**Data Analytics with Python – Assignment 2**

**Group 6 Members:**

Ma Estela Arenas

Sean Howman

Yuxiao Liu

Feng Nie

**1. Background**

It has often been observed that energy consumption tends to be at its highest on days with hotter temperatures. In this project, our group will develop models that predict the maximum daily energy usage and pricing category based on provided weather data.

**2. Purpose**

These models can be used to predict likely energy demands based on a weather forecast, which can help energy companies understand plan for future usage, and help businesses plan when to conduct energy-intensive operations.

**3. Process**

**3.1 Data Mining**

Two datasets were provided, one for weather data with 243 rows and 21 columns, which has blanks and columns with both float and string data types. Demand usage given is within the 30-minute time interval daily. The other dataset contains the price-demand data with 11,664 rows and 4 columns. Date range used in this project is between 1st of January and 31st of August 2021.

**3.2 Data Cleaning**

SQL and Python were used to wrangle and aggregate the datasets. SQL was chosen for its ease of use while Python enabled us to apply our learnings in this course.

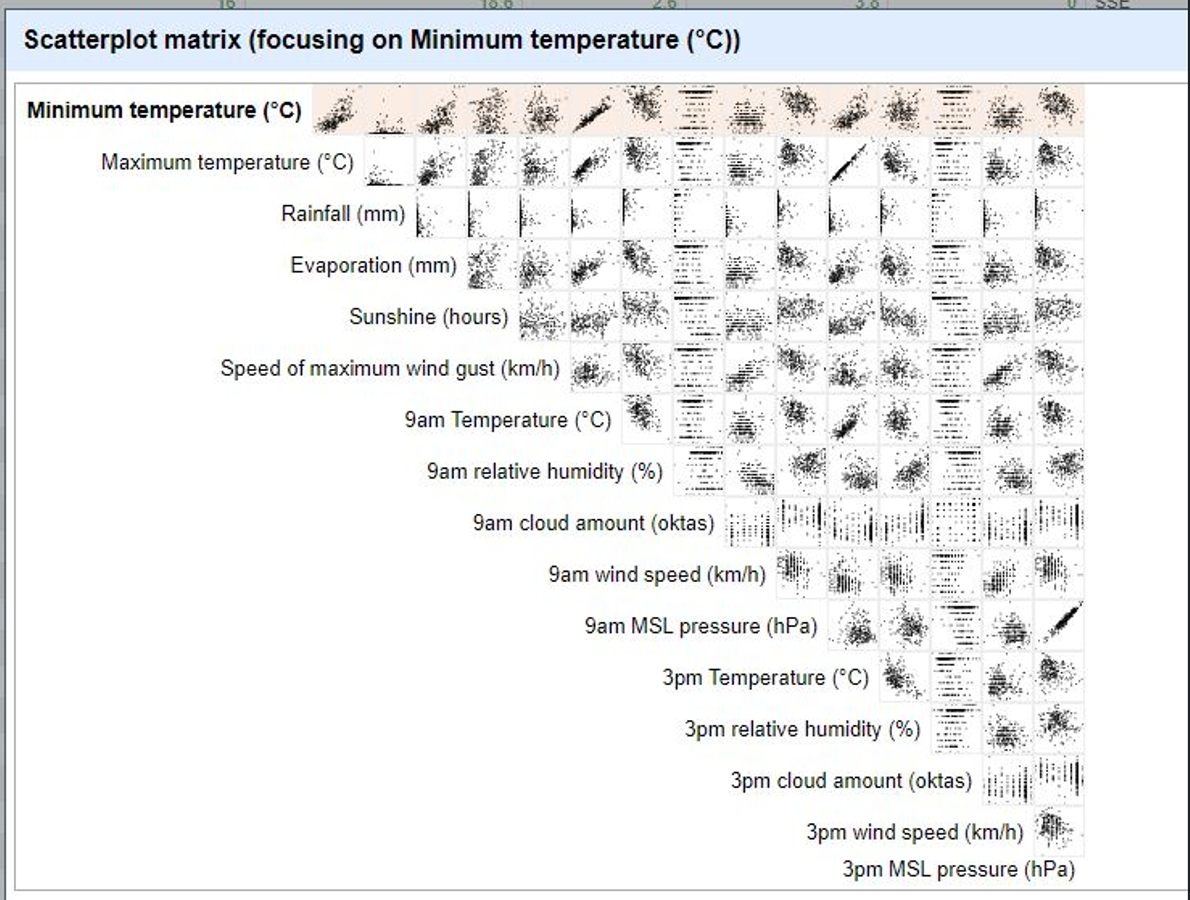
With the price-demand dataset, we’ve deleted the column for the region, grouped by date and selected the highest demand and price category. We’ve modified the date format to match the one in the weather dataset.

With the weather dataset, we’ve transformed necessary columns from value to numeric, from value to text, and filled in blanks. We’ve produced numeric facets and scatterplot facets for all numeric columns, to explore blanks, outliers and non-numeric data. Also, to highlight correlation of each feature with other features, in order to explore data and also to ascertain which features could be imputed using a simple linear relationship with other features.

After cleaning the data, we have combined the price-demand dataset and weather dataset, with the date as the common feature.

**3.3 Data Exploration**

We’ve selected features by observing scatterplots in OpenRefine, removing features that appear to be correlated with each other. We’ve also removed wind direction features as they are discrete features. Even if we change it to other data formats it will not add value to our model and by domain knowledge, we think they are unnecessary. We also chose minimum temperature as it works better than any other temperature features.



**3.4 Model Building and Prediction**

**3.4.1 Linear Regression Model**

The goal of this model is to predict the maximum daily energy usage based on provided weather data. The output is expected to be a numerical data thus we will be using linear regression to build our model.

Our assumptions for using linear regression are:

* The dependent variable is numerical.
* The independent variable is numerical.
* There is a linear relationship between the dependent and independent variables.
* There are no significant outliers.
* There is independence of observations.
* The data shows homoscedasticity, which is where the variances along the line of best fit remain similar as you move along the line

In order to create the model using linear regression algorithm, we did the following:

1. Import required libraries.
2. Load the combined data set.
3. Check the correlation between the highest demand with the numerical weather features, pick the Pearson correlation coefficient over 0.1 or lower than -0.1 ones.
4. Separate dataset into 70% train and 30% test parts.
5. Train the model and predict the result with test data.
6. Evaluate the result.

*<Below are for our notes only>*

*Linear regression fits a line to the data by finding the regression coefficient that results in the smallest MSE. The mean squared error (MSE) tells us how close a regression line is to a set of points. The lower the MSE, the better the forecast. The smaller the MSE, the closer you are to finding the line of best fit.*

*R-Squared (R² or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).*

* *A value of 1 would indicate that all of the variation is explained by our model, by the change in the independent variable (X). Therefore, there is no other error other than what was explained behind changes in X. Perfect linear relationship between X and Y: 100% of the variation in Y is explained by variation in X*
* *A value between 0 and 1 Weaker linear relationships between X and Y: Some but not all of the variation in Y is explained by variation in X*
* *A value of 0 indicates that none of the variation that we’re seeing in variable y is explainable by x, our model is not very useful at all for understanding was seen in y, no linear relationship between X and Y: The value of Y does not depend on X.*

**3.4.2 Classification Model**

The goal of this model is to predict the maximum price category based on provided weather data. The output is expected to be categorical data thus we will be using either decision tree or k nearest neighbor (KNN) to build our model. Our group has decided to try both algorithms and see later on which one of them will produce a better model.

Now that the data is totally prepared, the classifier is instantiated and the model is fit onto the data. The criterion chosen for this classifier is entropy. Once our model fits the data, we tried predicting values using the classifier model. This is often done in order to perform an unbiased evaluation and get the accuracy score of the model. We’ve done parameter tuning to select the best model.

**3.4.2.1 Decision Tree**

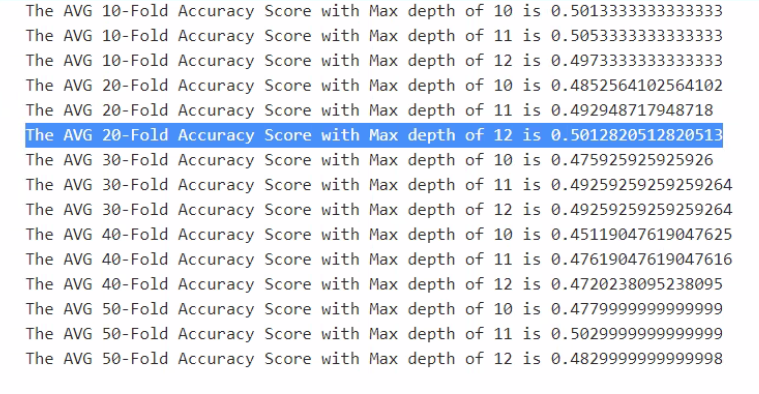
In order to create the model using the Decision Tree algorithm, we did the following:

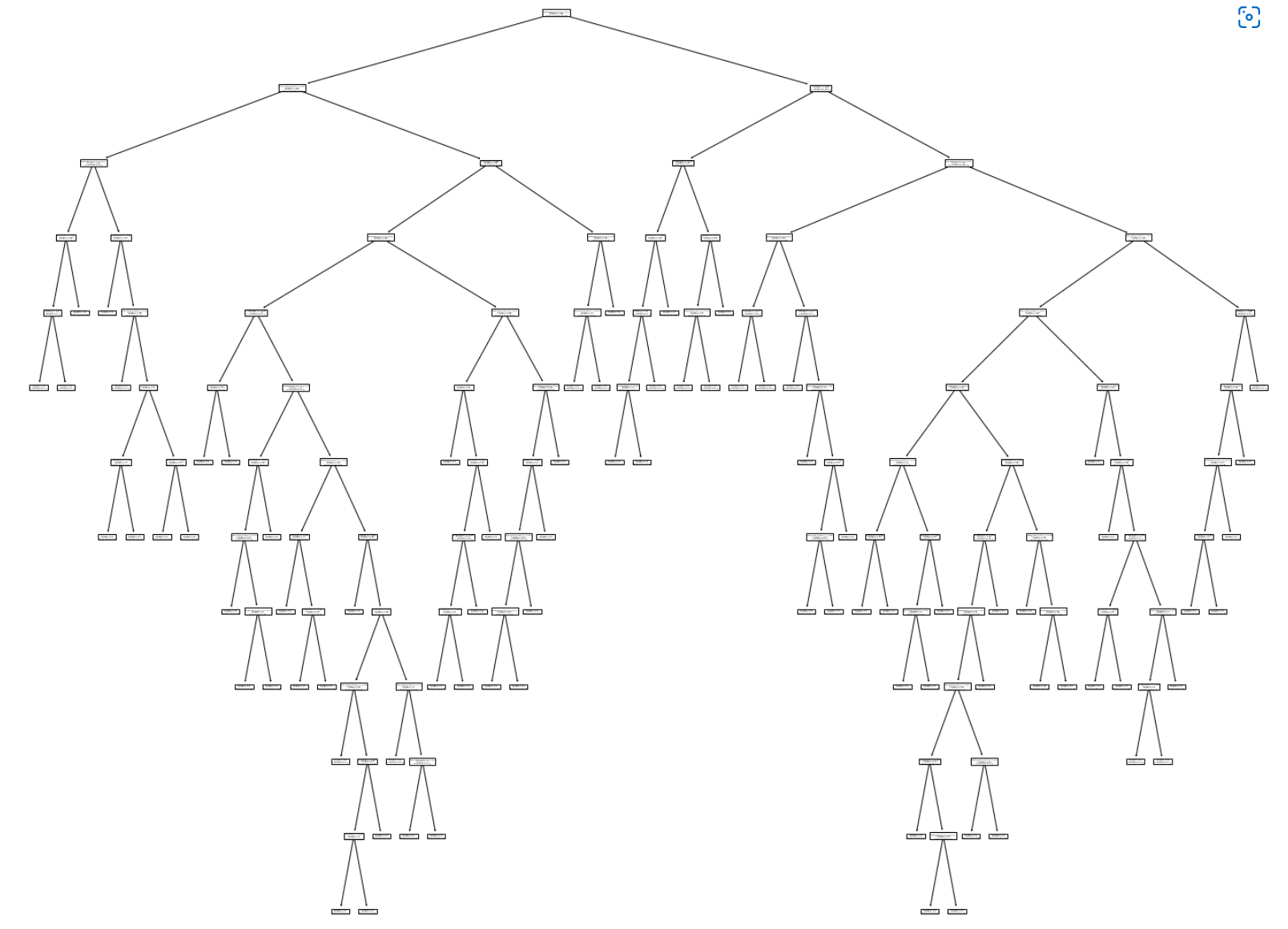
1. Import the required libraries.
2. Load the combined data set.
3. Select the features.

We’ve chosen minimum temperature, rainfall, evaporation, sunshine and max wind speed to be the features that will predict the maximum price category. These features were selected because they improve our model’s performance, they are correlated with the class label, they are dependent of the class label and they are not correlated with the other features

1. Separate dataset into 70% train and 30% test parts.
2. Train the model and predict the result with test data.

We’ve tried doing k-fold cross validation with a variety of k-folds and tree depths, with accuracy scores as shown below. We have observed that the higher the k-fold value there were fewer sample in each fold and the accuracy score is better. We have chosen the k value with the highest accuracy score. With that, We’ve used k-fold cross validation with 20 k-folds and maximum depth of 12.





1. Evaluate the result.

Based on the results of the Decision Tree algorithm, an accuracy score of approximately 50% indicates that our model can somewhat predict the maximum price category using the selected features, namely, minimum temperature, rainfall, evaporation, sunshine and max wind speed. We’ve experimented by using chi-square test to check the independence of pairs of the features but this resulted to a lower accuracy score. Using different random states, tree depths, k-folds and train/test splits will result to different accuracy scores.

**3.4.2.2 K-nearest neighbour (KNN)**

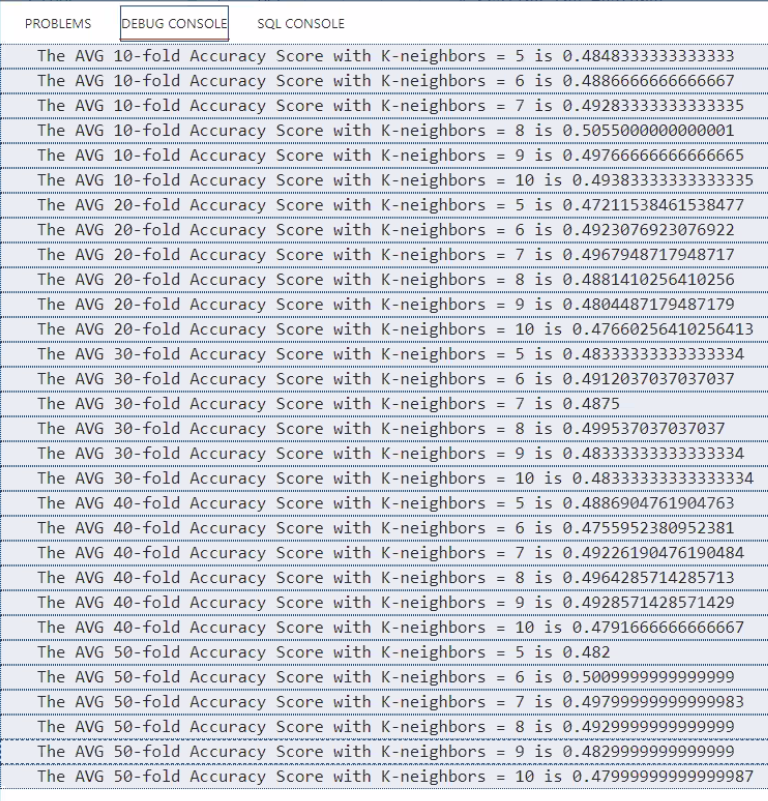
In order to create the model using the KNN algorithm, we did the following:

1. Import the required libraries.
2. Load the combined data set.
3. Select the features.

We’ve chosen minimum temperature, rainfall, evaporation, sunshine and max wind speed to be the features that will predict the maximum price category. These features were selected because they improve our model’s performance, they are correlated with the class label, they are dependent of the class label and they are not correlated with the other features.

1. Separate dataset into 70% train and 30% test parts.
2. Train the model and predict the result with test data.

We’ve tried doing k-fold cross validation with a variety of k-folds, with accuracy scores as shown below. We think that if k was too small it’s sensitive to noise points while if it’s too big the neighbourhood may include points from other classes, thus we decided to do 10-fold cross validation with a k value of 8. We have observed that the higher the k-fold value there were fewer sample in each fold and the accuracy score is better. We have chosen the k value with the highest accuracy score.



1. Evaluate the result.

Based on the results of the KNN algorithm, an accuracy score of approximately 50% indicates that our model can somewhat predict the maximum price category using the selected features, namely, minimum temperature, rainfall, evaporation, sunshine and max wind speed. We’ve experimented by using chi-square test to check the independence of pairs of the features but this resulted to a lower accuracy score. Using different random states, k neighbour values, k-folds and train/test splits will result to different accuracy scores.

**4. Result**

Given the results of our maximum demand prediction model, the underlying relationship between the selected features and demand is linear.

Based on the results of our maximum price category prediction model, we therefore conclude that a change in the selected features will result in a change in the price category. Both Decision Tree and KNN produced similar results to validate this.

The models can be improved by feature selection. Also, using different random states, k neighbour values, k-folds and train/test splits will result to different accuracy scores.

**5. Discussion**

**5.1. What wrangling and aggregation methods have you applied? Why have you chosen these methods over other alternatives?**

SQL and Python were used to wrangle and aggregate the datasets. SQL was chosen for its ease of use while Python enabled us to practice what we’ve learned in the subject.

With the price-demand dataset, we’ve deleted the column for the region, grouped by date and picked the highest demand and price category. We’ve modified the date format to match the one in the weather dataset. We have selected the daily maximum demand and price category.

With the weather dataset, we’ve transformed necessary columns from value to numeric, from value to text, and filled in blanks. We’ve produced numeric facets and scatterplot facets for all numeric columns, to explore blanks, outliers and non-numeric data. Also, to highlight correlation of each feature with other features, in order to explore data and also to ascertain which features could be imputed using a simple linear relationship with other features.

After cleaning the data, we have combined the price-demand dataset and weather dataset, with the date as the common feature.

**5.2. How have you gone about building your models and how do your models work?**

**5.3. How effective are your models? How have you evaluated this?**

**5.4. What insights can you draw from your analysis? For example, which input variables are most valuable for predicting energy usage/price?**

**5.5. Why are your results significant and valuable?**

**5.6. What are the limitations of your results and how can the project be improved for future?**

The models can be improved by feature selection. Also, using different random states, k neighbour values, k-folds and train/test splits will result to different accuracy scores.

**GROUP NOTES (Please remove this portion before submitting it)**

Kindly double check the items highlighted in yellow, please update as needed.

I think we can get rid of part 5 – Discussions as it has already been discussed in parts 1 – 4.