**FUNDAMENTALS OF MACHINE LEARNING**

**(FINAL PROJECT REPORT)**

**By: Shivani Haridas Pitla**

* **EXECUTIVE SUMMARY:**

Despite being a relatively safe and clean form of energy when it is used, electricity generation and transmission have an impact on the environment. The primary motive of any business is to maximize its profits with minimum cost. I would like to propose a model which is profitable as well as sustainable. In this analysis, I would like to predict which group of fossil fuels would produce high heat that would further be used to produce power with least cost as well as least or moderate emissions of combustion byproducts. K-means clustering is used to analyze the problem. By using silhouette method 7 clusters where formed.

* **INTRODUCTION:**

The dataset used for the analysis is about the generation of power in United States. This dataset is taken from source called “Public Utility Data Liberation (PUDL) project”.

The dataset contains 608,565 rows and 20 variables. Among these variables I have selected 7 variables those are: Fuel Group Code, Fuel Received Units, Fuel MMBtu per unit, Sulphur Content, Ash Content, Mercury Content, Fuel cost per mmbtu.

**The following actions are taken to address the issue based on data:**

1. Selecting the required variables
2. Imputing the variables with NA values
3. Sampling and partitioning the data
4. Data Normalization
5. Finding the Optimal “K” using silhouette method
6. K-means algorithm is used for clustering

**silhouette method**: The silhouette value predicts an object's cohesion with its own cluster in comparison to other clusters. The silhouette's value ranges from [1, -1], with a high value indicating that the object is well matched to its own cluster and a low value indicating that it is poorly matched to nearby clusters.

**K-means algorithm:** K-means is a technique for clustering data points that uses unsupervised learning. By minimizing the variance within each cluster, the algorithm iteratively divides the data points into K clusters.

I am using K-means for clustering as it is simple, flexible, efficient and it is highly suitable for large datasets.

* **PROBLEM STATEMENT:**

Which fuel group produces more heat with least cost also considering least emission of the substances that occur in combustion gases when these fuels are burned.

* **ANALYSIS AND DISCUSSION:**

By using “Silhouette method I got K=7 that is 7 clusters.

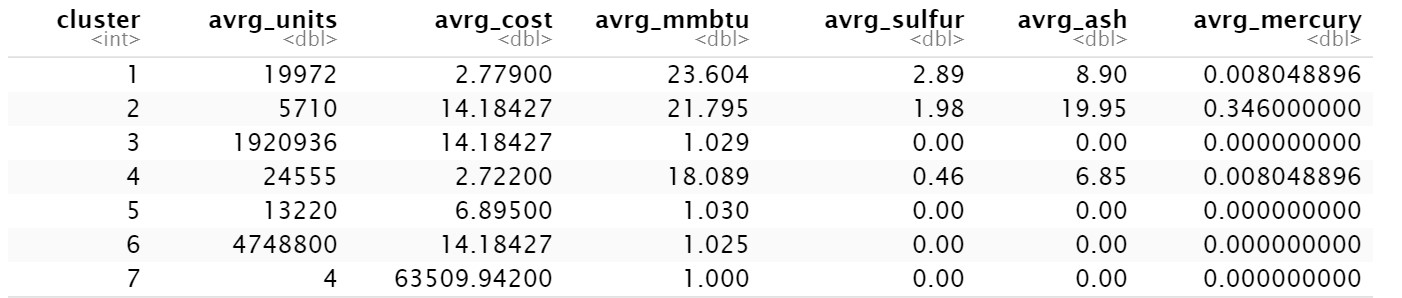
Among these 7 clusters, I selected cluster 4 as the best segmentation as it has good units of heat generated by burning 436 average units of natural gas and 2 average units of other gas. This cluster is also environmentally friendly and safe when it comes to human health as it produces very less amount of sulfur, ash, and mercury.

* **APPENDIX:**

Chart, line chart

Description automatically generatedChart

Description automatically generated



**References:**

[**https://laptrinhx.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters3695116625/#:~:text=Elbow%20and%20Silhouette%20methods%20are%20used%20to%20find,a%20better%20method%20compared%20to%20the%20Elbow%20method**](https://laptrinhx.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters3695116625/#:~:text=Elbow%20and%20Silhouette%20methods%20are%20used%20to%20find,a%20better%20method%20compared%20to%20the%20Elbow%20method)**.**

[**https://www.w3schools.com/python/python\_ml\_k-means.asp**](https://www.w3schools.com/python/python_ml_k-means.asp)