**Group 4 Assignment 2 Report**

BY

Sining Bao

Benjamin Wang

Pamela Yang

DATA0006

The University of Melbourne

10 December 2023

ACKNOWLEDGEMENT

We express our sincere gratitude to Dr. Geela Chee for her invaluable guidance and educational support in the realm of Python data analytics over the past 12 weeks. Additionally, we extend our appreciation to Mr. Daniel Pham for his facilitation of workshops and his insightful responses to our queries. Their expertise and commitment greatly contributed to our learning experience.

Table of Content

1. \*\*Executive Summary:\*\*

- Brief overview of the analysis, key findings, and recommendations.

- Targeted at non-technical stakeholders.

2. \*\*Introduction:\*\*

- Background information on the purpose of the analysis.

- Definition of the problem or question being addressed.

3. \*\*Data Collection:\*\*

- Description of the dataset(s) used.

- Data sources, collection methods, and any data preprocessing steps.

4. \*\*Exploratory Data Analysis (EDA):\*\*

- Summary statistics, distributions, and visualizations.

- Identification of patterns, trends, outliers, and correlations.

5. \*\*Data Cleaning and Preprocessing:\*\*

- Handling missing values, outliers, and duplicates.

- Feature engineering and transformation.

- Changing Data Types:

* Converted data types of columns using pd.to\_datetime() and astype().
* Converted “Calm” to “0” in the column of wind direction.

Merge the data:

Use the “left on” to merge two tables to one.

Grouping and Aggregating:

Grouped data by specific columns and aggregated information using functions like groupby().

Feature Selection:

Selected specific columns or features for analysis.

Data Visualization:

Used matplotlib for scatter plots and other visualizations.

Machine Learning:

Utilized machine learning models such as KNN, DecisionTreeRegression for predictive modeling.

Error Handling:

Addressed errors related to DataFrame operations and model evaluations.

Binning:

By dividing continuous data into several discrete intervals bins, the complexity of the data could be simplified.

The choice of methods may vary depending on the specific requirement and characteristics of the dataset, general knowledge(“Calm”), etc. Understanding the problem context and explore different approaches will help find the most suitable solution.

Assessment Q1: What data cleaning methods have you applied? Why have you chosen these methods

over other alternatives? Give examples to support your chosen methods.

6. \*\*Methodology:\*\*

- Explanation of the analysis approach and techniques used.

- Justification for the chosen methods.

Assessment Q2:Explain the process of building your best model. How did you select the features included in the model? How does your model work?

**Define the Problem:**

Review the dataset, understand the type of task, define it's classification, regression, clustering, etc.

**Data Import:**

Import the dataset for data cleaning and preprocessing.

**Data Cleaning and Preprocessing:**

Handle missing values, outliers, and any data quality issues.

Preprocess the data, which may include normalization, encoding categorical variables, etc.

**Visualize data:**

Understand the characteristics of your data through visualization and statistical analysis.

**Feature Selection:**

Identify and select relevant features that contribute to the predictive power of the model.

Use techniques like correlation analysis, recursive feature elimination, or domain knowledge to guide feature selection.

**Split the Data:**

Split the dataset into training and testing sets.

**Iterative Refinement to find a better model:**

Run the machine learning algorithms based on the dataset build. If the model performance is not satisfactory, revisit earlier steps. Adjust features, try different models, or collect more data.

7. \*\*Analysis and Results:\*\*

- Detailed presentation of analysis results.

- Statistical tests, machine learning models, or any other relevant methods.

- Interpretation of results and insights gained.

Assessment Q3: How effective is your model? How have you evaluated this?

Assessment Q4: What insights about weather and/or daily energy usage can you draw from your analysis? Discuss any significant results.

Depending on the graphics:

1. The correlation between temperature and energy demand, especially the maximum temperature.
2. The humidity influences energy demands.
3. With the season of the dataset is for summer of Victoria (Nov 2022 to Apr 2023), the rainfall and windspeed don’t impact the energy a lot.

8. \*\*Discussion:\*\*

- Interpretation of findings in the context of the problem.

- Addressing limitations and potential sources of bias.

- Comparisons with existing literature or benchmarks.

Assessment Q5: What are the limitations of your results?

9. \*\*Conclusion:\*\*

- Summary of key findings and their implications.

- Recommendations for further research or actions.

10. \*\*Visualizations and Tables:\*\*

- Well-labeled charts, graphs, and tables that support the analysis.

- Captions and explanations for each visualization.

11. \*\*Appendix:\*\*

- Additional details, code snippets, or supplementary analyses.

- Any technical information not included in the main report.

12. \*\*References:\*\*

- Citations for datasets, methodologies, or external sources.

## Executive Summary

## Introduction

## Data Collection

This study incorporates two datasets for analysis:

* weather.csv

This file encompasses twenty essential weather indicators recorded daily for the city of Melbourne spanning the period from November 2022 to April 2023. The data has been meticulously extracted from the Bureau of Meteorology.

* Price\_and\_demand.csv

Within this file, one can find energy price and demand data for the state of Victoria, captured at a half-hour interval throughout the timeframe from November 2022 to April 2023. The information has been sourced from the Australian Energy Market Operator, ensuring reliability and accuracy in the dataset.

## Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted primarily through Multivariate Graphical Analysis based upon background hypotheses regarding energy consumption.

The primary hypothesis was based on the background that most industries’ primary drivers for energy demand were Heating, Ventilation, and Air Conditioning (HVAC) equipment. HVAC demand, in turn, is hypothesized to be driven by temperature variations that fall outside typically ‘comfortable’ area.

## Data Cleaning and Preprocessing

The predominant purpose of data cleaning the input data sets has been to ensure that data can be consistently passed through relevant models and processed for predictive purposes. Concurrently, data cleaning and preprocessing were also conducted with the aim of reducing any potential bias caused by excessive processes.

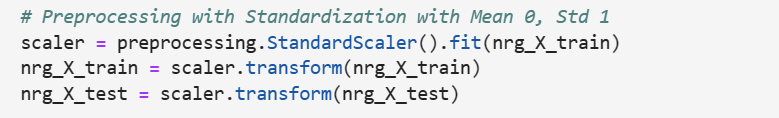
Data Cleaning: Deletion of Irrelevant or Missing Data

The preliminary data cleaning method employed with respect to the “Weather” input data has been the removal of missing data and irrelevant data.

Several features were noted to be completely empty in the original csv file, making their value to predictive modeling functionally irrelevant. As such, deletion has been applied to entirely remove the “Evaporation (mm)”, “Sunshine (mm)” and “9am cloud amount (oktas)” features. This has been completed through the *.dropna* method for features or columns with entirely empty datasets.

Subsequently, listwise deletion has also been applied to certain data points with predominantly incomplete or missing data to ensure that complete variable data sets were analyzed for correlation purposes. This, most notably, has been applied to the weather data point pertaining to April 24, 2024, in which 13 of 22 features were empty. The *.isnull* method has been applied to remove data points where a desired feature may return an empty value. As such, it only applies to empty data points when a desired feature has been selected.

##### Pre-Processing: Scalers

Additional preprocessing methods were utilized for regression analysis through the implementation of scalers to normalize the datasets across a mean of 0 and a standard deviation of 1 to ensure that the data was not distorted by units of measurement and scale.

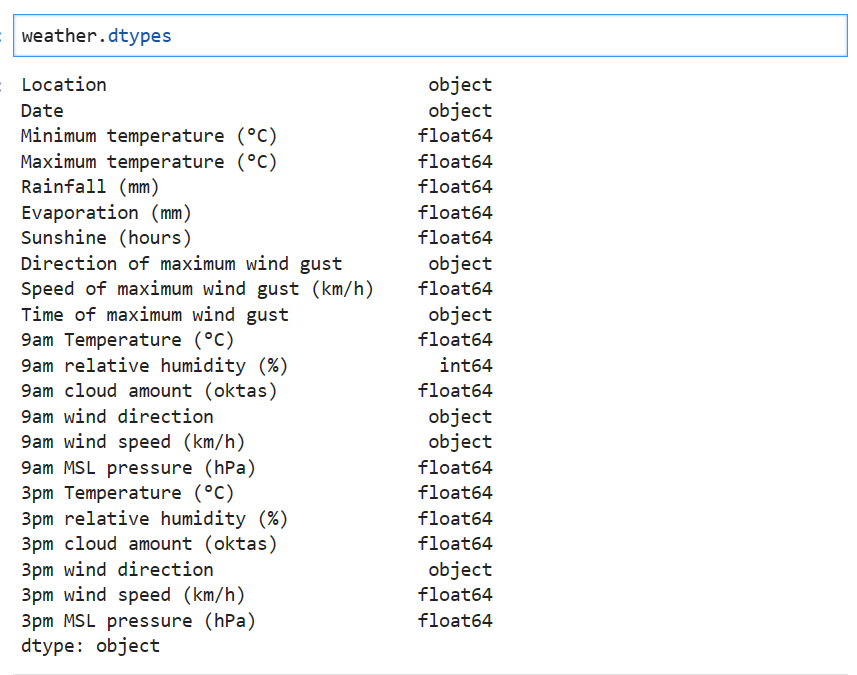
##### Pre-Processing: Imputation

The removal of missing or irrelevant data was completed only in situations where the majority of its relevant values were missing. This was done so in preference over the application of mean imputation to minimize the potential for distortion to the standard deviation.

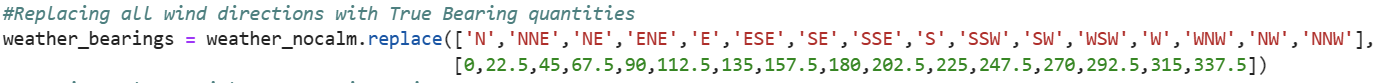
Functionally, for the weather dataset, imputation was essentially irrelevant as listwise deletion removed all empty data points.

Data Cleaning: Data Type Conversion

Additional steps were taken to ensure all values were numerical float values. This process has been conducted for the purpose of ensuring that all values are suitable for analysis and modeling. As most features were already float values (see below), a viable data type for predictive analysis and modeling, the remaining ‘object’ variables could easily be utilized for future modeling.



All relevant features were also converted to numerical quantities for the same purpose. An example of this is the conversion of “Direction of maximum wind gust” to its respective True Bearing Angle (see below).



This was done to further ensure that numerical data could be effectively passed through any predictive model based on numerical inputs.

For the Energy Pricing and Demand dataset, it was deemed that minimal data cleaning and processing as required

## Methodology

- Explanation of the analysis approach and techniques used.

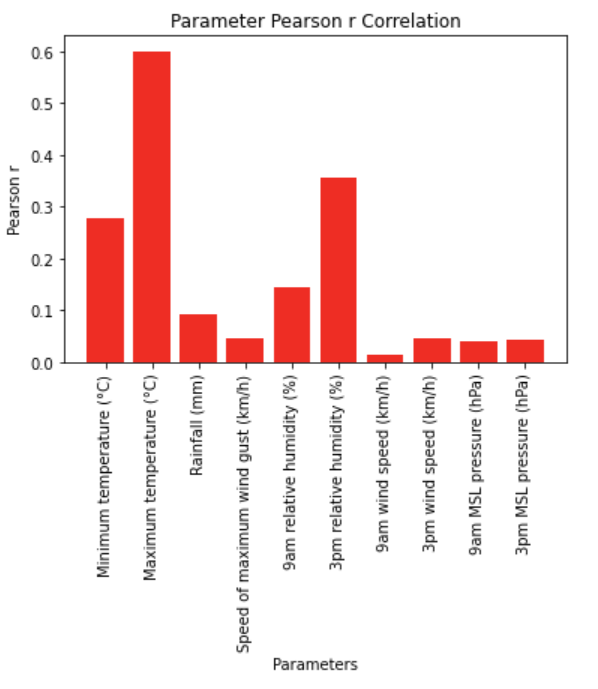
- Justification for the chosen methods.

Assessment Q2:Explain the process of building your best model. How did you select the features included in the model? How does your model work?

In order to develop a model which effectively predict maximum daily energy use, This study applies Linear Regression, Decision Tree Regression and K Nearest Neighbour algorithm to analyze, train and test the data.

### Pearson *r*

The Pearson *r* coefficient was calculated for all features with respect to Maximum Energy Demand. This enabled the identification of key variables with the greatest degree of correlation for predicting and modeling

.

As illustrated (see above), “Maximum Temperature” was noted to be the most linearly correlated with Maximum Energy Demand. For the purposes of Regression Modelling, the Pearson Correlation coefficient, therefore, also served as a prioritizer of independent / predictor variables to be passed through each regression model.

### 4.1 Linear Regression Analysis

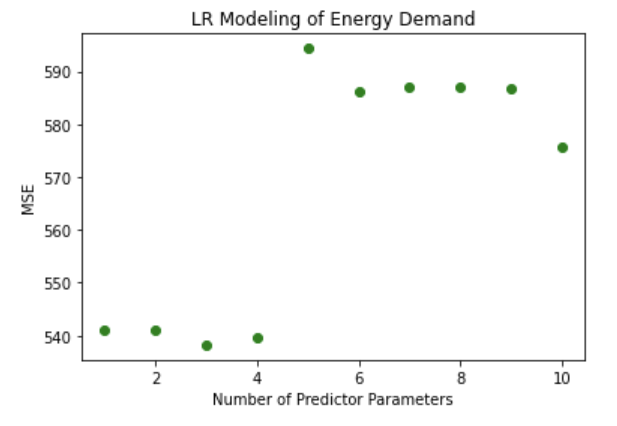
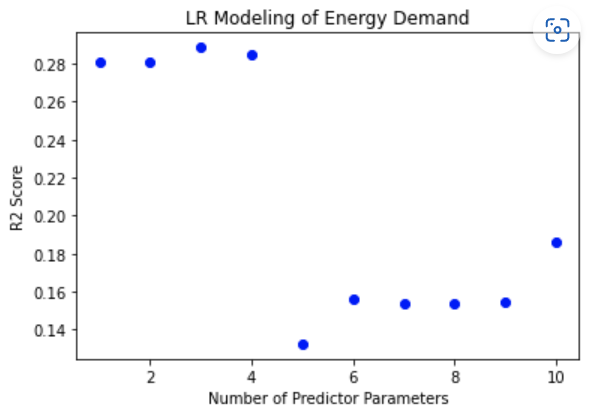
In light of the continuous nature of the target feature, which pertains to the maximum energy demand, a comprehensive analysis employing Multiple Linear Regression has been conducted. The primary objective is to discern and comprehend the linear relationship between the independent variables (weather indicators) and the dependent variable (maximum energy demand).

To ensure a robust evaluation, the dataset has been meticulously divided into two distinct sets: an 80% training set and a 20% test set. Subsequent to model fitting and prediction, the efficacy of the linear regression model has been meticulously assessed using key performance metrics, namely R-squared and Root Mean Squared Error. These metrics serve as pivotal indicators for gauging the effectiveness and appropriateness of the linear regression modeling approach in the context of this study.

#### Number of Predictor Variables vs Linear Regression R2 and MSE

Based on the ranking of ten predictor variables from highest Pearson *r* coefficients to lowest, the *r2* score of each Linear Regression model was plotted against the number of predictor variables provided.

For Linear Regression Modeling of Energy Demand, it can be seen that utilizing the following four independent variables yielded the highest available Linear Regression *r2* score (see below).

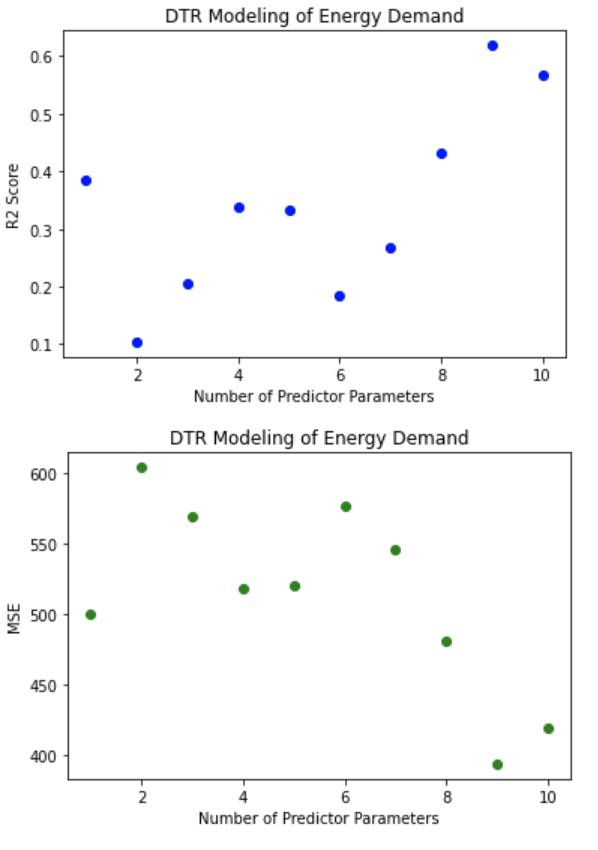
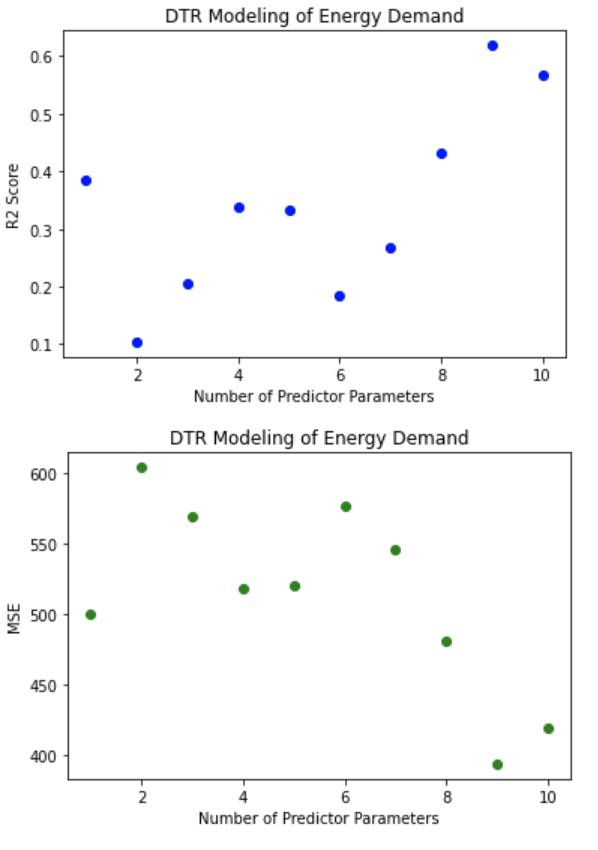


### 4.2 Decision Tree Regression Analysis

#### Number of Predictor Variables vs Decision Tree Regression R2 and MSE

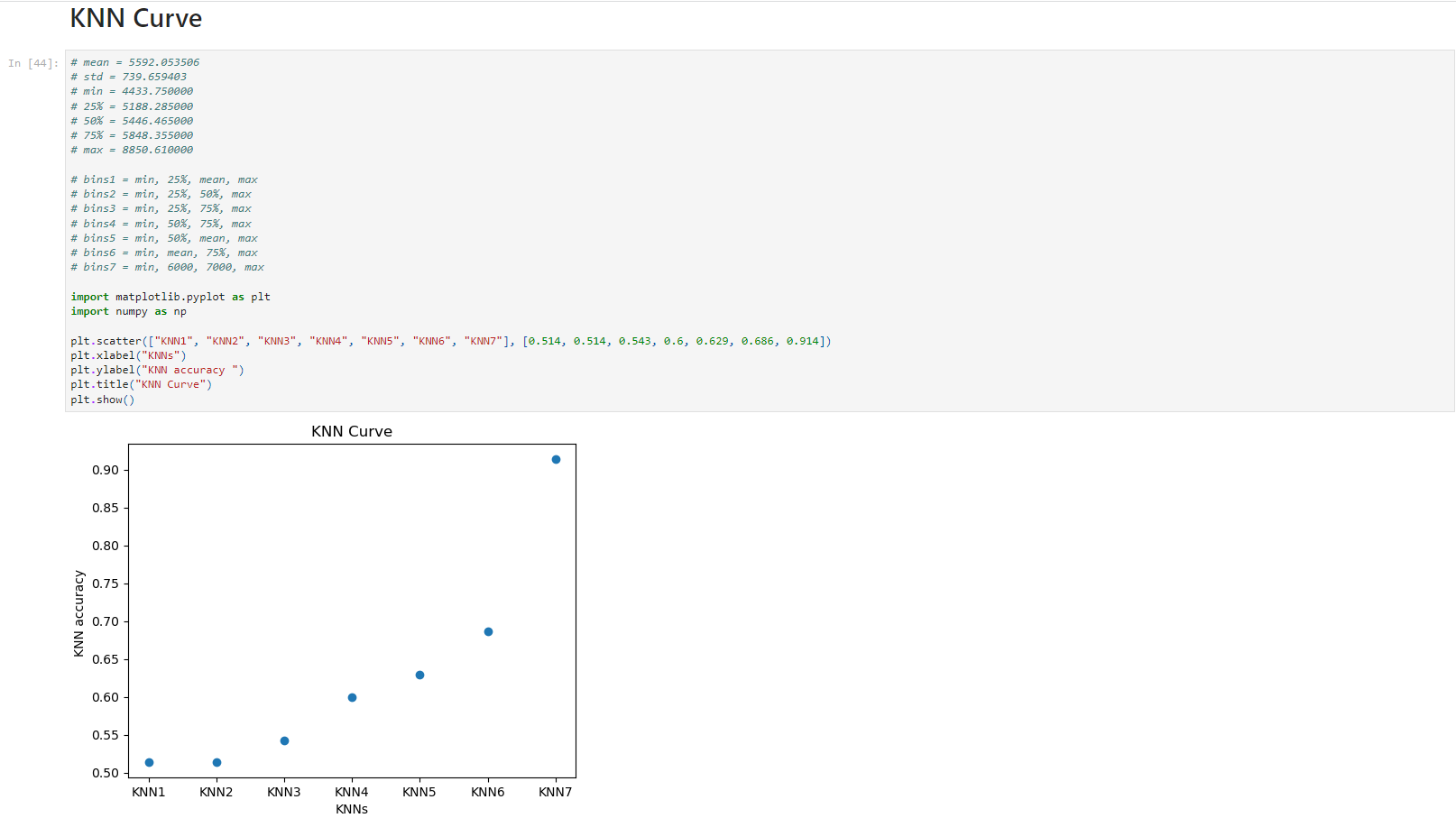
Similarly to the approach conducted with Linear Regression, the *r2* score of each Decision Tree Regression model was plotted against the number of predictor variables inputted.

Unlike Linear Regression, however, DTR Modelling appeared to increase in R2 score with a greater quantity of predictor variables. As such, it was deemed that the best approach to Decision Tree Regression was to utilize all ten selected variables.

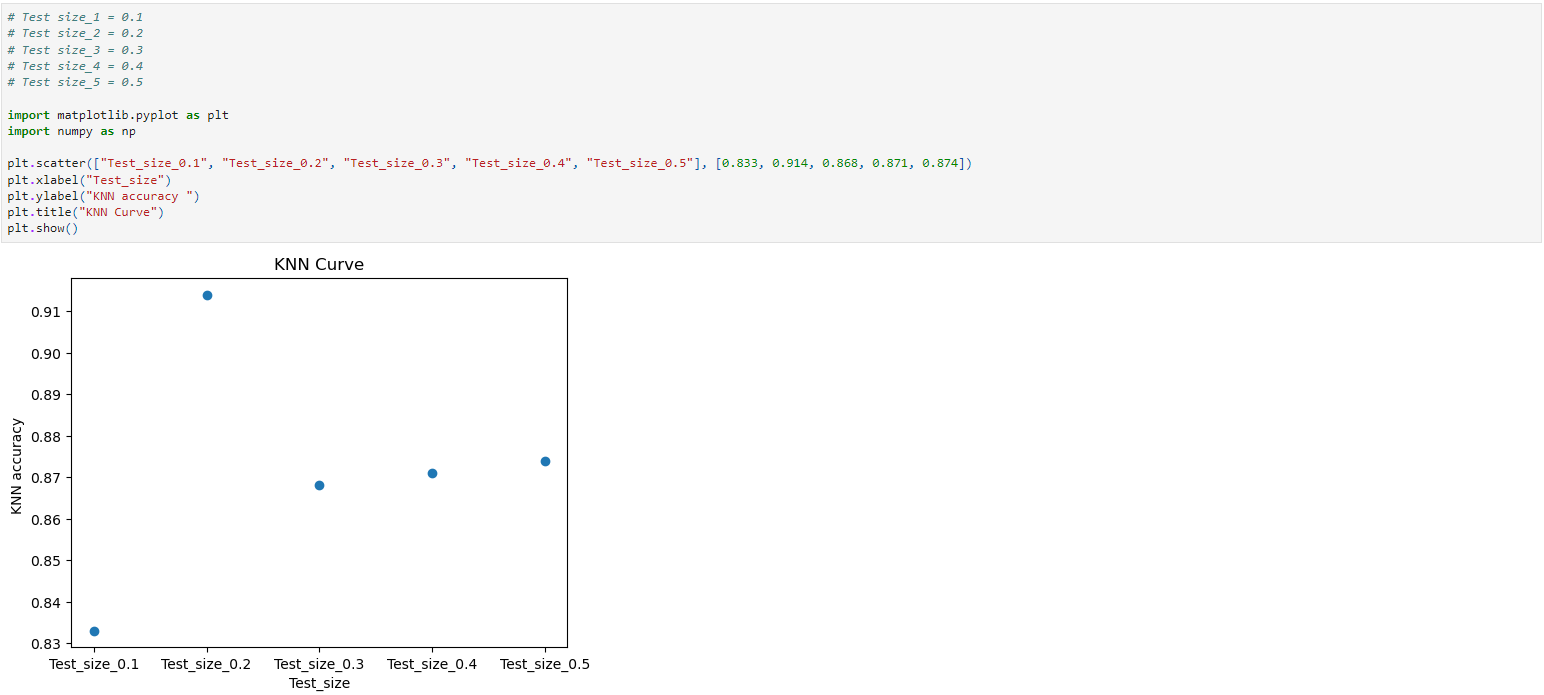


### 4.3 K Nearest Neighbour Analysis

The numerical variations in KNN accuracy are related to the selected range of categories. When attempting to change different numerical intervals, different KNN accuracy values can be obtained.



During prediction, the variation in test size also affects KNN accuracy. Under the same category, changing the size of the test set yields the following data.



The magnitude of KNN accuracy is also influenced by the features involved in prediction. Adding or removing relevant features will affect the numerical value of KNN accuracy. The difference lies in the fact that features with higher correlation will have a more significant impact, while features with lower correlation will have a relatively smaller impact.

## Analysis and Results

## Discussion

## Conclusion

The maximum energy demand is a continuous dataset, and theoretically, using regression algorithms would have better results. However, we utilized both discrete and continuous algorithms to predict the maximum energy demand. Surprisingly, the results showed that the discrete algorithm achieved better accuracy. My understanding of this outcome is that the data obtained contains a significant amount of noise. The missing columns in the data have a substantial impact on the final results, leading to these findings.

## Visualizations and Tables

Visualization and tables play a very intuitive role in data analysis. Observing the relationship between two sets of data through graphical representations allows for a preliminary assessment of the correlation between known and predicted data.

## Appendix

## References