CLUSTERING COUNTRIES
BASED ON
SOCIO-ECONOMIC AND HEALTH FACTORS

Project report

By

RAKESH KUMAR DONGARI - N254V945

SHIVAJI REDDY SAMA - F454W858

KISHORE POOJARI - G397M563

ABSTRACT:

Our project objective is to cluster the countries which are in direst need of financial aid by given a number of socio-economic factors from the countries data by using clustering techniques. The motivation of our project is to help countries in poverty and people of backward countries with basic amenities and relief by identifying which countries are in direst need of financial aid by using various clustering techniques such as K-Means clustering and Hierarchical clustering. And by using these techniques we are able to identify the countries which are in direst need of financial aid such as Barundi, Liberia, Congo, Dem. Rep, Madagascar, Mozambique.

INTRODUCTION:

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

Our objective is to analyze the given data among 167 countries, and we need to cluster the countries which are in direst need of financial aid using some socio-economic and health factors that determine the overall development of the country. These are performed by clustering methods like k-means clustering and hierarchical clustering. The 10 socio-economic factors provided in the dataset are country, child mortality, exports, imports, health, income, inflation, life expectancy, total fertility and GDP. By using these various socio-economic factors we have to cluster the countries.

DATASET AND FEATURES:

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 newborn 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.

DATA PRE-PROCESSING:

1. Missing Value Handling (If applicable)
2. Type Casting (If applicable)
3. removing unnecessary rows/columns (through missing value handling and correlation)
4. Handling Outliers
5. Feature Scaling (Standardization)

From primary observation of the dataset, we found out that

- There is only 1 categorical column here, i.e. ‘country’
- Rest all columns are numeric
- There are no missing values in the data, hence missing value handling is not required.
- All columns have correct datatypes, hence type casting is not required.
- ‘exports’, ‘health’, ‘imports’ are given in percentage of gdpp. These features would be converted to their actual values.
- Outliers are handled by performing flooring and capping.

HOPKINS TEST:

The Hopkins statistic is a way of measuring the cluster tendency of a data set, in other words: “how well the data can be clustered”. It belongs to the family of sparse sampling tests. It acts as a statistical hypothesis test where the null hypothesis is that the data is generated by a Poisson point process and are thus uniformly randomly distributed.

A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.

High value of Hopkins Statistics implements that dataset has high tendency to cluster. We have got a Hopkins score of 89.3 on our dataset.

METHODS:

1. K-Means Clustering:

K-Means clustering is an unsupervised machine learning algorithm that divides the given data into the given number of clusters. Here, the “K” is the given number of predefined clusters, that need to be created.

It is a centroid-based algorithm in which each cluster is associated with a centroid. The main idea is to reduce the distance between the data points and their respective cluster centroid.

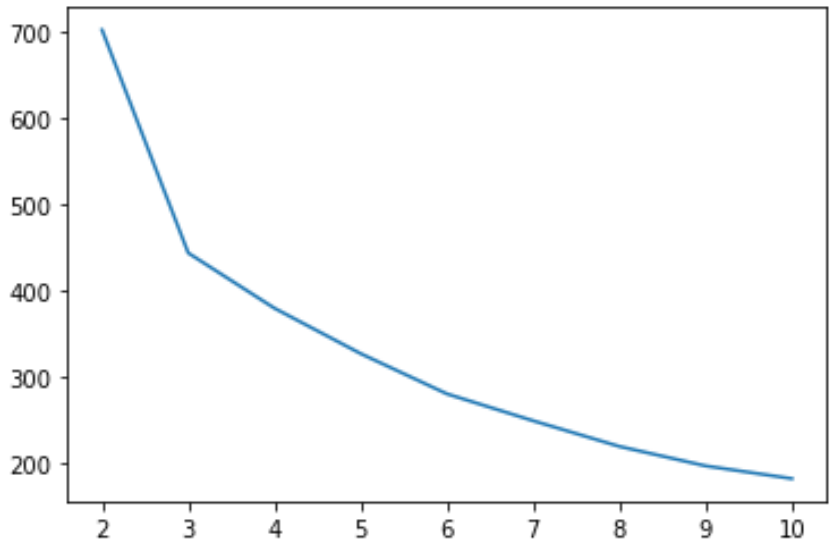
The algorithm takes raw unlabeled data as an input and divides the dataset into clusters and the process is repeated until the best clusters are found.

K-Means is very easy and simple to implement. It is highly scalable, can be applied to both small and large datasets. There is, however, a problem with choosing the number of clusters or K. Also, with the increase in dimensions, stability decreases. But overall K Means is a simple and robust algorithm that makes clustering very easy.

To Choose the value of k, there are two methods,

1. Elbow curve:

Elbow method is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the elbow point.



ELBOW CURVE

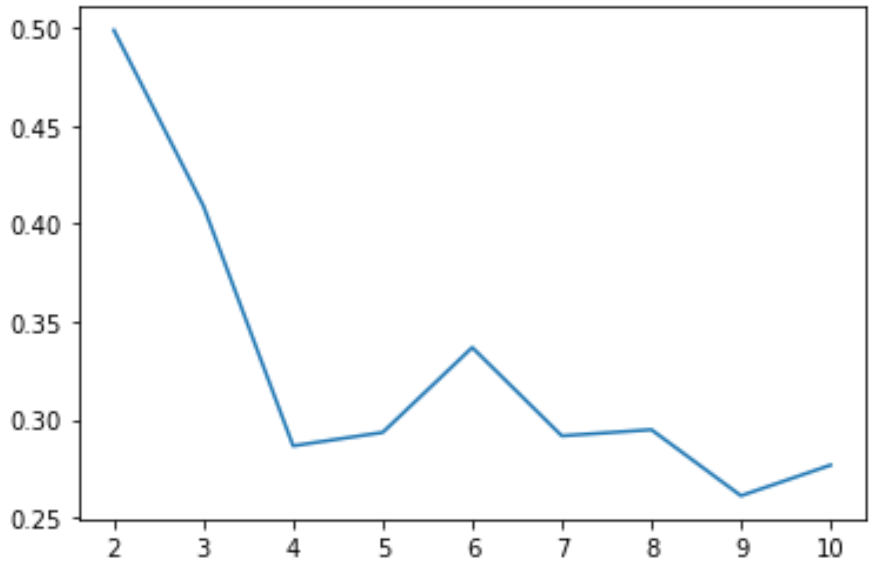
From the elbow curve above we can see that we have a breakpoint at 3, So we ca n go for the value of k with 3 i.e k=3.

2. Silhouette score:

$$\text{silhouette score} = p - q / \max(p, q)$$

p is the mean distance to the points in the nearest cluster that the data point is no t a part of
q is the mean intra-cluster distance to all the points in its own cluster.

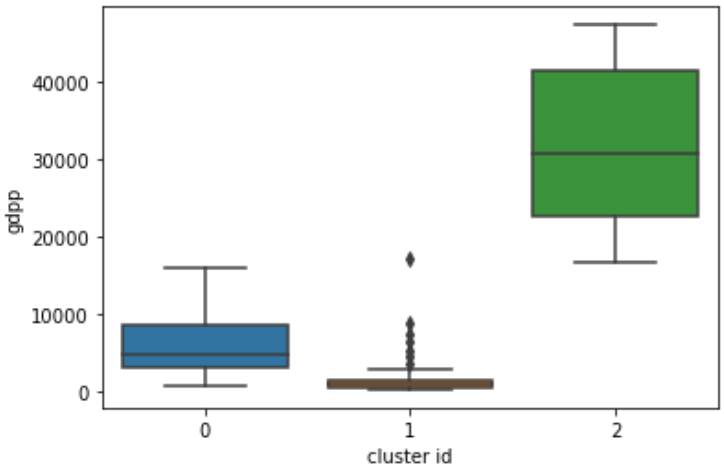
- The value of the silhouette score range lies between -1 to 1.
- A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
- A score closer to -1 indicates that the data point is not similar to the data p oints in its cluster.

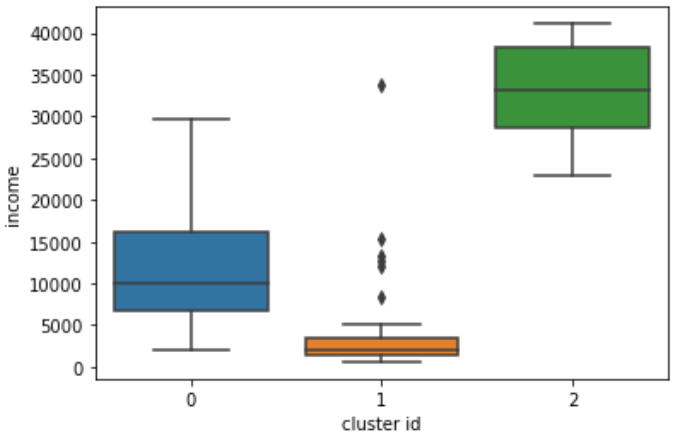


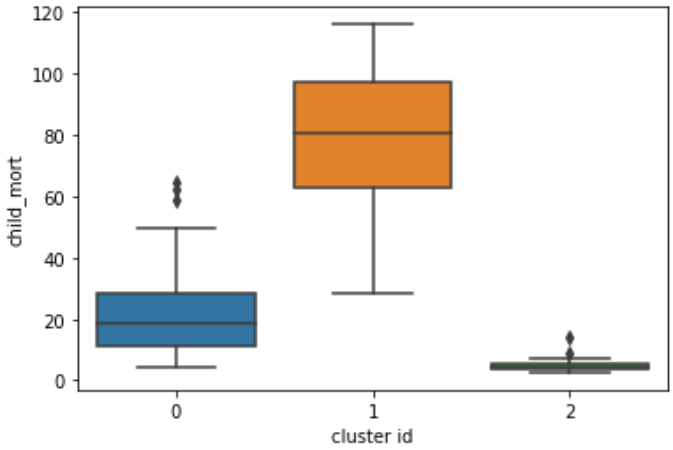
Silhouette Score

From the plot of silhouette score above, we can see that we have the maximum at 2, and next maximum is at 3. So, we can go with 3.

VISUALIZING THE CLUSTERS:







CLUSTER PROFILING:

From cluster profiling in K- means clustering we can see that :

- Cluster 0 is having medium income, GDP and child mortality
- Cluster 1 is having very Low income, very Low GDP but High child mortality
- Cluster 2 is having the High income, High GDP and very Low child mortality

We saw in cluster profiling that cluster 1 is having low income, low GDP and High Child Mortality. So, we can say that countries under cluster 1 are in need of aid.

After sorting the countries according to low GDP, low income and high child mortality rate in cluster 1, the 5 countries which are in direst need of aid are:

1. Barundi
2. Liberia
3. Congo, Dem. Rep.
4. Madagascar
5. Mozambique

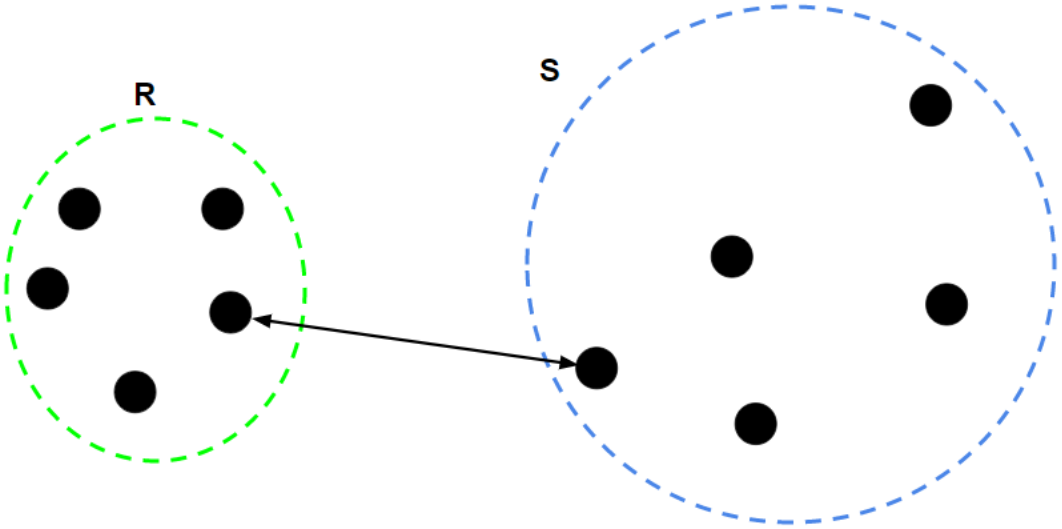
HEIRARCHICAL CLUSTERING:

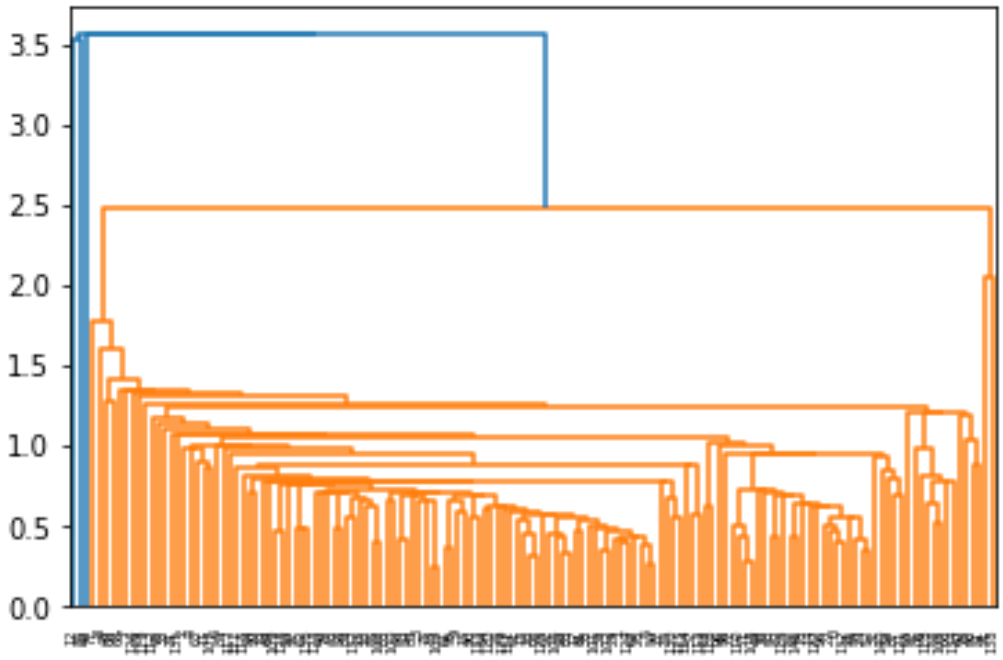
The Hierarchical clustering is a commonly used text clustering method, which can generate hierarchical nested classes. It clusters similar instances in a group by using similarities of them. This requires the use of a similarity (distance) measure which is generally Euclidean measure in general, and cosine similarity for documents. Therefore, a similarity (distance) matrix of instances has to be created before running the method. Hierarchical clustering can be categorized into two; agglomerative (bottom-up) and divisive (top-down) clustering which are explained in [2] with the names AGNES, and DIANA.

There are two linkages used in hierarchical clustering

Single Linkage:

In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two closest points



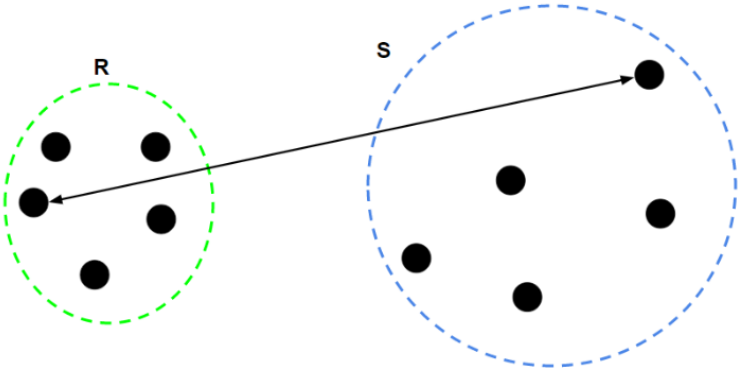


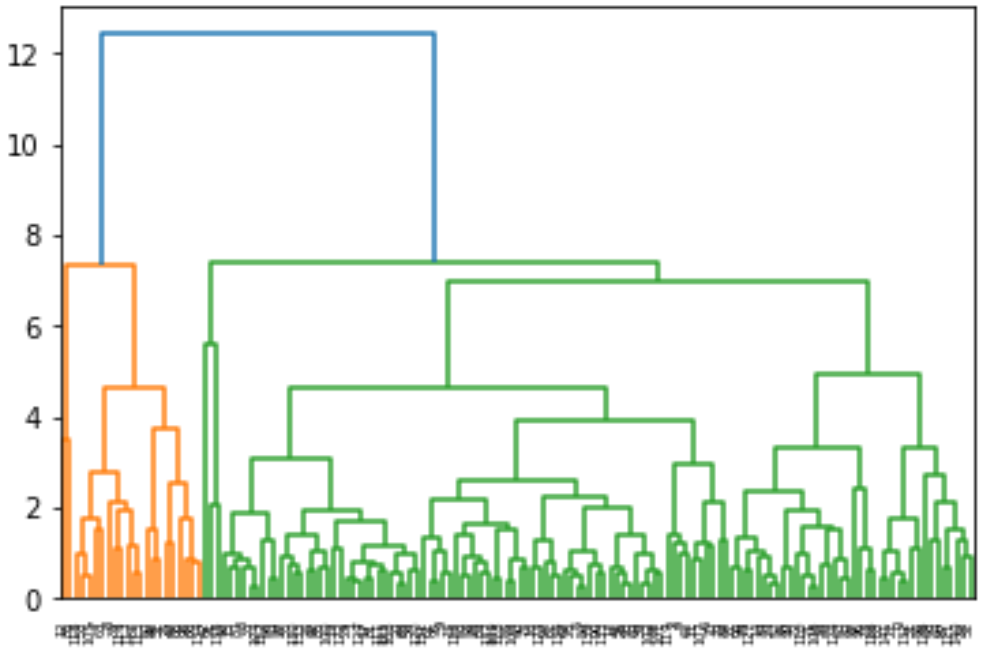
Single Linkage

We clearly see, single linkage doesn't produce a good enough result for us to analyse the clusters. Hence, we need to go ahead and utilise the complete linkage method and then analyse the clusters once again.

Complete Linkage:

In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two furthest points.

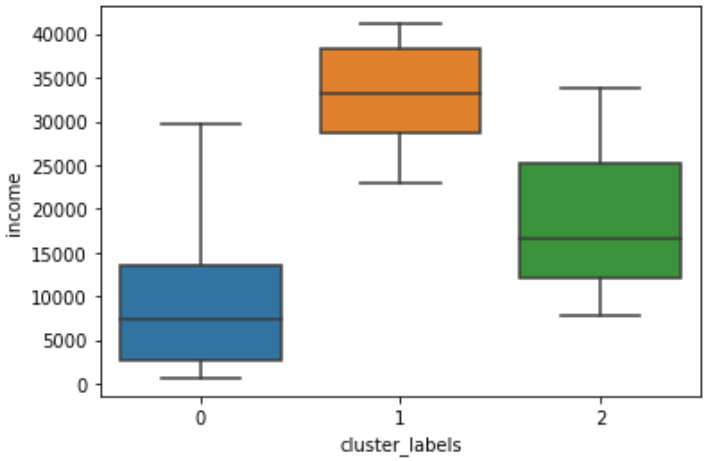


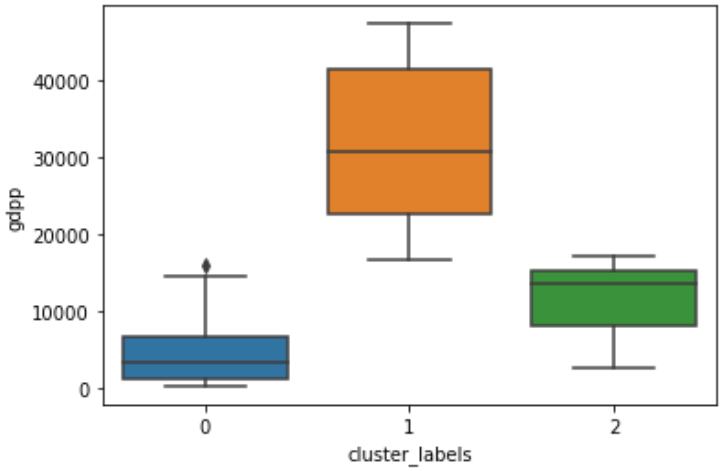


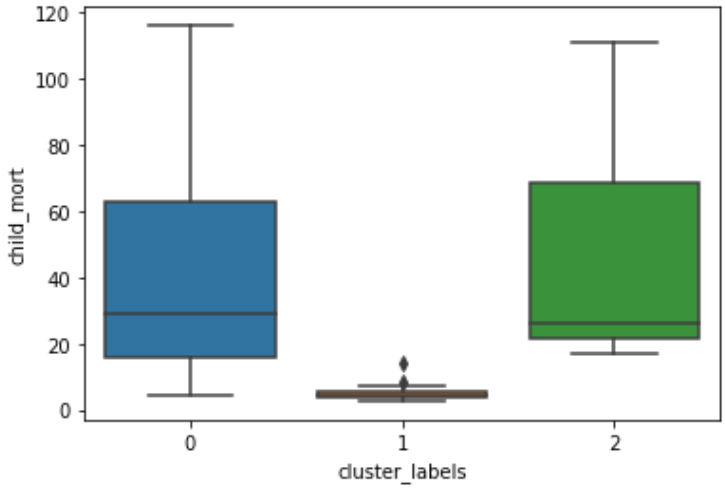
Complete Linkage

The dendrogram generate from Hierarchical Clustering with complete linkage shows 3 clusters distinctly. Hence 3 clusters are chosen.

VISUALIZING THE CLUSTERS:







CLUSTER PROFILING:

Cluster 0 have very low GDP, cluster 2 have high GDP and cluster 1 has moderte GDP

From cluster profiling using hierarchical clustering we can see that:

- Cluster 0 is having the High child mortality, low GDP, and low income.
- Cluster 1 is having Low child mortality, High income, and High GDP
- Cluster 2 is having very High child mortality, high income, and GDP

We saw in cluster profiling that cluster 0 is having low income, low GDP, and High Child Mortality. So, we can say that countries under cluster 0 are in need of aid.

After sorting the countries according to low GDP, low income and high child mortality rate in cluster 0, the 5 countries which are in direst need of aid are:

1. Barundi
2. Liberia
3. Congo, Dem. Rep.
4. Madagascar
5. Mozambique

CONCLUSIONS:

We have analyzed both K-means and Hierarchical clustering and found clusters formed in both are not identical. From K means clustering we got better clusters compared to Hierarchical clustering. The clusters formed in Hierarchical clustering are not great.

Cluster 1 in K-means is the better cluster we got with High child mortality, low income, and low GDP. So, we will proceed with the clusters formed by K-means and based on the information provided by the final clusters we will deduce the final list of countries which are in need of aid.

Final list of the 5 countries we got are:

1.Burundi

2.Liberia

3.Congo, Dem. Rep.

4.Madagascar

5.Mozambique