consumption which were the stored in an SQL Lite database.

The “athome” variable was calculated using “pings”. Defined as “Query (another computer on a network) to determine whether there is a connection to it” (Oxford Dictionaries | English, 2018) which would administer how many specified mobile phones were connected to the internet at a given time. To make this calculation more accurate the use of push notifications was also used.

This dataset was chosen because of its simple precise nature. It consisted of limited but necessary variables. Another reason of preference for this this dataset was its temporal proximity. The lifestyles of people today largely differ from the way of living of people as recent as 5-10 years ago. Factors such as weather, routine, job, lifestyle, activity, insulation and occupancy etc are all factors that influence the consumption of electricity. They are also factors that tend to differ which makes older datasets, if not outdated then at least, irrelevant. Leaving recent data being more relatable and accurate in its ability to determine anomalies in the modern-day electricity consumption patterns.

**3.1.4 Dataset split (Testing and training)**

The split for the training and testing data is an area worthy of attention. It was advised in the end of term presentation feedback to ensure that the split wasn’t right down the middle or that the training included all the dates of January and some of February and testing data of just February but rather make the split more random and distributed.

The percentage of the split was also a concern and there is debate when it comes to a fair value. A research journal paper similar in terms of data mining methodology split its data to an 80:20 – training to testing ratio (Sundar, Latha and Chandra, 2012). It is considered a fair split, whereas in some cases and scenarios even a split of 70:30 is justified. But then again, these split ratios are also highly dependent on the size of the data set and refinery capability of an algorithm.

The final percentage split that was adapted for this research was 67:33 training to testing ratio. Given the total 36210 instances, it divided into a seemingly healthy split of 24300 learnable instances to 11970 predicting/testing instances. The split randomness of the split was also ensured by the python built-in data split function having the ability to add a “random\_state” which would shuffle the instances before splitting based on the percentage provided. As shown in the figure below, the first 20 instances of both the training testing data (along with their serial numbers) was random.

|  |  |
| --- | --- |
| **Training dataset (67%)** | **Testing dataset (33%)** |
|  |  |

Table 3. The training and testing data split

\*The numbers on the right-hand side is the additional field that was added to the data set to serialise the headings of the power and ‘athome’ variable readings. Although the first 20 values only aren’t sufficient to display the randomness in the collection and split of the data it is still apparent that the data was not collected serially.

This part of the code is responsible for the split. The split ratio and the randomness of the split. The sklearn library has a function “model\_selection**.**train\_test\_split” which takes in attributes such as the output variables, the size of the testing data and the random state.

X\_train**,** X\_test**,** Y\_train**,** Y\_test **=** sklearn**.**model\_selection**.**train\_test\_split**(**X**,** y**,** test\_size **=** 0.33**,** random\_state **=** 4**)**

**3.2 Environment**

**3.2.1 Tools**

* **Programming Language:** Python 3.6
* **Integrated Development Environment (IDE):** Spyder 3.2.6

Spyder is an open source IDE designed to solemnly work with Python. It enhances the language’s ability given its many beneficial characteristics like syntax highlighting, debugging support, variable exploration, GUI based editor for most data structures, step-by-step execution, ability to run parts of code at a time, code prediction and completion, built-in editor and various other productivity tools. Its IPython (Qt) console makes for an able prompt data visualisation and visual communication area. in the form of graphs, tables and the display of regular outputs (Vaingast, 2014)

Like most IDE’s for python, it tackles the linear nature and inflexibility of the regular Python programme interface and provides a more participating candidate as opposed to the stand-alone Python programme.

The IDE “Rodeo” was initially considered, but it was found to be inflexible and caused conflicts in the existing programmes. The library import function was unable to detect to the location of the module where they were downloaded.

* **Dataset (.csv file):** Microsoft Excel
* **Statistics computation:** Manual

**3.2.2 Libraries** (Abraham et al., 2014)

* **Scikit learn**

Scikit-learn is a Python module that includes a wide range of state-of-the-art machine learning algorithms for supervised and unsupervised problems. This package emphasises ease of use, performance and documentation while focusing on bringing machine learning to non-experts using a general-purpose high-level language. It has minimal dependencies and is distributed under the simplified BSD license, encouraging its use in both academic and commercial settings (Varoquaux et al., 2015).

* **Numpy**

In the Python world, NumPy arrays are the standard representation for numerical data and enable efficient implementation of numerical computations in a high-level language. Performance of computations are improved through techniques such as vectorizing calculations, avoiding copying data in memory, and minimizing operation counts (Van Der Walt, Colbert and Varoquaux, 2011).

* **Pandas**

An open source, BSD-licensed library for the Python Language which provides rich data structures and functions which makes working with structured data fast, east and expressive. It enables Python to be a powerful and productive data analysis environment. (McKinney, n.d.)

* **Matplotlib**

A 2D graphics package used by Python for application development, interactive scripting and publication-quality image generation across user interfaces and operating systems (Hunter, 2007).

**3.3 Algorithms**

**3.3.1 Regression & Classification**

The basic difference between the said algorithm is that the outputs and behaviour of the algorithms. The aim of regression models is to discover relationships between response ‘Y’ and explanatory variable ‘X’. Ultimately to make predictions of the variable Y based on the observation of the input ‘X’ variable. (Seber and Lee, 2003). It finds the similarity of the input testing dataset to predict an attribute. The attribute prediction is “measured” based on a similarity function. In the regression model the output is an actual independent value. The value is calculated based on the average of its given nearest neighbours.

The classifier model, as the name implies, “classifies” or “categorizes” its predicted output in a defined set of possible outputs based on the similarity function. The output for a classification algorithm takes the value of the nearest neighbour most popular and in majority. This is known as majority vote (James, 1998).

**Example:**

Figure 1. Example supposed scatter for a data set.

**Classification method:**

Referring to the visual aid in Fig 1 above and in the case of *classification*, in a k-NN algorithm with a defined training dataset of 10 instances and 4 possible outputs: either ‘Blue’, ‘Red’, ‘Green’ or ‘Yellow’, If the family of a testing instance (Bigger, black dot) needs to be determined in a situation where k = 3, It is visible, in Fig 1, that out of the 3 closest points to the testing instance (black dot) on the graph, 2 are green and one is red hence the testing instance will be “classified” into the “Green” family based on the “Majority Vote” phenomenon discussed in chapter 2 – Literature Review.

**Regression Method:**

When it comes to regression, it becomes a little more complex than that. Regression will also produce a result that ultimately crowns the testing instance to belong to the green family but it technique of determination is different.

First the algorithm will assign the 3 possible instances (Blue, green, yellow and red) numeric values:

|  |  |  |
| --- | --- | --- |
| **Color** | | **Assigned Numeric Value** |
| Blue |  | 1 |
| Red |  | 2 |
| Green |  | 3 |
| Yellow |  | 4 |
| Black |  | To be predicted |

Table 4. Regressor, classifier example colour assigning

Given that k = 3, and that as visible in the graph out of the 3 closest points are 2 green and a red dot the algorithm will place the testing instance using the average of the nearest neighbours using their numeric values.

Nearest neighbours = [Green, Green, Red]

Numeric Values = [3, 3, 2]

Testing instance:

The produced value is closest to the numeric Value 3 and hence the testing instance is computed to belong in the green family.

Although the actual execution of the algorithms is far more complex and doesn’t rely on numeric assigning, the motive for the above example was to, in layman terms, explain the difference in operation of the two algorithms. In reality, the outcomes aren’t assigned numeric values rather the similarity function uses values based on the Euclidean distance of the testing instance to first determine the nearest neighbours, then compute the testing instances illegibility in every possible family.

**Actual case:**

Keeping in mind the example for the nearest neighbour models above, in the case of the electricity dataset being used for this project, the training and input data set being ‘kwh’ (power) and ‘athome’ (Number of occupants of the household), the possible classes for classification for output data becomes 0, 1 and 2 from the “athome” variable. Based on the similarity function, the algorithm ‘classifies’ every instance of the testing dataset into the three possible classes.

The regression model, calculates the average of the 3 closest neighbours to the testing instances and produces a float value which can be rounded off to produce a prediction of the class the testing instance might belong to.

**3.3.2 k-NN Algorithm**

In all ‘k’ type algorithms where the value of a certain variable is user defined (Like number of clusters ‘k’ in k-Means algorithm), the value of ‘k’ becomes a cause for concern.

Ultimately the value of ‘k’ (nearest neighbour to be taken into consideration) in both the models was assigned the value of 3 and kept constant for both the algorithms to maintain a fair environment as the algorithms were to be compared and the value of k can influence the performance of the algorithm drastically. However, this number wasn’t established right away.

As mentioned earlier Chapter 2 - Literature Review, the debate on the value of ‘k’ in k-NN algorithms and the research surrounding it being vague and contradictory. Since the lack of a bespoke determination and work done surrounding the optimal value of ‘k’ for electricity datasets, additional extensive measures were taken to determine which k value is custom best suited to the electricity data set.

Code was adapted to run the k-NN regression algorithm for a total of 53 different Values for “k” ranging between 1 and 156 (Square root of training data set) with increments of 3. The code would, as a result, calculate the precision of the predicted data, accuracy of the result based on correct and incorrect predictions and mean square error for every given value of “k”. There were built-in functions for measuring the precision, and mean square error, but the accuracy was computed based on code that was written to compare the testing dataset with the predicted data set to determine the number of correctly identified instances. Accuracy was then calculated using the formula:

**Accuracy Percentage Formula:**

The total values of ‘k’ used were 53, but the table below highlights

**The accuracy variation for said values of k were as follows (based on both distance and uniform):**

|  |  |  |  |
| --- | --- | --- | --- |
| **Values of ‘k’** | **Accuracy (%)** | **Precision score** | **Mean Square Error** |
| 1 | 96.75 | 0.968 | 0.0345 |
| 3 | 90.21 | 0.966 | 0.0246 |
| 9 | 74.77 | 0.964 | 0.0287 |
| 99 | 19.18 | 0.938 | 0.0571 |
| 156 - Square root of training data set | 9.87 | 0.930 | 0.0662 |

Table 5. Variance of performance with changing values of 'k'

The values of “k” and their resulting accuracies, precision and Mean Square Error (MSE) values were stored in lists.

Since the values of k considered were 53 and the linear representation of it in a table wouldn’t have aided, the said data was plotted in graph format to visually display the findings for a superior understanding of the effect of the values of “k”:

**The Graph for Values of “k” against Accuracies is as follows:**

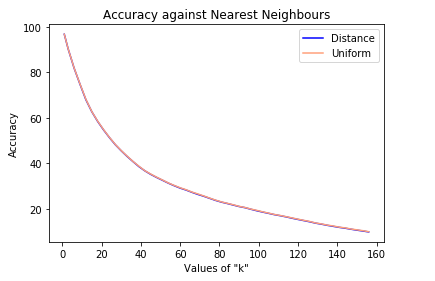


Figure 2. Accuracy against values of 'k'

There is a visible decrease in the accuracy of prediction with increasing members considered during the categorization and placement of a test dataset instance.

Although the highest percentage of accuracy visibly being achieved at K = 1, the realisation that a number so small may and is capable of being influenced by noise and hence increasing the probability of inaccuracy incorrect prediction, it was decided to find middle ground and settle for K = 3. Because it seemed like a reasonable number and it seemed to be better suited to the given dataset.

**The Graph for Values of “k” against Precision score is as follows:**

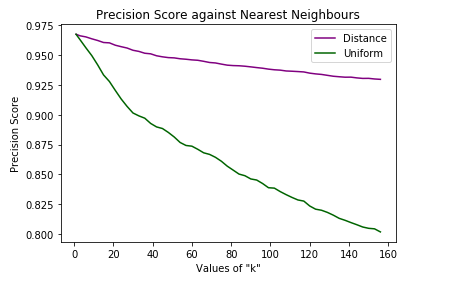


Figure 3. Precision Score against values of 'k'

**The Graph for Values of “k” against Mean Square Error is as follows:**

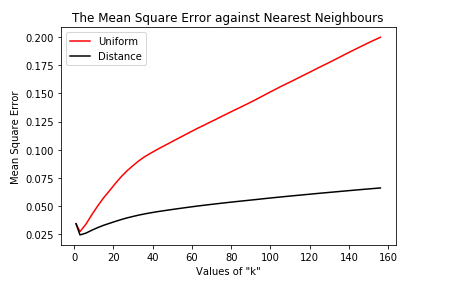


Figure 4. Mean square error against values of 'k'

Despite the research that has been carried out for the sole purpose of figuring out a suitable value for the number of nearest neighbours taken into consideration, during this research through personal experience it was explored that the value of ‘k’ is dependent on the type of dataset and shouldn’t be merged with existing notions of it working best when equal to the square root of the training dataset or some constant number. There is a reason that the value of nearest neighbours is not predefined and remains flexible as different types of datasets might call for different considerations. And the random assigning of different values for it reveals which one works best and gives an acceptable result.

All in all, it makes sense that a small value for ‘k’ might make the algorithm too influenced by individual cases whereas on the other hand, larger values may include too much of a deviation and complexity in a non-complex situation ultimately producing a prediction that is in either case not the most reliable and accurate as it can be.

However, there is obvious sense in the fact that it shouldn’t be too small either, but then again there is a pattern that shows that keeping it too large can cause deviations and the majority vote can get influenced given if the data, being worked with, is noisy. Having said that, increasing the value of K also saw a visible constant being reached, although the overall accuracy of the system came down arguably significantly, at a point increasing the value of ‘k’ would produce accuracy very close to that for a closer k value and the difference would not be substantial. This showed that whilst accuracy was being compromised, a large value for the number of nearest neighbours would also make the system less prone to be persuaded by noise and irregularities in data.

**Algorithm Implementation**

The determination of ‘k’ was the principle errand for both the k-NN algorithms (Regression and Classification). But the actual implementation of the algorithm was straight forward. Having said that, the outputs from the regression model were floats as it is the resulting calculation of an average of nearest neighbours and division often leads to a decimal number. This algorithm posed a challenge in the classification of correctly and wrongly determined instances. The result had to be “round up” using the built-in python function for exploitation.

**3.3.3 Linear Regression**

Implementation, problem with the detection of 0, critical reflection

**3.3.4 Neural Network Classification**

Unlike the other algorithms implemented in this research, the neural network model had a flexibility margin and capability to be refined and hence better the results. Due to the composition of the classification algorithm to include loss functions, activation function, dropout, number of layers, number of iterations and batch size, which are all variable factors they heavily influence the results and the overall performance of the algorithm.

Changing these variables and trying different combination of values for each lead to interesting discoveries.

NUMBER OF LAYERS, NUMBER OF NEURONS, critical reflection

**Amidst other factors, the most important dynamic factors were:**

**Loss function:** 'categorical\_crossentropy'

**Activation Function:** ‘relu’

**Epochs:** 100

**Batch Size:** 32

Batch size is the split or number of data instances taken at a time to train the system for one go or iteration. For a given dataset of 100 instances with a batch size of ‘5’ will take 4 iterations of 20 instances every time to train the model rather than process and train all 100 instances in one go.

**Dropout:**

Dropout is a technique for addressing overfitting. Its purpose is to randomly drop units (along with their connections) from the neural network during training. This prevents units from co-adapting too much (Srivastava, 2014).

Since no background knowledge or experience with the algorithm was at hand, the sole method of refining the results were to alter the variables and functions randomly with hopes of incurring a better result. The ‘AccuStats’ self-written function was used to measure the correct instances after every change made to the variables and the algorithm run. It is safe to say that not all possible combinations were tested but the mentioned variable values above gave the best result in comparison and to the best of knowledge. It was noted that a significant change in results was observed when Epochs was increased. The adjustment of others didn’t provoke any notable consequences.

**Initial vs final results**

Although not drastic, changing epoch from 20 to 100, reduced error, courtesy of the loss function given a chance of more iterations to work over. The initial accuracy ((Correct determined instances/total instances) \* 100) was 48. Increasing the epoch to 100 significantly increased algorithms runtime but yielded a better result of 58% accuracy. Logically it would be perceived that further increase of epoch would harvest better results but it only increased program runtime and produced a result that was 0.0002% worse. This could mean various things such as the data being noisy and unrefined, the number of layers in the neural network to be insufficient or more than etc.

**3.4 Statistical Analysis**

There are many procedures at hand to determine the success of algorithms in reaching a given target. The algorithms can be tested on their ability using markers such as speed of computability, solution quality, simulation, run time, errors encountered, optimisation etc. Many of these units are computable within the scope of code and the language being worked with. With an import of a library and the administration of parameters to a function and running the said code, these units of measure can be determined with the click of a button.

Such methods of algorithm analysis can be computed in one line of code. They are in most cases considered sufficient but it can be argued that they tend to be incomprehensive and in some cases, rendered meagre. In a professional environment however, there is a call for a far more encyclopaedic analysis especially in areas where a heavy part of research is formulated of a comparative statistical analysis.

Upon meeting with statistical expertise for mentoring regarding the most viable possible method that could be devoted to the investigation, Chi squared error was initially considered but on further description of the nature of the system and its outputs, it was rendered unprofitable and precision recall and F score was suggested instead.

Hence, this dissertation makes use of many notable techniques and manual mathematical calculations in collaboration with the results acquired from the algorithms and code to put forward a reasonable and thorough analysis of the algorithms that were worked with.

The statistical techniques and calculations applied to the results from each algorithm are as follows:

* F1 score
* True positive rate
* False negative rate
* True negative rate
* False positive rate
* Prevalence
* Positive predictive value precision
* False discovery rate
* Accuracy
* False omission rate
* Negative Predictive Value
* Positive likelihood ratio
* Negative likelihood ratio

*\*Some of the above are already listed in Chapter 2 – Literature Review, but as a recap and for additional definitions and formulae are mentioned again.*

**3.4.1 Test Statistics**

The statistic calculations mentioned above use “test cases” which are mentioned below. They are variables based on the results from the prediction algorithms (Groups.bme.gatech.edu, 2018).

**The test cases are as follows:**

* **True positives (TP)**

The algorithm correctly predicts the presence of a person as positive. Meaning predicting that a person is home when they in fact are.

* **False Positives (FP)**

The algorithm incorrectly predicts the presence of a person as positive. Meaning predicting that a person is home when they in fact aren’t.

* **True Negatives (TN)**

The algorithm correctly predicts the presence of a person as negative. Meaning predicting that a person is not home when they in fact aren’t.

* **False Negatives (FN)**

The algorithm incorrectly predicts the presence of a person as negative. Meaning predicting that a person is not home when they in fact are.

* **Condition Positives (CP)**

The total instances in the data where the person is home, where person occupancy or presence is positive.

* **Condition Negatives (CN)**

The total instances in the data where the person is not home, where person occupancy or presence is negative.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Positive | Negative |
|  |  | Actual data (+)  (Person(s) home) | Actual data (-)  (Person(s) not home) |
| Positive | Predicted Occupancy (+)  (Person(s) home) | **TP** | **FP** |
| Negative | Predicted Occupancy (-)  (Person(s) not home) | **FN** | **TN** |

Table 6. Prediction Classification

**3.4.2 Calculation of Test Statistics (TP, FP, TN, FN):**

A function was written specifically to classify and count the occurrences of the test statistics for all the algorithm. The function took as input parameters the prediction list and testing data list then compared and classified them accordingly to compute the corresponding test statistics which were then printed by the function.

A prediction was classified into test cases based on the confusion matrix and definition of the test cases above. The algorithm responsible for this calculation can be found in Appendix A – Figure

**3.4.3 Definition and formulae of Statistical computations**

Following is a brief explanation and formulas of the above-mentioned methods to test the prediction algorithms:

**F1 score – (Primary Method)**

Also known as “F score” or “F measure”, this calculation measures the accuracy of a subject (in our case the predictive algorithm). It can be explained as the weighted average between the precision and recall of the subject where precision is the regularity with which retrieved findings or predictions are relevant or ‘correct’, and is properly a form of Accuracy, and recall is defined as the regularity with which germane findings are simply ‘recalled’ by the algorithm,

An F score intends to combine these two into a single measure to search ‘effectiveness’. Its best possible value is 1, denoting a highly efficient and accurate subject which could not be bettered and worst possible outcome of 0, signifying no room for possibility of having performed more poorly.

<https://arxiv.org/ftp/arxiv/papers/1503/1503.06410.pdf>

The actual formula based on the definition of F1 score is as follows:

But through methods of substitution, a more direct and applicable formula for it can be derived:

**Positive Predictive Value (PPV) - Precision**

A second more broken down name for precision which makes the formula more understandable. Defined as the probability that instances where the positive was predicted (person is home) is true. (*Positive And Negative Predictive Value* 2018)

Formula:

**True positive rate (Sensitivity) - Recall**

Also referred to as the sensitivity or “hit rate” of a subject. As the name suggests, measures the proportion of positives identified correctly. Example the proportion of times the algorithm predicted occupancy of the household as positive (Someone home) when there was someone home. (*Sensitivity And Specificity* 2018)

Formula:

Since False Negatives are ultimately all the instances that are true, total positives = True Positives + False Negatives

**False positive rate**

Is defined as the proportion of negatives that return positive outcomes. The probability of a positive prediction result produced given that the condition being looked for is not present (negative). the expectancy of the false positive ratio. It is calculated as the ratio between number of negative events incorrectly classified as positive to the total number of negative events. (*False Positive Rate* 2018)

Formula:

Since False Positives are ultimately all the instances that are negatives, total negatives = True Negatives + False Positives

**True negative rate (Specificity)**

Also referred to as the specificity or “miss rate” of a subject. As the name suggests, measures the proportion of negatives identified correctly. Example the proportion of times the algorithm predicted occupancy of the household as negative (No one home) when there was no one home. (*Sensitivity And Specificity* 2018)

Formula:

Since False Positives are ultimately all the instances that are negatives, total negatives = True Negatives + False Positives

**False negative rate**

Clearly complimentary to “False Positive Rate”. Defined as the proportion of positives that return negative outcomes. The probability of a negative prediction result produced given that the condition being looked for is present (positive). (*False Positives And False Negatives* 2018)

Formula:

Since False Negatives are ultimately all the instances that are true, total positives = True Positives + False Negatives

**Prevalence**

Is defined as the proportion of a specific condition to be true. In our case of household occupancy, it will be defined as the fraction of times when a person is truly home to the total number of readings. Although not a unit of measurement of the accuracy of an algorithm, this measurement will not remain same for every algorithm and only signifies the popularity of occupancy.

Formula:

**Negative Predictive Value (NPV)**

Defined as the probability that instances where the negative was predicted (person is not home) is true. (*Positive And Negative Predictive Value* 2018)

Formula:

**False discovery rate**

Defined as the percentage of false predictions in a given set of predictions. (*False Discovery Rate | Columbia University Mailman School Of Public Health* 2018)

Formula:

**Accuracy**

Generally defined as the calibre to which the outcome of an assessment, calculation or measurement is precise or correct. In a statistical environment, error is taken into consideration. In order to achieve a high accuracy both precision and correctness of a measurement need to be high.

Formula:

**False omission rate**

False Omission Rate a measure of False Negatives that are incorrectly rejected. It is a compliment of the Negative Predictive Value.

Formula:

**CHAPTER 2**

**ANALYSIS**

**4.1 Visual Aids**

Theoretical analysis of the algorithms performance can bode well in situations where they are being thoroughly explained since they tend to be lengthy and hard to absorb and differentiate between. Visually presented data in collaboration with notation however, can be better absorbed and understood.

As this section heavily refers on visual representation, to gain a better insight to the following notation please refer to the figures in section 1 of Appendix A.

The figures display the graph for the Predicted ‘athome’ variable against the actual testing set for each algorithm. For the k-NN algorithms (Regression and classification) as seen in Figure 8, 9, 10 and 11 the predictions lie within the definition of the actual data set for the classification algorithm whereas in the case of regression, there is a possibility to predict any float value between or relatively close to the definition of the actual data (0 - 2).

In the case of classification however, the only possible outcome is one of the three, 0, 1 or 2. Hence the scatter graph differs from that of the k-NN regressor as seen in Figure 9. There is the occasional variance and randomness in the graph which is expected by the mean calculating algorithm k-NN regression and the military overlapping of the predicted data within the definition of the actual data scatter in the case of classification as seen in Figure 11.

Since the plot was plotted using ‘scatter’ for actual data and ‘plot’ for the predicted, the predictions are displayed using one conjoined line which is why it takes over and overlaps most of the visible graph. Although inefficient the ‘plot’ method was chosen over the ‘scatter’ because the complete overlapping in the scatter graph failed to show the actual data set scatter beneath.

The inability to view the actual data set, although, convoluted to interpret, also stands as a testament to the k-NN classifications efficient predictive ability.

**4.1.5 Notation**

**The test cases classification index for each algorithm:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **TP** | **FP** | **TN** | **FN** |
| **k-NN Regression** | 8028 | 64 | 3537 | 41 |
| **k-NN Classification** | 7998 | 71 | 3530 | 61 |
| **Neural Network** | 5406 | 2182 | 1419 | 947 |
| **Linear Regression** | 3437 | 3601 | 0 | 0 |

**4.2 Statistical Analysis Report:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | k-NN Regressor | k-NN Classifier | Linear Regressor | Neural Network regressor |
| False omission rate | **0.011** | **0.017** | **0.000** | **0.425** |
| False discovery rate | **0.008** | **0.009** | **0.511** | **0.295** |
| True negative rate (Specificity) | **0.982** | **0.980** | **0.000** | **0.360** |
| False positive rate | **0.018** | **0.020** | **0.000** | **0.640** |
| False negative rate | **0.005** | **0.007** | **0.000** | **0.148** |
| Negative Predictive Value (NPV) | **0.989** | **0.983** | **0.000** | **0.575** |
| Precision (Positive Predictive Value) | **0.992** | **0.991** | **1.000** | **0.705** |
| Recall (True positive rate /Sensitivity) | **0.995** | **0.992** | **1.000** | **0.852** |
| Accuracy | **0.991** | **0.989** | **0.488** | **0.670** |
| F1 Score | **0.994** | **0.992** | **0.656** | **0.772** |

**4.3 Comparison & explanation of the results.**

F1 score was used as the primary measure of the compared algorithms ability however, the other computed measures were of equal, if not paramount, significance as they measure the ability of the algorithms in different areas.

The entrant Algorithm ranking (Best to Worst):

1. **k-NN Regressor**
2. **k-NN Classifier**
3. **Neural network Classifier**
4. **Linear Regression**

The k-NN regressor algorithm performed admirably in the prediction of the household occupancy. With an F measure of 0.994 which demonstrates its precision and recall being laudable. The k-NN classifier followed with a very close F measure of 0.992 and the neural network classifier and linear regression with significantly varying and poor F measures of 0.772 and 0.656 respectively.

The structural and theoretical similarity in the k-NN regressor and classifier reflects in its results which are close. Regardless of F measure, even in other stats k-NN regressor algorithm displays superiority like accuracy(precision), true negative rate and false discovery rate. Other stats calculations would have been bagged by the said algorithm too if it wasn’t for the errored calculations from the linear regressor due to its inability to recognise 0’s which caused the False negative and true negative test cases for the algorithm to equal 0 hence altering the course of its analysis and producing faulty results.

That is where the failure of the linear regressor originates from. The only two variables in the output of the defeated algorithm were 1’s and 2’s.

The discrepancies between the winning algorithm and FILL IN THE ANALYSIS

**CHAPTER 2**

**CONCLUSION**

A relatively imperative objective of this research was the establishment of one data analysis algorithm as superior to the rest, not entirely but, in the specific scenario it was adapted into. Having said that, amongst the various algorithms used k-NN regressor performed best with an accuracy and precision value of 0.991 and 0.992 respectively and finally down to the primary statistical measure F Score a sweeping value of 0.994. meaning it would correctly predict the occupancy of the house 99.1% of the times and wrongly determines it less than 9 out of 994 times.

The findings of these research have positive educational repercussions and if they don’t push the boundaries of knowledge they have at least stood up to standards and to the borders of knowledge in terms of comparison between two, of each, regression and classification models written with the intent of prediction.

Researches regarding algorithms have been carried out immeasurable number or times across the world but keeping in sight the related work that was sought after for the population of section ADD literature review and all that lead up to it, it is safe to say that the findings of this research are unique and next to none in terms of the composition, application, choice of algorithm and statistical analysis of their performance. Individually countless researches have been done in these areas but this research work incorporates all these

These wrong determinations lie False Negatives. In the case of our project False Negatives can be defined as the set of those predictions that wrongly predicted a person not being home when the person, as per testing data, was available at home. This is a clear auditee. Given the possibility of determining auditees from sufficiently good results – the system can be used to identify problems – In more specificity the problem of someone being home but not using their electricity like they usually do perhaps because they have likely fallen victim to the unfortunate occurrences that come with independent living amongst the senior and disabled. This clearly highlights the tree of events starting from electricity data analysis to getting the said parties the help they need. This, in turn, proves and answers the research question “Can data analysis be used to get the elderly and disabled the help that they need”. Yes. In fact, it can. It is safe to say, with slight maturing of this dissertation work, data analysis can successfully be used, in collaboration with external real-time data readings and a little furtherment and linking of the written code for this research, to help getting the reliant the help they need.

The k-NN regressor algorithm taking the lead came as a surprise given its float mean output and its need to be rounded up whereas the classifier model has definitive outcomes it was expected to be more efficient.

The complexity, numerous variables and combinations of the neural network played a role in its refinement otherwise given its popularity amongst electricity data based prediction system, it was certain that it would have precedence over the other models.

On the other hand, linear regression algorithm lived up to its expectations and there was a reason for its unpopularity in electricity data analysis. EXPAND

**CHAPTER 7**

**REFLECTION**

The fact that machine learning, statistics and a critical evaluation and execution of existing and self-implemented work were all things that I haven’t done before yet expected of me in this dissertation are what I take pride in most. The principle concepts as listed above, where more than foreign to me.

Understanding the dissertation and what is expected of us, data collection, primary research and literature reviews was the most challenging part of this project. The sections such as literature review, methodology, analysis, results and conclusions had straight forward names and explanations that seemed to be understandable. We were told what was expected of us in each section and their implementations but it was the actual implementation that was challenging. Related work was to be read and analysed in the literature review but exactly what areas where crucial to discuss, what information had to be in the summary and what was unnecessary. Although the initial understanding was quite clear it wasn’t until I got to typing did I realise I had so many questions and concerns. They were handled and rectified in due time. One very important thing that helped with this would be the early submission of the Project Proposal which had a mini Literature Review, abstract and client motivation which helped getting us in the mindset of understanding what and how the dissertation worked. Although not much from the initial project proposal was made use of in the final project hand in, it was still of immense importance when it came to grasp the idea of the professionalism and formal nature of the report.

**Dataset:**

Projects that involve data mining often struggle with finding the perfect data set. I was lucky in terms that the I wanted a minimalist data set. An extremely simple one with very little and necessary attributes only. The struggle came in choosing between the two datasets that I narrowed it down to. One dataset was a public one from Kaggle. It included the required data but had other additional data and that too of 30 different households in a neighbourhood. This data was comprehensive yet a bit too intricate for the intentions of my project. Another aspect of this data set was that power reading from the household were taken every 30 minutes which made it undesirable to be worked with given the urgency and demand of my project application.

The other dataset that I chose came from my supervisor’s personal household. It changed the course of my ethics application and some formalities and consent forms had to be fulfilled but the dataset was perfect when it came to the requisites of the project.

The splitting and usage of the “datetime” attribute was not impossible but laborious. And given a situation where data had to be well broken down, the “datetime” field was left as it is as a reference and another serialised attribute by the name “unit” was added to overcome the struggles of manipulating the datetime field within python. Yes, even after the implementation of nearly 5 workable datamining algorithms I struggled with manipulation of data.

The importance of the occupancy prediction weighed out the patterns in electricity data usage based on the nature of days and the types of days (for example Mondays, weekdays, weekends etc). It is crucial when you think of it, it is crucial to help refine the system but in my opinion as a starting system even a false negative works just as well to indicate that there is room for concern.

**Implementation/Algorithms:**

The implementation of the machine learning algorithms seemed to be frightening. And it was at first. There was a learning curve for the understanding of how they worked. Small things such the training set, the testing set, supervised and unsupervised learning, different types and techniques to do the same thing. The availability of so many different techniques is what encouraged me to try out different ones and statistically analyse which one worked best. I was initially supposed to only implement k-Means clustering because theoretically it was the most understandable because of its graphical logic and it seemed to be the easiest. I had jumped at the mere thought of being expected to implement a something as complex sounding as a neural network. However, in the ended I ended up not even delivering a functional clustering algorithm but 4 different algorithms instead.

**Dropout of cluster algorithm:**

The coding and actual implementation of the code for the k- NN regression and linear regression didn’t take as long as understanding their concepts did. Similarly, for Neural networks, which was one of the more complex algorithms which differed most from the rest that were implemented.

Although all algorithms followed a different pattern, they were basically connected from the roots. Their basic logic of determination and prediction was very similar and once that basic concept of understanding machine learning and supervised learning was understood, the algorithms started to make sense.

Statistical analysis was a huge part of this research, it was the ending note and the determination of one victorious algorithm. Due to the requirement of rigour, a statistician from the sigma mathematics support session was sought after for expert opinion on the matter. The one-to-one meeting proved to be more confusing than helpful as the proposed methodology by the expert for the statistical analysis was “chi-square error” in python. Which has little to nothing to do with test cases (tp, fp, tn, tp)

However, after the further explanation of the system and the ability to calculate the true negatives and false positives etc and mention of precision, recall and f measure, I was informed that I was on the right track and f score was the way to go. I then shared material on f score that I had found earlier to double check and get validation over before I chose it as my primary method of algorithm analysis.

The most challenging part of the dissertation was the sections. Specifically, methodology and literature review. But the thing about literature review was that as you would read and read on related work your mind will pick up on the tone and structure of journal articles, researches and dissertations.so despite it posing off as a challenge, the more literature was read, the clearer the structure and understanding of the repot got.

**Presentation feedback:**

I was initially stressing a bit too much about the word limit and had a percentage plan of how many words would go into each section which my supervisor picked up on and advised me to take it easy and not be too bothered about it.

I had changed the whole course of the project up at this point, as opposed the initial plan to do a clustering algorithm only, I was now doing 4 different algorithms with the addition of a statistical analysis to compare them.

I was briefed on the necessity of having to perform a “rigorous statistical analysis”. Not just python built in functions like mean square error or precision score. The use of test cases (TP, FP, FN, TN) to calculate precision and recall then ultimately F score to crown an algorithm.

Additional feedback that was given was as follows:

* The need of a personal reflection on the writing experience what went right? What went wrong? Etc.
* Write-up stance to be formal and in third person. Avoiding citations from websites and blogs.
* To clearly highlight the intended application and the actual implementation of the project as the project title and explanation was prone to confusion as to what exactly was being implemented.
* Methodology to not include definitions and explanations. It is a section that has more to do with implementation of the project than definition.

The given feedback was recorded on my personal mobile phone voice recorder with the consent and awareness of the present parties (Supervisor: David Croft and Second Marker: Matthew England)

**CHAPTER 7**

**PROJECT MANAGEMENT**

**Project Plan**

This research took inspiration from iterative Agile techniques as its basic execution plan. Sprints every week were planned, executed and sought after. The techniques applied to the dissertation were inculcated in me during the entire course of this degree. Project, and time management has played a vital role in the success of this dissertation. The initial breakdown of tasks wasn’t ideal and exhaustive. It lacked many areas that showed themselves later at execution stage. The iterative adaptation and delivery of the project made it possible to go back to these areas and work on them accordingly.

Even in coding, after research based on the algorithms was carried out which formulated most of the literature review, all the different algorithms (Neural networks, k-Means clustering, Linear regression and k-NN models) were implemented and written from the bottom up, meaning their most basic initial stage and worked up in the form of reasonable increments. This helped breaking down a seemingly large problem in to smaller SMART (Specific, Measurable, Achievable, Relevant, Time limited) goals and tasks.

**Task breakdown example:**

|  |  |
| --- | --- |
| **Sequential Technique** | **Iterative process (adapted technique)** |
| 1. Implement the data mining algorithm to predict occupancy of a house | 1. Import resources    1. Import Dataset    2. Import relevant libraries    3. Extract variables from dataset    4. Write the basic algorithm |

Table 7. Example of task breakdown, sequential VS iterative process

**\*** For instance, the above table shows the implementation of the algorithms using a sequential/un-planned technique as opposed to the adapted iterative technique with measurable distribution of work load.

**Project Schedule:**

**Initial Gantt Chart:**

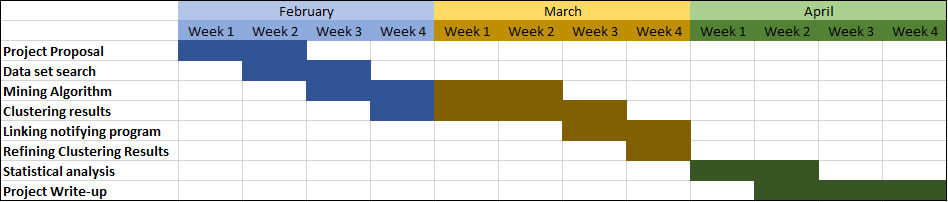


Figure 5. Initial Gantt chart

The initial Gantt Chart included a list of tasks and the time they would take (in weeks). The tasks were heavily underestimated in terms of the time they would require and their breakdown. There was potential in the tasks to be broken down much further to make them more workable.

**Final Gant Chart:**

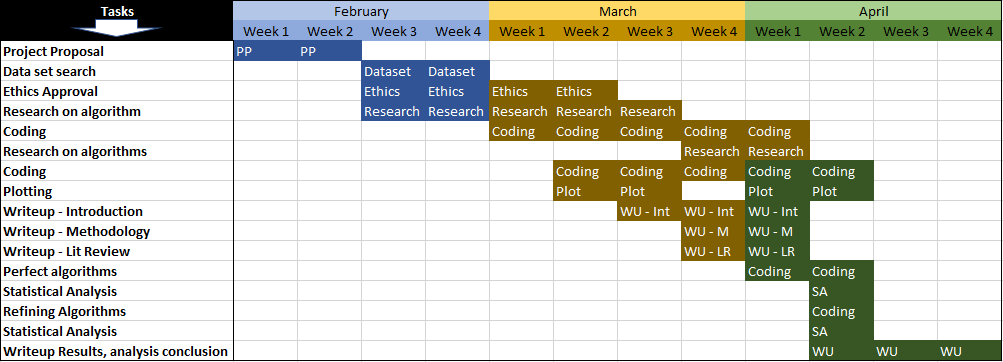


Figure 6. Final stages Gantt chart

The underestimation of tasks in the initial Gantt chart is evident in comparison to the final one. The drawbacks of conventions sequential management techniques are far too many and its inflexibility renders projects unsuccessful and failing in terms of time budget etc.

**Social, Legal and Ethical considerations:**

Ethical implications of this system. Ability to detect people’s presence if fallen into the wrong hands, this program can have severe ethical implications as it can be used to determine that a person isn’t home and as a result burgle their house etc.

**Risk Management:**

There are numerous paths that can be taken into consideration, to make a project less prone to error, such as the following:

**Avoidance of Risk**

The simplest way to abstain from getting involved into an identified risk is to try and avoid it altogether. This was done by going down paths that were socially and ethically least expensive. For instance, the data set that was used of the research was strictly kept to as minimalist as possible. It contained no sensitive personal information and a legal professional signed consent of the involved personalities to ensure awareness and involvement in the project. Risk was also avoided when a dataset was discarded due to its extensive information regarding area, house ID’s and social security numbers of residents which could have proven to be in breach of a data protection law.

**Risk Mitigation**

After taking steps to avoid the possibility to attract risk, the chances and effects of risk can be minimised or reduced as well. Risk was minimised in the project by task breakdown and divide and conquer method. Breaking down problems and tasks into more measurable ones. If too much was taken on at once, the risk of losing understanding and likeliness of something going wrong increases. This was prevented by the incorporation of sprints and increments as seen in Quality assurance section.

**Quality Assurance:**

**Sprint backlog – burn down chart – timing report**

**Supervisor meeting logs:**

**First meeting (Group) – 29-12-17, EC1-01**Detailed Structure of the dissertation and key areas - Topic discussed – Research areas and topics to research on listed – Project proposal explained – Sample project proposals emailed – for a clearer idea of what was expected of us.

**Second Meeting (Extended topic of research) – 11-01-18, Techno Centre**further explanation and breakdown, narrowing down the problem before breaking off for break.

**Third meeting (Project proposal -Literature Review) – 06 -02-18, William Morris:**Questions related to the project proposal content. How to address that were still uncertain like dataset and algorithms.

**Fourth meeting (Final Project Proposal Draft) 07-02-18, ASG-31:**Project proposal drafts and queries sent and feedback given via emails. Literature review in-text citations and client motivation area asked to be updated and pointers given. Structure redefined inform of questions that needed to be answered for example: Who is the target audience? how will the proposed system be beneficial to them etc.

**Fifth meeting (Ethics, Data set) – 20-02-18, ECM-15:**Ethics application submission discussed and submitted. The initial public Kaggle based dataset of 30 households discussed and played around with. The option to use Davis’s electricity dataset arose and ultimately chosen. The dataset was received via email. The ethics application was updated as a public open source data was not being used. A consent form between the owner and researcher written and ethics updated.

**Sixth meeting (Dataset Methodology) – 07-03-18, ASG-31:**The collection of the dataset explained. The manual techniques and the equipment required for the collection of the electricity consumption data as well as the pings that recorded the availability of the person at home. Ethics consent form designed discussed and the relevant elements pondered over.

**Seventh meeting (Clustering, k-Means Algorithm) – 22-03-18, Planet Earth:**The split in the dataset discussed. The need for the split to be randomly picked not containing consecutive instances in the training data (70%) and the remaining consecutive occurrences to formulate the testing set (30%). The k-Means algorithm and its nature discussed. The fact that the number of clusters (K) needs to be user provided.

**Eighth meeting (Deviation, Qualitative and Quantitative, Stats, Algorithms) – 27-03-18, EC Building:**Explanation that the understanding of the data machine learning had led me to understand that clustering alone is a stand-alone method to plot data based on similarity and trends and an additional methodology needs to be implemented toclassify an input testing data for prediction. A method such as nearest neighbour incorporated for prediction. The possibility of applying more than one algorithm and performing a statistical analysis to prove which works best in the given scenario. The possible ways to analyse the results, precision, accuracy, recall, f1 score etc discussed. The importance of having to meet with a statistician expert to agree upon an analysis methodology.

**Ninth meeting (End of term Presentation) – 29-03-18, William Morris:  
Attendees:** David Croft (Supervisor), Matthew England (Second Marker), Self.

All the progress in numerical values. 20% of the writeup along with nearly 45% of the coding done. What was left to be done, what were the problems encountered. Dissertation writeup breakdown, Word limit breakdown, project explanation from the bottom up. Its application and methodologies discussed.

Feedback from the presentation recorded on phone using voice recorder with consent of all present parties *(More on the feedback from the presentation in the reflection section)*

**Eleventh meeting (Draft check, run through) – 17-04-18, EC Building:**Limited feedback given at this stage as per university regulations. Suggestions and pointers as to what section might need another look etc. Structure, reflection and analysis section queries cleared out. Code unrequired but maintained in GitHub. Something of immense importance to be added to report itself otherwise snippets of relevant information to be added to appendix and referred to in the section of choice.

**Twelfth meeting (Draft check, run through) – 25-04-18, EC Building:**Analysis report, table of F-measure and statistical calculations. Feedback to keep it to the point and explain the findings and statistical calculations. Weekly log report, project management, analysis and project proposal, appendix, declaration of originality structure and placement discussed.

****

**Signed:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
David Croft   
(Supervisor)

**Dated:** \_\_\_\_\_\_\_\_\_\_\_\_\_28-04-18\_\_\_\_\_\_\_\_\_\_\_

**APPENDIX A**

**(Code and related material)**

**Function written to calculate the correct predictions then calculate percentage of accuracy:**

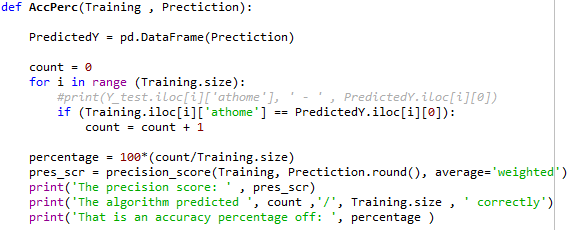


Figure 7. The Accuracy calculation function

**Function written to calculate the test cases for stats calculations:**



**Results from the algorithms. Plots of Predicted VS Actual data**

**k-NN Regressor:**

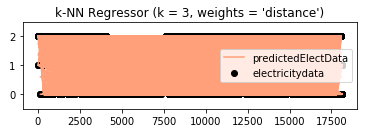


Figure 8. k-NN regression actual VS predicted data (plot)

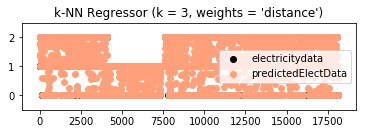


Figure 9. k-NN regression actual VS predicted data (scatter)

**k-NN Classifier:**

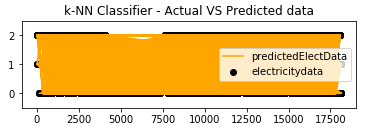


Figure 10. k-NN classification actual VS predicted data (plot)

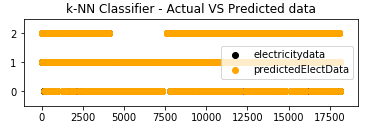


Figure 11. k-NN classification actual VS predicted data (scatter)

**Linear regressor:**

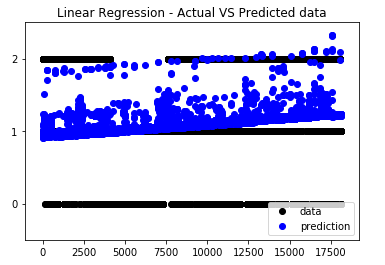


Figure 12. Linear regression actual VS predicted data

**Neural Network Classifier:**

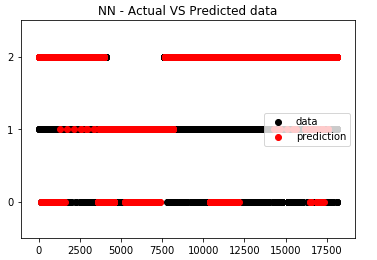


Figure 13. Neural Network Classifier actual VS predicted data

**The test cases distribution per algorithm:**

Figure 14. The test case distribution per algorithm

**The test case distribution:**

|  |  |
| --- | --- |
|  |  |
|  |  |

**APPENDIX C**

**PROJECT PROPOSAL**

**300/303COM Detailed Project Proposal**

|  |  |
| --- | --- |
| First Name: | SHAHZEB |
| Last Name: | DAWOOD |
| Student Number: | 7015511 |
| Supervisor: | DAVID CROFT |

## Section one: Defining your research Project

**1.1 Detailed research question**

**Help:** Your detailed research question is the statement of a problem within the computing domain which you will address in your project. Refining the research question involves narrowing down an initial question until it is answerable using a primary research method(s) that you will conduct during the time of your project. The refined research question must not be so general that it is answerable with a yes or no answer. It must not be so broad that you would be unable to achieve a solution during your project. The key to this is BEING SPECIFIC: Narrow down the method or technology you will use, narrow down the group that the question refers to (localize a general question) If the project is still ‘too big’, can you think of a way to work on a part of the problem? Avoid using words that cannot be measured, by you, without a huge research budget e.g. 'effects on society', 'effects on business'. *Example:* The initial question "Does cloud computing effect business" needs narrowing down *(for a start the answer is yes) W*hat is meant by cloud computing? Or 'effect'? Or 'business', in this question? Refining this first question will involve narrowing it down to something you, personally, can measure. A refined version of this question might be: "Does implementing a cloud based voting system improve the speed of decision making in a small company in Coventry?" This refined question is implementable: You can now identify a small company to work with, document their current decision-making processes, implement a cloud based voting system, compare decision making speeds over a limited time period (say 1 month) and evaluate your findings. *A small piece of genuinely new knowledge is produced.*

|  |
| --- |
| *“Can data analysis of electricity usage in the home be effective in identifying circumstances in which elderly or disabled residents may require aid”* |

**1.2 Keywords**

**Help:** Include up to 6 keywords separated by a semi-colon; what keywords are appropriate to describe your project in an online database like Google Scholar? Keywords should include the general research area and the specific technologies you will be working with. *Example.* A project that proposes a novel way of visualising large amounts of twitter feed data may have the keywords: Data visualisation; twitter; hashtags; database design; graphics libraries. For further help, take a look at the ACM keywords list http://www.computer.org/portal/web/publications/acmtaxonomy

|  |
| --- |
| Electricity Usage Trends; Clustering; Data mining; Data Analysis; Predictive algorithms; Household |

**1.3 Project title**

**Help:** The project title is a statement based on your detailed research question. For example, the research question *'to what extent does a mobile application reduce the number of errors made in class registers at Coventry University in comparison to current paper based registers'* may be stated in the project title*: "A Wi-Fi driven mobile application for large group registers using iBeacons".*

|  |
| --- |
| *“The use of data analytics to identify circumstances in which elderly or disabled residents may require aid through changes in electricity usage patterns.”* |

**1.4 Client, Audience and Motivation:**

**Help:** Why is this project important? To whom is this project important? A research project must address a research question that generates a small piece of new knowledge. This new knowledge must be important to a named group or to a specific client (such as a company, an academic audience, policy makers, people with disabilities) to make it worthwhile carrying out. This is the ***motivation*** for your project. In this section you should address who will benefit from your findings and how they will benefit. Example: If you intend to demonstrate that a mobile application that automates class registers at Coventry University will be more efficient than paper based registers - the group who would be interested in knowing/applying these findings would be both academic and administrative staff at Coventry University and they would benefit by time saved and a reduction in their administrative workload. If you are making a business case for an organization explain how the organisation will benefit from your findings.

|  |
| --- |
| In a research survey carried out by the NHS it was revealed that around one in three adults over the age of 65 are susceptible to fall over at least once a year (NHS, 2015). Around half of whom are likely to fall more frequently. Another research concluded that two thirds of those who fall again in the next 6 months (Senior Health & Wellness Blog, 2018). In the UK alone falling being the cause of death is the most common unfortunate occurrence amongst seniors aged 75 and over (Age UK, 2017).  The application of this research project will be for the elderly and disabled people who reside on their own. Problems like getting around the house and doing daily chores etc can become a struggle at their fragile age and accidents often happen. As mentioned in the stats above, they can fall over or hurt themselves and with no one around, this can escalate into a grave matter.  Many steps have been taken to prevent such events from happening but most of the times they require additional expensive equipment like infrared or heat sensors in homes to monitor movement of the residents. However, electricity usage data is readily available for most households and the required data is being recorded but going to waste (the implementation of smart meters will later ensure the recording and storage of the consumption data for every household). Trends and patterns from this data can be a less expensive alternative solution to the same problem.  Electricity consumption data collected over a specific period can be analysed to derive trends which can then be compared to everyday usage to categorise the data as either normal or abnormal.  Abnormalities being anything like an electricity spike not being shown roughly at around 7pm when the inhabitant uses the kettle for their evening tea.  Cost of the physical persons going around  When abnormalities as such are visible, the system can choose to notify a user of the unusual occurrence in the data pattern.  In simple terms, this project aspires to implement a system that is capable of noticing abnormalities and deviations from normal trends in electricity consumption entries. |

**1.5 Primary Research Plan**

**Help:** This is the plan as to how you will go about answering your detailed research question - It must include a primary research method (an extended literature review is not an acceptable primary method). Think and plan logically. Primary methods may include experiments, applications or software demonstrators, process models, surveys, analysis of generated data …  
  
Example: In the class register example above "to what extent does a mobile application reduce the number of errors made in class registers at Coventry University in comparison to current paper based registers" - the research plan may involve: 1) Collecting and analysing paper based registers in a given class on five occasions. 2) Identifying the error rate average on these occasions 3) Designing and implementing a mobile application that automatically records attendance in class. 4) Deploying the application in the class on five occasions. 5) Identifying the error rate average of the mobile application on these occasions. 6) Comparison of data and summary of findings.

|  |
| --- |
| This research will use iterative Agile techniques as its basic execution plan. Sprints every 2 weeks or so will be allotted to minimum functionality at first which will then be refined and further worked on every consecutive sprint.   1. Identify and choose an electricity consumption data set. 2. Identifying data analysis methodology such as clustering, neural networks etc. 3. Perform data analytics on the data set to extract trends and patterns 4. Refine results though false positives and false negatives 5. Implementation of data learning algorithm and software to notify user or system upon detection of an abnormal instance of data. 6. Apply statistical analysis on the results to identify the reliability of the system 7. Optional (depending on time): Conduct a survey on the willingness of people for the implementation of the system |

This is the end of section one.

## Section Two: abstract and Literature review

**2.1 Abstract**

**Help:** An abstract is a short summary of a research project that enables other researchers to know if your report or research paper is relevant to them without reading the whole report. It is usually written retrospectively so that it can include findings and results. It is fully expected that you will rewrite your abstract when you come to write your final paper. For now, you should write an abstract of about 250 words that define the project described in section one. Before writing your abstract you MUST read some abstracts from conference or journal papers on *Google Scholar* or from *portal.acm.org* (to understand their style) and then provide your own abstract that outlines what your question is and what you 'did' to answer it.

|  |
| --- |
| The worldwide widespread of smart meters means more understandable & workable data will be produced for every specific household. In the UK the smart meter bill was passed on the 28th of November 2017 and the implementation of which will lead to raw electricity consumption data being produced and hence used to measure usage more efficiently (Gov.uk, 2013). This data can be mined and analyzed for more than just producing efficient bills and smarter readings. This project intends to provide an achievable, affordable and a minimal change or resource implementation required solution for people by analyzing their electricity data and extracting specific trends and electricity consumption footprints. Every person has different routines and lifestyles which impact and reflect a different electricity usage trend. Learning those trends can be useful to make a judgement or sense if something is unusual.  This will be done by using data mining techniques such as Clustering and regression. These techniques will be applied on pre-existing open source electricity usage data sets, the results from which will be used to investigate if a deviation from norm has been experienced. Where norm is a customized trend for every household. As an extension to that, the program will choose to notify someone if any unusual electricity consumption behavior occurs that is a deviation enough from norm to be considered suspicious. |

**2.2 Initial/Mini Literature Review (500 words – 750 words)**

**Help:** A literature review is a select analysis of current existing research which is relevant to your topic, showing how it relates to your investigation. It explains and justifies how your investigation may help answer some of the questions or gaps in this area of research. A literature review is not a straightforward summary of everything you have read on the topic and it is not a chronological description of what was discovered in your field. Use your literature review to:

• compare and contrast different authors' views on an issue  
• criticise aspects of methodology, note areas in which authors are in disagreement  
• highlight exemplary studies  
• highlight gaps in research  
• show how your study relates to previous studies

|  |
| --- |
| There is plethora of research work that contributes to the area of computer science that deals with data mining and analysis, but I could find only a handful of work that directly associates my research. The literature that I chose adopts different methods of data analysis to interpret and analyse the fitting of their work in a specific application.    A research extract from a book analysed electricity data of a household to extract its specific trends which were then, along with the number of inhabitants of the household, used as input variables to a fuzzy model, to forecast how probable it was for a house apparatus to be used or started within the next 60 seconds (Zhu, 2016). The electricity data set collection for this research was long term and collected over a period of 12 months. The research relates to the “Electricity data analysis” side of my research. A case-study on fuzzy tools was presented along with neural network based algorithm was presented to forecast the home electricity consumption 24 hours ahead of time.  Researches relating to Anomaly detection hugely relate to my project. The second part of my research is to detect, effectively, occurrences and instances of data that can be considered as anomalous. A survey-based journal article spoke in depth about the different kinds of outliers in data analysis and the challenges faced in isolating a given anomaly from normal regions (Chandola, Banerjee and Kumar, 2009). The research will particularly prove helpful in narrowing down concepts and techniques most effective in successful outlier isolation. In addition to that, it also spoke about techniques for Anomaly detection such as Neural networks, rule based, Bayesian Network, Support vector, Clustering. Nearest neighbour based techniques etc.  Another research relating to anomaly detection was more practical and aimed to detect abnormalities in 2 transformers of the UK based on data and variables such as temperature, vibration, moisture, load current etc between an aged and a new transformer (Catterson, McArthur and Moss, 2010). This paper concluded that the Conditional Anomaly Detection (CAD) technique can be used in the online monitoring of the transformers.  A research specifically relating to electricity and anomaly detection with extensive results in the form of figures, diagrams and tables carries out data analysis on electricity consumption data using 2 experimenting approaches namely statistical approach and clustering (Jakkula and Cook, 2010). It concluded that the wrong identification of an outlier ratio in Clustering:Statistical approach was 1:43. Meaning, in this given instance, clustering would be inaccurate 2.3% of the time whereas Statistical approach would wrongly identify an outlier 97% of the time.  Existing work done related to my research work takes things as far as analysing electricity usage data to either extract trends or test different data analytical techniques to state which one works better. My project intends to take it a step further, by using that analysed data and link it to an algorithm for notification purposes. The Anomaly Survey research discussed entirely different application areas of anomaly detection but none of them came close to the intended work and application of this project which makes me confident that the deliverable is new and hence will contribute to a novel application of outlier detection.  In addition to that, the anomaly detection survey spoke of the many techniques available for the data analysis and the results of the Electricity anomaly detection research stands to show that clustering is a viable data analysis technique for electricity consumption data. |

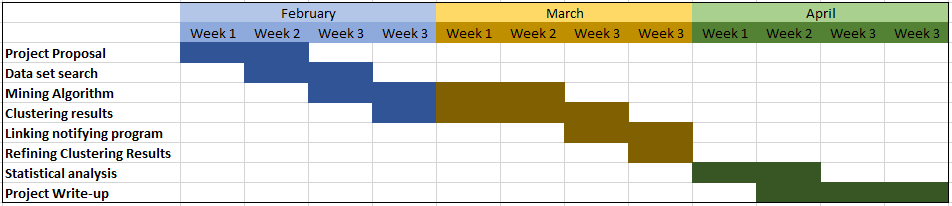
**2.3 Bibliography (key texts for your literature review)**

**Help:** Please provide references, in correct Harvard style, for at least three key texts that have informed your literature review. If you are implementing an application, select texts which demonstrate how other researchers have tackled similar implementations? The references should be recent and sufficiently technical or academic. Your markers will be looking for you to identify technical reports, conference papers, journal papers, and recent text books. Avoid *Wikipedia* entries, newspaper reports that do not cite sources, and general or introductory texts.

|  |
| --- |
| Age UK (2017) *Briefing: Health and Care of Older People in England 2017*, UK: Understanding Society.  Catterson, V., McArthur, S. and Moss, G. (2010). Online Conditional Anomaly Detection in Multivariate Data for Transformer Monitoring. *IEEE Transactions on Power Delivery*, 25(4), pp.2556-2564.  Chandola, V., Banerjee, A. and Kumar, V. (2009). Anomaly detection. *ACM Computing Surveys*, 41(3), pp.1-58.  Gov.uk. (2013). *Smart meters: a guide - GOV.UK*. [online] Available at: https://www.gov.uk/guidance/smart-meters-how-they-work [Accessed 9 Feb. 2018].  Jakkula, V. and Cook, D. (2010). Outlier Detection in Smart Environment Structured Power Datasets. *2010 Sixth International Conference on Intelligent Environments*.  NHS, 2015. *Falls*, NHS. Available at: https://www.nhs.uk/conditions/falls/ [Accessed February 9, 2018].  Senior Health & Wellness Blog. (2018). [Blog] *10 Statistics About Elderly Falls*. Available at: http://shellpoint.org/blog/2012/08/13/10-shocking-statistics-about-elderly-falls/ [Accessed 3 Feb. 2018].  Zhu, Q. (2016). *Complex system modelling and control through intelligent soft computations*. [S.l.]: Springer International Pu, pp.437-467. |

This is the end of SECTION TWO

Time Management Plan – Gantt Chart



\*The above is pre-mediated forecast plan for the breakdown and effective execution of the research project.

Detailed Project Proposal Grading Form

**The grade sheets for marking the 300COM / 303COM Detailed project proposal are attached on the next page.**

**Grading Notes:**

The proposal is marked out of 20 divided into 10 marks for the quality***, achievability and level of challenge demonstrated by the student's research question and proposed primary method* of solution generation** and 10 marks for the ***thoroughness of the proposa***l.

Modal grading: In awarding marks please consider the following modal template:

|  |  |  |
| --- | --- | --- |
|  | **Research question and primary research method in relation to learning outcomes** | **Thoroughness of the proposal.** |
| **>70%** | A well-considered project proposal that fully satisfies the Learning outcomes for which there is a succinct and focused aim with an associated project  A question or hypothesis that is well above norm for final-year undergraduate project level (approaching Masters level for >80%);  The project involves improving or developing a complex programme, tool, application or the enhancement of a theory or methodology or their application in a new context.  The project demonstrates a high degree of innovation and creativity | All fields completed demonstrating a clear blueprint for the research process and includes the necessary information with respect to the research question.  Research methods are well-considered with clear reasoning for choice of those methods over others;  A clear justification of the need for the project in relation to client or audience.  Projects proposals involving 'business case' reports clearly identify the organisation involved and consider how the case will be evaluated.  A sound grasp of the means of evidence by which the conduct and management of the project may be judged. |
| **Threshold (40%)** | A proposal that identifies an activity with some consideration of a broader context.  A research question which lacks enough substance, context and scope to allow for depth of analysis, but which is marginally acceptable against a threshold for final year undergraduate projects;  A primary method(s) which only just relates to the production of an appropriate solution to the research question. | Completion of sections is cursory or minimal with some cohesiveness and contextualisation.  Sections demonstrate some understanding of the research process involved which loosely links with idea outlined (key question, method, audience);  Research methods are discussed but demonstrate little consideration as to whether they are the most appropriate and lack refinement and further detail.  Identification of some methods of evidence for conduct and management of the project but unclear thinking about planning for reflection or accounting for conduct. |

**REFERENCES**

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B. and Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8.

Age UK (2017) *Briefing: Health and Care of Older People in England 2017*, UK: Understanding Society.

Alpaydin, E. (2014) Introduction To Machine Learning. Cambridge, Mass.: MIT Press

Al-Qahtani, F. (2013). Multivariate k-Nearest Neighbour Regression for Time Series data - a novel Algorithm for Forecasting UK Electricity Demand. In: *Proceedings of International Joint Conference on Neural Networks*.

Basheer, A., Ali Abbadi, M. and Awad, G. (2014). Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach. *(IJCSIS) International Journal of Computer Science and Information Security*, 12(8).

Brownlee, J. (2016) *Supervised And Unsupervised Machine Learning Algorithms*.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Ph. D. Aston University.

Catalão, J., Mariano, S., Mendes, V. and Ferreira, L. (2007). Short-term electricity prices forecasting in a competitive market: A neural network approach. *Electric Power Systems Research*, 7(15), pp.1297-1304.

Catterson, V., McArthur, S. and Moss, G. (2010). Online Conditional Anomaly Detection in Multivariate Data for Transformer Monitoring. *IEEE Transactions on Power Delivery*, 25(4), pp.2556-2564.

Chandola, V., Banerjee, A. and Kumar, V. (2009). Anomaly detection. *ACM Computing Surveys*, 41(3), pp.1-58.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). Deep learning. Cambridge, Mass.: The MIT Press, pp.136-150.

Gov.uk. (2013). *Smart meters: a guide - GOV.UK*. [online] Available at: https://www.gov.uk/guidance/smart-meters-how-they-work [Accessed 9 Feb. 2018].

Groups.bme.gatech.edu. (2018). [online] Available at: https://groups.bme.gatech.edu/groups/biml/resources/useful\_documents/Test\_Statistics.pdf [Accessed 23 Apr. 2018].

Haque, T. and Kashtiban, A. (2005). Application of Neural Networks in Power Systems; A Review. In: *World Academy of Science, Engineering and Technology*.

Hough, A. (2010) "The Telegraph". *More People 'Fear Losing Independence In Old Age Than Death'*

Hunter, J. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), pp.90-95.

IOT based smart energy monitoring. (2018). *International Research Journal of Engineering and Technology,* 5(3).

Jakkula, V. and Cook, D. (2010). Outlier Detection in Smart Environment Structured Power Datasets. *2010 Sixth International Conference on Intelligent Environments*.

James, G. (1998). Majority vote classifiers: Theory and application. Ph. D. Stanford University.

McKinney, W. (n.d.). Python for data analysis. pp.4-7.

Montgomery, D., Peck, E. and Vining, G. (2007). Student solutions manual to accompany introduction to linear regression analysis. Hoboken, N.J: Wiley-Interscience, pp.12-13.

Navot, A., Shpigelman, L., Tishby, N. and Vaadia, E. (2014). *Nearest Neighbor Based Feature Selection for Regression and its Application to Neural Activity*. Undergraduate. The Hebrew University Jerusalem.

NHS, 2015. *Falls*, NHS. Available at: https://www.nhs.uk/conditions/falls/ [Accessed February 9, 2018].

Oxford Dictionaries | English. (2018). *ping | Definition of ping in English by Oxford Dictionaries*. [online] Available at: https://en.oxforddictionaries.com/definition/ping [Accessed 15 Mar. 2018].

Palaniappan, S. and Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. 2008 IEEE/ACS International Conference on Computer Systems and Applications.

Raspberry Pi. (2018). *What is a Raspberry Pi?*. [online] Available at: https://www.raspberrypi.org/help/what-%20is-a-raspberry-pi/ [Accessed 15 Apr. 2018].

Reuk.co.uk. (2018). *Flashing LED On Electricity Meter | REUK.co.uk*. [online] Available at: http://www.reuk.co.uk/wordpress/solar/flashing-led-on-electricity-meter/ [Accessed 16 Apr. 2018].

Seber, G. and Lee, A. (2003). Linear regression analysis. Hoboken, N.J.: Wiley-Interscience, p.391.

Senior Health & Wellness Blog. (2018). [Blog] *10 Statistics About Elderly Falls*. Available at: http://shellpoint.org/blog/2012/08/13/10-shocking-statistics-about-elderly-falls/ [Accessed 3 Feb. 2018].

Sokolova, M., Japkowicz, N. and Szpakowicz, S. (2006). Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. *Lecture Notes in Computer Science*, pp.1015-1021.

Srivastava, N. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15.

Staff, O. (2018) *Comparing Costs For In-Home Care, Nursing Homes, Assisted Living And Adult Day Care | Ourparents Resources*

Suguna, N. (2010). An Improved k-Nearest Neighbor Classification Using Genetic Algorithm. *IJCSI International Journal of Computer Science Issues*, 7(4), pp.18-21.

Sundar, A., Latha, P. and Chandra, R. (2012). Performance analysis of classification data mining techniques over heart disease database. *International Journal of Engineering science & Advanced Technology (IJESAT)*, 2(3), pp.470 – 478.

Vaideek, K. and Vasundhara, K. (2011). A study on development of Smart Electricity Meter based on IoT by using NODEMCU. *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering (IJIREEICE )*, 5(11).

Vaingast, S. (2014). *Beginning Python visualization*. New York, NY: Apress, p.42.

Van Der Walt, S., Colbert, S. and Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*, 13(2), pp.22-30.

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F. and Mueller, A. (2015). Scikit-learn. *GetMobile: Mobile Computing and Communications*, 19(1), pp.29-33.

Yu, Y., Chen, L., Sun, F. and Wu, C. (2007). Neural-network based analysis and prediction of a compressor’s characteristic performance map. *Applied Energy*, 84(1), pp.48-55

Zhu, Q. (2016). *Complex system modelling and control through intelligent soft computations*. [S.l.]: Springer International Pu, pp.437-467.