FINAL PROJECT

OBJECTIVES OF THE PROJECT

The aim of this project is to classify countries based on socio-economic and health factors that significantly influence overall national development.

The results of this project will provide philanthropic non-governmental organizations (NGOs) with valuable insights to allocate their funds strategically and effectively target countries facing the most acute need for aid.

DATA

Column Name	Description
country	Name of the country
child_mort	Death of children under 5 years of age per 1000 live births
exports	Exports of goods and services per capita. Given as %age of the GDP per capita
health	Total health spending per capita. Given as %age of GDP per capita
imports	Imports of goods and services per capita. Given as %age of the GDP per capita
Income	Net income per person
Inflation	The measurement of the annual growth rate of the Total GDP
life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same
total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

Source of Data: https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data

CLUSTERING MODELS

K-MEANS MODEL

- The elbow method was used to determine the number of clusters
- The KneeLocator function was used to confirm the optimal number of clusters
- Based on the elbow method and kneeLocator function, the selected optimal number of clusters is 4

- The KNN algorithm used is as follows;

```
KMeans(n_clusters =4, max_iter = 500,  init = 'k-means++', random_state=
42)
```

- Countries were grouped into four (4) categories based on their developmental levels: 'Very High', 'High', 'Low', and 'Very Low'. For instance, nations demonstrating strong positive development indicators such as high GDP per capita, elevated income, increased life expectancy, and low child mortality rates were classified as 'Very High'. Conversely, countries exhibiting weaker positive development indicators, including low GDP per capita, reduced income, shorter life expectancy, and high child mortality rates, were categorized as 'Very Low'.
- The following table presents a cross-tabulation detailing the average socio-economic and health indicators across different clusters

Development Status	Child mortality	exports	health Spending	imports	income	inflation	Life expectancy	Total fertility	GDP per capital
High	5.18	46.12	9.09	40.58	44,021.88	2.51	80.08	1.79	42,118.75
Low	21.69	41.07	6.20	47.91	12,671.41	7.61	72.87	2.30	6,519.55
Very High	4.13	176.00	6.79	156.67	64,033.33	2.47	81.43	1.38	57,566.67
Very Low	92.96	29.15	6.39	42.32	3,942.40	12.02	59.19	5.01	1,922.38

AGGLOMERATIVE CLUSTERING

- The Agglomerative Clustering algorithm was used with the intention to create 4 clusters using the 'ward' linkage method while computing the full hierarchical tree of clusters, providing an in-depth view of the clustering structure.
- The following algorithm was used;

```
AgglomerativeClustering(n_clusters=4, linkage='ward', compute_full_tree=True)
```

- The clusters were renamed using a similar approach as applied in the KMeans Clustering
- A cross tabulation of the average socio economic and health indicators across the clusters are as follows;

Development Status	Child mortality	exports	health Spending	imports	income	inflation	Life expectancy	Total fertility	GDP per capital
Very Low	92.96	29.15	6.39	42.32	3,942.40	12.02	59.19	5.01	1,922.38
Low	21.69	41.07	6.20	47.91	12,671.41	7.61	72.87	2.30	6,519.55
High	5.18	46.12	9.09	40.58	44,021.88	2.51	80.08	1.79	42,118.75
Very High	4.13	176.00	6.79	156.67	64,033.33	2.47	81.43	1.38	57,566.67

Mean Shift Clustering

- Mean Shift clustering algorithm was used for the unsupervised learning tasks, specifically for grouping data points into clusters based on their density. The bandwidth parameter in MeanShift determines the size of the region to consider when estimating the density around a data point. When MeanShift(bandwidth=None) is used, it means that the bandwidth parameter is set to None, allowing the algorithm to automatically estimate the bandwidth using a heuristic approach. The following algorithims were used to fit the data;

```
MeanShift(bandwidth=None)
```

- The Mean Shift Clustering algorithm gave 7 clusters. I renamed the clusters based on the level of development as follows;
    - Very High
    - High
    - Moderately High
    - Average
    - Moderately Low
    - Low
    - Very Low
- A cross tabulation of the average socio economic and health indicators across the clusters are as follows;

Development Status	Child mortality	exports	health Spending	imports	income	inflation	Life expectancy	Total fertility	GDP per capital
Very Low	153.34	24.02	7.92	54.04	1,583.60	7.04	47.52	4.73	714.80
Low	130.00	25.30	5.07	17.40	5,150.00	104.00	60.50	5.84	2,330.00
Moderately Low	35.58	38.25	6.84	44.94	14,831.34	7.29	70.88	2.93	11,149.08
Average	5.50	128.00	8.92	120.25	37,000.00	0.30	80.35	1.71	34,900.00
Moderately High	2.80	200.00	3.96	174.00	72,100.00	(0.05)	82.70	1.15	46,600.00
High	8.38	59.03	4.19	27.68	85,775.00	10.21	78.95	2.02	57,975.00
Very High	2.80	175.00	7.77	142.00	91,700.00	3.62	81.30	1.63	105,000.00

Selecting the Best Model

The silhouette score was used to select the best model. The silhouette score is useful for comparing different clustering algorithms or selecting the optimal number of clusters in algorithms like KMeans, Hierarchical Clustering, Mean Shift, etc. It helps in assessing the goodness of the clustering result without ground truth labels. A score close to +1 indicates that the sample is well-clustered and lies far from neighboring clusters.

A score close to 0 indicates overlapping clusters or that the sample is close to the decision boundary between clusters.

A score close to -1 suggests that the sample might have been assigned to the wrong cluster

The results of the silhouette scores is as follows;

Model	Silhouette Score
KMeans	0.295952
Agglomerative Clustering	0.248119
Mean Shift	0.232216

Based on the Sihouette Score, the best model is KMeans Clustering.

Key Findings

- The KMeans Clustering identified a group of 47 economies with the worst socio-economic and health issues.
- These countries are in the most acute need for aid

- Philanthropic non-governmental organizations (NGOs) may allocate their funds strategically and effectively to these countries to optimize impact..

LIMITATION OF THE MODEL AND FUTURE REVISION

Flaw: KMeans assumes that clusters have equal variance and size, which might not hold in real-world scenarios.

Action: To consider using other algorithms like Gaussian Mixture Models (GMM), which relax these assumptions, or use scaled data to mitigate this issue.