**FUNDAMENTALS OF MACHINE LEARNING**

**FINAL PROJECT**

**POWER PLANT-BASED PROJECT OF THE US GOVERNMENT**

The United States uses many different energy sources and technologies to produce electricity where the major categories are fossil fuels (coal, natural gas, petroleum) and generate electricity with steam turbines. Generally, some types of power plants use more electricity to operate than they generate, even for electricity storage, and therefore may have negative net generation on a monthly or annual basis. From the study of data that has been given by the US government, I understood that there are different variables to be considered while calculating the operating costs of fuel, to explore the operating costs of individual power plants, and see how fuel costs impact the viability of different types of generation.  It can highlight the competitiveness of renewable electricity in the market today.  It can show how the generation mix of different utilities has evolved over time, and how the usage of individual power plants has changed as fuel prices have changed and more renewable generation has been brought online. Many questions were raised when I study the data on fuel, for example: how types of fuel affect the increase in the generation of electricity, how fuel contents mix will have an impact on both sustainability and territory, and how fuel prices have a huge impact on the economy/market. Hence, I have analyzed the data using certain algorithms to know what variables are altering the price of fuel.

**Introduction:**

Analyzing data is done by checking all the possibilities.

* The data contains 30 variables and 608,565 rows where it contains many NA values, and blank spaces hence it contains a lot of noisy data, so to get accurate results I have done data cleaning by replacing empty spaces with “0” and omitted NA values.
* Data preparation: then I filtered the data and chose the required columns which can affect the objective using corrplot, which gives me an idea of how correlated each variable I have. A sample is chosen from the whole data because we cannot get accurate results on big data, it is split into training and testing to avoid overfitting. Then normalized numerical data to remove anomalies that are caused because of the transitive dependency.
* Clustering analysis is done to group similar data points in large datasets. Here, I have used K-means clustering because of a large set of data.

Clusters are formed by giving a random k value and checked for the elbow method and silhouette method to find the best optimal k value.

From my observations, k=2 is the best k value, which means that 2 clusters are formed based on the similarity of different variables.

**Problem statement:**

what types of fuels are affecting the operating costs.

How do fuel contents like sulfur, ash, and chlorine affect the heat produced to generate electricity?

What is the relation between fuel received and fuel cost?

What kind of impact will fuels have on the environment?

**Analysis and discussion:**

From the observations, we can tell that k=2 is the optimum k value, and it is verified with the WSS and silhouette method. Generally, the elbow method is affected by outliers, we can choose the k value by comparing both methods.

In the elbow method, we got k=3 but whereas in the silhouette method we got k=2, as it is not affected by outliers, we can consider k=2 as the best k value. Now that we have determined the most optimal number of clusters, we can create a cluster plot using K-means to better illustrate the clustering. The graph below shows our three clusters obtained using K-means clustering.

Chart

Description automatically generated with low confidence

The two clusters above are formed considering fuel cost as the major aspect. These clusters were created using all numerical data from the dataset. We can observe that cluster 1 is larger than cluster 2

. Table

Description automatically generated with medium confidence

From the means values of the clusters, we can observe that cluster 2 produces more heat than cluster 1 but the cost and contents like sulfur and ash are less in cluster 1 which is beneficial both economically and ecologically.

Actually to confirm whether our prediction on clusters is accurate are not, I have done a linear regression model on test data to predict the clusters, hence, from the results, it is clearly shown that I have considered the data correctly. After predicting I observed that my assumption in choosing k is accurate.

Cluster 1: SPIC AND SPAN- less pollutant with good fuel-receiving units

Cluster 2: CURDDY- contains a high number of pollutants that can affect the environment.

**Summary:**

From the observation, we can conclude that from fuel types like coal, oil, and natural gas, coal was used effectively but we know there is a scarcity of coal for future generations, so natural gas is a better option as it is readily available. Even though it is a little bit costlier than coal, natural gas use is a better option, and in the US 31% of the fuel type used is natural gas which highlights the competitiveness in the market on renewable energy.

We know that if cost is less, supply is more, hence if fuel received units is more then there is low cost on fuels. Fuels contents like sulfur release sulfur oxides while producing electricity which leads to global warming. So, avoiding more sulfur and content is good when considering the fuel type. As coal is one major reason for power generation that leads to carbon emissions

From our analysis, we know that power plants that use less fuel to produce more heat to generate electricity are chosen as best, so the power plants in cluster 1 are better, and the firm can suggest to the US government that these are well-operated power plants.