Student Name: Aniruddha Joshi

Data Set: <https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data/data>

**Milestone 3**

**Project Topic:** Clustering Countries for Targeted Socio-Economic and Health Interventions

**Business Problem:** Many countries face significant challenges in meeting the socio-economic and health needs of their populations. Limited resources necessitate a targeted approach to development and funding. This project aims to categorize countries based on key indicators to identify specific needs and prioritize interventions in areas such as infrastructure, health, and education.

**Background History**

The unequal distribution of resources and development worldwide has caused major differences in quality of life, health, and economic opportunities. This project aims to cluster countries for specific socio-economic and health interventions.

**Data Explanation**

**Unsupervised Learning on Country Data (Kaggle):** is the reference data set selected for the project. The dataset contains around 167 records of country information such as child mortality, exports, health, import, income, inflation, life expectancy, gdp etc.

1. **Link:** <https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data/data>
2. **Format:** the data format is CSV file
3. **Contents:** upto 10 variables, such as child mortality, GDP per capita, life expectancy, literacy rate, access to clean water

**Data Preparation**

The data preparation involved the following steps:

1. Using Pandas and NumPy in Python to clean the dataset, removing invalid or incomplete rows.
2. Filling missing data with mean or average values.
3. Converting categorical data to one-hot encoding as required.
4. Converting the data to the appropriate datatypes.
5. Dropping NA records.

The result of the data set after cleaning:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Column | Total Count | Missing Count | Missing (%) | NA Count | NA (%) | Data Type | Unique Values | Outlier Count | Minimum | Maximum | Skewness |
| country | 167 | 0 | 0.0 | 0 | 0.0 | category | 167 | NaN | NaN | NaN | NaN |
| child\_mort | 167 | 0 | 0.0 | 0 | 0.0 | float64 | 139 | 4.0 | 2.6000 | 208.00 | 1.450774 |
| exports | 167 | 0 | 0.0 | 0 | 0.0 | float64 | 147 | 5.0 | 0.1090 | 200.00 | 2.445824 |
| health | 167 | 0 | 0.0 | 0 | 0.0 | float64 | 147 | 2.0 | 1.8100 | 17.90 | 0.705746 |
| imports | 167 | 0 | 0.0 | 0 | 0.0 | float64 | 151 | 4.0 | 0.0659 | 174.00 | 1.905276 |
| income | 167 | 0 | 0.0 | 0 | 0.0 | float64 | 156 | 8.0 | 609.0000 | 125000.00 | 2.231480 |
| inflation | 167 | 0 | 0.0 | 0 | 0.0 | float64 | 156 | 5.0 | -4.2100 | 104.00 | 5.154049 |
| life\_expec | 167 | 0 | 0.0 | 0 | 0.0 | float64 | 127 | 3.0 | 32.1000 | 82.80 | -0.970996 |
| total\_fer | 167 | 0 | 0.0 | 0 | 0.0 | float64 | 138 | 1.0 | 1.1500 | 7.49 | 0.967092 |
| gdpp | 167 | 0 | 0.0 | 0 | 0.0 | float64 | 157 | 25.0 | 231.0000 | 105000.00 | 2.218051 |

**Methods:**

The project employs a combination of data exploration, feature engineering, and machine learning modelling

1. **Data Exploration:**

Visualized the data using Matplotlib, Plotly, and Seaborn to identify relationships. Scatter plots were used to explore correlations, histograms to understand distributions. A correlation matrix was calculated and visualized as a heatmap. World maps were created to visualize child mortality and GDP. Refer below charts.

A screen shot of a graph

Description automatically generated

Figure Heatmap and correlation matrix

A map of the world

Description automatically generated

Figure Child mortality and GDP of the countries

A screenshot of a graph

Description automatically generated

Figure Correlation graph between each variable

1. **Feature Engineering:**

Reviewing and potentially removing unrelated columns based on a cross-reference matrix review. The project uses columns including 'country', 'child\_mort', 'exports','health','imports','income','inflation','life\_expec','total\_fer','gdpp'.

1. **Machine Learning Modeling:**

Trained the model using k-means clustering to categorize countries. The optimal number of clusters was evaluated using the elbow method and silhouette analysis. Scikit-learn and TensorFlow libraries were considered for model training.

**Analysis**

The analysis involved visualizing the data to identify relationships and patterns using scatter plots, histograms, and world plots. A correlation matrix was generated and visualized as a heatmap. World maps were created to visualize child mortality and GDP. To standardize the data, the Standard Scaler was used, scaling features to have a mean of 0 and a variance of 1, ensuring all features contributed equally to the analysis.

The elbow method and silhouette analysis were applied to determine the optimal number of clusters for k-means clustering. The elbow method involved plotting the within-cluster sum of squares against the number of clusters, identifying the point where adding more clusters did not significantly reduce the sum of squares, which was found to be 4. Silhouette analysis, which measures how similar an object is to its own cluster compared to other clusters, further validated 4 as the optimal number of clusters.

A graph with a blue line

Description automatically generated

Figure Elbow method used to identify optimal number of clusters

A white sheet of paper with numbers and lines

Description automatically generated

Figure Silhouette Score for different values of k

Finally, a world map was created to visualize countries grouped based on these clusters, highlighting similarities and differences among countries concerning the analyzed variables. This clustering revealed meaningful patterns and groups.

**A map of the world

Description automatically generated**

Figure World map showing the country clusters

**Conclusion**

The project aims to categorize countries based on socio-economic and health indicators. The clustering model effectively groups countries with similar characteristics. The project found that k=4 provides the best clustering results based on the elbow method and silhouette analysis. The graph shows more funds are needed in the in Asia and south America region.

**Assumptions**

The project makes below assumptions:

1. The dataset is complete and accurate.
2. The dataset is representative of the global population and contains minimal biases.

**Limitations**

Below are some of the limitations:

1. The small sample size of 167 records and 10 columns.
2. The dataset may not fully represent countries with limited data collection.
3. The selected indicators might not completely represent the complexity of socio-economic and health needs within each country.

**Challenges**

Challenges faced during selection of K value. Used different methods elbow and Silhouette to identify the selection of K.

**Future Uses/Additional Applications**

Expanding the dataset to include more countries and more variables and applying the clustering model to other socio-economic and health datasets. The model could be adapted to cluster regions within a country, allowing for more targeted interventions at a sub-national level.

**Recommendations**

Recommendations based on the project findings could include specific interventions for countries based on their cluster categorization such as Infrastructure development, Health initiatives and Educational programs

**Implementation Plan**

Based on the findings, recommendations and clustering the funds can be diverted accordingly. Currently cluster 4 shows vast majority of the countries that needs funding. Those can be further categorized based on further data collection and recategorization.

**Ethical Considerations:**

1. **Data Bias**: Acknowledge that the data may reflect historical inequalities or biases in data collection methods.
2. **Representation**: Ensure that the dataset includes a diverse representation of countries from different regions and income levels.

**Challenges/Issues:**

1. **Data Cleaning and feature:** As there are more around 167 data records based on country the clustering algorithm may have face challenges.
2. **Model Selection:** testing and finding the accuracy of the model on the low volume of data seems challenging.

**References:**

1. (Kokkula, n.d.)Unsupervised Learning on Country Data
2. (Tanja A. J. Houweling, n.d.) Socio-economic inequalities in childhood mortality in low- and middle-income countries: a review of the international evidence
3. (Rifat Atun, n.d.) A systematic review of the evidence on integration of targeted health interventions into health systems