Country Clustering Analysis Using Unsupervised ML

Temple University

Principles of Data Science

CIS 3715

Final Project Report

Fuad Hassan

April 17, 2024

## Introduction:

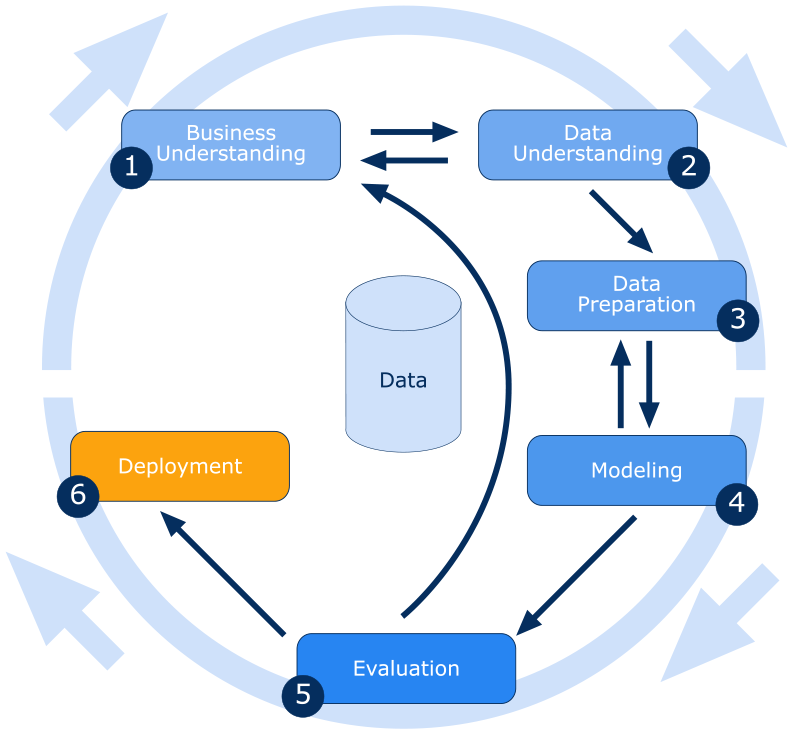
With so many disasters happening all around the world, it’s important for us to know where aid should go first. There are countries that are self-sustained and there are other countries who rely on funds to be sent. So how can we determine what countries need aid first? HELP international makes a different the lives of everyday people across underdeveloped regions. It’s important for us to categorize countries based on socio economic and health factors to assess their overall development. Through this final project I hope to sort countries in 4 categories, Highly Developed, Upper-Middle Developed, Lower-Middle Developed, and Least developed countries based on features like Child mortality rate, Import / Exports of goods, Total health spending, Net income per person, life expectancy, and the GDP per capita. Overall aiming to help countries that need the most aid.

## Problem Statement:

HELP International have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. So, CEO has to make decision to choose the countries that are in the direst need of aid. Hence, your Job as a Data scientist is to categories the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

## Approach:

I have used a descriptive analytic approach to classify countries into a high number of distinct, non-overlapping categories. My method involves using the [Cross-Industry Standard Process for Data Mining (CRISP-DM),](https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining) which includes the following:



CRISP-DM breaks the process of data mining into six major phases:

* **Business Understanding**
  + For us, our goal was to help Non-profits and NGOs leverage the power of data for social good and humanitarian causes.
* **Data Understanding**
  + Data cleaning and exploratory data analysis.
* **Data Preparation**
  + Transform data into a usable dataset for modeling.
* **Modeling**
  + To choose the best model, I have applied multiple algorithms to compare for the best result. For that I have followed
    - Performed Principal Component Analysis.
    - Performed feature selection based on boxplot analysis
* **Evaluation**
  + Clustering models were tested using visually inspecting the model results and Inertia
* **Deployment**
  + I have deployed a webpage that includes interactive graphs from a .ipynb file converted to HTML. Users can explore and interact with these graphs to gain a better understanding of the data

## Data Understanding and Preparation:

The dataset used for this analysis is available on Kaggle and is titled "Unsupervised Learning on Country Data". The dataset contains information about various countries and includes the following columns:

* country: Name of the country
* child\_mort: Child mortality rate (per 1000 live births)
* exports: Exports of goods and services (% of GDP)
* health: Total health spending (% of GDP)
* imports: Imports of goods and services (% of GDP)
* income: Net income per person
* inflation: The measurement of the annual growth rate of the Total GDP
* life\_expec: Life expectancy at birth (in years)
* total\_fer: The number of children that would be born to each woman if the current age-fertility rates remain the same
* gdpp: Gross Domestic Product per capital

### Sample of Data:

A screenshot of a computer screen

Description automatically generated

### Data Description:

A screenshot of a graph

Description automatically generated


From the description I was able to extract following data:

* The average GDP for the whole is 12964 USD
* The average Income is 17145 USD which is half of the top 25%
* The average Life expectancy is 71
* Median fertility rate is 2.4 children and average 2.9

### Data Outlier:

All features, except for ‘Life\_Expec’, exhibit outliers on the higher end of the data distribution. After verifying the data for these countries online, I concluded that the outliers are valid observations, so I retained them in the dataset.

### Exploratory Data Analysis:

I analyzed data distribution and skewness as well as features’ correlation.

Pairplots - Variable distribution and pairwise correlation:

A chart of blue and white graphs

Description automatically generated with medium confidence

#### Skewness:

* "child\_mort" skew: 1.45. The variable is **NOT** normally distributed.
* "exports" skew: 2.45. The variable is **NOT** normally distributed.
* "health" skew: 0.71. The Variable is **normally** distributed
* "imports" skew: 1.91. The variable is **NOT** normally distributed.
* "income" skew: 2.23. The variable is **NOT** normally distributed.
* "inflation" skew: 5.15. The variable is **NOT** normally distributed.
* "life\_expec" skew: -0.97. The variable is **NOT** normally distributed.
* "total\_fer" skew: 0.97. The variable is **NOT** normally distributed.
* "gdpp" skew: 2.22. The variable is **NOT** normally distributed.

#### Correlation:

A screenshot of a graph

Description automatically generated

### Observations and Explanations:

*[NOTE]: Close to 1: Strong positive relationship. Close to -1: Strong negative relationship. Close to 0: Weak or no relationship.*

* **child\_mort (Child Mortality):**
  + **With income:** Correlation of -0.524 means that higher income is linked to lower child mortality.
  + **With life\_expec:** Correlation of -0.886 shows that higher child mortality is linked to lower life expectancy.
  + **With total\_fer:** Correlation of 0.848 suggests that higher child mortality is linked to higher total fertility rates.
* **exports:**
  + **With imports:** Correlation of 0.737 means higher exports are linked to higher imports.
  + **With gdpp:** Correlation of 0.418 means higher exports are linked to higher GDP per capita.
* **health:**
  + **With gdpp:** Correlation of 0.346 shows a positive relationship between health investment and GDP per capita.
* **income:**
  + **With gdpp:** Correlation of 0.896 indicates a strong positive link between income and GDP per capita.
  + **With life\_expec:** Correlation of 0.612 suggests a positive link between income and life expectancy.
* **inflation:**
  + **With income:** Correlation of -0.148 shows a weak negative relationship between inflation and income.
  + **With gdpp:** Correlation of -0.222 shows a weak negative relationship between inflation and GDP per capita.
* **life\_expec (Life Expectancy):**
  + **With gdpp:** Correlation of 0.600 suggests that higher life expectancy is linked to higher GDP per capita.
* **total\_fer (Total Fertility Rate):**
  + **With gdpp:** Correlation of -0.455 suggests higher fertility rates are linked to lower GDP per capita.

Overall, the chart suggests strong relationships between child mortality, life expectancy, income, and GDP per capita. This means that higher income and GDP per capita often lead to better health outcomes like lower child mortality and higher life expectancy.

### Data Transformation:

Normality isn't necessary for clustering algorithms, but a normally distributed dataset can improve results because outliers can impact cluster centroids. To encourage normality and strengthen variable correlations, I applied the following transformations to the data:

* **Logarithmic Transformation**
* **Square Root Transformation**
* [**Box-Cox Transformation**](https://en.wikipedia.org/wiki/Power_transform#Box%E2%80%93Cox_transformation)

And Box-Cox Transformation provides us the best transformation with most normal distribution.

After **Box-Cox Transformation:**

* "child\_mort" skew: -0.0. The Variable is **normally** distributed
* "exports" skew: 0.19. The Variable is **normally** distributed
* "health" skew: -0.01. The Variable is **normally** distributed
* "imports" skew: 0.27. The Variable is **normally** distributed
* "income" skew: -0.04. The Variable is **normally** distributed
* "inflation" skew: 0.34. The Variable is **normally** distributed
* "life\_expec" skew: -0.19. The Variable is **normally** distributed
* "total\_fer" skew: 0.09. The Variable is **normally** distributed
* "gdpp" skew: 0.0. The Variable is **normally** distributed

## Model Development:

For this project I have not dropped any features, I wanted to keep as much as info in the original form. I have total of 10 features. I applied the K-Means algorithm to the dataset, using a range of cluster counts from 2 to 10. I used the k-means++ initialization method for more efficient starting centroids. For each value of K (number of clusters) I recorded the Inertia, which was the total sum of intracluster distances (squared distances from each data point to its cluster centroid)

A graph with blue and orange lines and dots

Description automatically generated

My goal is to cluster the data in 4 different cluster. As we can see in the graph, we do have a flattish line in 3 and 5. Therefore, we I’ll continue with our initial plan of having 4 cluster.

### K-mean

A screenshot of a computer code

Description automatically generated

Pairplot - Number of K-Means clusters (using all data features): 4

A chart of different colored dots

Description automatically generated with medium confidence

After using K mean we get:

|  |  |
| --- | --- |
| Cluster | Size |
| 0 | 40 |
| 1 | 39 |
| 2 | 37 |
| 3 | 51 |

### PCA

A screenshot of a computer

Description automatically generated

A screen shot of a computer screen

Description automatically generated

After PCA our cluster stays the same

|  |  |
| --- | --- |
| Cluster | Size |
| 0 | 40 |
| 1 | 39 |
| 2 | 37 |
| 3 | 51 |

### Result and Clusters description:

The 4 clusters have approximately same size: there’s no major outliers when it comes to the number of observations per cluster.

A blue square with white text

Description automatically generated

Cluster 0: This cluster include the countries with the highest average values in Income, Life Expectancy, GDP per capita, and lowest average values in Child Mortality and Total Fertility. Countries that belong to the Cluster 0 are the **Highly Developed**

A yellow and blue graph

Description automatically generated with medium confidence

Cluster 1: The highest average values in Child Mortality and Total Fertility as well as the smallest average values in Income, Life Expectancy, and GDP per capita. Countries that belong to the Cluster 1 are the **Least Developed** countries

A green and blue compass

Description automatically generated

Cluster 2: Countries with Income, Life Expectancy, GDP per capita mean values below the average, and Child Mortality and Total Fertility mean values above the average. Countries that belong to the Cluster 2 are the **Lower-Middle Developed** countries

A red and blue map

Description automatically generated

Cluster 3: These countries have lower average values than the Highly Developed observations in Income, Life Expectancy, GDP per capita but still above the overall average. They also have higher average values than the cluster 1 countries in Child Mortality and Total Fertility, however below the dataset average. Countries that belong to the Cluster 3 are the **Upper-Middle Developed** countries.

#### 

#### Main insights:

1. The clusters identified not only reveal patterns in the features used for training the clustering algorithm but also provide insights into features such as Exports, Imports, and Inflation that were not directly used in the training process.
2. There is a trend indicating that higher development in a country correlates with lower inflation rates.
3. The Least Developed Countries have significantly higher child mortality rates compared to other clusters, with 95 children under 5 years old per 1,000 live births versus the population average of 38 children per 1,000 live births. Additionally, these countries also have a higher total fertility rate of 5.12 births per woman, in contrast to the population average of 2.95 births per woman.

These findings highlight a world marked by significant disparities in living standards and well-being, where prosperity and high-quality living conditions are still limited to a select few. NGOs should prioritize their efforts in aiding the Least Developed Countries by working to reduce child mortality rates and support economic growth policies that will increase GDPP and income levels.

#### Acknowledgements

* ROHAN KOKKULA, Kaggle, for sharing the Data set
* GitHub Repo from Username: [Iron486](https://github.com/Iron486) from the inspersion

Links

* [GitHub Repository](https://github.com/fuadh246/Global-Insights-Clustering)
* [HTML format Jupiter Notebook with interactive graph](https://fuadh246.github.io/Global-Insights-Clustering/model/Cluster.html)