



Energy Consumption Prediction Using Data Mining Models

AIDI 2005-01: Capstone Term II

Submitted by: **Group 7**

- ❑ Bavithra Ganesan (100900119)
- ❑ Jasmeet Kaur (100881373)
- ❑ Pritesh Dalal (100872247)
- ❑ Rutvik Shah (100886648)

Problem statement

This study aims to investigate the viability of applying data mining algorithms for energy consumption forecasting in industrial settings, specifically in the context of a South Korean steel manufacturing facility. The study uses cutting-edge methods including machine learning algorithms and artificial neural networks to forecast daily energy use patterns. The main objective of this study is to assess the precision and efficacy of these models in predicting energy consumption and to find possible areas where energy may be saved, operational efficiency can be increased, and useful insights can be provided for practical applications



Dataset Description

- ❑ **The data was obtained from the South Korean company DAEWOO steel.co. Ltd in Gwanyang.**
- ❑ **It consists of :**
 - Date feature
 - seven numerical attributes: Industry energy consumption (target variable), lagging current power factor, Lagging current reactive factor, Leading current reactive factor, Leading current power factor, Number of seconds from midnight, CO2
 - Three categorical attributes like Week status, Day of Week, and Load Type.



Exploratory Data Analysis

❑ **Chart 1** shows that 'Light_Load' has highest frequency (51.58%) among different types of loads, followed by 'Medium_Load' (27.67%) and 'Maximum_Load' (20.75%) has lowest frequency.

❑ **Chart 2** : The first plot shows that weekdays have higher frequency than weekends. The second plot shows that: 'Light_Load' frequency is high on weekends and constant for weekdays and 'Medium_Load' frequency and 'Maximum_Load' frequency are low on weekends (lowest on Sundays) and constant for weekdays.

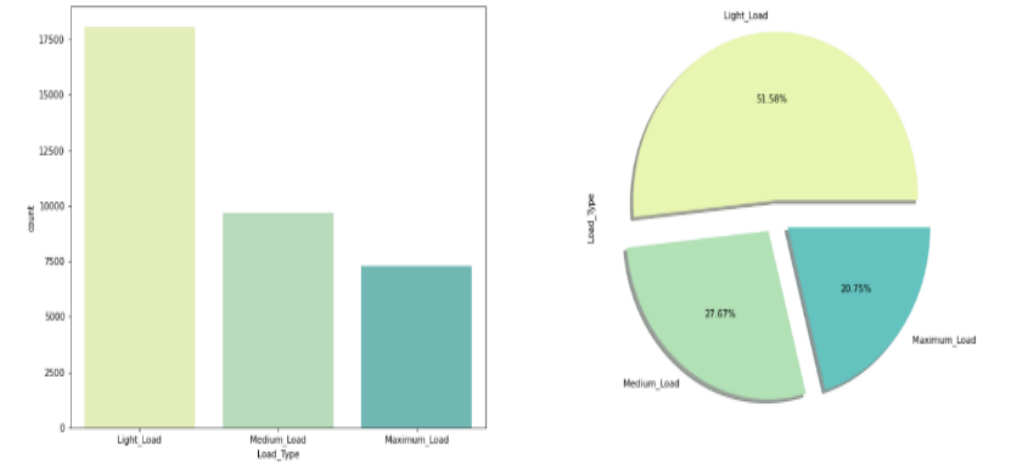


Chart 1

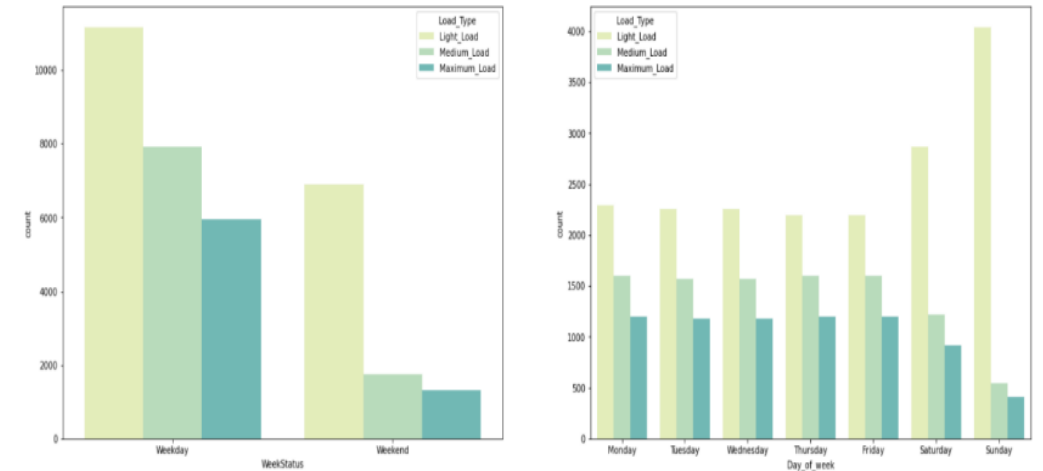
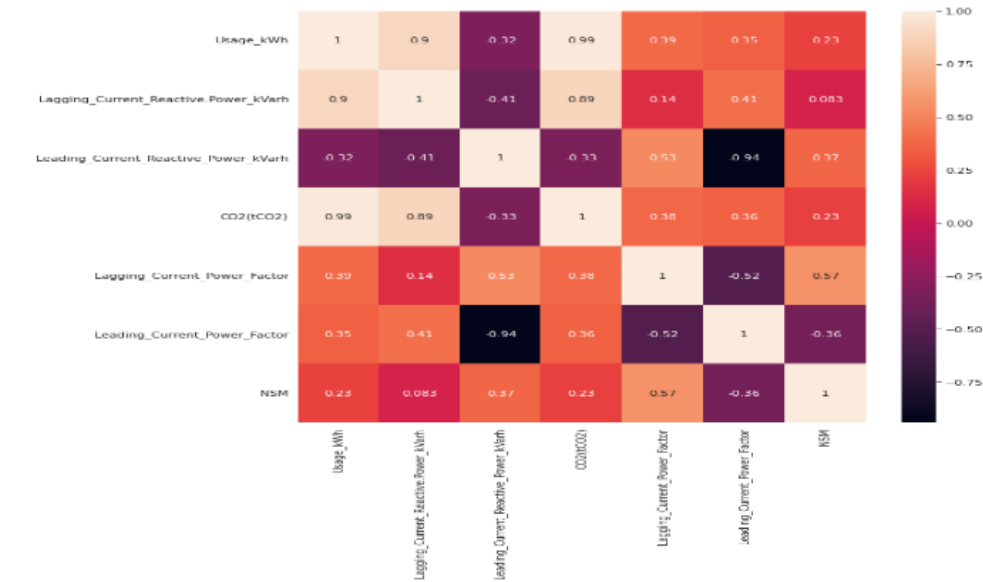


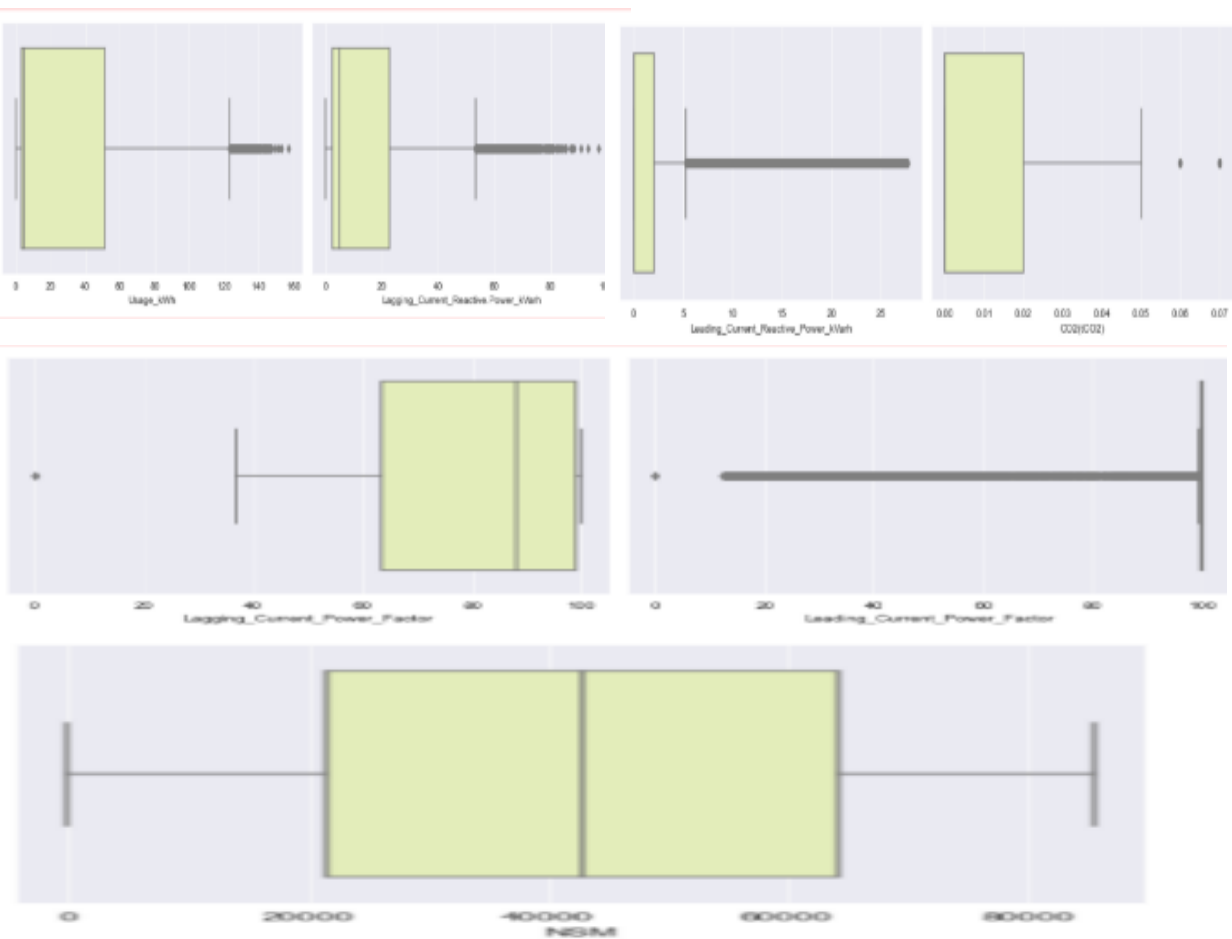
Chart 2

Chart 3



From the **correlation graph**, it can be observed that lagging current reactive power is directly proportional to usage, leading to greater energy consumption. This usage is also highly correlated with tCO2, resulting in increased emission of CO2 that is detrimental to the environment. Receiving more power from the source, lagging reactive power has a high correlation with tCO2, thus causing a rise in CO2 emission.


Chart 4



In the above **box plots**, it can see that there are outliers in all features except NSM.

Outlier Detection and Management

Interquartile Range (IQR) is a method for identifying outliers by comparing the upper and lower quartiles of a dataset. Data points that fall outside of a certain range are considered outliers.



MICE imputation is an effective method for dealing with missing data, including outliers. It can capture complex relationships between variables and produce accurate imputations for missing data by employing a series of regression models.



Iterative Imputer is used as a method for reducing disparities in a dataset.



Data Preparation

Feature Engineering

- **Label Encoder** is used for the process of converting categorical variables into distinct integer values.
- It will enable the models to capture patterns and relationships in data, thereby improving their predictive power.

Feature Scaling

- In order to standardize and convert numerical data so that it has a mean of zero and a standard deviation of one, a **standard scalar** is a data preprocessing technique.
- To increase the efficiency and precision of machine learning algorithms, this method is frequently utilized.

Feature Selection

- **SelectKBest** is a machine learning feature selection strategy that chooses the K-best features from a dataset using statistical testing.
- Each feature is analyzed and given a score, and the K characteristics with the highest scores are chosen as the model's most pertinent features.
- This method aids in decreasing the dataset's dimensionality and enhances the functionality of machine learning models.



Model Training and Evaluation

Linear Regressor

- ❑ The link between a dependent variable and one or more independent variables can be modelled statistically using linear regression.

	MODEL TYPE	DATASET TYPE	R-2 SCORE	MEAN SQUARED ERROR	MEAN ABSOLUTE ERROR
0	Linear Regression MODEL after HYPER PARAMETER ...	Training Dataset	0.981203	19.261986	2.481332
1	Linear Regression MODEL after HYPER PARAMETER ...	Test Dataset	0.984596	16.070619	2.438943

- ❑ Though R2 score is greater than 0.98 for both model's training and testing dataset which indicates strong relationship between the independent and dependent variables in the model , MSE values is high for both phases. We conclude that this model is not suitable for our dataset .

Random Forest Regressor

Random Forest Regressor is a well-known machine learning algorithm that combines several decision trees to form a powerful regression model capable of making accurate predictions on a wide range of data.

	MODEL TYPE	DATASET TYPE	R-2 SCORE	MEAN SQUARED ERROR	MEAN ABSOLUTE ERROR
0	RFR MODEL after HYPER PARAMETER OPTIMIZATION	Training Dataset	0.992581	7.602452	1.596773
1	RFR MODEL after HYPER PARAMETER OPTIMIZATION	Test Dataset	0.990845	9.550621	1.656967

We can observe that R2 scores greater than 0.99 for both models' training and test datasets indicate a strong relationship between the independent and dependent variables in the model as well as MSE score which is comparatively less than Linear Regressor model , also the difference between training and testing is less makes it a suitable model.



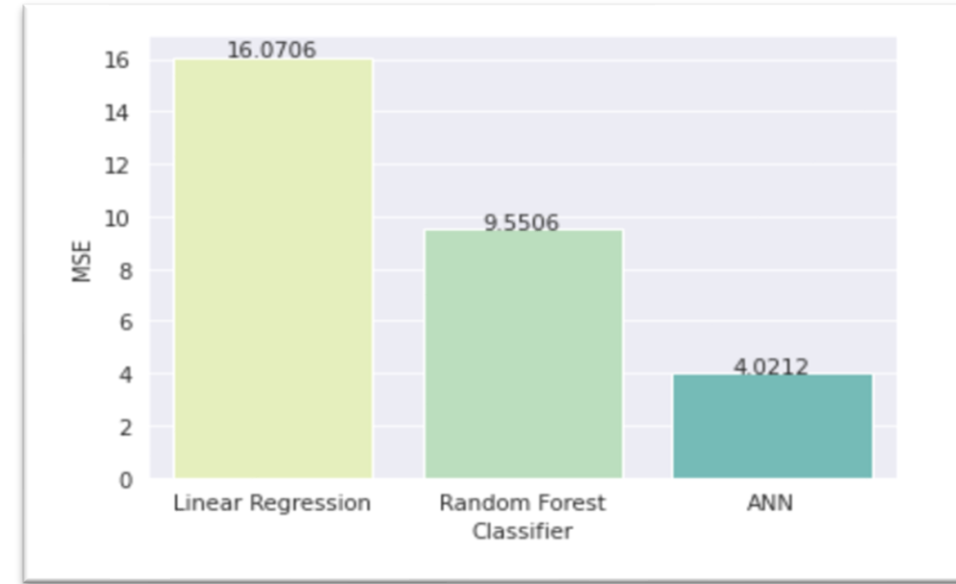
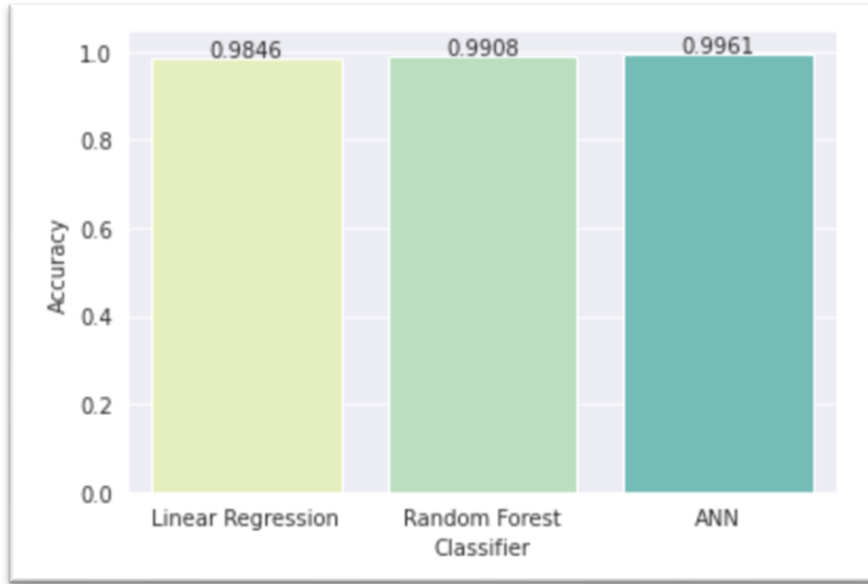
Artificial Neural Network using Regression

An artificial neural network (ANN) is a particular kind of machine learning model that takes its cues from the structure and operation of the human brain. It is made up of a network of artificial neurons or interconnected nodes that can process and transfer information.

	MODEL TYPE	DATASET TYPE	R-2 SCORE	MEAN SQUARED ERROR	MEAN ABSOLUTE ERROR
0	ANN MODEL after HYPER PARAMETER OPTIMIZATION	Training Dataset	0.996849	3.734037	0.954219
1	ANN MODEL after HYPER PARAMETER OPTIMIZATION	Test Dataset	0.995509	5.190266	0.960304

We can conclude from the above table R2 score is 0.99 for both training and testing , also MSE score is lesser than the other model which makes it an ideal choice for energy consumption prediction.

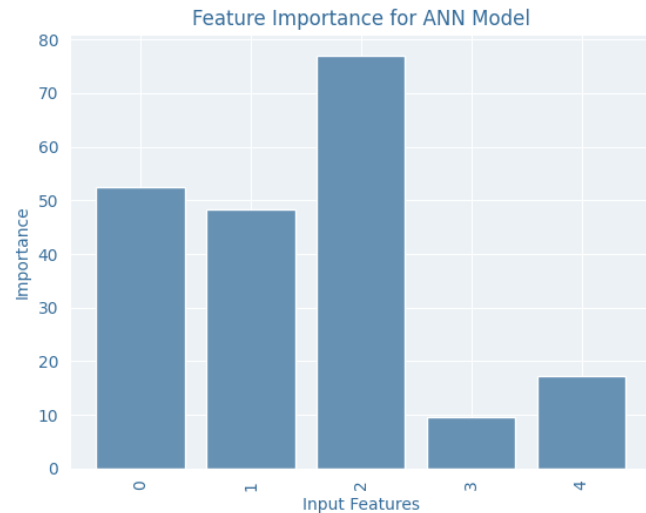
Model Comparison



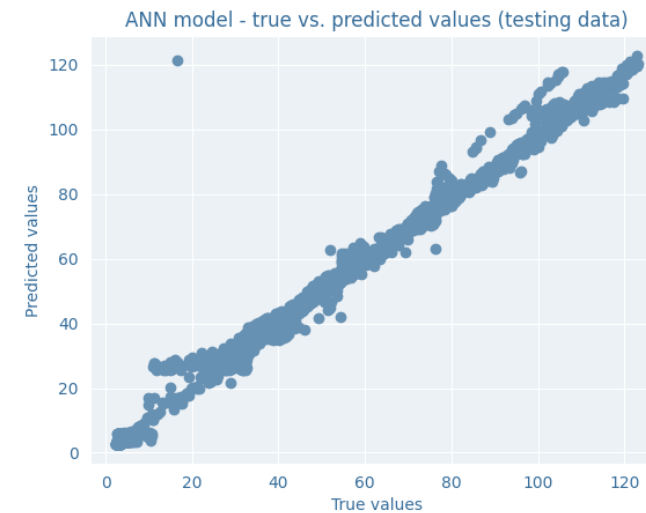
- ❑ From the above graphs, it can be seen that ANN, has highest R2 score and lowest MSE among all models.
- ❑ The difference in other metric MSE points out that ANN is better than other two.
- ❑ There is no overfitting of data as explained above in modelling section.
- ❑ It can be concluded that ANN performs better than other models and predict the energy consumption precisely.

Analyzing Best Performing Algorithm

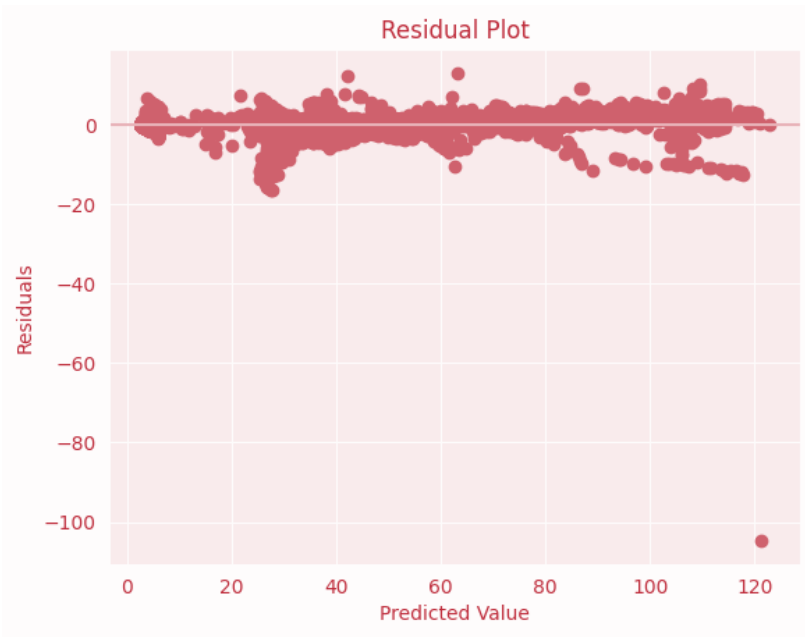
- ❑ Best performing algorithm – ANN
- ❑ Best Hyperparameters - {'dropout rate': 0.0, 'hidden layers': 3, 'l2': 0.0, 'learning rate': 0.01, 'neurons': 64}



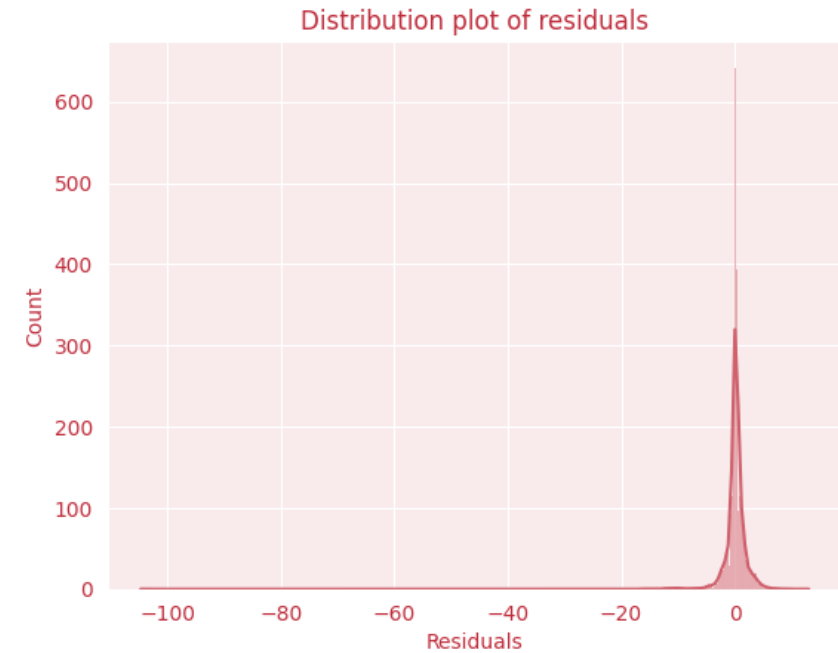
Feature 0 depicts (Lagging Current Reactive Power), *Feature 1* depicts (CO2(tCO2)), *Feature 2* depicts (Lagging Current Power Factor), *Feature 3* depicts (Week Status), and *Feature 4* depicts (Load Type)



The scatter plot of true vs. predicted values reveals a clear linear relationship between the two, showing that the model predicts the target variable accurately. The points are near to the diagonal line, indicating that the anticipated values are accurate



The scatter plot of predicted values vs residuals does not reveal any discernible pattern, and the residual values are randomly distributed around the horizontal line at $y=0$. There is no evident funnel shape or heteroscedasticity in the scatter plot, showing that the model performs consistently across the range of anticipated values.



The distribution plot of residuals shows it is good to use ANN model as the distribution plot of residuals have a bell-shaped curve, indicating that the residuals are normally distributed around zero. There are no obvious skewness or outliers in the plot

Key Insights

- ❑ The research indicated that ANN performs better than other models and predict the energy consumption accurately.
- ❑ The use of ANN model to predict energy consumption can be applied to any industry and can help reduce energy waste, optimize production, and minimize costs.
- ❑ ANN model can identify patterns and anomalies in energy consumption data and alert operators to take corrective action.
- ❑ This can help reduce energy consumption, improve energy efficiency, and reduce costs.





Future Directions

- ❑ Expanding the analysis to more locations and business sectors to evaluate how well the data mining models can be applied broadly.
- ❑ Analysing the underlying causes of energy use more thoroughly to find areas where energy can be saved, and efficiency can be increased.
- ❑ Analysing the effectiveness of the data mining models across extended time spans to determine their applicability in actual energy management scenarios.



THANK YOU