Clustering the Countries by using Unsupervised Learning for HELP International

By: Hari Priya Ramamoorthy

Objective:

To categorise the countries using socio-economic and health factors that determine the overall development of the country.

Dataset Details:

Kaggle Dataset - **https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data/data**

Dataset used has 10 columns and description of each is as follows.

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.

health - Total health spending per capita.

imports - Imports of goods and services 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.

What is Done Here?

1. Visualize power of different unsupervised clustering algorithms by applying them on the country dataset mentioned.

2. Algorithms Used: K-Means Clustering, Agglomerative Clustering, DBSCAN Clustering
3. PCA

EDA Observations:

1. There is no non-null data. 9 numeric and 1 string(country name) column.
2. child_mort, exports, imports, income, inflation, gdpp - seems to have large difference between 75% percentile and max value. it looks like these features are right skewed.
3. country feature is identical value, cant be considered as categorical as there is no multiple entries. so, this particular feature might not be helpful for the modeling. but, we shall use for EDA.
4. Based on Correlation values, we see that child_mort, life_expec and total_fer are highly correlated. Also, income and gdpp are positively correlated. whereas life_expec and child_mort are highly negative correlated.
5. Based on the observations, it's important to transform skewed data and scale all columns to normalise them.

ML Clustering:

1. For the clustering problem, to decide how many number of clusters consider, used a most popular elbow method Based on Elbow method, there should be 3 clusters.
2. Clustered the countries using K-Means Clustering, Agglomerative Clustering, DBSCAN Clustering
3. It is observed that Country having high child-mortality, low GDP per capita and low inflation(The measurement of the annual growth rate of the Total GDP) is a under-developing country Country having low child-mortality, high gdpp and high infaltion is the developed country

Feature Importance with PCA

We've been able to identify some patterns in the data and group countries into 3 clusters. However, we should not rely solely on this result to make the recommendation of countries that should receive funding. There are a few alternatives to explore before we can make this recommendation.

The implementation of a clustering model in this case did not bring up patters that we might have not found otherwise, in a way, it only confirmed general knowledge of intuition about this topic. The clustering can be considered as a preprocessing step and further analysis is required.

Conclusion:

In this notebook I have:

1. Classified countries using different unsupervised learning algorithms like K-Means, DBSCAN and Hierarchical clustering
2. Identified that features like GDP, Health, Income, Child mortality rate contribute more to the classification of clusters
3. To confirm that, used PCA Technique. Dropped highly correlated features and checked the error rate of the model on k-means algorithm. There was no much change in the SSE and cluster needed(refer graph).