Measure energy consumption using machine learning

Project Title: Measure energy consumption
Phase 3: Development part

Introduction:

The prediction method will employ 3 machine learning algorithms which are k-NN, SVM and ANN. Feature attributes for this prediction will use electrical power data consisting of power factor, voltage and current, in which the demand would be the targeted output. The prediction modeling will be conducted inside Microsoft Azure Machine Learning Studio (AzureML) utilising R programming language (Caret Package). Microsoft Azure has been chosen as a platform in this research based on the literature reviewed in the previous section. Before model training and testing, the raw data will initially be analysed and pre-processed to reduce the complexity of the model training and to manage any missing data. Finally, each model will be evaluated using validation metrics. Consequently, the energy consumption prediction framework will consist of four parts, which are:

Step 1
Normality testing of dataset

Step 2
Data pre-processing

Step 3
Model development (training)

Step 4
Model evaluation (test)

In this research data analysis, a normality testing of the dataset for each tenant was conducted to determine the dataset distribution. This process was orchestrated by identifying the skewness and kurtosis of the dataset. Normality testing is significant for model development as it usually assumes that the dataset was normally distributed. Based on the background research by Mishra et al. (2019), it was found that normality testing is ignorable if the sample size exceeds 100. However, understanding of the dataset distribution could provide a consequential analysis for the result of the prediction. On the ground of statistical analysis, skewness is defined as the measure of irregular probability distribution around the mean value whereas kurtosis is a quantification of the distribution peakness. The formula for skewness and kurtosis is as shown in equations (1), (2), respectively

where N is the total number of hours, xi is power consumption, and i is hour of the day.

data pre-processing:

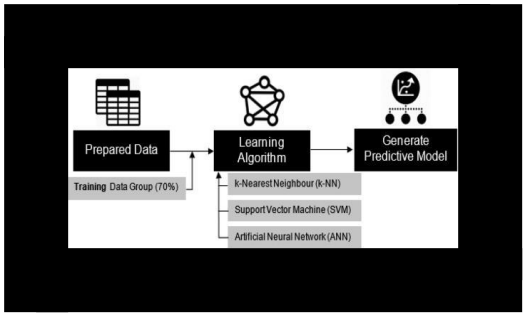
- For this analysis, AzureML Summarize Data module was utilised to present the normality testing numerically and graphically. This module in AzureML is used to create a set of standard statistical measures that describe each column in the input table.
- The preliminary process in machine learning includes data pre-processing for the preparation of data, and it usually consumes much time and computational power. This

process is required as the dataset usually consists of missing value and an inconsistent scale of value between features (Fontama et al., 2014). In this research, the data was pre-processed using mechanics of imputation of missing data and standardisation in which the former utilised Azure ML proprietary Clean Missing Data module while the latter was done using Caret R Package. For Clean Missing Data module in AzureML, it was used to remove, replace, or infer missing values. This module supports multiple types of operations for cleaning missing values including replacing missing values with a placeholder, mean, or other value; completely removing rows and columns that have missing values; or inferring values based on statistical methods.

model development:

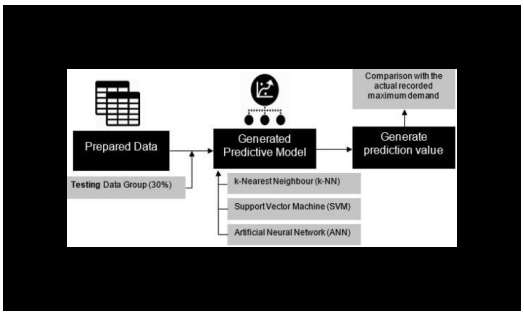
- This research used a supervised machine learning methodology to predict energy consumption. After data was prepared, it was then inputted into the learning algorithm. Different feature combinations were fed into the algorithm to generate a candidate for the predictive model. Before using the data to create and train the model, data partitioning was done to separate the data into two groups – a training group and a testing group.

- The predictive modeling for this research used a classification method to predict discrete variables instead of regressive prediction. As Azure ML does not have k-Nearest Neighbour and Artificial Neural Network for classification, the modeling function in Caret R package was utilised for all prediction to ensure uniform execution. Three types of machine learning algorithm were used for this research which were Artificial Neural Network (ANN-MLP), k-Nearest Neighbour (k-NN), and Support Vector Machine (SVM-RBF). Fig. 1 below shows the process after the data preparation until the generation of the predictive m



Model evaluation:

1. Before inputting the data to the machine learning algorithm, the data was partitioned into two groups whereby 70% of the dataset was used for training and the other 30% was partitioned as testing data groups. The training groups of data were used to train each machine learning algorithm and generate a predictive model that could output value that matches with the recorded maximum demand data while the rest of the data was held back to be used to test the trained predictive model. The process is as illustrated in figure.



Necessary step to follow:

1. Import the necessary libraries:
   python
   import pandas as pd
   from sklearn.model_selection import train_test_split
   from sklearn.preprocessing import StandardScaler

2. Load the dataset into a pandas DataFrame:
   python
   data = pd.read_csv('energy_consumption.csv')

3. Split the dataset into input features (X) and target variable (y):
   python
   X = data.drop('energy_consumption', axis=1)
   y = data['energy_consumption']

4. Split the data into training and testing sets:
   python
   X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

5. Perform any necessary preprocessing steps on the input features:
   - Handle missing values: If there are any missing values in the dataset, you can use methods like dropping rows/columns or imputing values.
   - Encode categorical variables: If there are categorical variables in the dataset, you may need to encode them using techniques like one-hot encoding or label encoding.
   - Scale numerical variables: It's often beneficial to scale numerical variables to a standard range using techniques like standardization or normalization.

6. Apply preprocessing transformations to both training and testing sets:
   python
   scaler = StandardScaler()

```
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Now, you have the preprocessed dataset ready to be used for training machine learning models to measure energy consumption.

Conclusion:

Focusing on the objective of this research, three supervised machine learning prediction methods namely k-Nearest Neighbour, Support Vector Machine with Radial Basis Function kernel, and Artificial Neural Network with Multilayer Perceptron model, were chosen as the algorithm for the predictive model. These methods were successfully compared in terms of their resultant structure and prediction performance. The consequence of the model training and testing shows that each method performed differently for every tenant. SVM method shows the most promising result, whereby it managed to be the best method for 2 tenants which were Tenant A1 and Tenant A2, with RMSE valued at 4.7506789 and 3.5898263, respectively.