The Prediction and Optimisation of Smart Energy Usage through Machine Learning Recommendations

MSc Research Project
Data Analytics

Mark McGrane
Student ID: x19140606

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Mark McGrane
Student ID:	x19140606
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	16/08/2021
Project Title:	The Prediction and Optimisation of Smart Energy Usage through Machine Learning Recommendations
Word Count:	XXX
Page Count:	30

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Mark McGrane
Date:	21st September 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

The Prediction and Optimisation of Smart Energy Usage through Machine Learning Recommendations

Mark McGrane
x19140606

Abstract

Even though renewable energy does not have the same limited supply as fossil fuels, it is still a commodity that needs efficient management to maintain an uninterrupted supply. The development of the Colour Code My Energy (CCME) Recommder algorithm provides personalised recommendations to users on how their current usage compares to an optimum value. Armed with this knowledge, the premise occupants can make instant adjustments and recalibrations of their habits when needed. Applying a Deep Neural Network(DNN), outcome predictions of the algorithm were retrospectively applied to the Hourly Usage Energy (HUE) dataset and demonstrated how the algorithm builds a knowledge base of best behaviours over time, with improvements on the quality of recommendations as it learns. Over time, the algorithm substantially increased its knowledgeable recommendations. Beginning with the ability to recommend 20% of reads, this increased to 80% by the end of the 3rd year. The DNN attained a peak accuracy of 0.98. Also explored within the project was the prediction of daily energy usage for a premise through multiple regression algorithms with Root Mean Squared Error (RMSE) scores of under 0.05 achieved in two out of the three models. Results of this nature facilitate energy efficiency at the consumer and supply level.

1 Introduction

Smart Technology has facilitated and enhanced day-to-day activities by extracting, transforming, integrating, and storing digitalised data. As the data collected is up to the present minute, our knowledge base now allows for increased and improved decision-making abilities in real-time. The development of data capture points within society, combined with high-speed networks, enables this rapid transportation and archiving of large data volumes, allowing the analysis of output from our actions on a more granular level than ever before. However, collection and storage are only the first steps to make the best use of the harvested data. Sophisticated data mining techniques are required to establish patterns, trends, and outliers, leading to a constructive analysis of past and present data diagnostics. Once the ascertaining of data history and standings is complete, it allows for future developments and the modelling of likely probabilities. Ultimately this means practical actions taken now can lead to tangible benefits later in time.

Smart Energy has provided the end-user with vast data concerning their historical consumption and likely future needs. Such information was previously only available when a monthly or bi-monthly bill would arrive. Even at that point, the bill would concentrate on aggregated consumption over a block period. As a result, there was no facility to isolate individual room and appliance usage at specific points of the day or week. However, there is now the potential to store and scrutinise this information to establish financially beneficial behaviours concerning energy usage.

With information and knowledge derived, the fundamental goal of the analysis is to optimise energy consumption within buildings without impacting the lifestyle of its inhabitants. For this reason, users must retain control over the system, and the implementation of any recommendations must be done manually as opposed to executed automatically. The efficiency of the present energy consumption level against the optimum value for the same set of circumstances is classified through the custom-developed recommender algorithm Colour Code My Energy (CCME). First, the CCME algorithm compares the optimum read (already on file for this scenario). Then, it assigns a colour coded ranking reflecting the status of the present read. The composition of a scenario is based on a myriad of inputs such as the building type, features combined with weather and time characteristics. Whilst there is a plethora of data available for the end-user to research, participation is more likely if the recommendations are straightforward and easy to access. The intention will be to activate one of seven colour-coded lights on a display panel. The colour of the light will symbolise how the building's present energy consumption level is, in comparison to what is on file as the ideal value in that specific scenario. Therefore the user is always aware of how their present behaviour compares to the optimum consumption level and has the opportunity to make instant adjustments if necessary.

Optimising the outputs will identify areas of potential energy waste and unnecessary financial loss due to excess usage and inefficient habits. In addition, continual and gradual adjusting of the recommendations over time to align present consumption closer to a baseline peak value is likely to encourage long term sustainable behaviour from the end consumer. The goal is not necessarily to reach that optimum target but rather to move in its direction. If the user is aware of poor consumption habits and can instantly adjust their behaviour, the establishment of savings and better traditions will inevitably follow.

1.1 Research Question, Project Objectives and Contributions

The research project is concerned with predicting and reaching optimum energy consumption levels for a diverse range of users based on their specific individual behaviours, characteristics and environments. While most of the research will concentrate on snapshots of how efficient the energy usage is within the present moment, there is also interest in predicting the cascading effect of how the daily usage will look based on one hours worth of consumption. Therefore the research question and sub research questions are as follows.

RQ; "How can energy usage recommendations be established according to weather patterns, building characteristics, and the previous consumption history of a premise to optimise both comfort levels and energy demand?"

Sub RQ 1; "Upon establishing recommendations, to what extent can predictions be made from deep learning techniques, so the end-user fully understands their consumption efficiency?"

Sub RQ 2; "How accurately can a day's worth of consumption be predicted for a premise based on an hourly read through machine learning regression techniques to enable the efficient management of supply?"

The project objectives are outlined in Table 1.

Table 1: Project Objectives

No.	Objective	Methods	Evaluation Metrics
1	Critically evaluate and analyse literature of forecasting, recommender systems and optimisation techniques within energy consumption.		
2	Identify, download and merge datasets of energy, weather usage and other relevant data.		
3	Create data visualisations and exploratory analysis. Enhance data by deriving supplementary values from originals fields.		Time Series Trends Correlation Plots Descriptive Statistics Heatmaps Histograms
4	Design and implement an external relational database model to serve as a structured backup.	PostgreSQL	Referential Integrity Read Data From Write Data To
5	Develop multiple regression machine learning models to predict the daily energy consumption of a house from an hourly energy read.	Random Forest Regressor Decision Tree Regressor KNN Regressor	MAPE MAE MSE RMSE
6	Develop the CCME recommender algorithm, which benchmarks a present read against a previously stored optimum value for the same circumstances.		
7	Create a neural network that can predict the recommendations of the algorithm.	DNN	Accuracy F1 Score Precision Recall ROC AUC
8	Analyse outputs and results.		
9	Compare and contrast results against existing external models.		

This project solves the research question and sub research questions by creating a diverse range of past and present descriptive and diagnostic visualisations. The establishment of recommendations derived from the CCME algorithm and accurate forecasts from the machine and deep learning models form the fundamental basis of the implemented solution. Enhanced and effective energy management through these predictions, recommendations, and optimisation reduces the financial cost to the end consumer, leads to more efficient management of energy for the supply grid and is more beneficial to the environment as a whole. A major contribution has been delivered that provides a unique approach of predicting and optimising energy through up to the minute recommendations, which keep the end-user in total control. A minor contribution has also been delivered, which provides a restructured relational data model of the dataset used in this study. These project contributions will supplement the existing body of knowledge in this area.

The remainder of this report is structured as Section 2 will provide a thorough critical review of literature already published in the area of smart energy predictions, recommendations and optimisations. Section 3 will provide a methodology design at a high level of how the project will be implemented. Section 4 will document the steps taken throughout the implementation as well as the challenges faced. Section 5 will provide a platform for discussion of results from the implementation, with comparisons made to other existing work in this area. Finally, 7 offers some suggestions for future improvements and concludes the project.

2 Related Work

The development of a solid methodology will begin following critical analysis of related work. This related work will concentrate on areas of research that have covered energy forecasting (both short and long term), energy-efficient recommender systems, the optimisation of energy consumption and the exploration of digital energy theft. These respective subgenres should provide the best learning insights, providing solid foundations for project development.

2.1 Short Term Energy Forecasting

Various studies have researched short-term forecasting with different approaches and various timelines used for the target prediction. Oprea and Bâra (2019) forecasts the consumption of electric energy for the following 24 hours while Khan et al. (2020) focuses on several different periods in the near future for a multitude of energy types. Short term forecasting can generally expect to attain a high level of accuracy as predictions are likely to mirror present trends. Khan et al. (2020) focuses on the forecasting of power through a fusion of 3 different machine learning approaches. As noted within the report, ”Combinations of prediction methods are receiving increasing attention”. Therefore ensemble learning is joining multiple different models to get a more robust learning outcome. This study presents a combination of CatBoost with both Support Vector Regressor(SVR) and Multilayer Perception(MLP). Each model is trained independently, with the results concatenated for a final forecasting figure. Oprea and Bâra (2019) incorporates a NoSQL database (Mongo DB) to merge energy consumption data from a residential building containing smart meters with weather patterns of the same timeframe. The study looks at implementing a feed-forward Artificial Neural Network(ANN), with the option of benchmarking against numerous other pre-existing algorithms. During the initial analysis of the data, the factors that influence consumption, such as weather characteristics, day of the week, and time of the year, are given a ranking number representing how significant they are in the amount of energy used within the building. Next, the application of

K means clustering partitions, customers into groups where the consumption levels are similar. Finally, the aggregation of data into 24-hour values means that the consumption for the building as a whole for the next 24 hours can be predicted.

It is interesting to note that whilst Oprea and Bâra (2019) analyses the building as a whole, Shapi et al. (2021) splits the study between 2 separate commercial type customers and also drills down into the data as opposed to aggregating it. Predicting energy consumption within a smart building begins with analysing the data collected from Internet of Things(IoT) meter sensors attached to electrical sockets. The data can then be stored and examined on a granular minute by minute level. The combination of K Nearest Neighbour(KNN), Support Vector Machine(SVM), and an ANN within the methodology allows the prediction of maximum demand based on electricity statistics, measuring the spread of the distribution, and making forecasts based on predictions on patterns mined within the data. The dataset is for a minimal amount of time and only encompasses from June 2018 until December 2018 (for two different tenants). Again this is in contrast to Oprea and Bâra (2019) whose study has a much more comprehensive dataset containing over 6 million rows of data for 114 New England based apartments with consumption readings also recorded on a minute per-minute basis.

Even allowing for a drill down to a specific 60-second timeframe, knowledge of seasonality trends and if the specific consumption is in keeping or out of sync with long term trends is not available to us in Shapi et al. (2021). Nevertheless, not all studies factor in characteristics (i.e. weather and building) and instead focus on a data-driven approach only as it argues that the embodiment of these features are already encapsulated within the smart meter data. One such research is Eneyew et al. (2020), which seek to forecast the next hour's consumption value through a DNN that comprises convolutional features and a wavelet transform. One of the methodology segments involves deploying a one-dimensional convolution model with max-pooling operations to isolate and identify features from the dataset provided. After fitting the model, the output layer will generate predictions of the consumption. Six years' worth of hourly energy reads is used within the study across ten different houses. Whilst some differences exist between the houses, there does not seem to be any extreme outliers. The study uses the Mean Absolute Percentage Error(MAPE) as a barometer of success through 3 interconnected evolutionary experiments. A 1-dimensional dense model established preliminary results. The application of stationary wavelet transformation within a Long Short-Term Memory(LSTM) convolutional model subsequently improved these results. Goyal et al. (2020) specialises their study in the impact systems such as heating, ventilation, and air conditioning (HVAC) have on the overall energy usage of a building and how the obligation to ensure there is adequate supply can place extreme pressure on the supply grid. For this reason, high accuracy forecasting is vital for the management of resources on an efficient level.

Using nonlinear regression and fuzzy c-means clustering, Chen et al. (2020) also seeks to predict the subsequent day's energy consumption whilst factoring in the trends which exist within the time-based characteristics. The study looks at how features diversify and which factors influence this. Whilst the hypothesis that a consistent energy pattern can play a crucial role in increasing the accuracy levels of predictions is stated clearly from the outset; it is unclear whether any new learnings are gained from this other than the applicable techniques themselves. As expected, incorporating weather statistics improves the extrapolations. Prediction windows are predetermined and cover what is described as "different typical seasons respectively". Revisiting the use of data-driven techniques seen in Eneyew et al. (2020), Yiyi et al. (2020) looks at predicting the daily usage of electricity in a residential dwelling for 17 months from 2013 to 2015. The homes involved have indoor sensors to record environmental settings to predict daily consumption through a 1-Dimensional Convoluted Neural Network(CNN) and an LSTM recurrent solution. Outliers were removed from the dataset with a Principal Component Analysis(PCA), reducing the number of independent variables used in the study to 16. By splitting the parameters between external weather variables and those belonging to the internal building, the study focuses on areas of particular interest. For example, the experiment selects only one house, and within that one house, certain rooms are excluded due to low occupancy throughout the days. Root Mean Square Error (RMSE) scores are impressive, but the study hones in on very niche scenarios such as "only bedrooms" and "data without bathroom inputs".

2.2 Energy Efficient Consumption Recommender Systems

Recommendations to adjust and reduce energy consumption will have different philosophical views regarding residential and commercial buildings. Whilst the comfort level is of paramount importance to the occupant of a residential building, the same is not necessarily true of a commercial premise. The goal of the recEnergy recommender system in research by Wei et al. (2020) is to sufficiently optimise energy consumption but retain human decision making within the process through deep level reinforcement learning. While the focus of the design and implementation is to reduce energy consumption in commercial buildings, actions that can achieve savings of high energy if implemented are recommended to end-users who then have the option to deploy or reject. Implementation of changes will not occur automatically. Instead, the hope is to mould human behaviour in the direction of more frugal habits. Slightly contrasting is the view that recommendations are likely to have a higher implementation success rate if they result from knowledge attained through learning and analysis of individual user behaviour defined within historical data. Mishra et al. (2020) develops a pattern mining algorithm that is specific to individual user behaviours intending to offer alternative recommendations. The establishment of these actions is from patterns found in newly created events against a suitable alternative. Statistical binning is incorporated to define single activities as granular level incidents combined into one "transaction" containing a sequence of many activities. Results of this type of experiment are fuzzy by nature, but 3 of the top 4 questions achieved a +80% acceptance when surveying 120 sample users over 15 days. The wording of some questions tends to lean towards a specific response, with terms such as "relevant", "fair", and "similar" all somewhat ambiguous by their nature.

Furthermore, a substantial breakdown of actions and adjusted recommendations does not accompany this paper. However, there is no doubting that both Wei et al. (2020) and Mishra et al. (2020) openly challenge the perception that the achieving of proper optimisation within energy is possible without altering the behaviour of its occupants as they are primarily responsible for a large portion of same. In turn, altered behaviour might not translate into savings if consumption is simply directed elsewhere or are not a realistic option.

Both Schweizer et al. (2015) and Alsalemi et al. (2021) focus their study over a short period of 4 weeks and achieved similar results with savings of approximately 7%

in energy usage. Schweizer et al. (2015) explores how a buildings energy usage history and individual needs can be analysed. The study does not include training and testing of data as part of this research. Instead, predictions are relayed personally to users, who rate how practical and informative those recommendations are. In order to accurately provide recommendations of the next best action to users, knowledge of historical power consumption is not enough by itself. Obtaining data is also vital so that explorations and discoveries can be made in terms of the order of events. Schweizer et al. (2015) designs and implements the Window Sliding with De-Duplication (WSDD) algorithm for the efficient mining of patterns within the user home, combined with equivalent energy consumption. Main comparisons with algorithms such as BIDE+ and PrefixSpan are limited to technical aspects such as run times and internal memory consumption. Benchmarking to other recommender systems is not done within this study. Instead, the report focuses on a recalibration of the suggestions but yielded no improvement on the rules, resulting in recommendations (23 accepted in phase 1, 17 accepted in phase 2). Alsalemi et al. (2021) encompasses a ”Micro-Moment” type event, which is used to track individual activities at the appliance level. Once established and grouped, multiple micro-moments can predict usage patterns and hence attempt to forecast individual energy requirements. Sensors record not only energy consumption but also particular weather characteristics. Whilst predicting energy consumption is the primary area of research, a follow-up analysis explores how the information gathered can be of tangible use in changing attitudes and reducing overall consumption. A rule-based recommender system marries information gathered about user data and the physical environmental snapshot at that moment in time. Subsequently, it then factors in details such as whether or not a room is occupied. Though deep learning achieves an accuracy of 99.95%, the timeframe of 4 weeks and the number of users does not provide a deep dive into the model long term. From an ethical point of view, There is acknowledgement within the current limitations of the project that the connection and transfer of data to the server are not secure. The lack of security puts the data in jeopardy and potentially highlights patterns in which the home is unoccupied to unauthorised individuals.

The non-linearity of relationships and how they factor in terms of statistical applications are addressed by Wang et al. (2020). The study explores those which specifically attempt to predict energy consumption. The information identified can allow users to save on their electricity consumption. From a dataset of 14 residential houses in Hong Kong over one year (2018-2019), the study seeks to marry the users' current and historical consumption profiles and rank their potential savings against others who are also taking part in the survey. As with Schweizer et al. (2015) and Alsalemi et al. (2021), a final average saving of a little over 7% is achieved, but extreme outliers exist at both ends of the scale, meaning whilst some houses see huge savings, others achieve minimal (if any) reductions in their consumption habits. The figure is not very surprising in many ways as over 40% of the households surveyed did not have children. Concurrently half started the study immediately, whilst half delayed their commencement in the scheme. Once pre-processed and standardised, the data has an SVR model applied. Results are positive, with the algorithm outperforming random forest and different types of neural networks. The family size of the household and whether or not children were present seemed to be the deciding factors on how the model performed overall.

2.3 Optimisation of Energy Consumption

Optimally adjusting consumer behaviour when factoring in extra fuzzy parameters, such as the external environment, is a significant challenge in making recommendations to lower consumption levels. Kontogiannis et al. (2021) explores the mechanisms in which this can be derived and presented to the end consumer. The study expands on previous research to generate rules through decision tree linearisation. It is interesting to note how the study picks up on a subtle issue not explored in many other pieces of literature. Implementing a model based solely on mathematics and logic excludes the specialised knowledge that only humans can attach to the input data and the interpretation of outputs.

For this reason, there is a requirement for a degree of fuzziness, which can handle the inconsistencies and flexibility of real-world scenarios. Ambiguity also arises when individuals describe an input in linguistic terms instead of statistical figures (e.g. cold/hot as opposed to the specific temperature)—establishing a rule base to allow for "fuzzification" of inputs and acts as a translator within the decision-making unit. The principals within this are a series of IF-THEN rules that establish a feature importance grade from the specific input. Each attribute will have linguistic terms attached to the variable as opposed to numeric values. The approach by Kontogiannis et al. (2021) is in contrast to Nazeriye et al. (2020) who seek to optimise energy usage by identifying specific appliances which are driving up consumption and can be adjusted. The paper demonstrates trends and patterns before applying a decisions trees model; the k means algorithm, and the differences after application. In conjunction with altering user behaviour, critical strategic decisions (of which can be argued would have a more significant impact) such as effectively insulating the property, regular service of boilers and the installation of cavity walls are also carried out within each of the homes. As the study is across many houses in England and Wales, the consumption data is normalised relative to the floor size of the property. Houses with patterns comparable to others are clustered together, with each group then being analysed in isolation. Once clustering is applied, the decision tree algorithm will assess the results of the different models.

While the most common techniques regarding the optimisation of energy primarily encapsulate appliance usage and central heating features, Raval et al. (2021) extends the level of analysis by exploring how the efficiency of power sockets and sensors themselves can be improved. Part of the hypothesis is that sensing features that traffic the data usurp and monopolise a vast amount of energy. The paper develops a suite of energy management features for the various IoT devices present within the building using a genetic algorithm to select the most efficient parameters, with MATLAB Simulink and OpenModelica used to simulate the results of the various scenarios. Through incorporating a feedback mechanism and reinforcement learning, phase 3 of the tests demonstrate impressive results in improving optimisation time, operating efficiency and consumption used. Pawar and Vittal (2017) also uses a specialised management system to monitor and optimise the power of a building. Once again, hardware is the aim of the study, with the smart technology concept expanded to encompass power sockets to estimate how similar/different the usage of specific appliances are. ZigBee acts as a digital mediator to determine whether sufficient power is available (in real-time) to approve a request when a user physically activates a socket. The study briefly mentions a "power negotiation algorithm", but the details of what steps are taken if the request is denied is absent from the methodology brief.

2.4 Long Term Energy Forecasting

The development of the Smart Energy infrastructure has allowed better recording and tracking of energy usage than ever before. Concurrently, there is now the potential for future simulations with greater accuracy due to the highly granular level of detail we now have. However, like all forecasting, accuracy decreases as the timespan into the future increases that the predictions are made. Nabavi et al. (2020) identify the need to predict residential and commercial energy demands for Iran by 2040. One of the drawbacks of long-term forecasting is the inability to accurately factor in a significant change within the social behaviour of society, such as the work from home shift during the COVID 19 pandemic and its subsequent instant impact on the energy supply grid. While the study encompasses input points such as population figures, gas/electricity price, and Gross Domestic Production(GDP), there is challenging to reproduce past circumstances into the future. The training/testing split is unusual, with 47 years between 1967 to 2014 used to train the model, but only the most recent years of 2015, 2016 and 2017 used to test the predictions. Nystrup et al. (2021) supplements the notion that present methods will not suffice in the area of long term forecasting and that an unsupervised learning approach using wavelet transform is used to develop eight unique and diverse load profiles. These profiles can then be analysed for the robustness and elasticity needed for long term forecasting through min-max scaling.

2.5 Digital Energy Theft

Throughout recent times, energy grids have developed their techniques to improve all aspects from production to power delivery. While there has been an enhancement in reliability and security, there are still issues to address in terms of cyber energy theft. Digitalisation has allowed the development of more sophisticated techniques to address inconsistent energy patterns. Syed et al. (2020) focuses on developing a DNN/LSTM which will identify anomalies within the electricity network. While the paper approaches this purely from the point of view of tracking theft of power through suspicious usage patterns, erroneous readings can also be something that can identify more benign issues such as a damaged meter sensor or software fault. A Chinese smart grid company provides the dataset used for this study. 8.5% of the customer base has been identified as committing energy theft from January 2014 to September 2016. The missing values make up 25% of the actual data. The results can identify significantly different trends and patterns between thieving and non-thieving customers with an accuracy level of 92%. The model itself is simple in its application as only consumption history is analysed through the data-driven approach. External datasets which are present in other studies, such as weather characteristics, are not factored in. While the study yields impressive results, it does not elaborate on how a customer who drastically reduces energy consumption for genuine reasons, such as when the house is vacant, is dealt with. Jindal et al. (2020) highlights the struggle of developing nations and their attempt to tackle this issue; much like Syed et al. (2020), this study also focuses on using data-driven techniques to detect theft of energy. Although the study mainly analyses central grid infrastructures, the application of such methods can easily be transferred to individual premises. One of the scenarios in the study takes a dataset with no record of energy theft and simulates abnormalities into the data to reduce consumption on critical appliances. As similar consumers can be grouped onto one profile based on similar historical patterns of usage, a sudden change of one portion of the data would indicate possible theft.

2.6 Gaps for Development and Conclusion

Present gaps from the literature reviewed would indicate scope exists to expand the types of recommendations provided to users. Keeping the user in total control and modelling recommendations based on the behaviour specific to that individual building is a cornerstone to be retained within this project; there needs to be a mechanism where the user gets constant information about the next best action. This message needs to be relayed clearly and unambiguously. One criticism of present systems is that most studies focus their communications on ad hoc transmissions or develop a structure where the user has to search for the information. Many of the projects reviewed seem to choose between combining external characteristics to the data or concentrating solely on a data-driven approach. It would be interesting to analyse, establish and verify if a hybrid of the two approaches is possible and successful. Short term predictions of the next day's energy are also somewhat flawed, considering there is the potential for the next day to have a different set of conditions to the present day. This is demonstrated twice every seven day period where today is a weekday and tomorrow is a weekend or vice versa. Therefore, it makes more sense to develop predictions based on the present day. Objective 1 has now been fully achieved.

3 Smart Energy Usage Methodology Approach and Design Specifications

3.1 Overview

The designing of a structured methodology enabled the production of tangible results and analysis upon the project's conclusion. A methodology was developed (See Figure 1) based loosely on both the CRoss Industry Standard Process for Data Mining (CRISP-DM) and Knowledge Discovery in Databases (KDD). The identification and collection of data led to the combining of multiple sources into a solitary primary dataset. This then allowed for the development of initial explorations and visualisations. Following this, machine learning models were created to attain high-quality predictions with the CCME recommender algorithm also established.

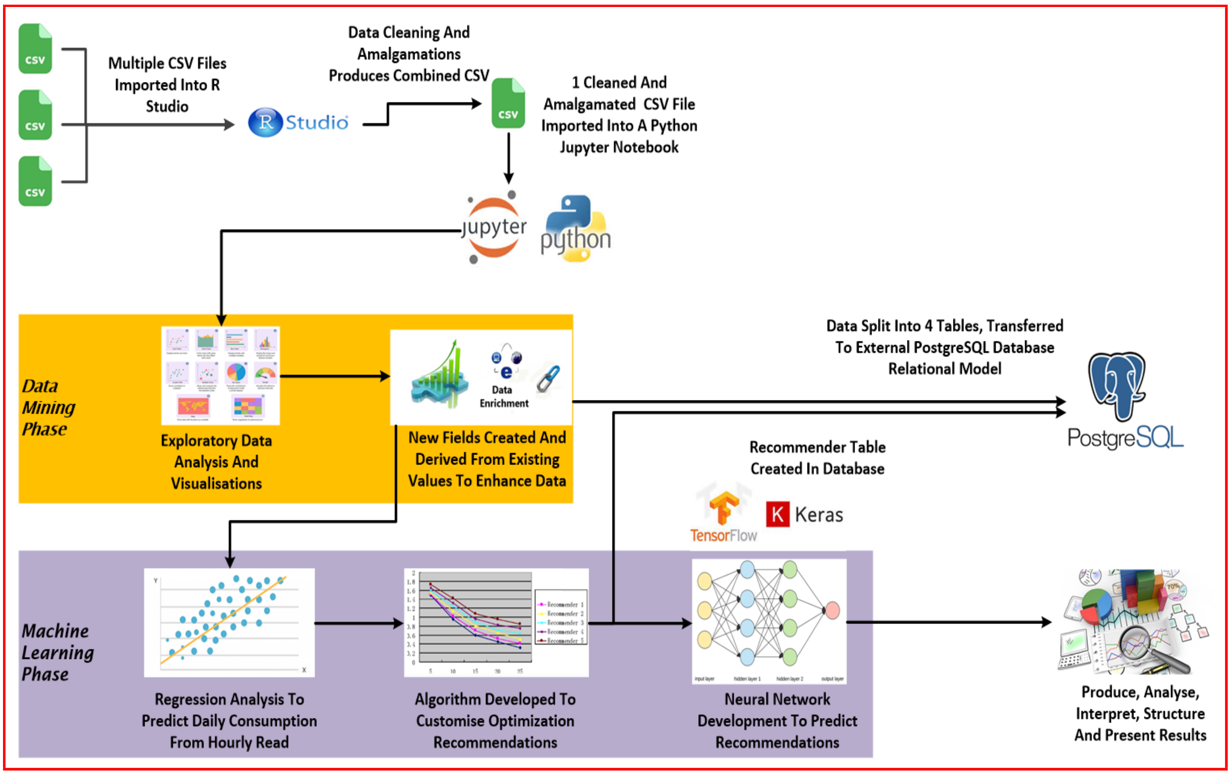

CSV
CSV
CSV
Multiple CSV Files Imported Into R Studio
R Studio
Data Cleaning And Amalgamations Produces Combined CSV
CSV
1 Cleaned And Amalgamated CSV File Imported Into A Python Jupyter Notebook
jupyter
python
Data Mining Phase
Exploratory Data Analysis And Visualisations
Data Enrichment
New Fields Created And Derived From Existing Values To Enhance Data
Data Split Into 4 Tables, Transferred To External PostgreSQL Database Relational Model
PostgreSQL
Recommender Table Created In Database
TensorFlow
Keras
Machine Learning Phase
Regression Analysis To Predict Daily Consumption From Hourly Read
Algorithm Developed To Customise Optimization Recommendations
Neural Network Development To Predict Recommendations
Produce, Analyse, Interpret, Structure And Present Results

Figure 1: Smart Energy Usage Methodoloy

3.2 Smart Energy Usage Data Pre-Procesing

Whilst section 2 of this project explored some pieces of research that concentrated on a small number of properties over a limited period, these studies can be susceptible to one-off events that can distort the regular patterns and trends. For this reason, there was the need for data to encompass multiple years and be combined from various sources to give the necessary inputs to attain solid learnings and conclusions.

The Hourly Usage Energy (HUE) dataset Makonin (2018) not only matched the requirements of the project but has also been proven beneficial in several other research assignments investigating and seeking to prove an array of null hypotheses Makonin (2019). Combining energy and weather statistics for seven years within the Canadian province of British Columbia, the data is of the most recent variety considering the oldest reads on file are no earlier than 2013. The individual household's privacy is fully protected with the removal of all personally identifiable data, but the dataset retains useful abstract information regarding the house type, facing direction, and HVAC appliances. In addition, individual households each retrieved their personal historical usage from their encrypted online portal, allowing for the donation of their energy usage. Overall the HUE dataset contains hourly readings for 28 houses and separate files for hourly weather statistics, bank holiday dates and details of the houses involved. All files (apart from the house characteristics) are in the format of Comma Separated Value(CSV) files.

Section 4.1 expands on the mechanisms and structure for combining the data frames and dealing with missing/bad data.

3.3 Smart Energy Usage Data Mining

Now that the data is combined, exploration could begin to discover trends and patterns. Establishing which visuals give the best insights into the learning of information is vital at this juncture. Histograms show the distribution of energy, temperature, humidity and pressure. Horizontal bar charts demonstrate a timeline of what period each house has documented data. Timelines of energy consumption individual to each house and each type of house compare and contrast outlines over time. Scatter plots demonstrate how each of the weather characteristics specifically influences energy usage. 3D and 4D plots allow for the impact of more than one independent variable. Other miscellaneous charts such as treemaps, heatmaps and pie charts display energy usage over various categorical values. Concerning each house, comparisons were made relating to their total, average and range of energy. This analysis established the first indication of which houses are operating efficiently and those that need improvement.

Date and timestamp fields often prove cumbersome values for the learning phase of machine models. Therefore, separating these into additional categories such as weekday/weekend, time of the day, and time of the year improves learning efficiency. Creating these additional categories also adds quality and depth to the dataset. The grouping of these extra data into categories yields more valuable information for visual demonstrations and machine learning. A field to record the daily usage is also introduced for each house. This field serves as the dependant variable for the machine learning regression models in the next phase.

3.4 Smart Energy Usage Prediction Models

The machine learning phase of the project was involved in establishing the findings concerning the research question and sub research questions. This included developing multiple regression models to see how accurately a house's daily energy can be predicted from knowing an hours' worth of consumption. This phase also establishes the CCME recommender algorithm, categorising the value of the present hour's energy read based on previous knowledge about the user. Finally, the creation of a deep neural network seeks to predict these recommendations accurately.

Regression analysis is a machine learning technique that involves predicting a continuous numeric value from multiple independent variables that will be either numerical or categorical. For this section of the project, the target is the prediction of the daily energy of a specific house from one hour's worth of consumption. The project will use the regression algorithms displayed in Figure 2 to ascertain which model gives the best result. **Decision Trees Regressor** constructs the model based on a tree structure in which the dataset is reduced continually into smaller subsections. **Random Forrest Regressor**constructs multiple individual decision trees on many subsets of the data. The algorithm uses averaging to improve the accuracy and reduce the risk of overfitting. **K Nearest Neighbour Regressor** works on the assumption that data points which exist in close vicinity are similar.

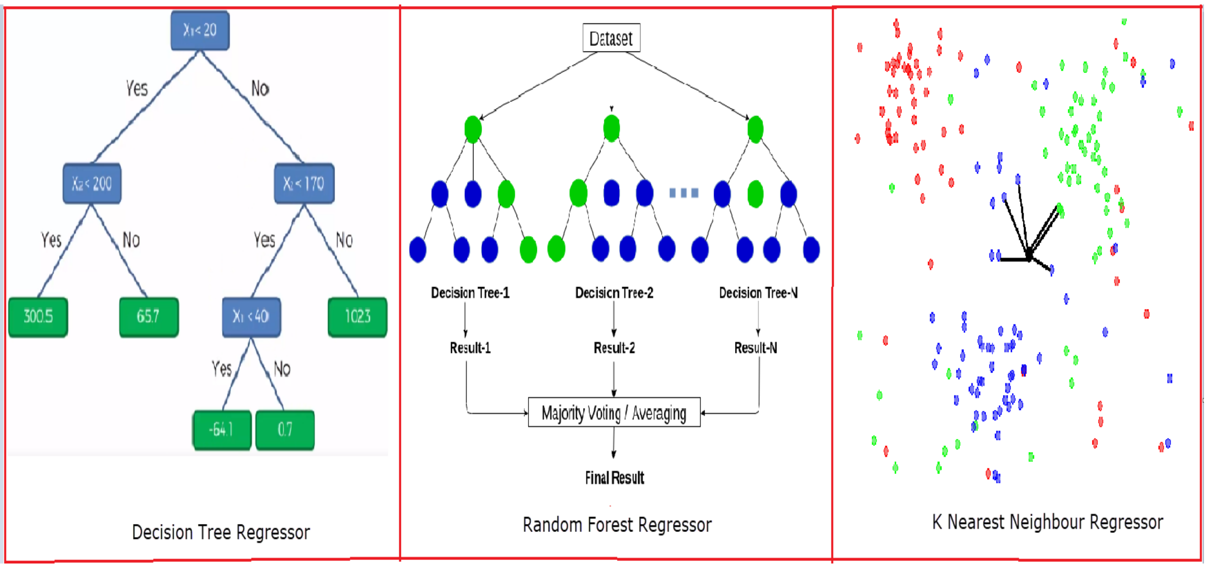


Figure 2: Regression Models

The final step of the methodology is developing and implementing a DNN so that predictions can be made of what the recommendations should be from the algorithm. As seen in Figure 3, the model has many hidden layers between inputs and outputs. Unlike the regression models to predict daily energy consumption, the DNN is categorical and will use the CCME Recommender algorithm colours as predictions for what it believes the energy reads should be.

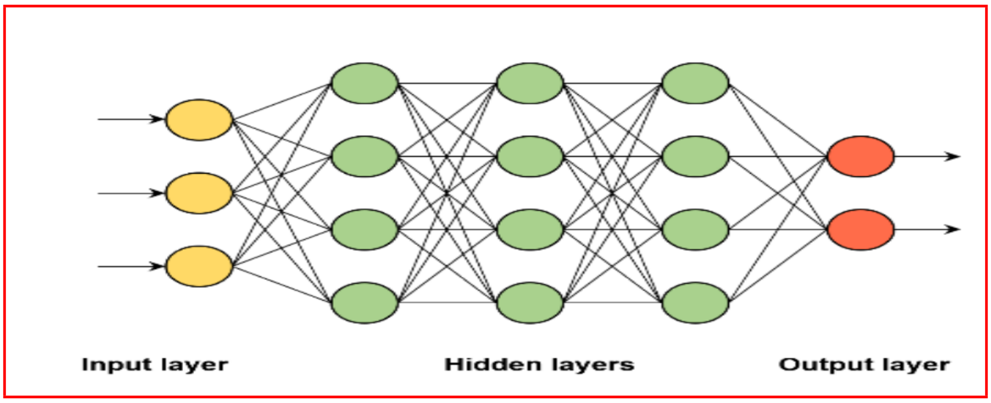


Figure 3: Deep Neural Network

A DNN is configured at this point as opposed to a regular machine learning classification model due to the ability of the DNN in its effectiveness to make decisions and not simply learn from the data.

4 Implementation of Prediction and Optimization of Smart Energy Usage

This implementation phase puts into practice the steps discussed and outlined in the section 3. Consolidation of the data was done in the pre-processing element, which allowed informative visualisations to represent the combined dataset. The transfer of the data to a PostgreSQL database provides a new viewpoint of the datasets whilst retaining the amalgamations of the pre-processing phase. Predictions of daily energy were established through multiple regression models. The CCME algorithm and DNN allow for the creation of recommendations for the end-user whilst also creating a roadmap to optimisation of energy and reductions regarding inefficiencies.

4.1 Data Pre-Procesing

After downloading the multitude of CSV files externally for the (HUE) dataset Makonin (2018), there was the need to combine these into a single shared data source for structured analysis to occur. R Studio is a dedicated Integration Development Environment (IDE) that allows direct data manipulation through the R language. In the pre-processing phase, R Studio was responsible for importing each user history consumption file. Each house file (House Number 1 – Number 28) has three columns detailing the date, hour and energy (kWh). In addition, an identifier was attached to each reading to ensure it was possible to trace it back to a specific house once merged. The conclusion of this phase resulted in an output of 1 solitary CSV file. Objective 2 has now been fully achieved.

House number 7 and number 15 were removed from the dataset and will not be considered part of the analysis. This is because House number 7 lacks details about its characteristics, while house 15 is from a different region to the other houses, so that it may be susceptible to different energy demands. This leaves 26 houses for investigation within the project.

The text file containing details about each house, such as facing direction, energy appliances, house type, and other information, was manually inputted into a separate data frame within R Studio. This data frame only contained 28 rows and needed specific customisations regarding details of the energy systems. In addition, the weather statistics were imported into a separate data frame with no alternations needed.

The quality of the data is very good overall. Details of how the limited number of missing values from the CSV files that were addressed are as follows

- **Missing Energy Values -** Any missing energy values were populated by grouping the combined dataset by house, then by date ascending. Once grouped, the missing value was filled by taking the next hourly read available. The belief was that energy consumption does not alter drastically from hour to hour.

- **Missing House Chars -** House 7 was the only house that didn't have descriptive details, and as such, the house and its read history were removed from the project.

- **Missing Weather Reading Values -** The weather file was missing 40 individual hours worth of data between 01-Jan-2012 and 13-May-2020. As this was minuscule, 40 shell rows were added to the file with NA values. Next, the NAs were updated similarly to the missing energy values (Taking the following data available after sorting the data frame from oldest to newest).

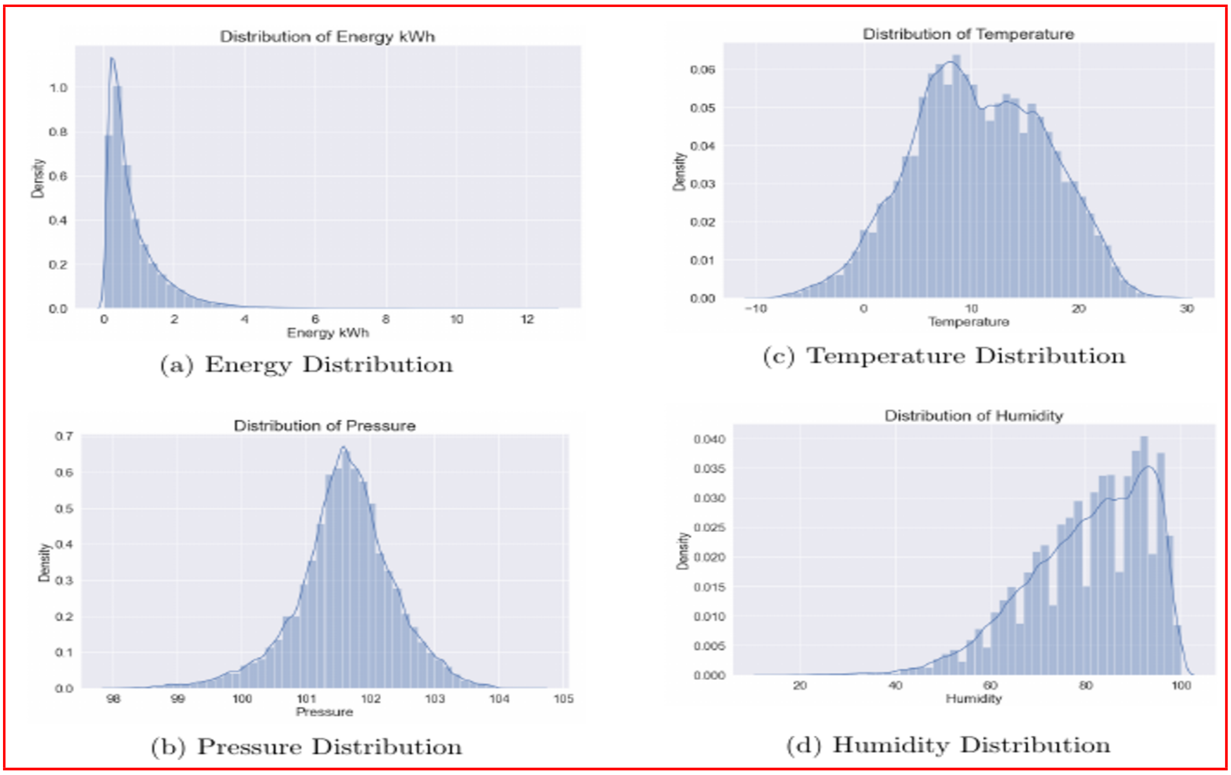
- **Missing Weather Descriptions -** The weather file is missing almost 45% of weather descriptions. Whilst not removed, the field was not used in any form of analysis or learning as the volume of missing data, and the loose category of the labels present within the field (i.e. Clear, Cloudy, Mainly Clear, Mostly Cloudy) was unlikely to make for accurate interpretations.

As a final step of the pre-processing phase, the three data frames are joined on common keys and outputted as one single CSV file to begin the next stage of the implementation.

4.2 Data Visualisation

Jupyter notebooks are used for the interrogation and visualisation of the data. This implementation stage involved putting structure to transform raw data into meaningful information to draw constructive conclusions. The open-source python language offers various libraries such as numpy, pandas, seaborn and matplotlib. After importing the amalgamated CSV file, updates are made to the data types to ensure numeric and categorical values are appropriately assigned.

Energy and the three weather categories (Pressure, Humidity and Temperature) are the primary continuous figures within the dataset, and a crucial first piece of exploratory analysis is to investigate the distribution trends of each. As can be seen in Figure 4 the distribution of energy is positively skewed; humidity is negatively skewed, with both temperature and pressure generally following a normal distribution trend.



(a) Energy Distribution

(c) Temperature Distribution

(b) Pressure Distribution

(d) Humidity Distribution

Figure 4: Distributions of Weather and Energy

This project seeks to derive learnings from energy concerning its distribution, prediction and optimisation. Hence, the next set of data visualisations investigate how the weather and HVAC appliances impact energy usage. From Figure 5, we see a negative correlation of energy concerning each of the three weather attributes. As any one of the weather values decreases, energy is predicted to rise. A similar pattern exists concerning the HVAC appliances, which would be expected.

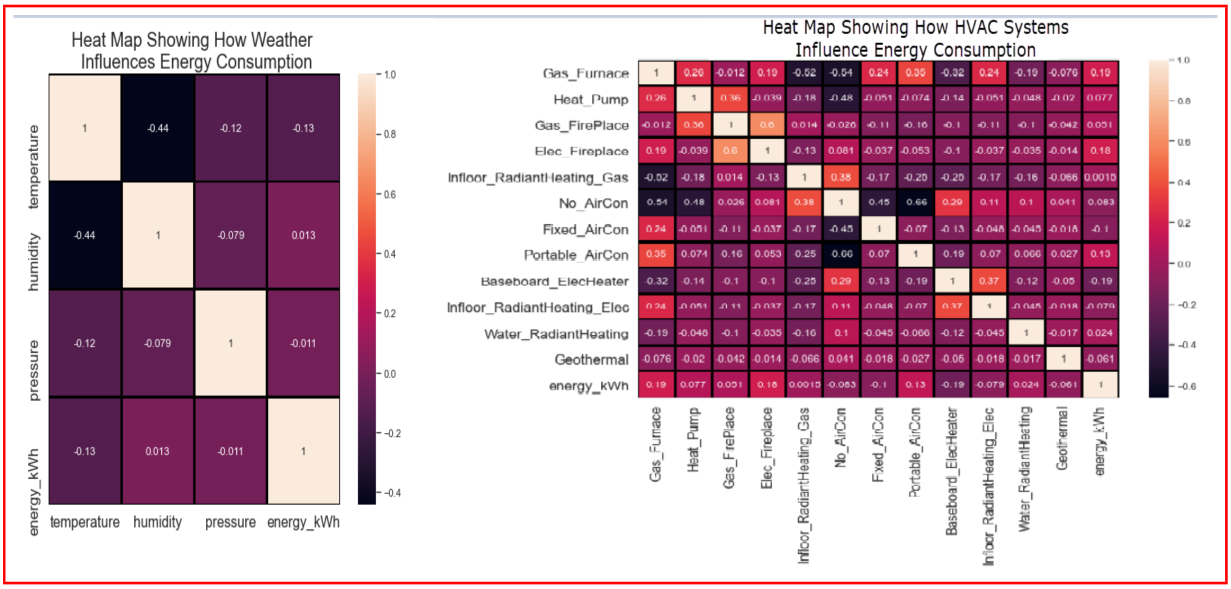


Figure 5: How Weather Attributes and HVAC Systems Influence Energy Consumption

Figure 6 explores energy trends and the effect one of the weather characteristics (temperature) has on a group of houses that are all labelled as of type ”character”. Whilst trends are primarily similar; the temperature does not negatively impact house eight as much compared to the other 3. Maximum energy does not go above 35 kWh vs 50-60kWh in the other houses. The regression line is also not as steep compared to houses 4 and 20, indicating less energy is consumed either by efficient behaviour or infrastructure within the home. Objective 3 has now been fully achieved.

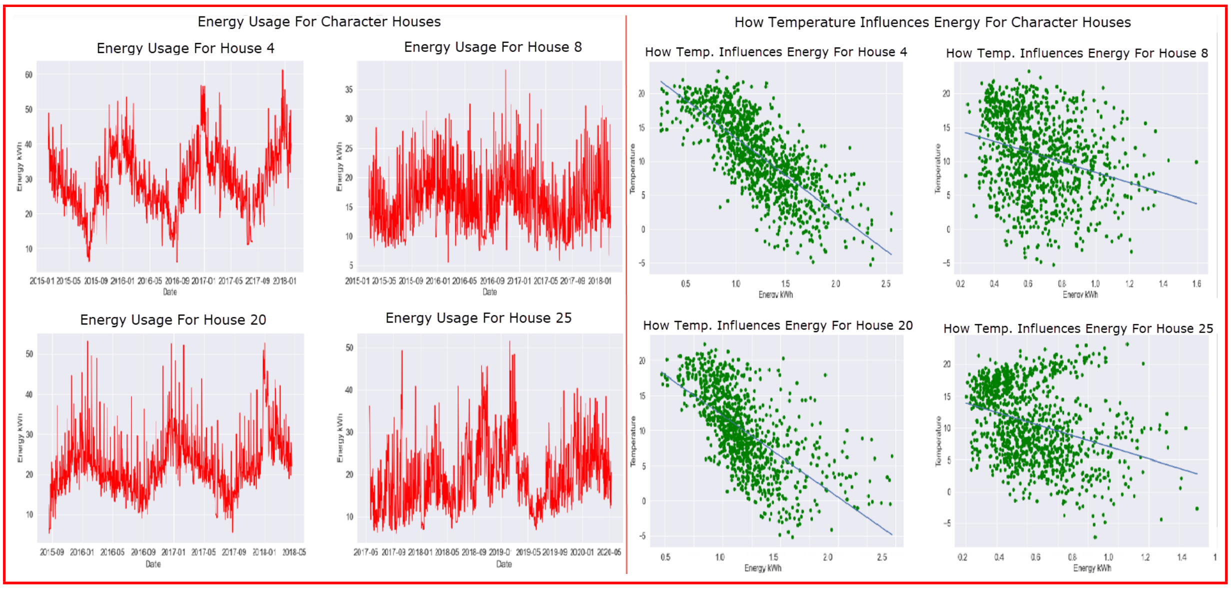


Figure 6: Energy Usage and Temperature Influence for Character Houses

4.3 Data Transfer

The data is now ready to pass to the machine learning element of the project. With no more fields likely to be created, this is the logical time to back up the data and make it available to external sources by transferring it to a PostgreSQL relational database. For the transfer, the data is divided into several separate entities. The four tables established are not dissimilar from the final data frames in R Studio before the merging. Although the data is segmented once more, the crucial difference is that the relational model retains the link to the amalgamated data frame through its primary and foreign keys to maintain referential integrity. This is displayed inFigure 7 below. Further interrogation of the data in Structured Query Language (SQL) queries will now be available to any number of authorised individuals. Psycopg was installed within the python environment to implement the database Application Programming Interface (API). Once installed, the ability to connect to PostgreSQL and store and retrieve data was possible. Objective 4 has now been fully achieved.

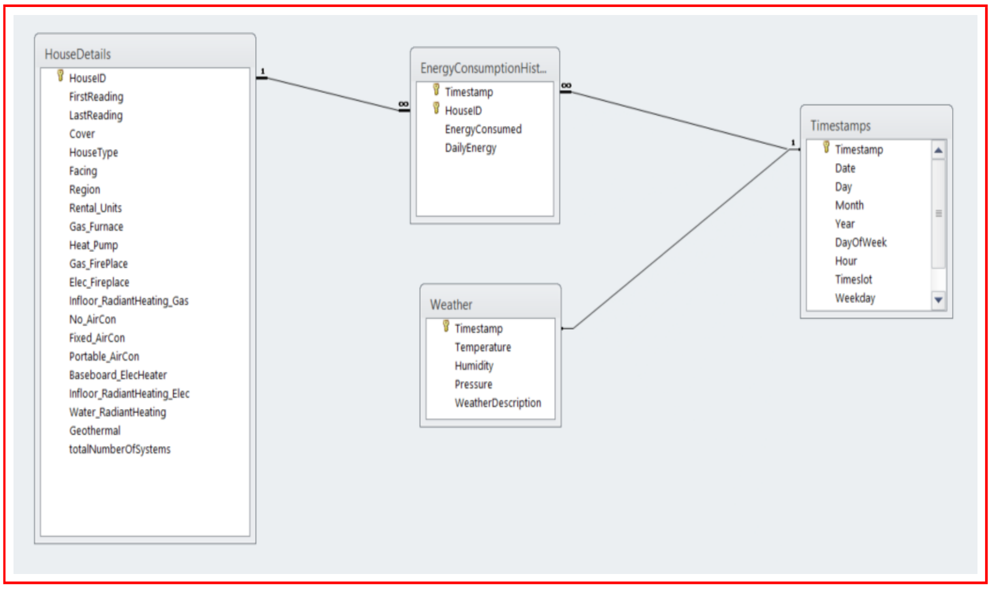


Figure 7: Relational Back Up

4.4 Implementation of Colour Code My Energy (CCME) Recommender Algorithm

The primary aim of the CCME Recommender algorithm will be to judge the present consumption read vs the optimum consumption read, which falls under the same scenario. The generation of a scenario accompanies the creation of every new energy read. A scenario comprises the house identifier, the weather characteristics, timeslot of the day and time of the year upon which the read is created. The algorithm then checks the present scenarios on file to see if this scenario already has an optimum value. Figure 8 outlines the workflow of the CCME Recommender algorithm.

If the algorithm does not find a match for this scenario, then no optimum value exists. Therefore, the present energy read will serve as the baseline recommendation for this scenario and is added to the recommendation table. This read will be categorised as a ”White”, indicating no previous ideal value existed before this read. However, if a match is found, the recommender algorithm compares the stored optimum value to the present read to classify the read. If the present read is lower than the read on the recommender table, the (now formally) optimum read will be deleted and replaced with the present read as the new optimum, leading to the categorisation of the as ”Green”.

If the present read has a higher energy consumption than the optimum read, the algorithm will classify how far off the optimum read is from the present read and categorise it accordingly as one of ”Blue”, ”Yellow”, ”Orange”, ”Red” or ”Black” depending on the distance between the two energy values.

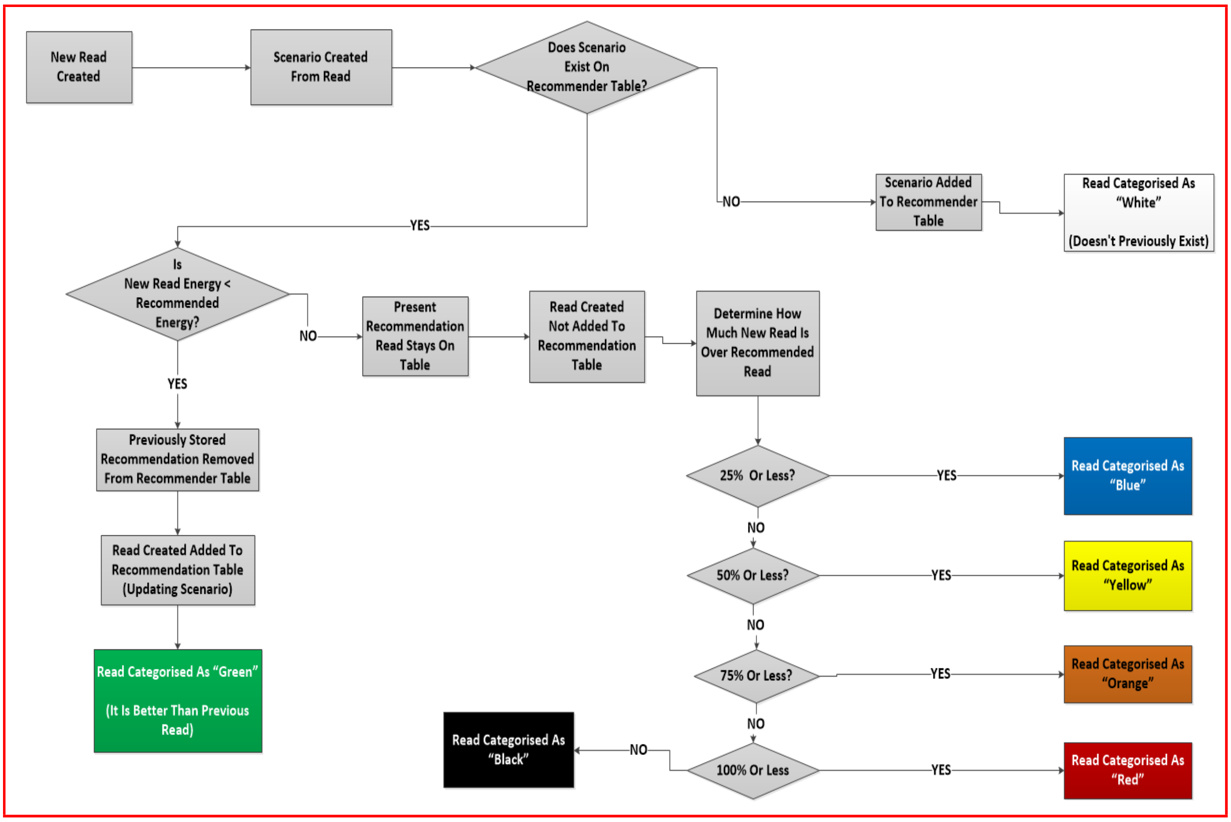

New Read Created
Scenario Created From Read
Does Scenario Exist On Recommender Table?
NO
Scenario Added To Recommender Table
Read Categorised As "White" (Doesn't Previously Exist)
YES
Is New Read Energy < Recommended Energy?
NO
Present Recommendation Read Stays On Table
Read Created Not Added To Recommendation Table
Determine How Much New Read Is Over Recommended Read
YES
Previously Stored Recommendation Removed From Recommender Table
Read Created Added To Recommendation Table (Updating Scenario)
Read Categorised As "Green" (It Is Better Than Previous Read)
25% Or Less?
YES
Read Categorised As "Blue"
NO
50% Or Less?
YES
Read Categorised As "Yellow"
NO
75% Or Less?
YES
Read Categorised As "Orange"
NO
100% Or Less
YES
Read Categorised As "Red"
NO
Read Categorised As "Black"

Figure 8: CCME Recommender Algorithm

The algorithm is developed in python over two dedicated files, which retrospectively fit recommendations based on the historical dataset. Objective 6 has now been fully achieved.

4.5 Implementation of Prediction Models

The machine and deep learning models followed a specific route, which starts with defining the dependent variables influencers and ends with producing results that may or may not be deemed satisfactory. Figure 9 outlines the detailed steps involved in this process. Both the machine and deep learnings models are attempting to predict a variable. In the case of the machine learning regression models, the goal is to forecast a numeric value as close as possible to a continuous figure. The deep learning model seeks to attach a label (from a predefined list) to a presented scenario.

As part of this, the respective datasets are split into one dependent variable (what the model is trying to predict) and several independent variables (the features used to predict this end value). For the regression models, the data is scaled (using the standard scaler) to values between 0 and 1, allowing the measurement of how far the regression line is from the data points.

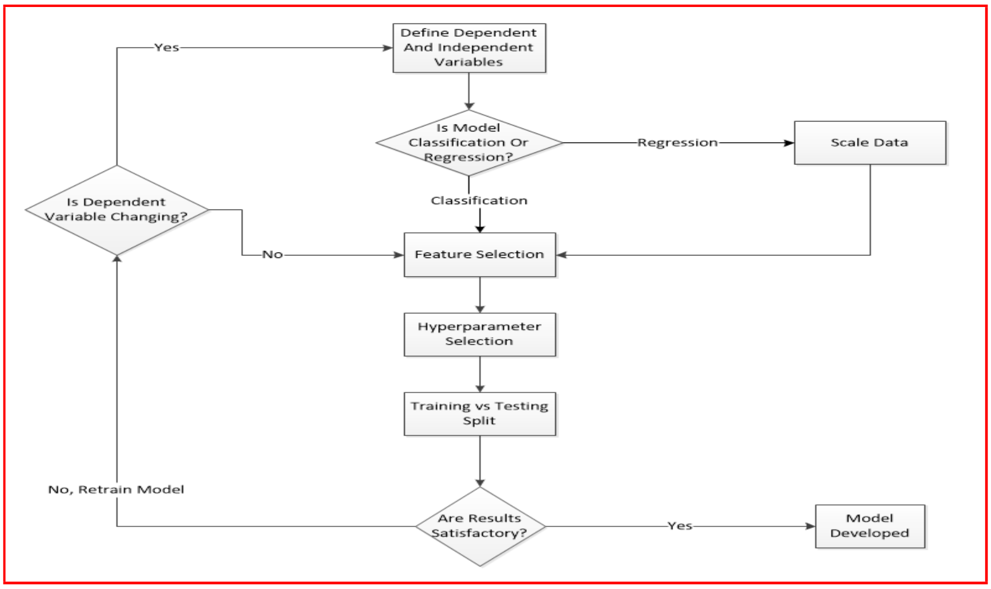

Define Dependent And Independent Variables
Yes
Is Model Classification Or Regression?
Regression
Scale Data
Is Dependent Variable Changing?
Classification
No
Feature Selection
Hyperparameter Selection
Training vs Testing Split
No, Retrain Model
Are Results Satisfactory?
Yes
Model Developed

Figure 9: How a Model is Trained

Not all variables in the data frame, which are independent, directly influence the behaviour of the dependent variable. This is particularly true for larger datasets where irrelevant features can lead to the machine mistaking irrelevant and noisy data for patterns if they are not removed.

As such, feature selection tools like Independent Component Analysis (ICA) and Principal Component Analysis (PCA) can distil the independent features from a total to an optimum value as part of the implementation phase. The use of PCA in reducing the independent category down to 23 from 45 for the deep learning model as shown in Figure 10. It also reduced the independent variables to 23 from 40 in the KNN regression model.

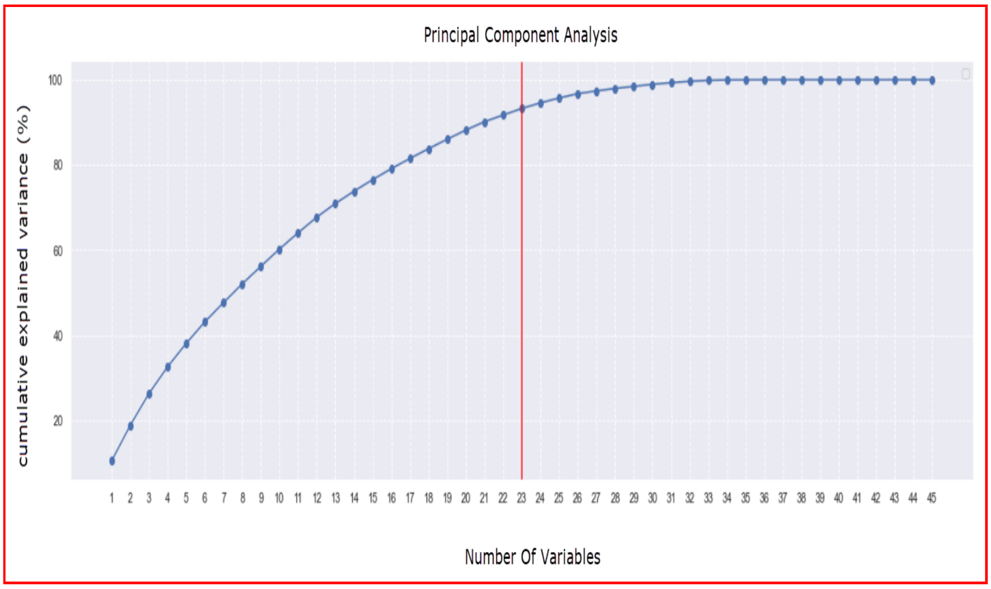


Figure 10: Principal Component Analysis

Hyperparameters (and their tuning) govern the model, especially concerning minimising loss whilst training. Choosing the hyperparameters is a specialised task and needed an intensive grid search to measure the value of all permutations. Table 2 displays optimum values for each of the models

Table 2: Optimum Hyperparamters

Model	**Optimum Hyperparamters**
KNN Regression Model	K = 3
Random Forrest Regression Model	bootstrap=False, maxDepth=None, maxFeatures=sqrt, minSamplesSplit=2, nEstimators=250
Decision Tree Regression Model	max-depth=5, max-features='auto', max-leaf-nodes=20, min-samples-leaf=2, min-weight-fraction-leaf=0.1, splitter='best'
Deep Learning Model	batchSize=16, optimizer=Adam, kernelInitializer=uniform, activation=relu, biasInitializer=randomNormal, dropoutLayer1=0, dropoutLayer2=0.6

The final step in training a model is to split the data into its training and testing set. The training phase involves exposing the machine to 75% of the dataset in which it learns how conclusions regarding the dependent variable are calculated. Then, the remaining 25% of the dataset is partitioned and is used to test the accuracy of the machine. The performance of predictions on this testing data is what verifies how well the model performs. Several distinct metrics are generated from the testing data. Depending on the performance, a retraining period from the point of data scaling onwards may be necessary. Objective 5 and Objective 7 have now been fully achieved.

5 Evaluation and Results of Developed Models

5.1 Regression Models for Smart Energy Predictions

Whilst the primary benefit of daily prediction is experienced by the consumer; there is an additional advantage of the supply grid knowing how much energy is needed for the remainder of the present 24 hour period. The following specific metrics will be used to measure the performance of the regression models.

- **Mean Squared Error (MSE)** The MSE informs how close a line of regression is to a particular set of points. It is calculated by squaring the differences between the points and the line.
- **Root Mean Squared Error (RMSE)** The RMSE is the standard deviation of the prediction errors of how far the data points are from the line of regression. It represents how strong the concentration of data points is around the best fitting line.
- **Mean Absolute Error (MAE)** The MAE is the most basic form of a regression error metric. The residuals for each data point are calculated, but only the absolute value is taken (i.e. the direction is not considered).
- **Mean Absolute Percentage Error (MAPE)** The MAPE is the average of absolute percentage errors of predictions.

Random Forest and Decision Trees provided the best outputs, as can be seen with the results in tableTable 3. A further breakdown of how the regression line is fitted between the actual and predicted values is displayed in Figure 11. It is noticeable how close the incorrect values are to the line of best fit. No outliers (severely inaccurate predictions) are present, making random forest and decision trees the most efficient regression models. By contrast, KNN achieves poor results when measured across the four metrics, with results depreciating as more neighbours are included in the model as shown in Figure 12. Objective 8 has now been partially achieved. Sub Research Question 2 has now been fully solved.

Table 3: Regressor Results

Metric	**Random Forest Regressor Results**	**Decision Tree Regressor Results**
MAPE	0.0046	0.0044
MAE	0.0007	0.0006
MSE	0.0019	0.0019
RMSE	0.0435	0.0436

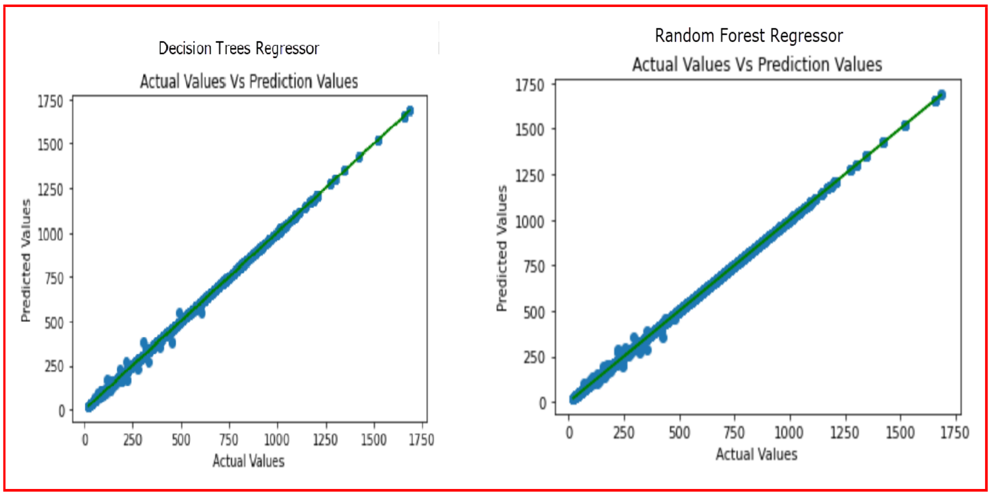


Figure 11: Decision Trees and Random Forest Actual Values vs Predicted Values

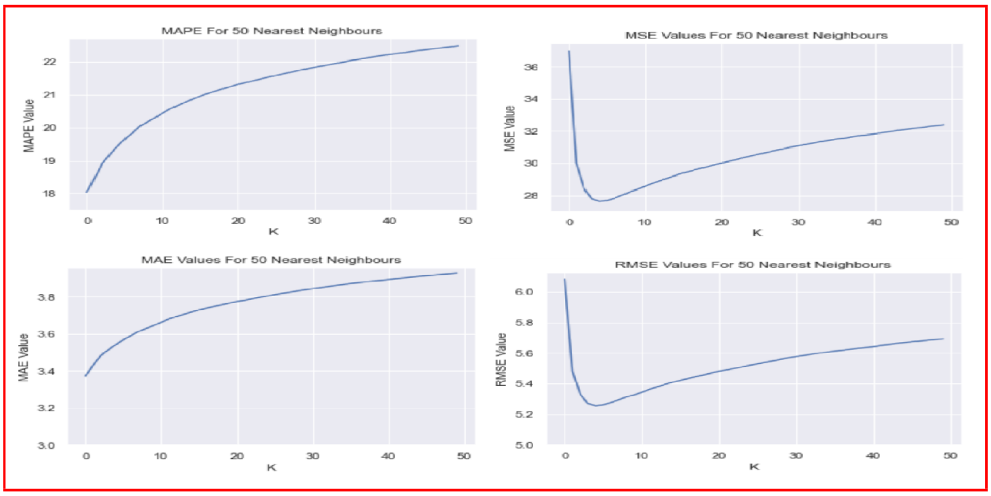


Figure 12: KNearest Neighbour Results over 50 Neighbours

5.2 Colour Code My Energy(CCME) Algorithm for Smart Energy Optimisations

It can be seen from Figure 13 that the model will offer minimal insights or valuable recommendations in the first 12 months as it is still building a learning base about the user. The difference between what the algorithm views as new reads vs reads it can match are converging but is still a considerable distance apart by the end of the first year. Section 7 will address future ideas as to how this gap can be bridged. Between the first and second years, we can see the knowledge and number of matched reads increase to a 50/50 or 60/40 split for the majority of this time. The knowledge base drastically improves in the 3rd year and reaches an 80/20 split towards the end of this period.

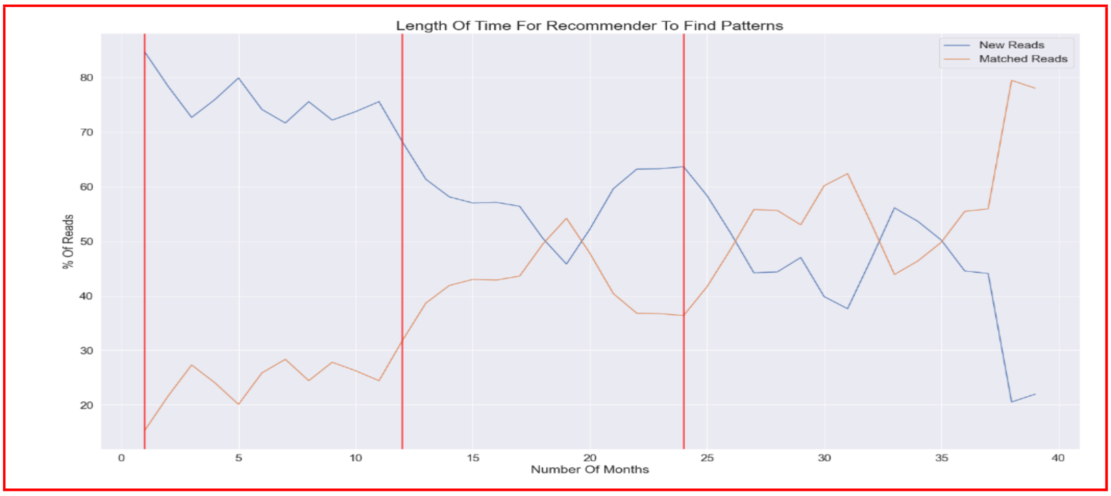


Figure 13: Time to Find Recommendations

A snapshot of how the algorithms developing knowledge is best demonstrated by analysing a specific house. Figure 14 compares an early month (month number 2) to an advanced month (month number 33) for house 9. The number of reads found scenario has increased from a little under 100 to almost five times this amount. The quality of the recommendations has also dramatically increased within this timespan as the CCME recommender has moved from labelling the majority of the reads labelled in the ”white” category (symbolising new/previously unrecorded scenarios) to providing critical recommendations as to how much excess energy is consumed whilst also updating the optimum values for any reads classified as ”green”. This visual also demonstrates that recommendations are not necessarily about hitting the optimum value. Allowing for a certain percentage of ”orange” reads to be upgraded to ”yellow” and subsequently ”yellow” read to be improved to ”blue”, overall efficiency improves without increasing the number of ”green” reads.

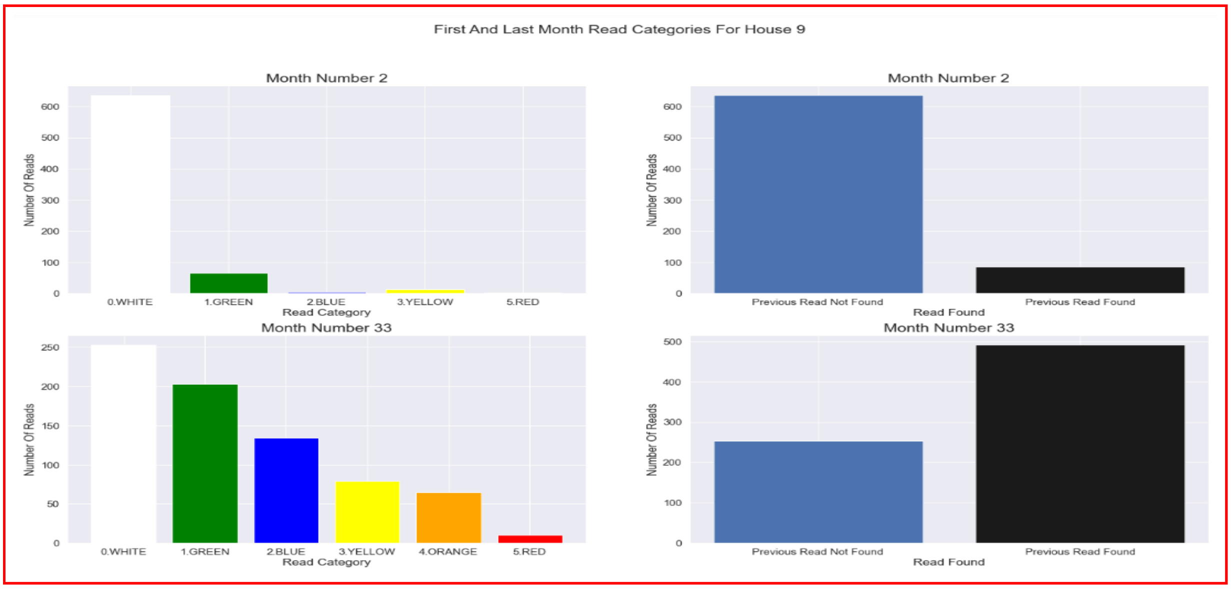


Figure 14: House 9 Statistics

Even though house number 9 achieves a high percentage of ”green” reads, there is much potential within months 14,15 and 16 and 26,27 and 28 to reduce consumption and move closer to the recommended level, as can be seen in Figure 15.

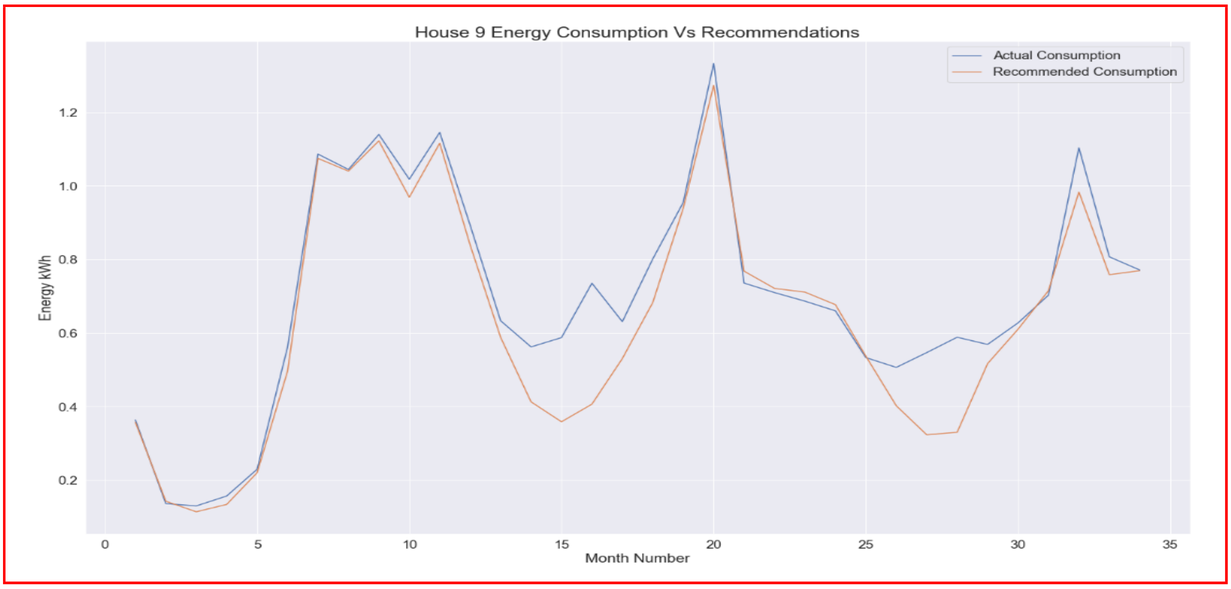


Figure 15: House 9 Statistics

As acknowledged, the CCME recommender will take over two years before it begins making real-time recommendations by classifying scenarios accurately. Nevertheless, as a starting point, Figure 16 demonstrates how there is the facility to enter historical data and apply the algorithm retrospectively so that users can be aware of instant areas to target concerning excess consumption. Once again focussing on house number 9, it can be seen that despite the number of green reads it attains, there is substantial work to be done to reduce wastage in season 2 (summertime) and, to a lesser extent, daily timeslot 4. Objective 8 has now been partially achieved. The Research Question has now been partially solved.

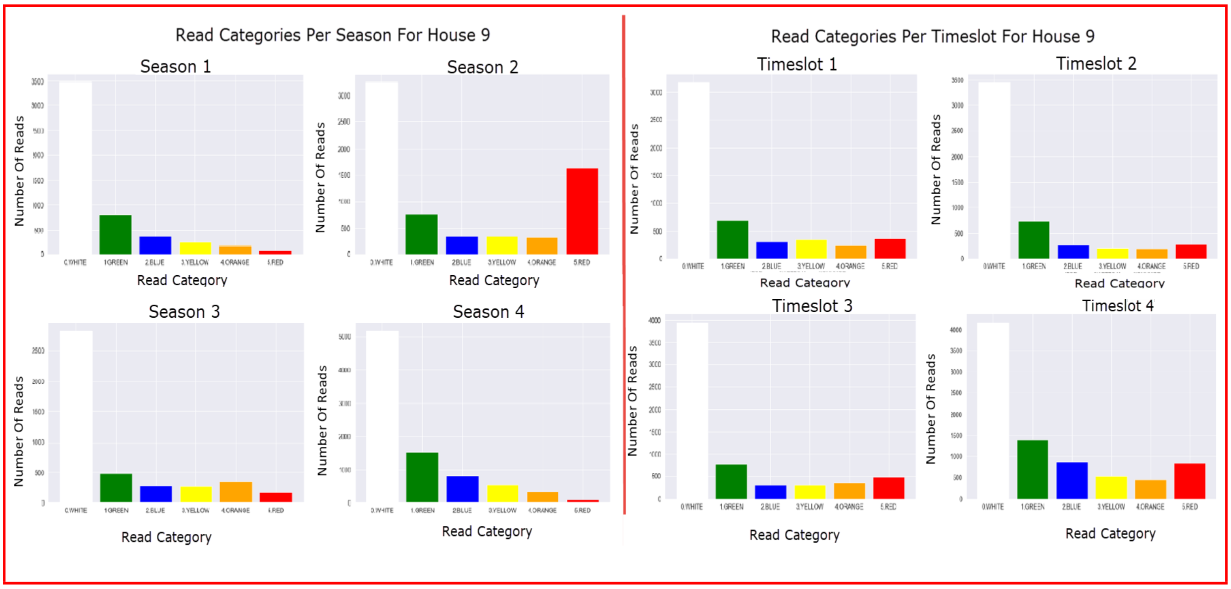


Figure 16: House 9 Retrospective Statistics

5.3 Deep Learning Model for Smart Energy Recommendations

The deep learning neural network seeks to predict what category should be attached to an energy read based on the testing set after a training period. The main metrics in which to measure success are the model accuracy and loss. Accuracy is the most basic metric within the classification set of evaluations and represents how many predictions the model got right as a percentage over the model's total predictions. The loss number represents the bad predictions and indicates how poorly the model predicts a single example. An epoch is an entire iteration of the training set. The batch size determines the number of records accessed. The batch size was set to 512, which resulted in 889 records used for validation during each epoch. Setting the number of epochs to be executed to 5000 with a proviso of early stopping was deemed the most suitable strategy as attempting to prescribe a definitive epoch quantity is a delicate matter. The model terminated at 180 epochs. As can be seen in Figure 17, the model achieved accuracy levels of just over 0.98 and a loss of 0.07 as its peak results.

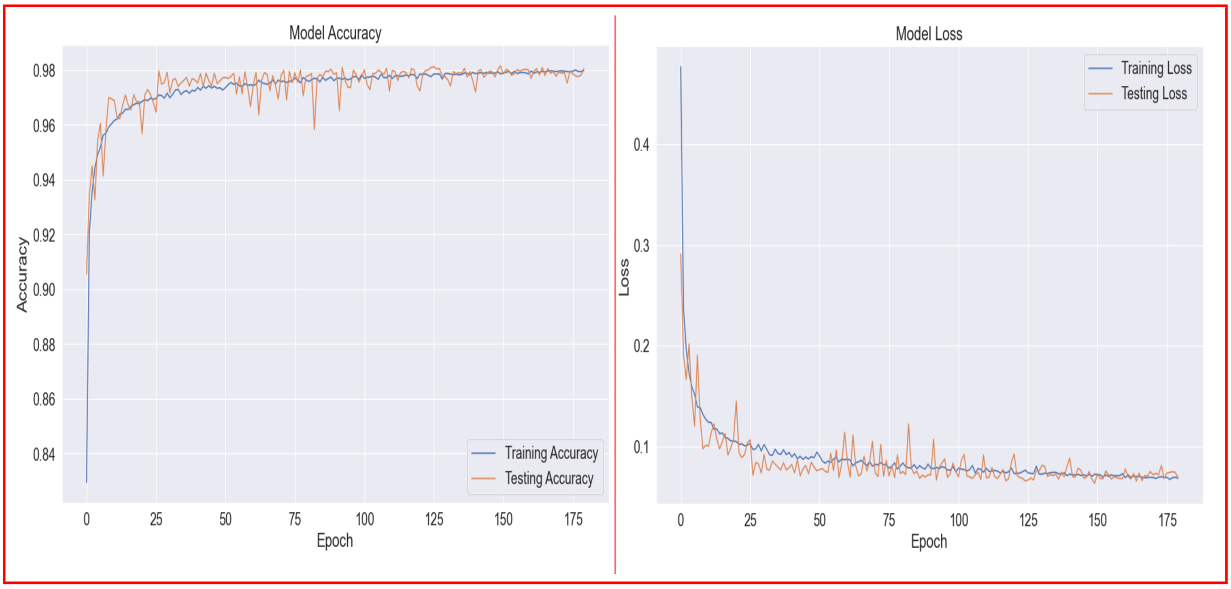
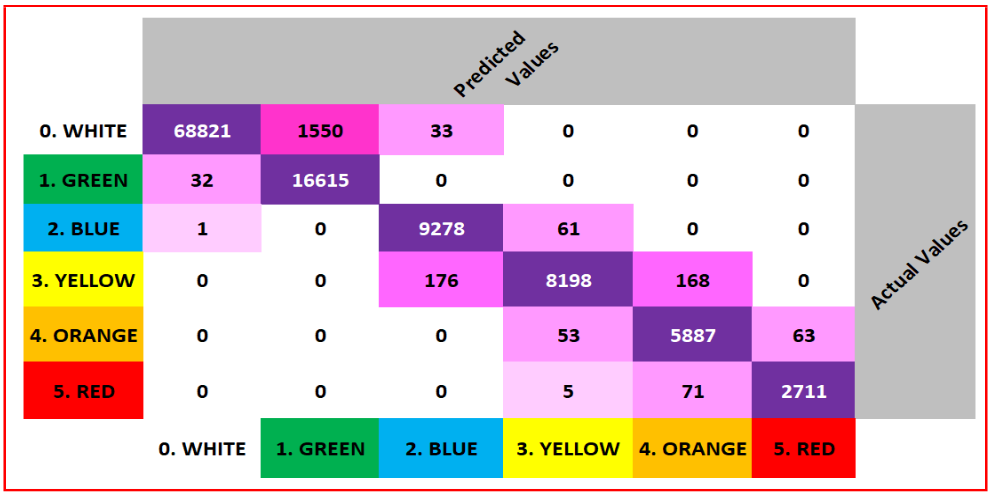


Figure 17: DNN Accuracy vs Loss

Moving onto a class by class analysis, the confusion matrix in figurename 18 to specifically track how the numbers between the prediction values and their actual values. It is a useful visual for honing in on where the incorrect predictions were classified. Whilst the frequency of incorrect predictions is on the low level, the 1550 cases where "green" is predicted when the reads were "white" might indicate a small level of retraining might be necessary. Precision calculates the amount of correct positive predictions the model has made and is defined as the predicted positive examples / the total number of positive examples predicted. The recall metric calculates the ratio of positive samples correctly classified as positive to the total number of positive samples. Thus, it is a measure of how well the model predicts positive samples. The F1 Score is a balance between precision and recall and needs to be considered concerning both of these metrics. Precision statistics were high, with the lowest number of 0.96 recorded in the "yellow" class, whilst Recall documented a slightly lower figure of 0.91 in the "green" category. Objective 8 has now been fully achieved. Sub Research Question 1 has now been fully solved. The Research Question has now been fully solved.



Predicted Values

Actual Values	0. WHITE	1. GREEN	2. BLUE	3. YELLOW	4. ORANGE	5. RED
0. WHITE	68821	1550	33	0	0	0
1. GREEN	32	16615	0	0	0	0
2. BLUE	1	0	9278	61	0	0
3. YELLOW	0	0	176	8198	168	0
4. ORANGE	0	0	0	53	5887	63
5. RED	0	0	0	5	71	2711

Figure 18: DNN Confusion Matrix

6 Discussion and Comparison

Comparing the CCME recommender to other literature pieces did not find anything remotely comparable. Nazeriye et al. (2020) who has the intention of optimising through analysis of specific appliances which are increasing consumption. Whilst this is interesting to see trends at a level as low as appliances, there is little room to optimise, other than not using the particular device. Both Raval et al. (2021) and Pawar and Vittal (2017) focus more on optimisation of hardware and infrastructure. It is interesting to note that whilst the papers looked at various aspects in isolation; no one piece sought to marry prediction, recommendations and optimisation within one body of work. It is difficult to imagine the three areas not interlinked, and as such, one or both of the alternative areas is counterproductive.

Whilst 7% is a figure of savings reached in studies of Wang et al. (2020), Schweizer et al. (2015) and Alsalemi et al. (2021), the CCME algorithm has not achieved higher, it must be countered that this was not a study in which the houses were aware of and provided updates of recommendations throughout their readings.

Oprea and Bâra (2019), Shapi et al. (2021) and Oprea and Bâra (2019) focus on short term predictions, but none of the three studies focuses on the present-day or achieves MAPE, MAE, RMSE or MSE scores in the region of the decision tree and random forest regressors developed within this study.

With these comparisons, Objective 9 has now been fully achieved.

7 Future Work and Conclusion

Although significant discoveries were made as part of the research in this project, some avenues can be explored concerning future work. For example, the machine learning regression algorithms generally yielded impressive figures. Still, KNN could be re-examined, perhaps as part of an ensemble learning technique in which multiple models are looked at together instead of in isolation.

Considering its infancy, there is scope for further developing the CCME recommender algorithm and extending it to make more realistic predictions and attain true optimisation. For example, the facility could connect to a recommender table for identical house types within a certain radius when installing into new homes. Using a median/mode value of like for like properties in the same scenario as its baseline recommendation instead of spending the first number of months or years establishing patterns for the new user would enable a quicker learning curve. The trade-off for this would be that recommendations will be made at a global/area level instead of that of the localised individual. Still, the amount of "white" coloured reads in the early months will be drastically reduced.

A subsequent and perhaps more challenging area for future development is realigning one off optimisations that serve as an unrealistic baseline figure. For example, imagining a situation where a house is unoccupied for some time, energy consumption will be at a bare minimum. It will be entered as the new recommended value for the scenarios encountered within this timespan. However, it is impracticable to retain this figure as the target to aim for and grade any reads from this point on as in the higher usage category when more normalised circumstances resume. Likewise, there needs to be an option to recalibrate what "normal" circumstances are when changes to what is normal occur. Examples are not limited to but include, during the school holidays or if an individual increases the number of days they work from home. In these circumstances, the energy consumption environment change meaning recommendations need a level of recalibration.

Acknowledgements

I would like to express my deepest and most sincere thanks to my Mum and Dad for their never-ending love, emotional and financial support. Without their ever-present encouragement, none of this would be possible.

I wish to thank Dr. Catherine Mulwa, who went above and beyond as project supervisor. Never letting standards drop and making sure my full potential was reached. She was a constant source of encouragement.

I am grateful to the National College of Ireland, all the staff and lecturers I have met throughout the two years of study in my Postgraduate Diploma and Masters in Data Analytics.

Finally, I am incredibly thankful to the many classmates who have become friends. Individuals who helped, motivated and inspired through the good and challenging times.

I dedicate this work to my cousin Michael.

References

Alsalemi, A., Himeur, Y., Bensaali, F., Amira, A., Sardianos, C., Chronis, C., Varlamis, I. and Dimitrakopoulos, G. (2021). A micro-moment system for domestic energy efficiency analysis, IEEE Systems Journal **15**(1): 1256–1263.

Chen, Y., Zhang, F. and Berardi, U. (2020). Day-ahead prediction of hourly subentry energy consumption in the building sector using pattern recognition algorithms, Energy **211**: 118530.

Eneyew, D. D., Capretz, M. A. M., Bitsuamlak, G. T. and Mir, S. (2020). Predicting residential energy consumption using wavelet decomposition with deep neural network, 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, Miami, FL, USA, pp. 895–900.

Goyal, M., Pandey, M. and Thakur, R. (2020). Exploratory analysis of machine learning techniques to predict energy efficiency in buildings, 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1033–1037.

Jindal, A., Schaeffer-Filho, A., Marnerides, A. K., Smith, P., Mauthe, A. and Granville, L. (2020). Tackling energy theft in smart grids through data-driven analysis, 2020 International Conference on Computing, Networking and Communications (ICNC), IEEE, Big Island, HI, USA, pp. 410–414.

Khan, P. W., Byun, Y.-C., Lee, S.-J., Kang, D.-H., Kang, J.-Y. and Park, H.-S. (2020). Machine learning-based approach to predict energy consumption of renewable and non-renewable power sources, Energies **13**(18).

Kontogiannis, D., Bargiotas, D. and Daskalopulu, A. (2021). Fuzzy control system for smart energy management in residential buildings based on environmental data, Energies **14**(3).

Makonin, S. (2018). HUE: The Hourly Usage of Energy Dataset for Buildings in British Columbia.
URL: *https://doi.org/10.7910/DVN/N3HGRN*

Makonin, S. (2019). Hue: The hourly usage of energy dataset for buildings in british columbia, Data in Brief **23**: 103744.
URL: *https://www.sciencedirect.com/science/article/pii/S2352340919300939*

Mishra, P., Gudla, S. K., ShanBhag, A. D. and Bose, J. (2020). Alternate action recommender system using recurrent patterns of smart home users, 2020 IEEE 17th Annual Consumer Communications Networking Conference (CCNC), IEEE, Las Vegas, NV, USA, pp. 1–6.

Nabavi, S. A., Aslani, A., Zaidan, M. A., Zandi, M., Mohammadi, S. and Hossein Motlagh, N. (2020). Machine learning modeling for energy consumption of residential and commercial sectors, Energies **13**(19).

Nazeriye, M., Haeri, A. and Martínez-Álvarez, F. (2020). Analysis of the impact of residential property and equipment on building energy efficiency and consumption—a data mining approach, Applied Sciences **10**(10).

Nystrup, P., Madsen, H., Blomgren, E. M. and de Zotti, G. (2021). Clustering commercial and industrial load patterns for long-term energy planning, Smart Energy **2**: 100010.

Oprea, S.-V. and Bâra, A. (2019). Machine learning algorithms for short-term load forecast in residential buildings using smart meters, sensors and big data solutions, IEEE Access **7**: 177874–177889.

Pawar, P. and Vittal, K. P. (2017). Design of smart socket for power optimization in home energy management system, 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT), pp. 1739–1744.

Raval, M., Bhardwaj, S., Aravelli, A., Dofe, J. and Gohel, H. (2021). Smart energy optimization for massive iot using artificial intelligence, Internet of Things **13**: 100354.

Schweizer, D., Zehnder, M., Wache, H., Witschel, H.-F., Zanatta, D. and Rodriguez, M. (2015). Using consumer behavior data to reduce energy consumption in smart homes: Applying machine learning to save energy without lowering comfort of inhabitants, 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), IEEE, Miami, Florida, USA, pp. 1123–1129.

Shapi, M., Ramli, N. A. and Awalin, L. (2021). Energy consumption prediction by using machine learning for smart building: Case study in malaysia, Developments in the Built Environment **5**: 100037.

Syed, D., Abu-Rub, H., S. Refaat, S. and Xie, L. (2020). Detection of energy theft in smart grids using electricity consumption patterns, 2020 IEEE International Conference on Big Data (Big Data), IEEE, Atlanta, GA, USA, pp. 4059–4064.

Wang, A., Lam, J., Song, S., Li, V. and Guo, P. (2020). Can smart energy information interventions help householders save electricity? a svr machine learning approach, Environmental Science & Policy **112**: 381–393.

Wei, P., Xia, S., Chen, R., Qian, J., Li, C. and Jiang, X. (2020). A deep-reinforcement-learning-based recommender system for occupant-driven energy optimization in commercial buildings, IEEE Internet of Things Journal **7**(7): 6402–6413.

Yiyi, C., Mitra, D. and Cetin, K. (2020). Data-driven energy prediction in residential buildings using lstm and 1-d cnn., ASHRAE Transactions **126**(2): 80 – 87.