## **Assignment 2 – Group 12**

Manish Hariyani, Henrique César, Li-chia Lo and Ben Kent

**GitHub repository link: https://github.com/manishvhariyani/Group12Assignment2**

**Overview of the Problem**

It is commonly assumed that energy consumption is correlated with weather conditions. However, weather conditions change frequently and contain various variables including temperature, rainfall, evaporation, and so on. Which variable or the combination of variables contribute to the electricity consumption and influence the price range? With two datasets of weather and price between 1st January and 31st August 2022, we aim to find out the determining factors to the daily electricity demand and predict future price range.

**Data Modeling Approach**

While conducting our analysis we have considered three approaches:

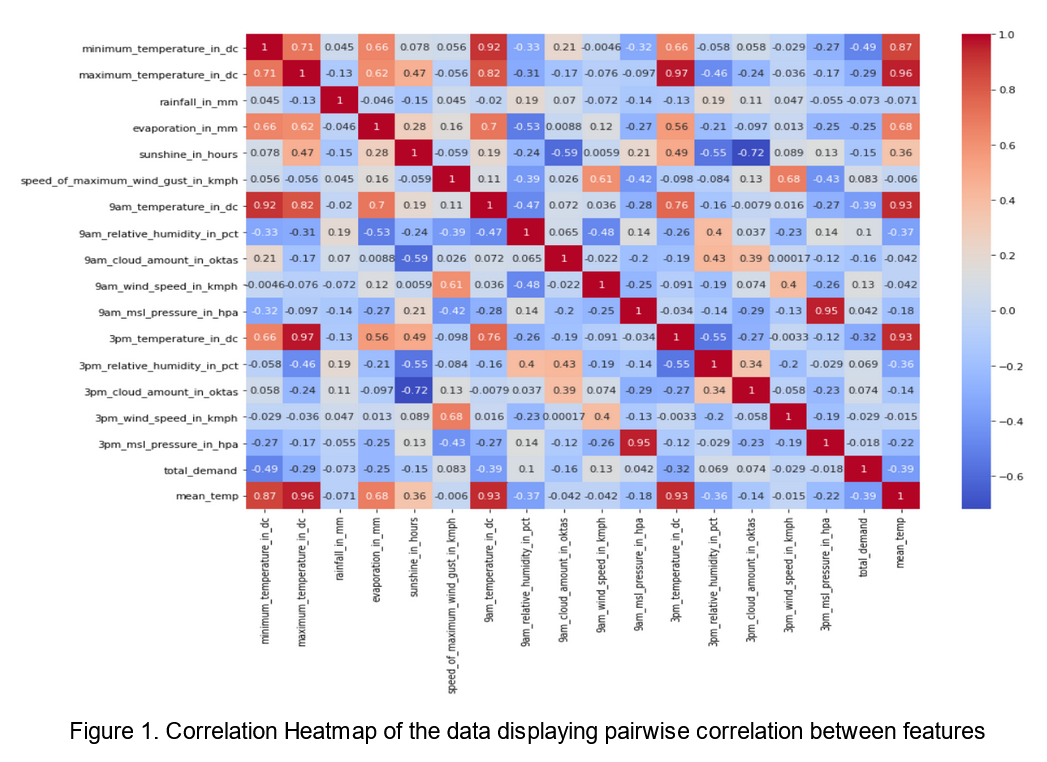
1. Dropping null values
2. Replacing null values with median for numerical data as mean values can be influenced by outliers
3. Replacing null values with mean using a simple imputer

We got similar results in all of the approaches. As our dataset is not large and there are not many null values.

In addition, all column names for the dataframe were changed to lowercase, whitespaces were replaced by underscore, symbols like '°c','km/h','%', were replaced with 'dc', 'kmph', 'pct', and symbols like ')',’(‘ were removed. This made the columns easily accessible.

We have two datasets where date is the common value. The price demand data has 48 values for each date ( a data point for every half an hour), so a group by method is used on the settlement\_date in price demand data, and ‘max’ is the aggregation method used for fetching the maximum total\_demand and price\_category values for that date. We have used inner join to merge both datasets as we didn’t require the rest of the data from price demand data.

We built several models, incorporating linear regression for continuous variable prediction, k-nn and decision tree for categorical variable prediction, and k-fold cross validation. Each model is discussed below.

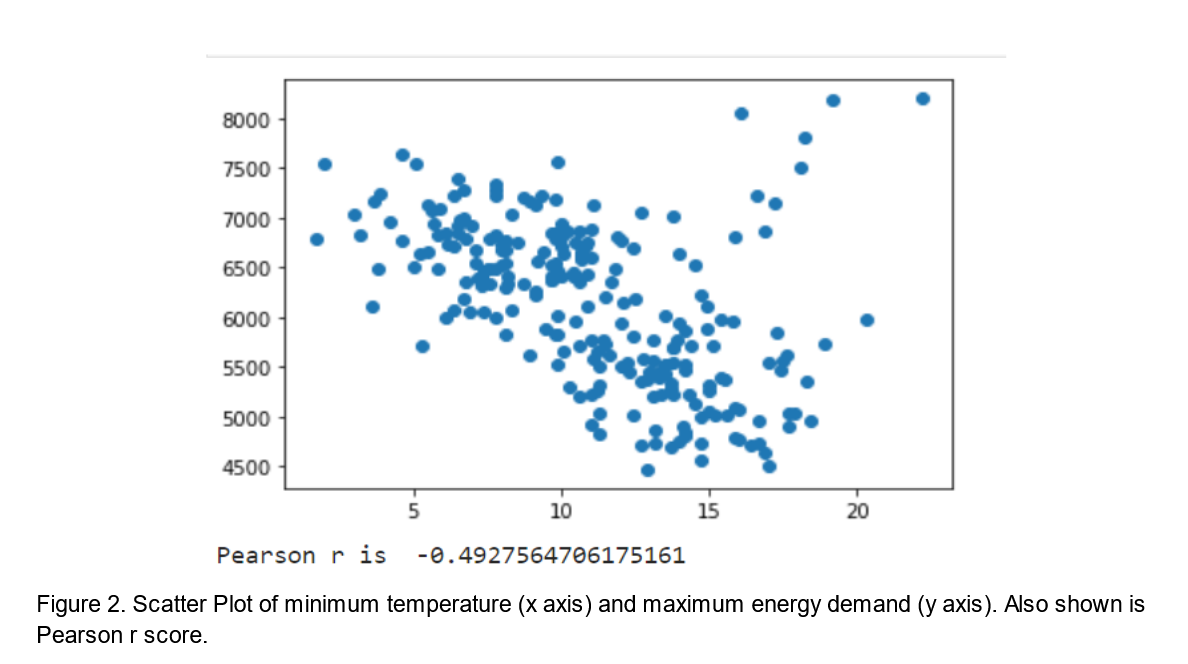


**Linear Regression**

The objective of our model is to predict daily total demand of energy (TD), which is our continuous dependent variable, using a selection of weather indicators. To identify the variables that would help us predict TD, we first conduct a Pearson’s correlation analysis across the whole set of indicators we have available in our dataset. The results of this correlation analysis can be visualized in the heatmap presented below (Figure 1).

As observed on the correlation heatmap above, TD has the highest correlation with minimum temperature. A correlation of -0.49 is a negative value which shows the probability that when the minimum temperature increases energy demand is likely to decrease.

TD is also related to other temperature variables: maximum temperature, 9am temperature and 3pm temperature by -0.29, -0.39 and -0.32 respectively, but the correlation is not high. Additionally there is some correlation with other variables such as, evaporation\_in\_mm(-0.25), sunshine\_in\_hours (-0.15) and 9am cloud amount(-0.16) which have been considered for analysis in some models.



As we can see in Figure 2 above, the relationship between minimum temperature and energy demand is likely to be linear. There are some outliers which show that a different relationship is possible, so we have also calculated the Normalized Mutual Information score for the relationship which is 0.35. Hence the negative linear relationship exists, which can be observed in the scatter plot above, and is also considered for further linear regression analysis.

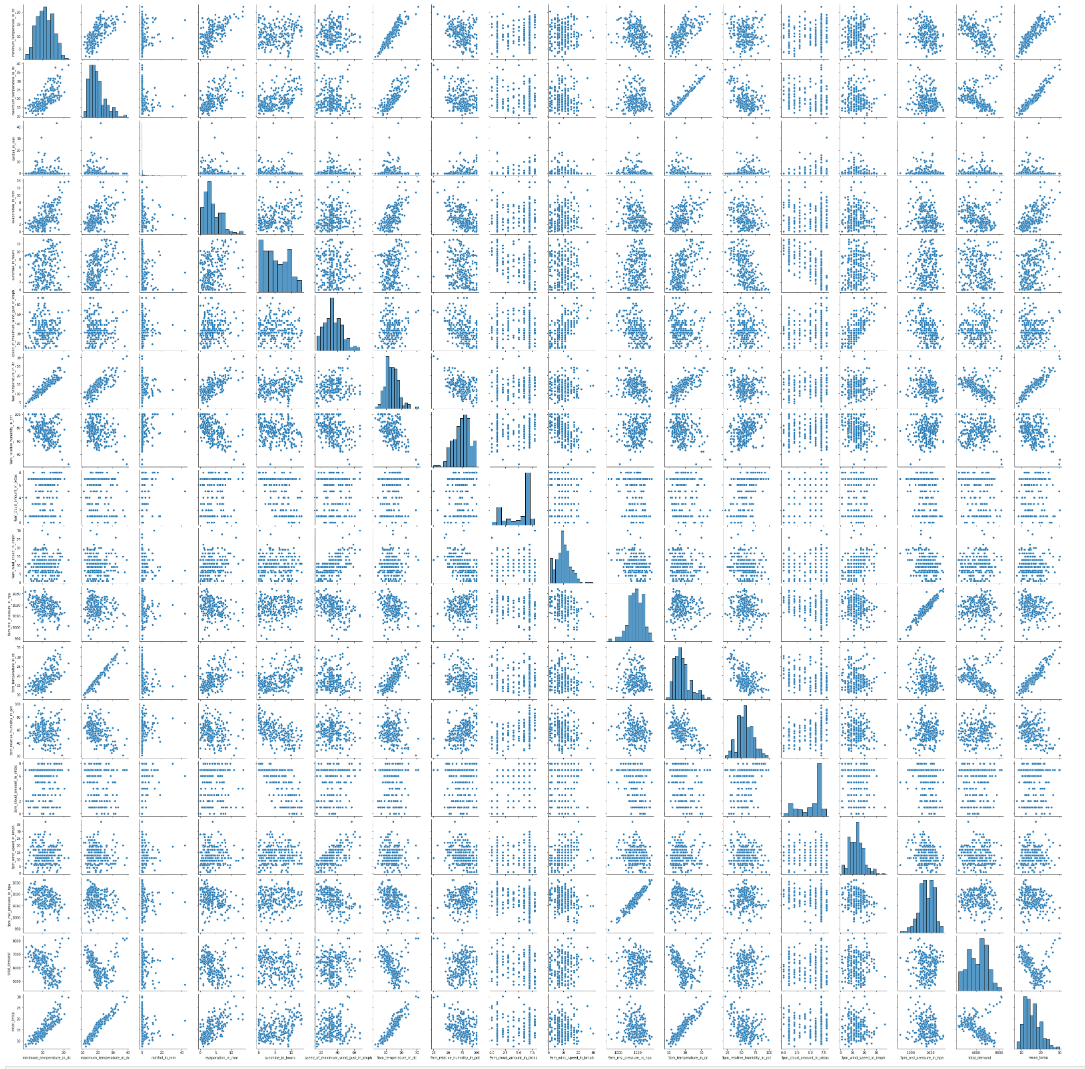
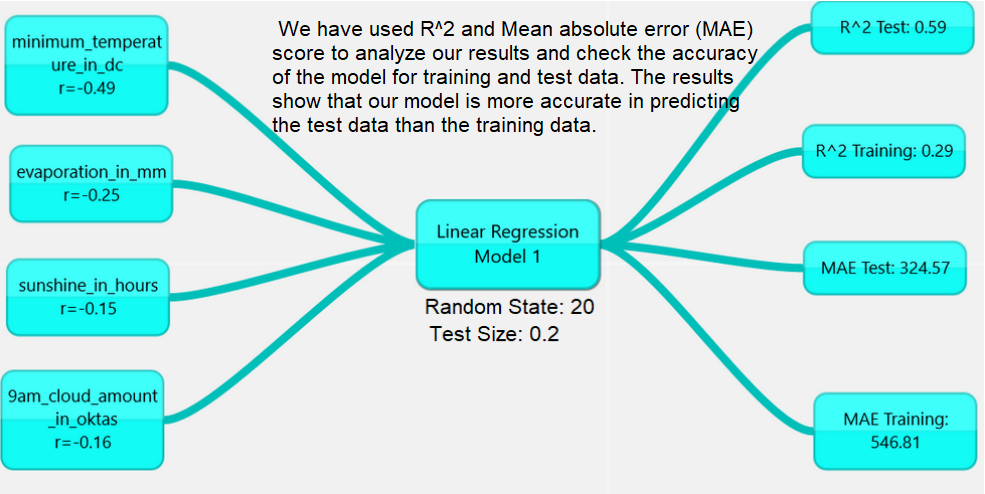
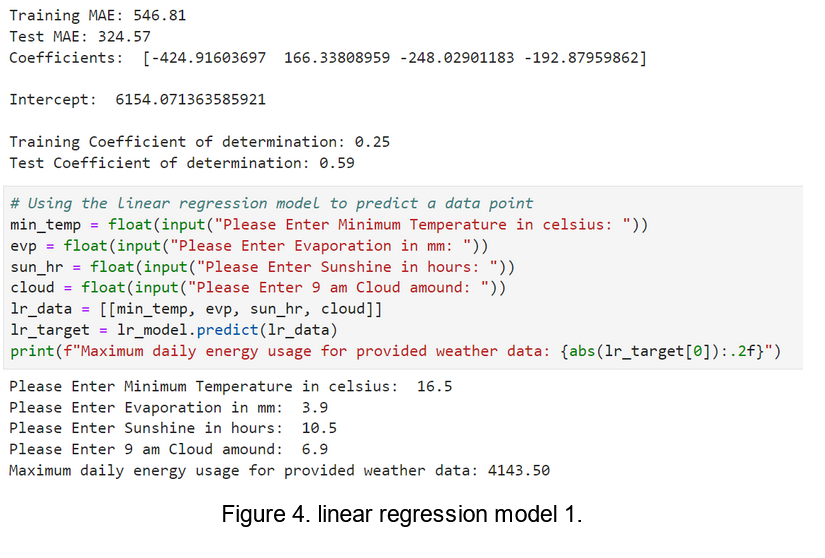


Figure 3. Pairplot showing pairwise scatter plot between variables.(Clear version available in source code file)

As shown in figure 3 a pairplot has been plotted for the entire dataframe, where pairwise relationships can be observed visually and compared with their Pearson r score from the heatmap in figure 1. Variables depicting visual linear relationship and a considerable correlation score, with the target variable, are highly considered for building a linear regression model.

**Observing the heatmap and pairplot we selected the following features for our model 1:**





The resulting model is presented in Equation 1 below.

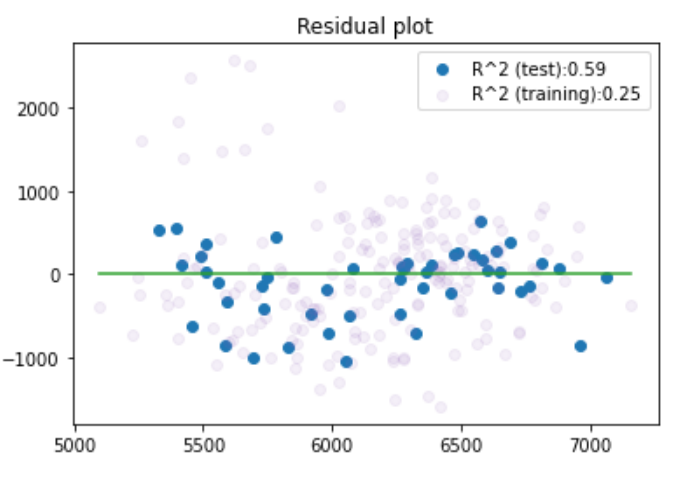
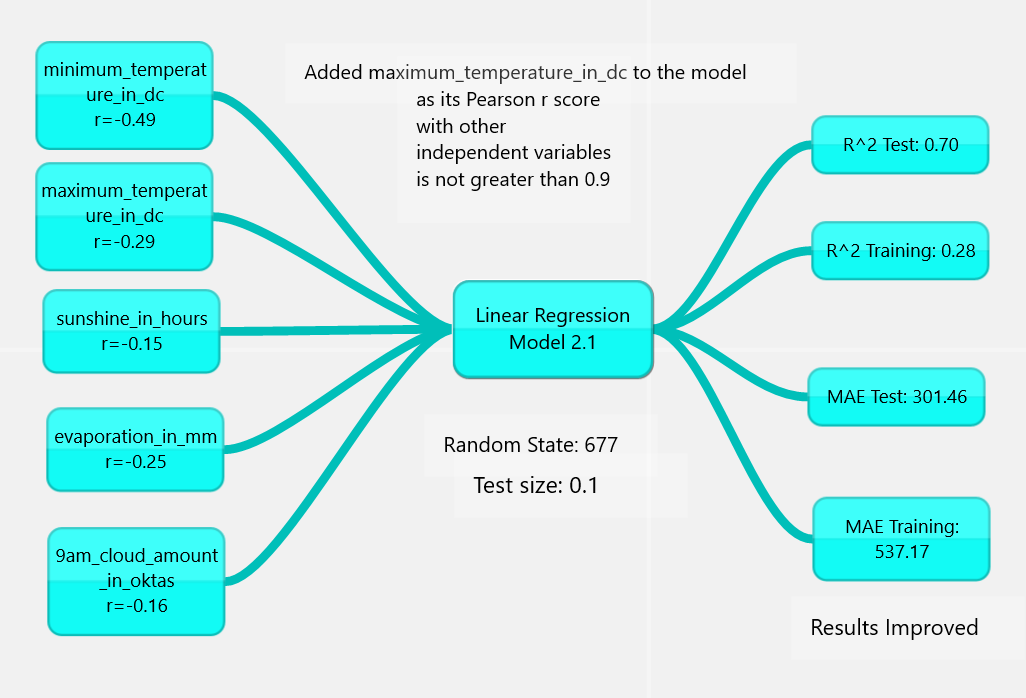
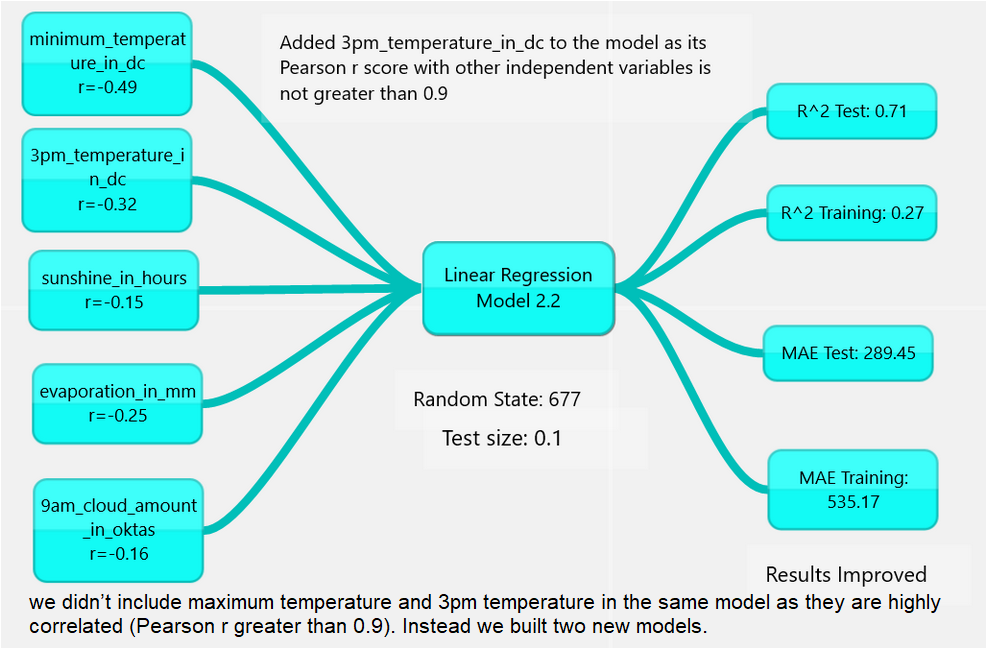


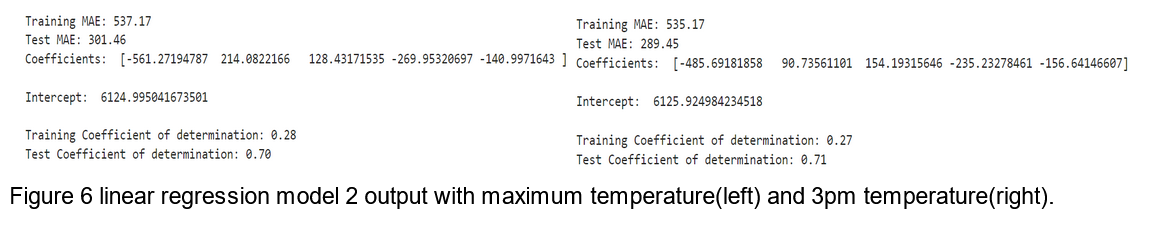
Figure 5 Residual plot of the linear regression model 1.

Residual plot above shows that apart from some outliers, the points are dispersed randomly around a horizontal line, showing that a linear regression model is suitable for this data.

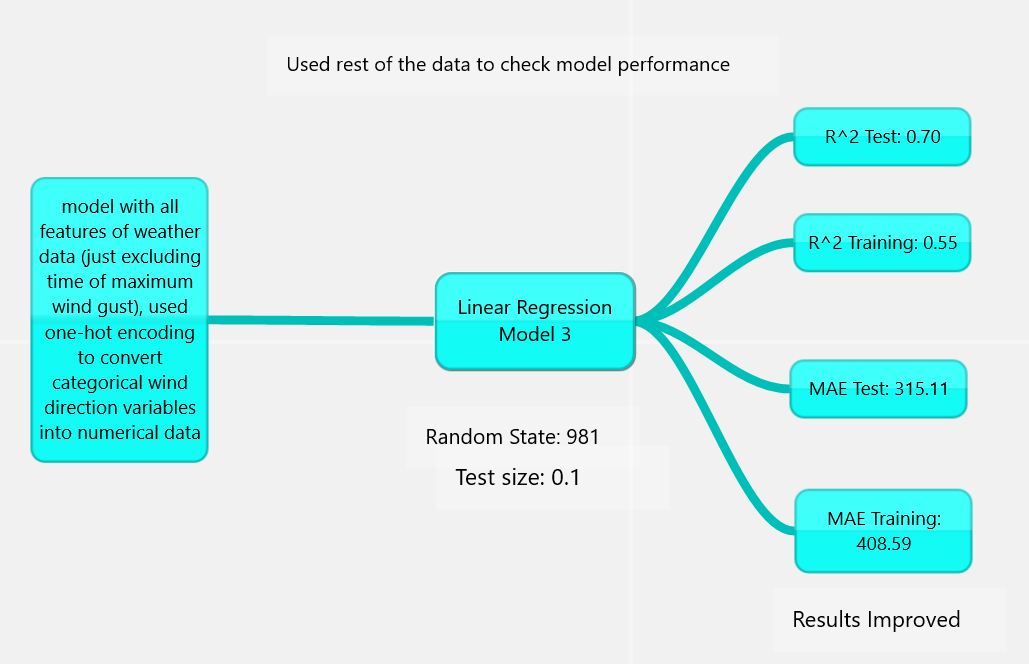
**Model 2:**

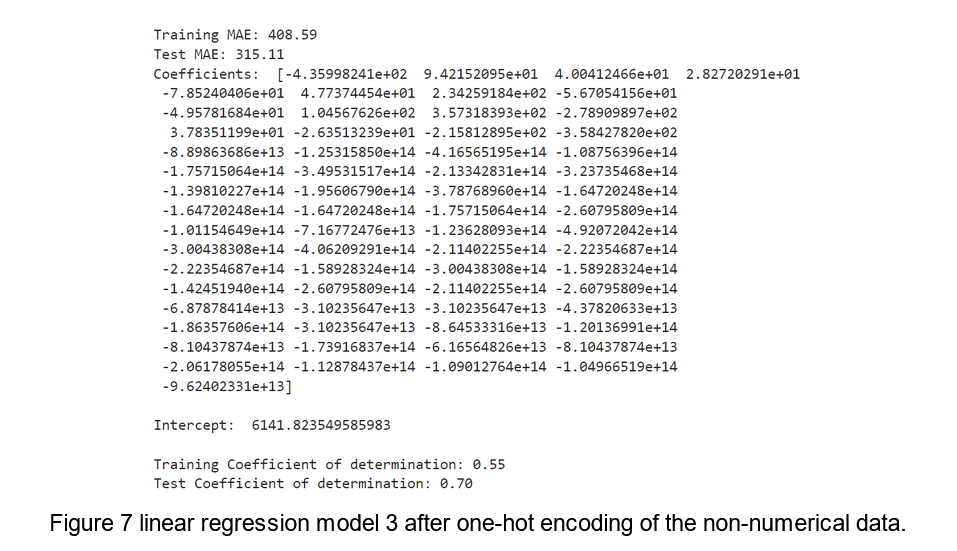




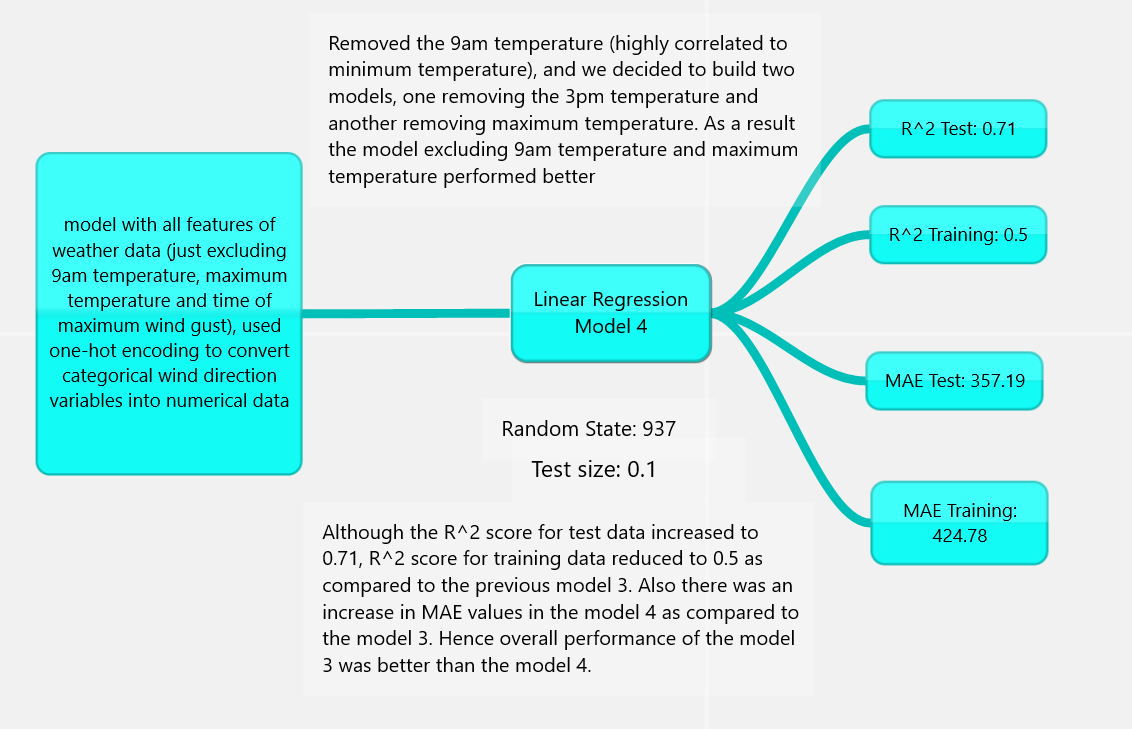


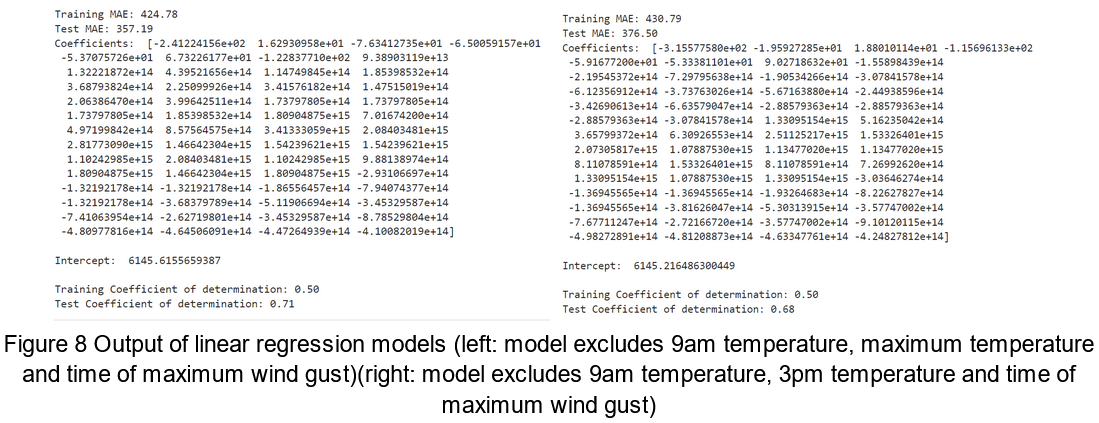
**Model 3:**



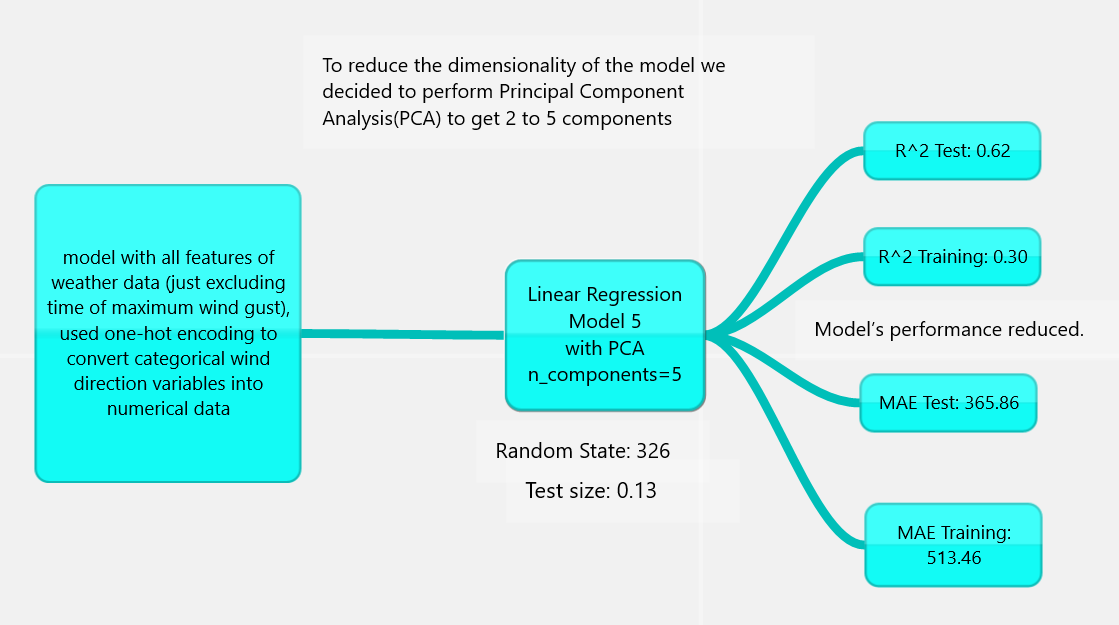


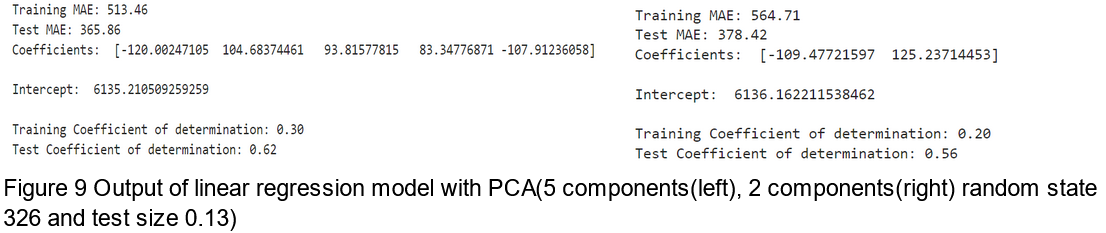
**Model 4:**





**Model 5:**





Models built with variables including minimum\_temperature, 3pm\_temperature/maximum\_temperature, sunshine\_in\_hours, evaporation\_in\_mm and 9am\_cloud\_amount have one of the best performances in predicting maximum energy demand for the day. Although common sense says that the energy consumption tends to be at its highest on days with hotter temperatures the model created contradicts this assumption, for example an increase of 1 Celsius degree resulted in a 424.91 decrease in energy consumption.

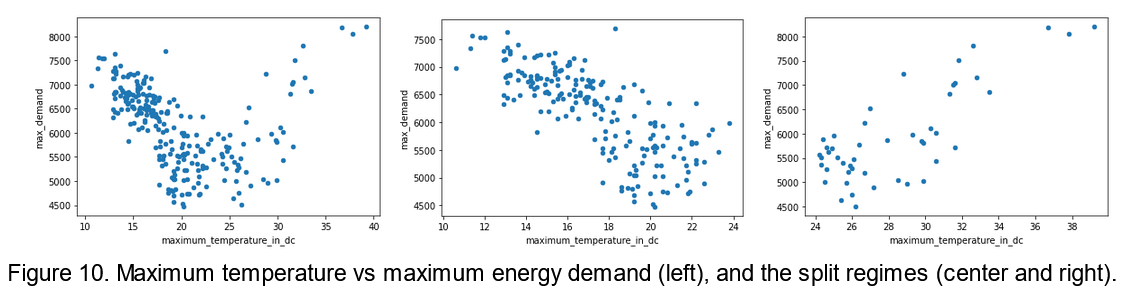
Lastly, about the limitations of the model we can infer that the r^2 could not entirely satisfy the necessity of the field of this area of study once it is just 59% accurate when predicting total demand. Also, the model does not predict the total demand of the days where the temperature is too high since it was considered as outliers.

Therefore, the model can be improved by transforming one or more variables using a “log” transformation. It could change the shape of the distribution reducing the distance between data points and providing a better sum of squared errors. Furthermore, a dataset with more observation points could provide a better model.

**Split Temperature Range Linear Regression**

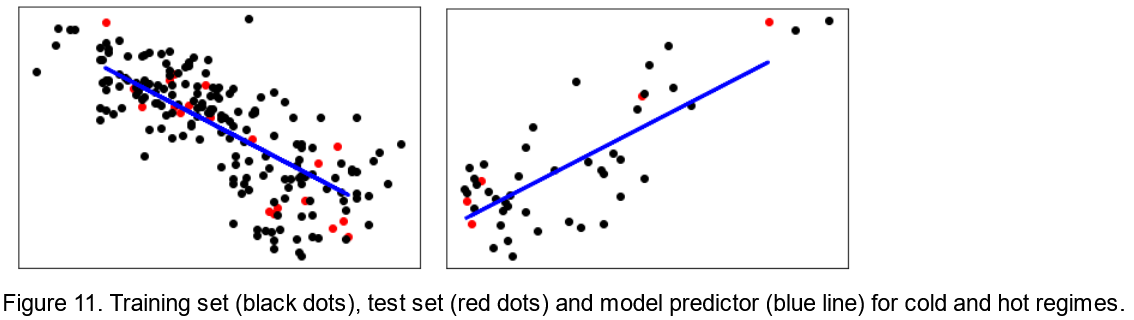
An alternative approach to above was to consider the main trends of electricity use vs weather patterns from our own experience. That is, energy use increases when the weather is very cold due to heating, and also increases on very hot days due to cooling. This is born out in the scatter plot of maximum temperature vs maximum energy demand.

To apply linear regression to these regimes, we split the data into cool (<24 degrees) and hot (≥ 24 degrees).

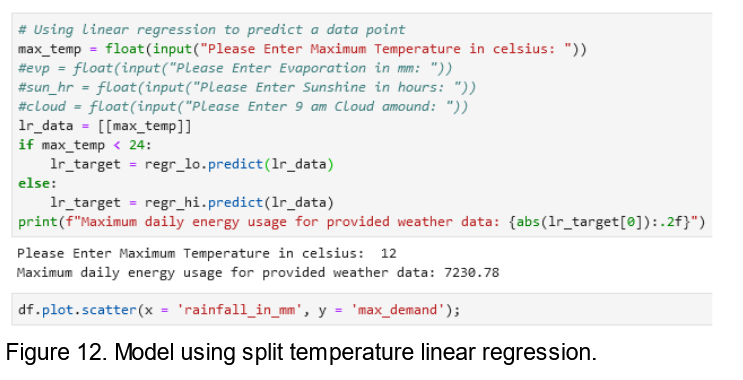


With split regimes, the data showed clear linear trends, with pearson’s r correlations of -0.76 and 0.79 respectively. We then based our model using only the maximum temperature feature.

Using a 0.9/0.1 train test split, we achieved linear regression models with coefficients of determination of 0.71 and 0.86 for the cold and hot regimes respectively.



We then used an if statement in our model predictor, to select the correct model based on used input.

****

This approach achieved a better performance than linear regression over the entire temperature range, and avoided the outliers that were present due to high energy usage on hot days in the earlier models.

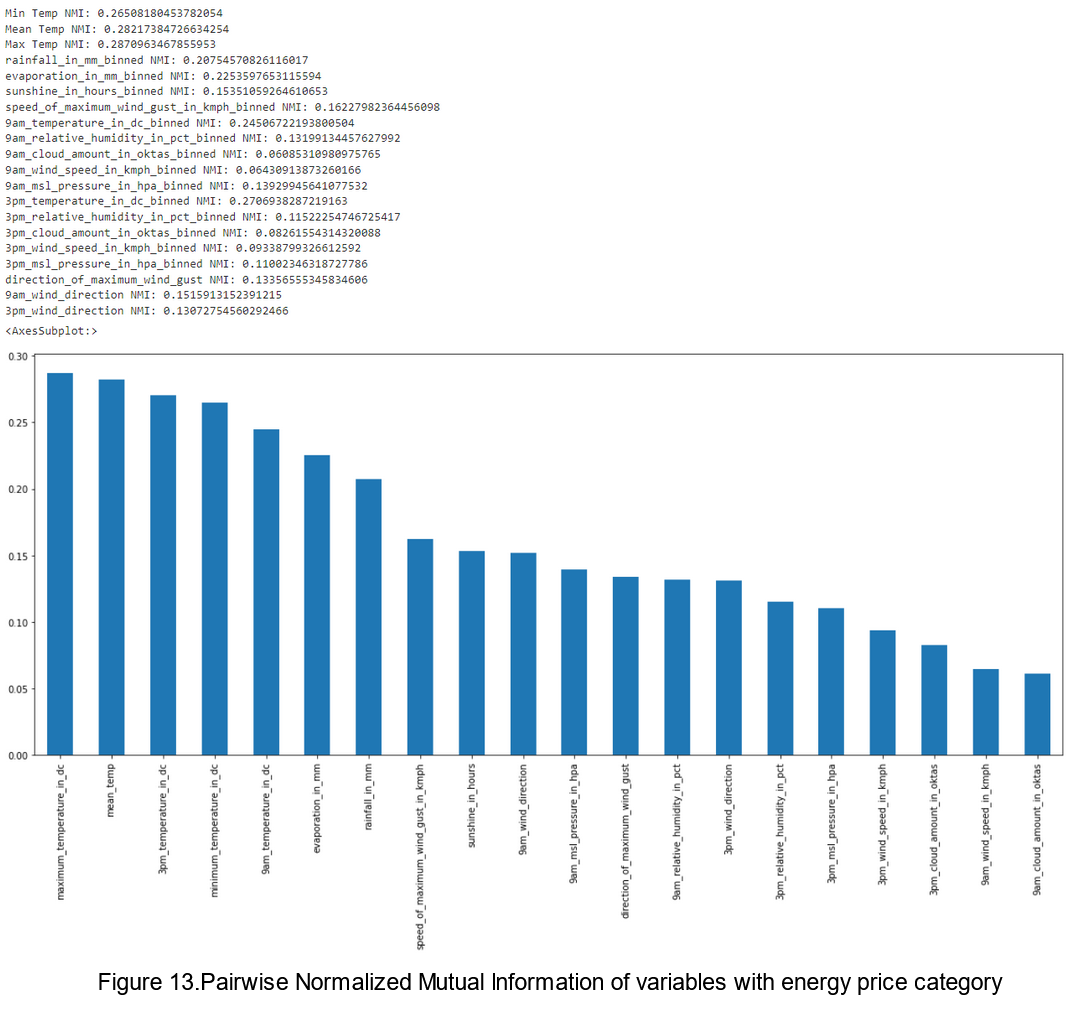
Maximum temperature was the most useful variable when considering a split cold/hot dataset. This achieved our highest accuracy in the prediction of energy usage.

Our results are significant in that it can be shown that a single variable can often be sufficient to create a model with satisfactory accuracy in prediction.

Firstly the data can only be predicted within the temperature range in the training dataset. To include a larger range, a larger dataset could be used which would include more extreme temperatures.

Improvements could be made by incorporating further variables in the model and performing multi-regression. However, we found through trial and error that a single variable produced satisfactory results.

**Classification of price category using k-NN**

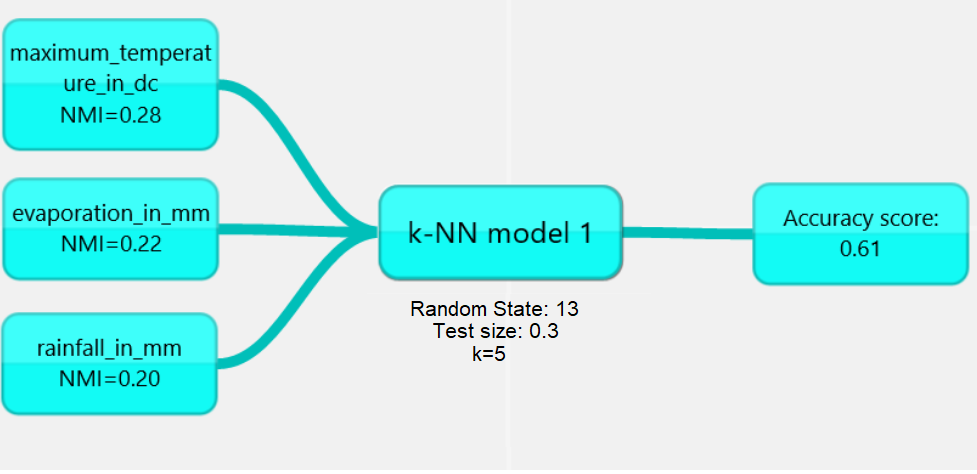


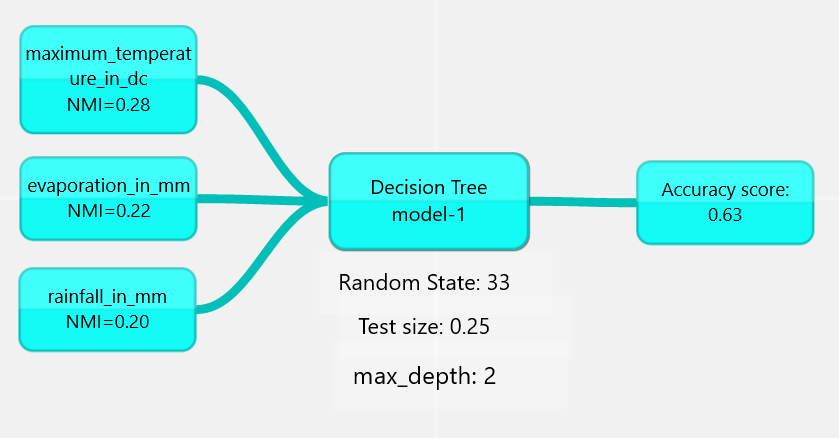
As our goal is to build a model for predicting the price category (categorical variable), the k-NN model is the best option for supporting the need for classifications with high accuracy. It is particularly useful when we have a relatively small and supervised dataset.

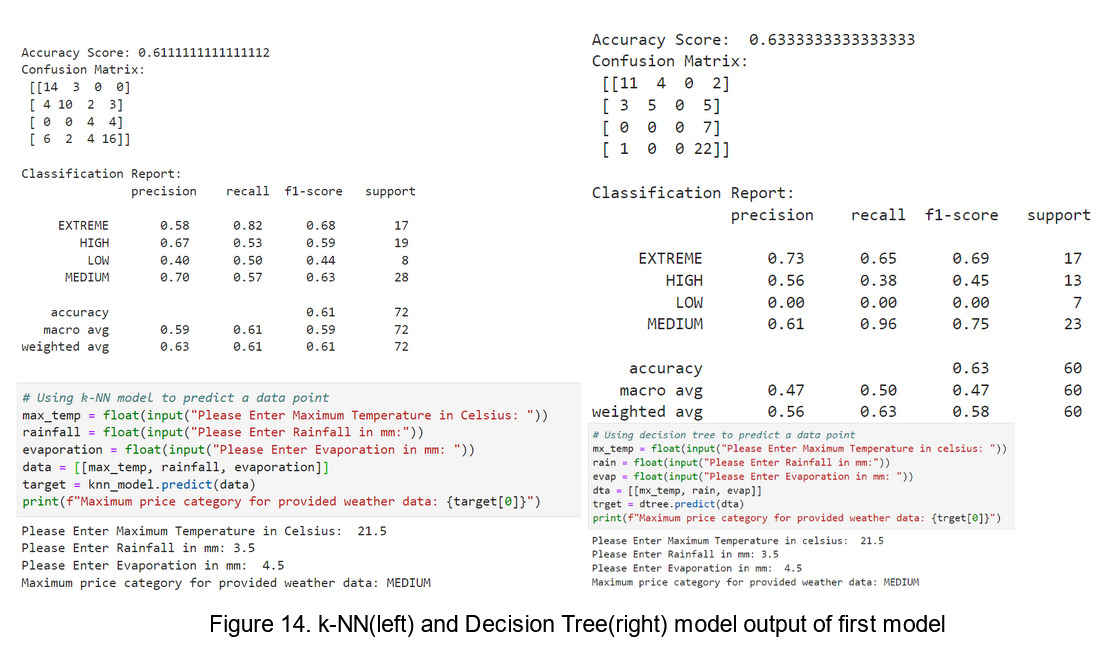
Figure 13 above depicts that maximum temperature has the highest normalized mutual information score 0.28, followed by 3pm temperature(0.27), minimum temperature(0.26) and 9 am temperature(0.24). Apart from these temperature variables, evaporation\_in\_mm(0.22), rainfall\_in\_mm(0.20), sunshine\_in\_hours(0.15) and categorical variables like 9am\_wind\_direction(0.15), 3pm\_wind\_direction(0.13) and direction\_of\_maximum\_wind\_gust(0.13) also provide some information about the target variable price\_category.

We believe the weather condition as an independent variable can be further broken down to temperature, rainfall, and evaporation. Although minimum temperature, 9am temperature and 3pm temperature have high NMI scores, they are highly correlated to maximum temperature, hence they were not considered as a feature of the initial model.

**Model 1:**







By entering the numbers into the three values, the model can generate a price category for us. In the testing case that we illustrate below, when the maximum temperature, rainfall, and evaporation are 21.5 Celsius degrees, 3.5 mm, and 4.5 mm, the model generates the Medium price range. After configuring several random state, test/train sizes and hyperparameter k values we came up with the following models:

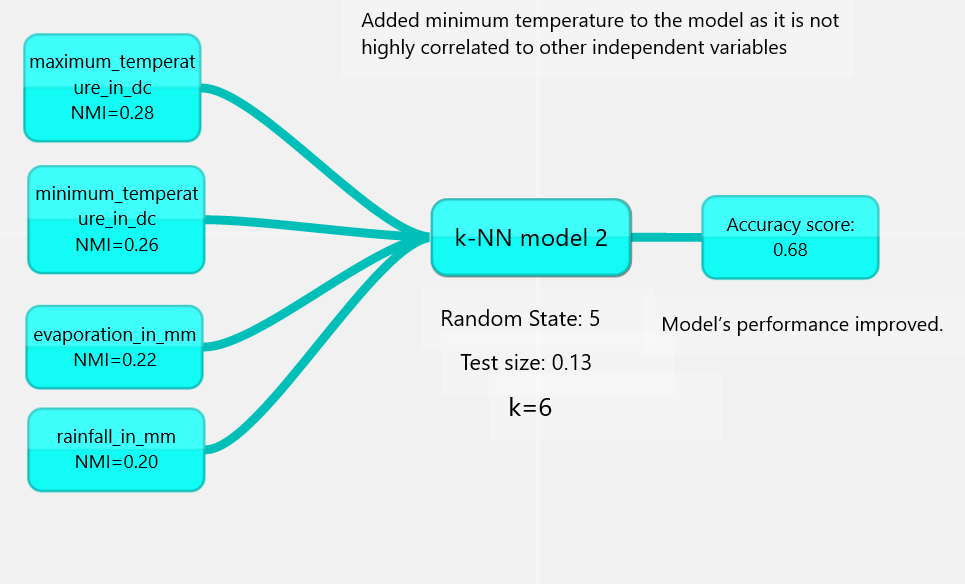
We have also used a confusion matrix and classification report to analyze our results and check the accuracy of the model for each class label.

This shows that the model is more biased in predicting EXTREME and MEDIUM values and makes more errors in predicting LOW class labels.

Although the Decision tree has a higher accuracy score, it couldn’t predict any of the LOW class labels correctly, which can be clearly seen in the confusion matrix and classification report above on the right side of figure 14. By increasing the max\_depth to 3 the accuracy score dropped below 0.5. Moreover the decision tree had inefficient results, and was unreliable in looping through various random states, test sizes and max\_depth values, with or without k fold cross validation.

Hence when building better models further, we proceeded with k-NN as our algorithm for classification.

**Model 2:**



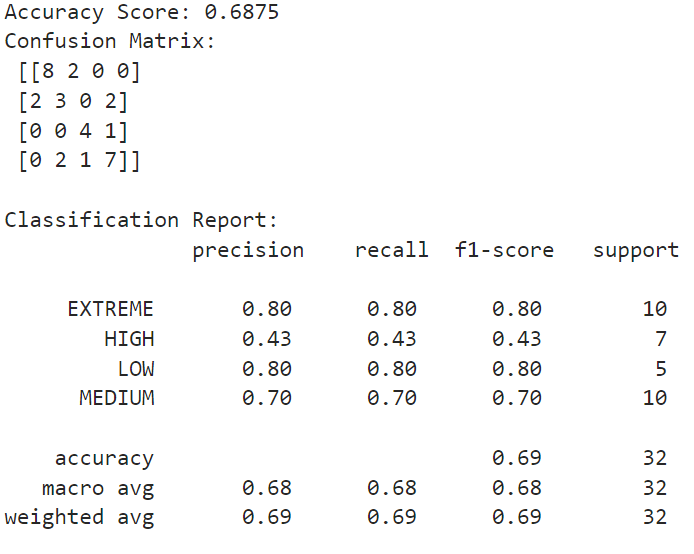
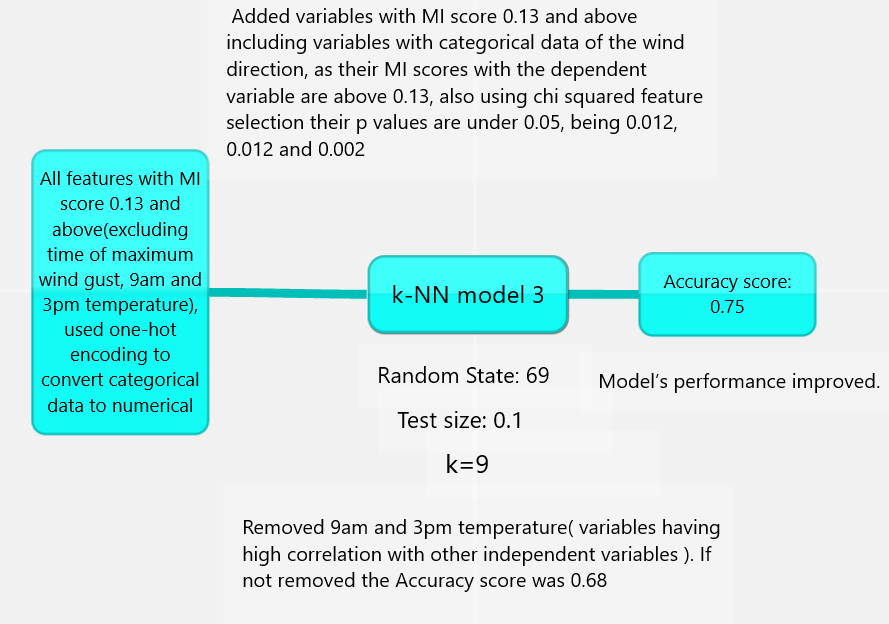
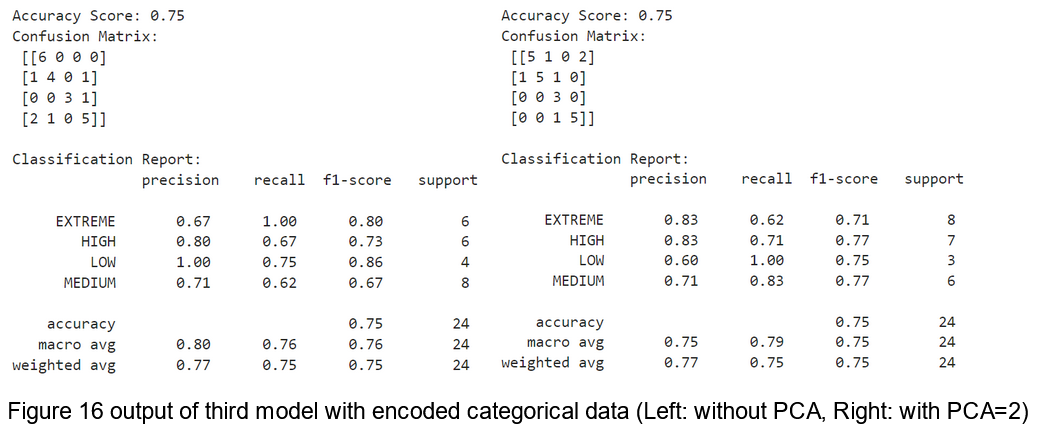


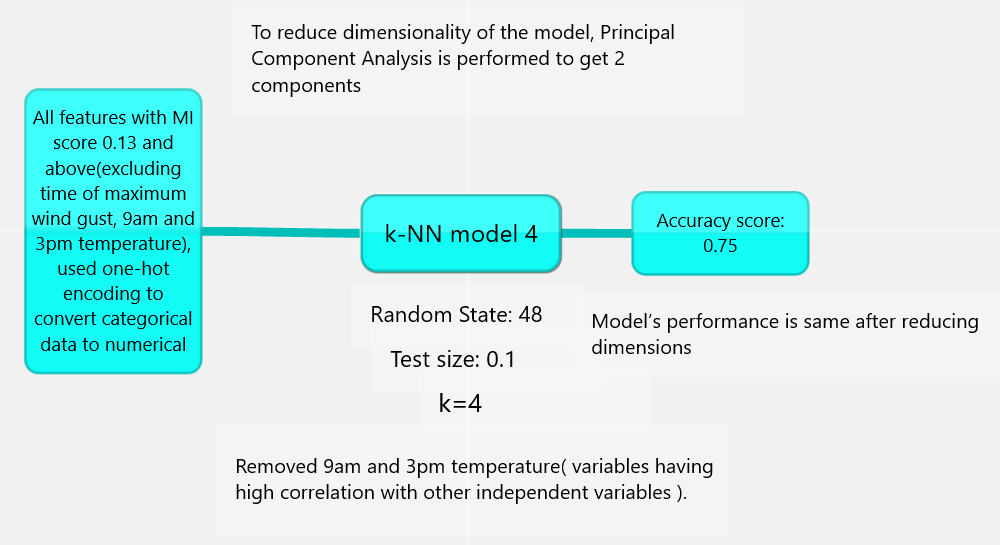
Figure 15 k-NN output of second model

**Model 3:**





**Model 4:**



To make the model better we reduced the dimensionality by reducing the number of features by performing PCA

Maximum and minimum temperature play a crucial role in our k-NN model. Nevertheless, other factors such as rainfall, evaporation and all variables with MI score 0.13 and above are equally important in predicting the price category, but compared with maximum temperature, they are functioning as contributing factors rather than determining factors.Overall the performance of k-NN models is average, data has a mediocre relationship with the target variables, data provided is not enough to reliably predict the target variable.

The testing scores show that our model has a high accuracy score of 0.75 in predicting price category in certain test cases. However all of these models were based on specific random states, test splits and k values. So we decided to cross validate the model performance. Using k fold cross validation with k=5 and k=10, the model performance is reduced between 0.47 to 0.53 for features used in k-NN model 1, 2 and 3, with and without PCA. This can be considered as an actual overall performance of the model, but our model starts an illuminating journey in exploring the connection between weather and electricity price.

Overall, the performance of both models could be described as average. The data does not have a strong relationship with the target variables. Data provided is not enough to reliably predict the target variables. Therefore, a more comprehensive dataset, over a greater timeframe could improve accuracy. For future development, random forest along with cross validation could be considered. A multilinear relationship is observed in predicting total\_demand. Hence Multilinear regression is required which is out of the scope of this report. Also a strong relationship is observed between target variables and the time of the day, such as morning around 6:30 to 9:30 am and evening 17:00 to 20:00 when the target variables are maximum. Time series analysis is out of the scope of this report and can be considered as a future prospect.

**Conclusion**

In summary, we have built five models of Linear regression, one model of Split Temperature Range Linear Regression, and four models of KNN Classification to achieve better results of predicting the maximum daily electricity demand and the price category. However, while our models show promising insights, the complexity of weather conditions tell us that our model is not 100% perfect. Collecting more data and exploring more useful algorithms are the future direction of our efforts.