Fundamental of Machine Learning

**Final Project**

horizontal line

# 

**Yousef Alomary**

**Spring 2023**

**Kent State University**

**Introduction:**

The data used in this code is related to fuel receipts and costs for power plants in the United States. The dataset is obtained from the Energy Information Administration (EIA) and includes information about the type of fuel used, its cost, and various other attributes such as energy content, sulfur content, and ash content. The code begins with data cleaning steps, including dropping unnecessary columns, removing duplicate rows, and sampling a small subset of the data for clustering analysis. Categorical variables are converted to factors, and numeric variables are scaled before performing k-means clustering to identify clusters based on similarities in fuel attributes. The optimal number of clusters is determined using the elbow method, and the cluster assignments are added to the data. This process can help identify patterns in fuel usage and costs and provide insights into fuel procurement strategies for power plants.

**Executive Summary:**

Our analysis of the fuel data for power generation in the US suggests that clustering based on "energy\_source\_code" could reveal valuable insights into the energy mix of different regions and the country as a whole. Understanding the composition of these clusters could provide information about the environmental impact and cost-effectiveness of different fuel types. Clustering on other variables such as "fuel\_mmbtu\_per\_unit" or "fuel\_cost\_per\_mmbtu" could provide further insights into the efficiency and cost-effectiveness of different fuel types. Our findings suggest that this information could be used to inform policy decisions related to energy production and consumption and help utilities make more informed decisions about fuel procurement and power plant operations.

**Problem statement:**

The objective of this study is to explore fuel data for power generation in the US and identify patterns in the types of fuels used, their energy efficiency, environmental impact, and cost-effectiveness. We will use clustering analysis to segment the data based on variables such as energy\_source\_code, fuel\_mmbtu\_per\_unit, sulfur\_content\_pct, ash\_content\_pct, and fuel\_cost\_per\_mmbtu. The questions we aim to answer are:

What are the main types of fuels used for power generation in the US, and how are they distributed across different regions?

How do different fuel types vary in terms of their energy efficiency, environmental impact, and cost-effectiveness?

Can we identify clusters of power plants with similar fuel characteristics, and what insights do these clusters provide for policy decisions and stakeholders in the power industry?

**Analysis and Discussion:**

The US power industry relies on a variety of fuel types for power generation, including coal, natural gas, nuclear energy, hydroelectric power, and renewable sources. These fuels are distributed differently across the country, with some regions relying more heavily on certain fuels than others. Understanding the distribution of fuel types can help identify regions that may be more vulnerable to changes in fuel prices or supply disruptions.

In terms of energy efficiency, environmental impact, and cost-effectiveness, different fuel types vary widely. Coal, for example, is generally less efficient and has a higher environmental impact than natural gas, but it is often cheaper. Nuclear energy is highly efficient but raises concerns about safety and nuclear waste disposal. Renewable sources such as wind and solar are cleaner but may be less reliable and more expensive than fossil fuels.

Clustering algorithms such as k-means can be used to identify clusters of power plants with similar fuel characteristics. These clusters can provide valuable insights for policy decisions and stakeholders in the power industry. For example, clustering could reveal that certain regions rely heavily on coal-fired power plants and have a higher environmental impact as a result. This information could be used to inform policy decisions related to reducing carbon emissions or transitioning to cleaner energy sources. Similarly, clustering could reveal that certain fuel types are more cost-effective or energy-efficient, which could be useful for utilities and other stakeholders in making decisions about fuel procurement and power plant operations.

In conclusion, understanding the distribution of fuel types for power generation in the US and their varying characteristics is crucial for making informed policy decisions and helping stakeholders in the power industry make strategic decisions. Clustering algorithms can provide valuable insights into fuel use patterns and efficiency, and should be used in conjunction with other data analysis techniques to develop a comprehensive understanding of the US power industry.

**Conclusions:**

Based on the analysis of the fuel data for power generation in the US, it can be concluded that clustering analysis provides valuable insights into the energy mix of different regions and the country as a whole. The clustering of power plants based on their fuel characteristics, such as energy\_source\_code, fuel\_mmbtu\_per\_unit, sulfur\_content\_pct, ash\_content\_pct, and fuel\_cost\_per\_mmbtu, can help identify patterns in fuel usage and costs. It can also provide insights into fuel procurement strategies for power plants, and inform policy decisions related to energy production and consumption. Assuming that the data used in this study is representative of the entire US power industry, the findings suggest that understanding the composition of fuel clusters is crucial for evaluating the environmental impact and cost-effectiveness of different fuel types. The clustering analysis reveals that the US power industry relies on a variety of fuel types, and their distribution varies across different regions. Coal-fired power plants are still prevalent in some regions, and they have a higher environmental impact than natural gas, nuclear energy, and renewable sources.

Final thoughts:

In conclusion, the US power industry faces a complex set of challenges related to fuel procurement, power generation, and environmental impact. Clustering analysis provides a useful tool for identifying fuel use patterns and efficiency, and the insights gained from this analysis can be used to inform policy decisions and stakeholders in the power industry. However, it is important to note that clustering analysis is just one of many data analysis techniques and should be used in conjunction with other methods to develop a comprehensive understanding of the US power industry. Therefore, further research and analysis are required to gain a deeper understanding of the US power industry and identify effective solutions to the challenges it faces.