**Executive Summary: COVID-19 Severity Index Clustering Analysis**

**Purpose of the Application –** This analysis aims to apply machine learning techniques to understand and categorize hospitals based on COVID-19 severity data. Using **K-Means clustering**, the application groups hospitals into distinct clusters according to the severity of COVID-19 cases observed over 7 days to uncover patterns that can inform decision-making, resource allocation, and healthcare response strategies.

**Dataset Overview -** The dataset used for this analysis is the **COVID-19 Severity Index** from the Registry of Open Data on AWS. Key columns include*Severity Scores*: 7 columns, each representing the severity of COVID-19 cases for each hospital over 7 days (from severity\_1-day to severity\_7-day).Also, *Hospital Data*: Includes hospital-specific information such as total\_deaths\_hospital, hospital\_name, countyname, statename, and geographical coordinates (latitude, longitude).

**Methodology**

1. **Data Preprocessing**: The severity scores over the 7 days were selected as key features for clustering. Missing data was imputed with the average value for each column to ensure a clean dataset. Also, the severity scores were standardized using **StandardScaler** to ensure all features had equal importance during the clustering process.
2. **Clustering with K-Means**: The **K-Means algorithm** was applied to group hospitals based on the severity scores. The optimal number of clusters was determined using the Elbow Method, which suggested three distinct clusters (Muhtasim, 2023).
3. **Dimensionality Reduction with PCA**: To better visualize the clusters, Principal Component Analysis (PCA) was used to reduce the data to two dimensions. This allowed us to plot the clusters and observe how hospitals grouped based on the severity scores.

**Results -** The analysis revealed three distinct clusters of hospitals based on their severity scores over 7 days:

* **Cluster 1**: Hospitals with consistently high severity scores across the seven days, indicating a high burden of COVID-19 cases.
* **Cluster 2**: Hospitals with moderate severity, showing fluctuations but generally lower severity levels than Cluster 1.
* **Cluster 3**: Hospitals with low severity scores, indicating relatively mild or controlled cases during the observed period.

**Insights and Conclusions**

* **Resource Allocation**: Hospitals in **Cluster 1**, with high severity, may require more immediate resources such as ICU beds, ventilators, and medical personnel. This understanding can help public health officials allocate resources more effectively.
* **Trend Analysis**: The clustering provides a clearer picture of how hospitals in different regions are experiencing COVID-19 outbreaks. Hospitals in **Cluster 2** and **Cluster 3** may be experiencing fewer severe cases and could serve as potential overflow sites or examples of effective management strategies.
* **PCA Visualization**: The PCA plot provides a clear, visual representation of how hospitals differ in COVID-19 severity. Hospitals in the same cluster tend to exhibit similar severity trends, which can inform decision-makers on regional response efforts (Jolliffe, 2002).
* **Next Steps**: Further analysis could explore the factors contributing to these severity patterns, such as hospital capacity, geographic location, or health interventions. This could lead to more targeted responses in areas with high severity.

**Conclusion**

This clustering analysis helps healthcare providers and policymakers understand the varying levels of COVID-19 severity across hospitals. By using machine learning techniques like K-Means and PCA, we can identify areas of concern and take action to address potential resource shortages, improve preparedness, and optimize the response to the ongoing pandemic.

**References**

AWS. (2024). *COVID-19 Severity Index Dataset*. Retrieved from <https://covid19-lake.s3.amazonaws.com/index.html>

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). Springer-Verlag. <https://doi.org/10.1007/b98835>

Muhtasim, Md. Abdul Masud. (2023). Clustering Countries on COVID-19 Data among Different Waves Using K-Means Clustering. <https://doi.org/10.4236/jcc.2023.117001>