**PROJECT REPORT**

**fundamentals of machine learning**

***BY :- NIKHIL REDDY ADDULA***

***Executive Summary:***

The objective of this project is to analyze various factors associated with power generation in the U.S.A This report is the result of a study for the programmatic review of the PUDL dataset, which contains a large amount of data about Power Generation in the United States and was published under the review of the U.S. Department of Energy's to provide their customers with affordable, dependable electricity with the least possible environmental impact. Using the K-means approach to evaluate and categorize the PUDL dataset, four clusters were created that comprised various results and provided an outcome indicating that using natural gas is the best option when both economic and environmental issues are taken into account. More study has been undertaken based on the average fuel cost, the number of fuel units received at power plants and the chemical composition of the fuel. Gas is the fuel type on which the government spends the most, while oil is the fuel type on which the government spends the least, according to the statistics. Despite having a lower per unit cost for MMBtu, coal is not widely used, partly because it contains contaminants such as ash, mercury, and sulfur.

***Introduction:***

The data has information related to the monthly Purchases and Contracts. It has more than 25 variables The missing values were computed by first converting them into the median and then preparing the data into 75%, which produced some outliers. The outliers were computed before scaling the necessary variables that were required for our project. taking 2% of the total data and removing the columns that weren't necessary.

***Problem Statement:***

* *what clusters are formed based on the algorithm used?*
* *Do clusters have minimal fuel cost?*
* *Do clusters have more ASH & SULFUR content?*
* *What is the Most Optimal cluster?*

***Analysis and Discussion:***

analyzing the data using machine learning I have used k means clustering for the data to find the best cluster that provides what we want to achieve with the given data I have used “wss” and “silhouette” and then determined the k value and the centre values as the k=4 which is the optimal outcome both the methods shows the same value of k.

***Conclusion:***

Cluster 1

This cluster receives fuel of 0.18177742. As they are receiving low fuel, sulphur & ash their heat content in fuel(fuel\_mmbtu) is also low (0.4458332). The fuel cost per MMBtu is higher(0.0048009553) than all 4 clusters formed. Due to the high cost of fuel per MMBtu, this Cluster is not a favoured one to suggest to the US government.

Cluster 2

This cluster receives fuel of 3.95554628 which is high than all the clusters. The heat content in the fuel is very very low at 0.1033619 compared to all 4 clusters. The fuel cost per MMBtu is lower(0.0014246208) than all 4 clusters formed. This cluster is also not a preferred one to recommend for us Government because of fuel MMBtu per unit.

Cluster 3

This cluster receives fuel of 0.05023818 which is minimal. Their heat content in the fuel is 2.3366447 which is good for the fuel received compared to the other 3 clusters. The fuel cost per MMBtu is also very good(0.0017970585) to the fuel received and the heat content.

This Cluster is the one that the US Government should be recommended since it takes all the variables, including fuel received, heat content, and fuel cost per MMBtu.

***Reference:***

<http://tiantiy.people.clemson.edu/blog/2019/MissingData/random_forest.html>

<https://stackoverflow.com/questions/54332680/multiple-imputation-in-r-using-missforest-on-categorical-variables>

<https://www.eia.gov/energyexplained/us-energy-facts/>

<https://catalyst.coop/pudl/>

<https://stats.stackexchange.com/questions/483222/how-to-use-missforest-in-r-for-test-data-imputation>