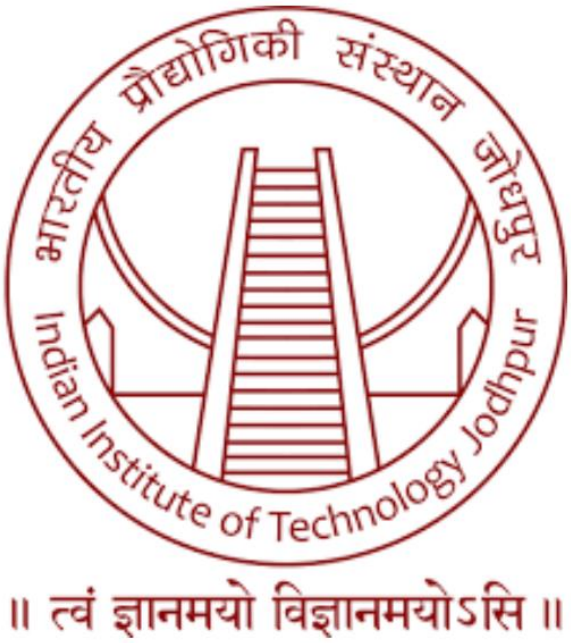
Pattern Recognition and Machine Learning

Indian Institute of Technology, Jodhpur

Minor Course Project



Project-5

"Determining the overall development of the country using the 'Country Data' set"

Dev Agarwal – B21CS023

Devang Shrivastava – B21CS024

Keshav Malani – B21CS038

Abstract

Help International NGO has raised $10 million to help various countries. The NGO's CEO must decide how to spend this money strategically and effectively. As a result, the CEO must make a choice on which countries require the most assistance. Hence Our job here is to categorize the given countries on various socio-economic and health factors that determine the overall development of the country.

We have achieved this goal by applying unsupervised machine learning methods to form different groups (categories or clusters). We have applied the KMeans and Hierarchal Clustering method on 4 different types of datasets to compare. We have finally tried to analyse the obtained results using decision trees and tried to demonstrate the effects of features in the obtained clustering.

Dataset

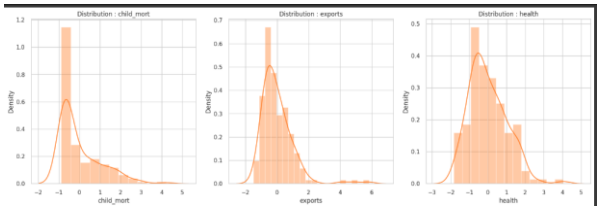
The main dataset contains 167 rows and 10 columns.

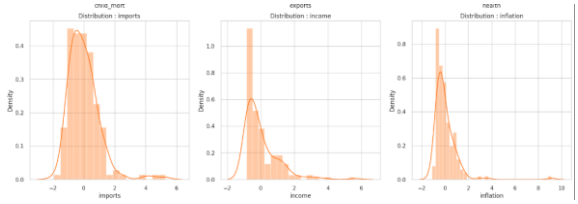
1. Country: Name of the given country.
2. Child Mortality: number of child deaths under the age of 5 per 1000 live births.
3. Exports: Exports of goods and services per capita. Given as %age of the GDP per capita
4. Health: Total health spending per capita. Given as %age of GDP per capita
5. Imports: Imports of goods and services per capita. Given as %age of the GDP per capita
6. Income: Net income per person
7. Inflation: The measurement of the annual growth rate of the Total GDP
8. Life Expectancy: The average number of years a new-born child would live if the current mortality patterns are to remain the same
9. Total Fertility Rate: The number of children that would be born to each woman if the current age-fertility rates remain the same.
10. GDPP: The GDP per capita. Calculated as the Total GDP divided by the total population.

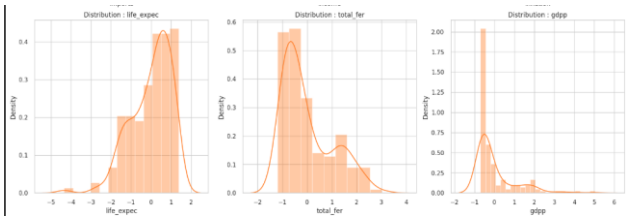
Pre-processing

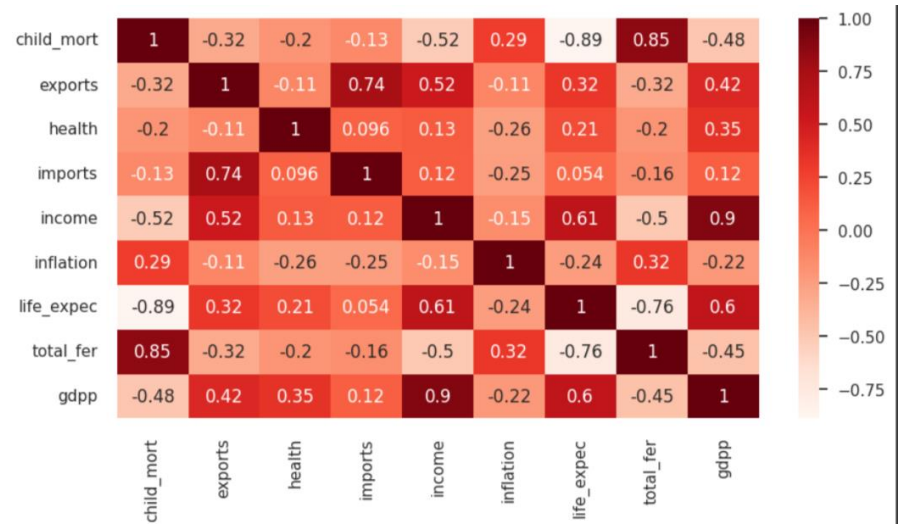
After preliminary data loading,

- we dropped the country feature and stored it.
- Standardized the remaining features and visualized their distribution as well as the heatmap.









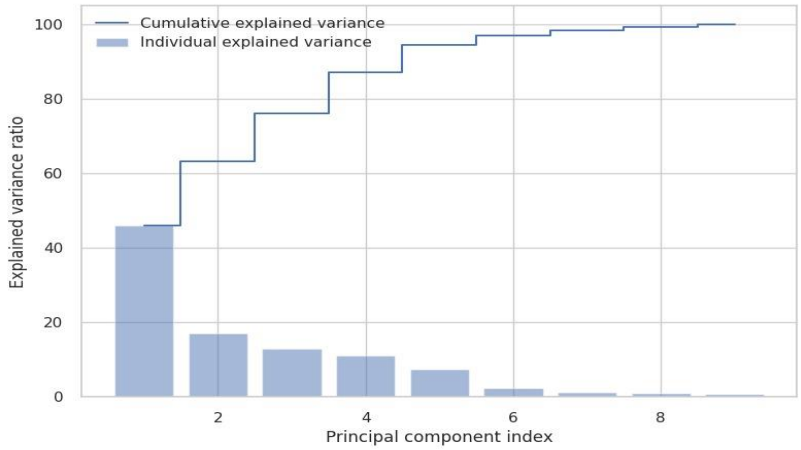
Observations:

- 4 features had near gaussian distribution whereas the remaining 5 had non-gaussian distribution.
- Gaussian distribution refers to the symmetric distribution of data.
- Few feature pairs had high correlation:
    - Life Expectancy vs child Mortality: -0.89
    - Total fertility vs child Mortality: 0.85
    - Total fertility vs Life expectancy: -0.76
    - Imports vs exports: 0.74
    - GDPP vs income: 0.9

Dimensionality Reduction

We have transformed the scaled data into 3 different data frames:

1. Transforming the scaled dataset by applying PCA without any feature selection.



    - By the graph we decide to use the number of principal components as 5 as 94.5% variance is covered

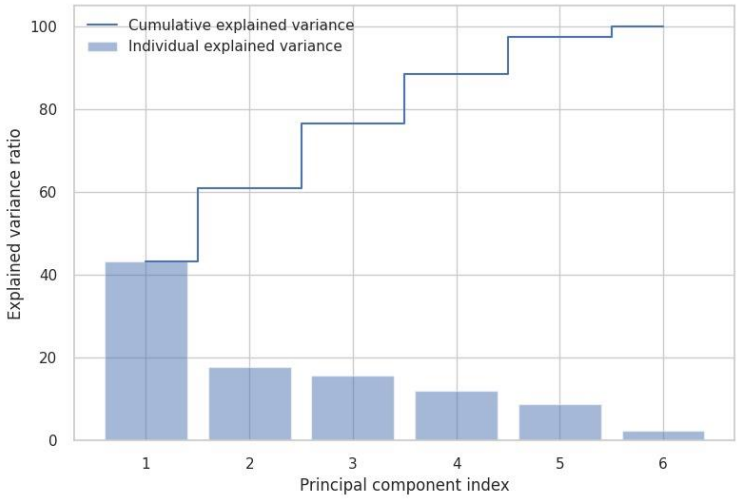
2. Transforming the dataset by doing feature selection. According to the heatmap of the dataset, we identified pairs of features that had high correlation:

- Health sector → Life Expectancy vs Child Mortality: -0.89
- Finance sector →GDPP vs Income: 0.9
- Trade sector → Imports vs Exports: 0.74
- And hence we chose to remove one feature each.

- Final Features:

    - Child mortality
    - Imports
    - GDPP
    - Health
    - Inflation
    - Total Fertility

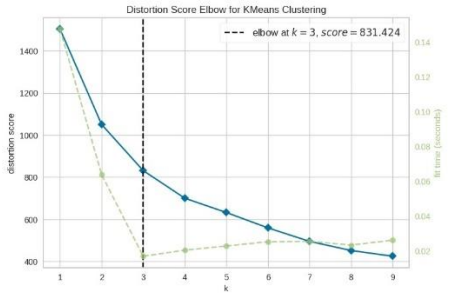
3. In the third transformation we have further applied PCA to these selected features.



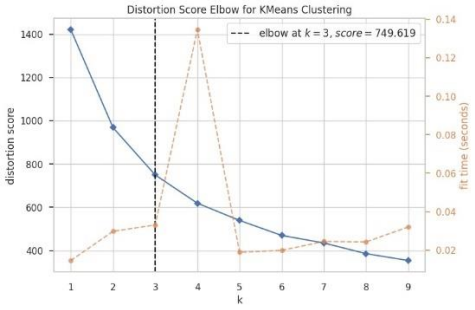
- By the graph we decide to use the number of principal components as 5 as 97.5% variance is covered.

KMeans

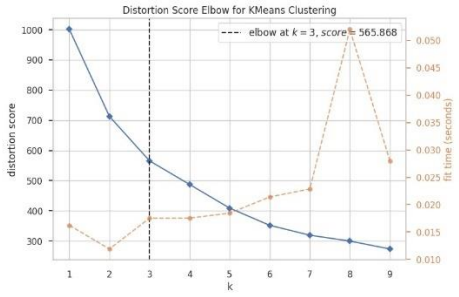
We plotted the graph of the distortion score vs the number of clusters. With the elbow method, we identified the best value of k in each case.



for normally scaled dataset
silhouette score:0.225

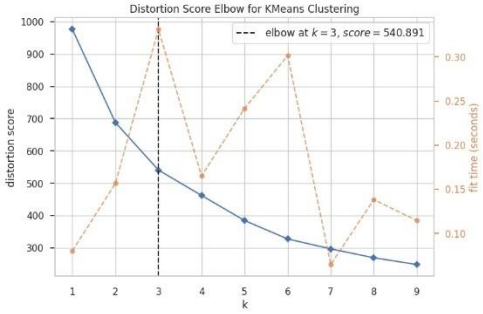


PCA on normally scaled dataset
silhouette score:0.283



Feature Selected dataset

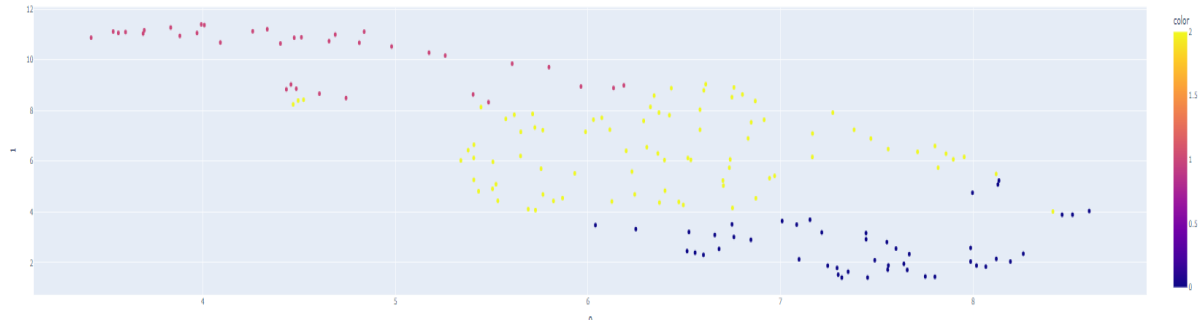
silhouette score:0.281



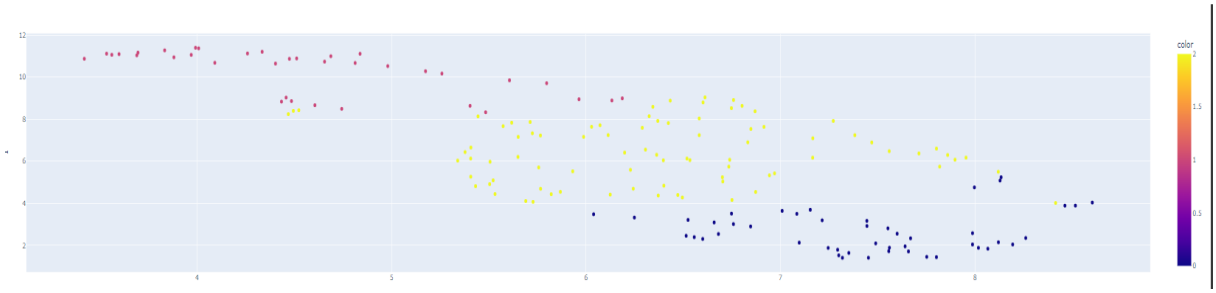
PCA on Feature Selected dataset

silhouette score: 0.2814

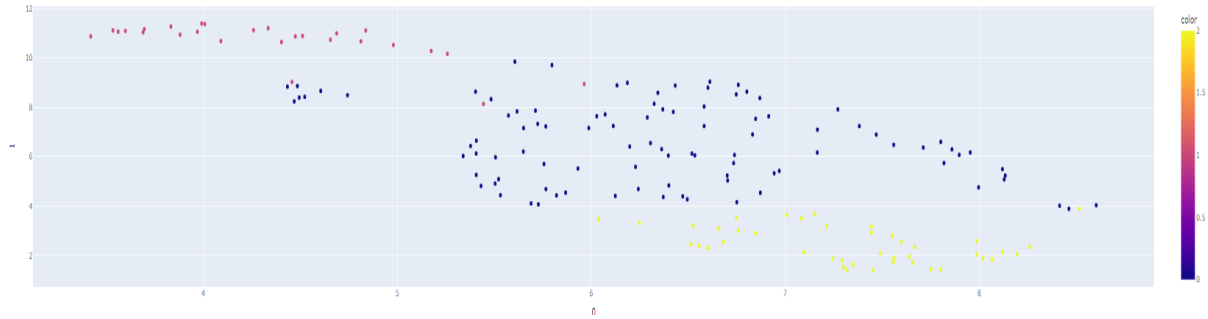
We then clustered the data according to the best k and found the respective silhouette score.



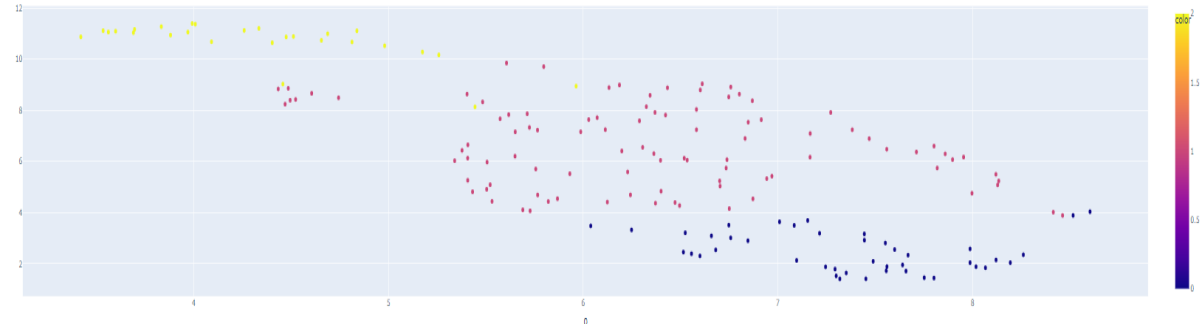
for normally scaled dataset



for normally scaled dataset: PCA applied



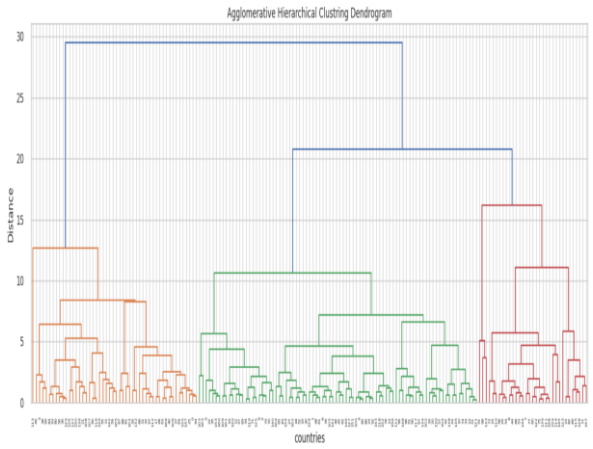
Feature selected dataset



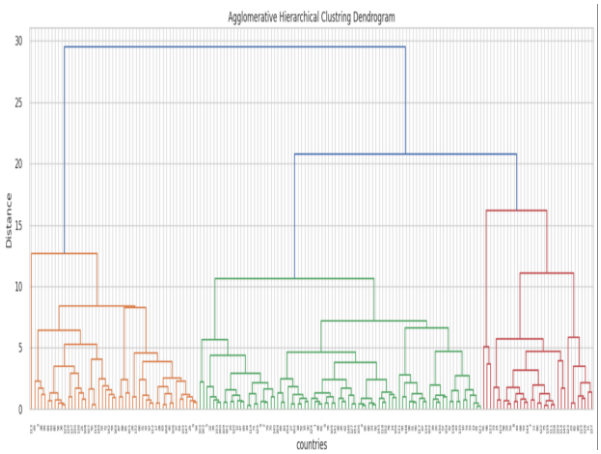
Feature selected dataset: PCA applied

Hierarchal Clustering

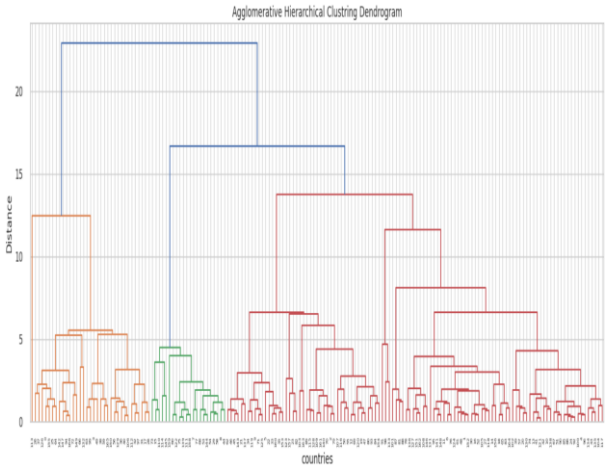
In the hierarchical clustering, we first plotted the dendrogram using the agglomerative approach. We then manually clustered the data for different numbers of clusters according to the dendrogram and picked out the best value according to the silhouette score. The best value k (no. of clusters) in each case was 3.



for normally scaled dataset
silhouette score:0.280

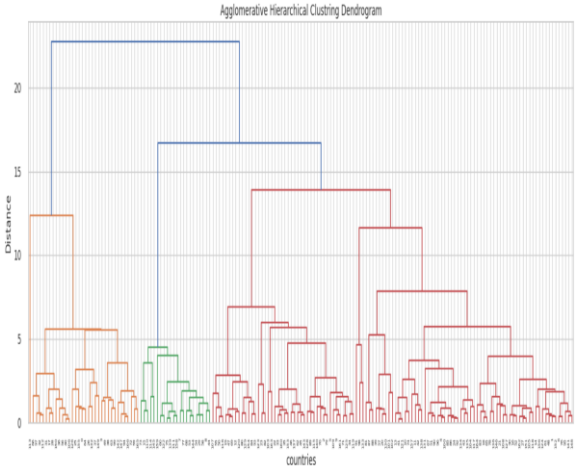


PCA on normally scaled dataset
silhouette score:0.245



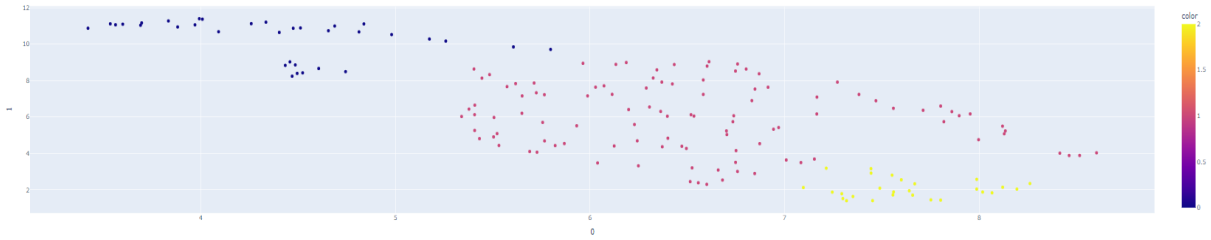
Feature Selected dataset

silhouette score:0.244

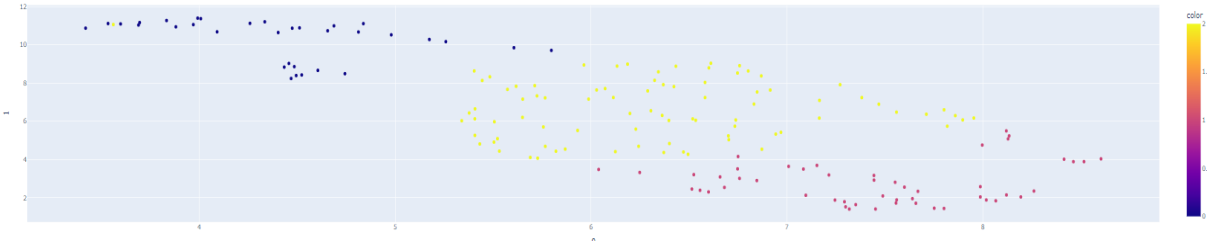


PCA on Feature Selected dataset

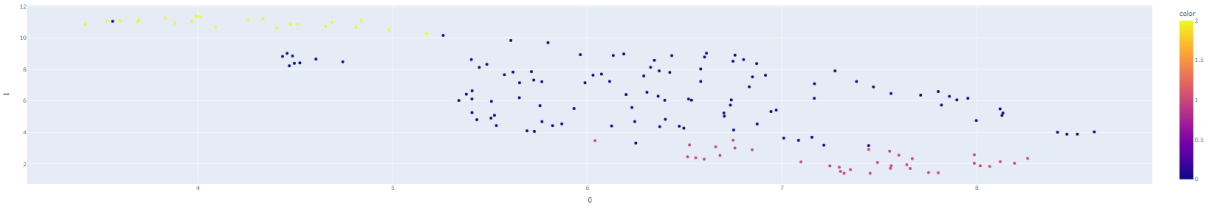
silhouette score: 0.241



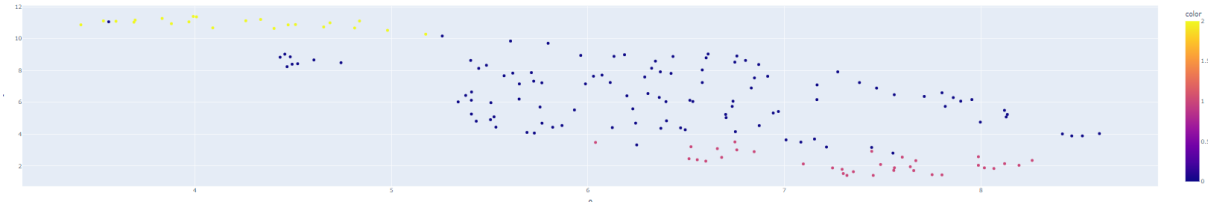
for normally scaled dataset



for normally scaled dataset: PCA applied



Feature selected dataset



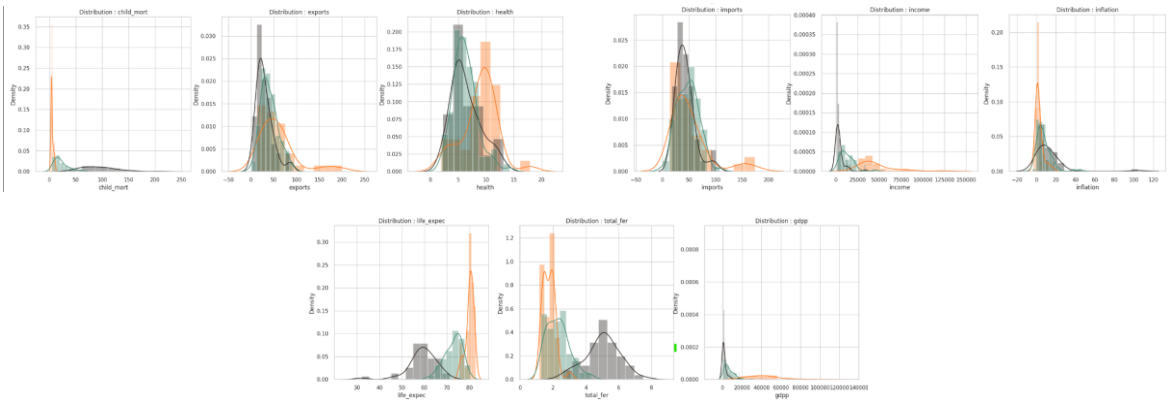
Feature selected dataset: PCA applied

Data Analysis

Following was the result of our project:

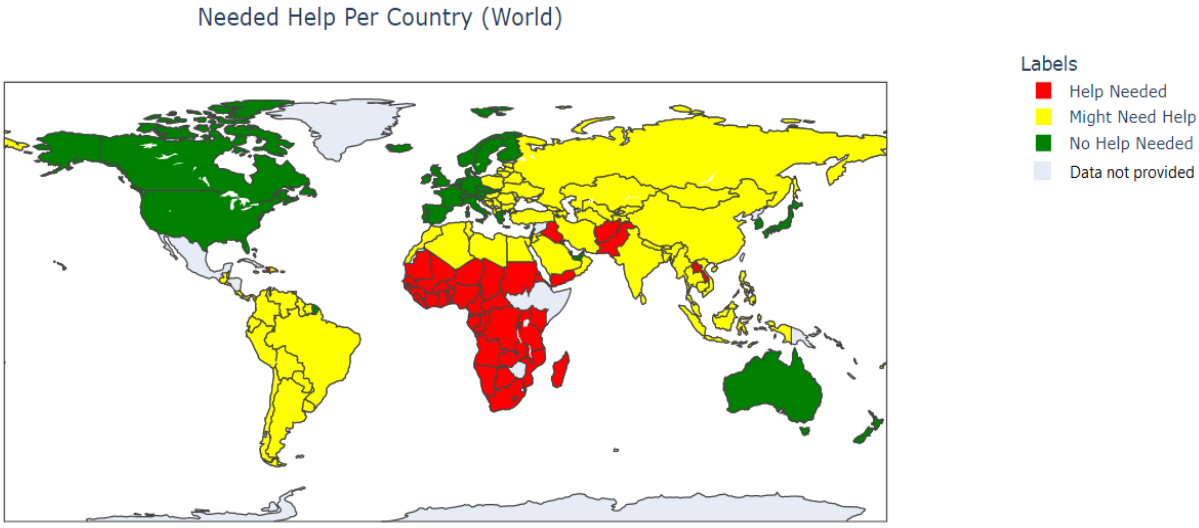
Dataset	Silhouette Score KMeans	Silhouette Score Hierarchal
Normally scaled	0.225	0.280
Normally scaled - PCA	0.283	0.245
Feature selected	0.281	0.244
Feature selected - PCA	0.2814	0.241

We observe that when we apply K-Means with the normally scaled – PCA applied data we observe the best clustering with a silhouette score: of 0.283.



Class-wise distribution with different features

Here is the respective categorization of countries:



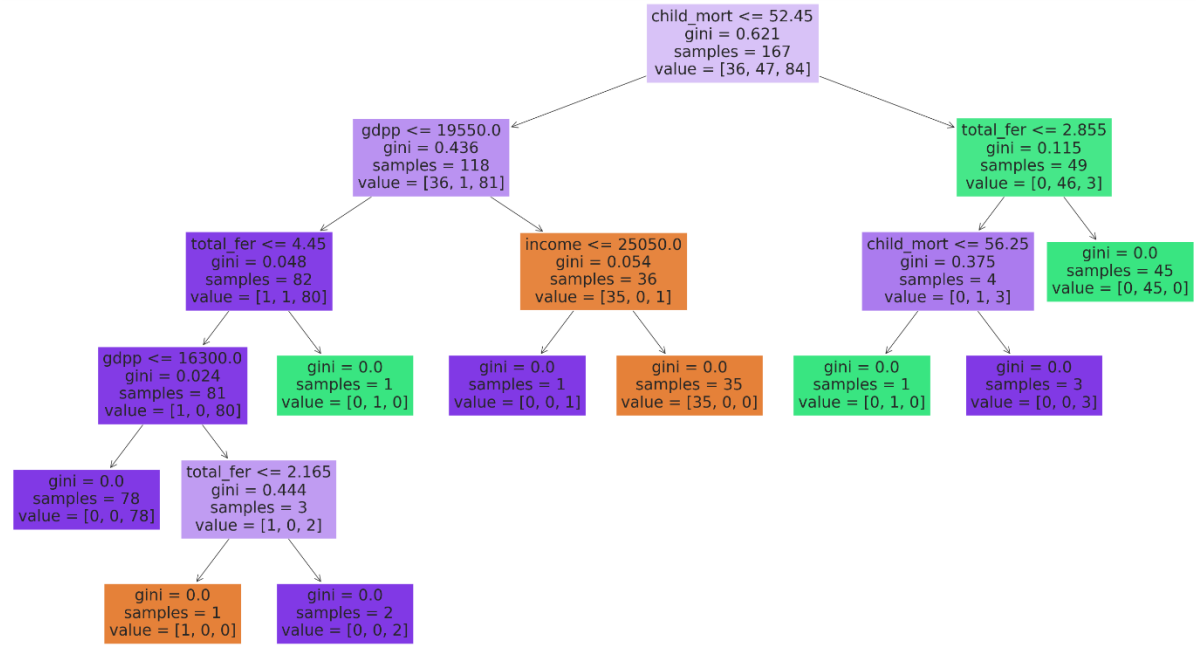
Total Countries: 167

Developed (No help needed): 36

Developing (Might need help): 84

Under Developed (Need help): 47

To understand the dependence of clusters on the features we applied decision tree learning once we had the final labelling (based on the best clusters). Here is the result:



From the above tree labeling, we observe that:

- We see that the decision tree has used the features-“Child Mortality, GDPP, Total Fertility, and Income” in classifying the data.
- This tells us that these features played an important role in differentiating the developed, developing, and underdeveloped countries
- For example, if the child mortality and the total fertility rate of a country are high, then there is a very high chance that it is an underdeveloped country and needs more help.
- We observe that the countries with high GDPP and income are very likely to be classified as developed countries which do not need any help.