Country Data Project

Group Members

Name	Roll Number
Shalin Jain	B21CS070
Sameer Sharma	B21CS066
Prakhar Gupta	B21AI027

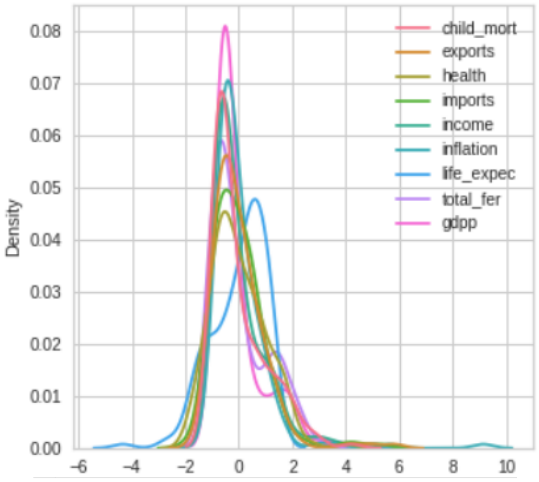
Task

Our objective is to categorize the countries using socio-economic and health factors that determine the overall development of the country.

Idea

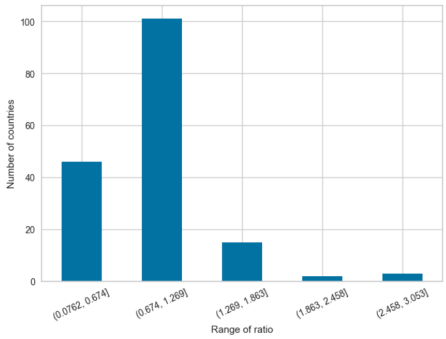
Since it is a problem where labels of the datasets are not given we have to apply unsupervised learning. We will use different clustering algorithms to cluster the data under different labels. We will use the following clustering algorithms for classification of the dataset.

1. KMeans Clustering
2. Hierarchical Clustering
3. DBSCAN Clustering
4. Gaussian Mixture Clustering



The reason for using the Gaussian Clustering algorithm is that the nature of the data as seen in the above graph is normal and since the number of countries is > 160 we can use Gaussian Mixture.
And the remaining are mostly popular algorithms used for clustering the data.

The bar plot is a plot of export-import ratio which shows the countries developing nature in the world market.



Observation we get from visualizing the data are as below: -

- Life expectancy is strictly inversely proportional to total fertility and child mortality while directly proportional to income. Thus child mortality and total fertility are directly proportional to each other. Exports are directly proportional to income and imports. income is directly proportional to life expectancy, exports while inversely proportional to child mortality. As we can see that income and gdpp are highly positively correlated, we can infer that countries with higher income, exports and imports usually have higher gdpp which tends to better life expectancy.
- Around 50% of the countries have equal amounts of exports and imports showing their developing nature. Countries with less exports but higher imports show their underdeveloped nature as depicted by the range of ratio (0.0762,0.674]. There are some countries showing their developed nature with higher export import ratio (>1).

Experiments or Process

Transformation of the data

We applied the PCA transformation of the data and by checking the cumulative_variance_ratio we saw that the optimal value of the n_components is 5. We used the obtained value of optimal components and transformed the dataset into a dataset with five feature columns as 'PC1','PC2','PC3','PC4' and 'PC5. For all the further processes we used this transformed dataset.

Clustering of the Data Points

We performed different clustering techniques mentioned above for clustering the data points.

1. KMeans Clustering
   For getting the optimal number of clusters we plotted the KElbowVisualize for silhouette score as well as the distortion score. We saw that the optimal number of clusters obtained is 4. We then trained the KMeans Model with n_clusters as 4 and obtained the following results: -

```
Value counts:
Class 0: 86     51.50%
Class 1: 47     28.14%
Class 2: 31     18.56%
Class 3: 3      1.80%
```

2. GaussianMixture Clustering
   For getting the optimal number of clusters we plotted the Bayesian information criterion against the number of clusters. We chose the minimum value of BIC for the optimal number of clusters. We then trained the GaussianMixture Model with n_clusters as 5 and obtained the following results: -

```
Value counts:
Class 0: 63     37.72%
Class 1: 52     31.14%
Class 2: 32     19.16%
Class 3: 3      1.80%
Class 4: 17     10.18%
```

3. Hierarchical Clustering
   For getting the optimal number of clusters we plotted the dendrogram and then performed the agglomerative clustering with the number of clusters as 3. The obtained results are as follows: -

```
Value counts:
Class 0: 33      19.76%
Class 1: 50      29.94%
Class 2: 84      50.30%
```

4. DBSCAN Clustering
   For getting the optimal value for epsilon for DBSCAN Clustering we used the Nearest Neighbour technique and obtained the value of epsilon as 1.5. We then trained the DBSCAN model and obtained only 2 labels and rest points as noise.

It will be unfair to distribute the money equally among the countries since they differ in their socio economic status. There are some countries which are nearly on the verge of becoming developing nations and don't need much support and funding when compared to rest. So, we are going to repeat the process of clusterring as we have done on the whole country dataset but this time we will use the dataset of countries that lie in the under developed cluster. Results for these experiments are in the results section.

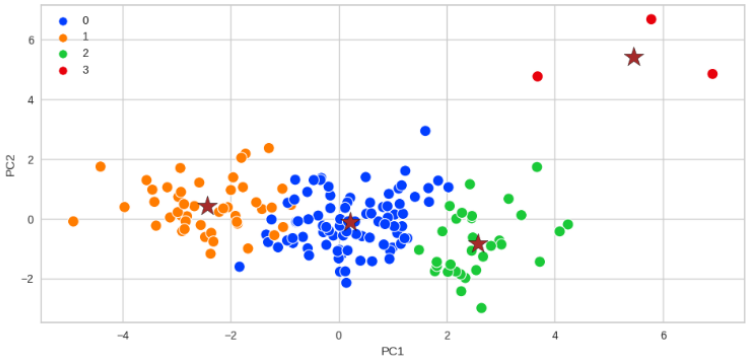
Results

Silhouette score for all the algorithms we implemented are as follows

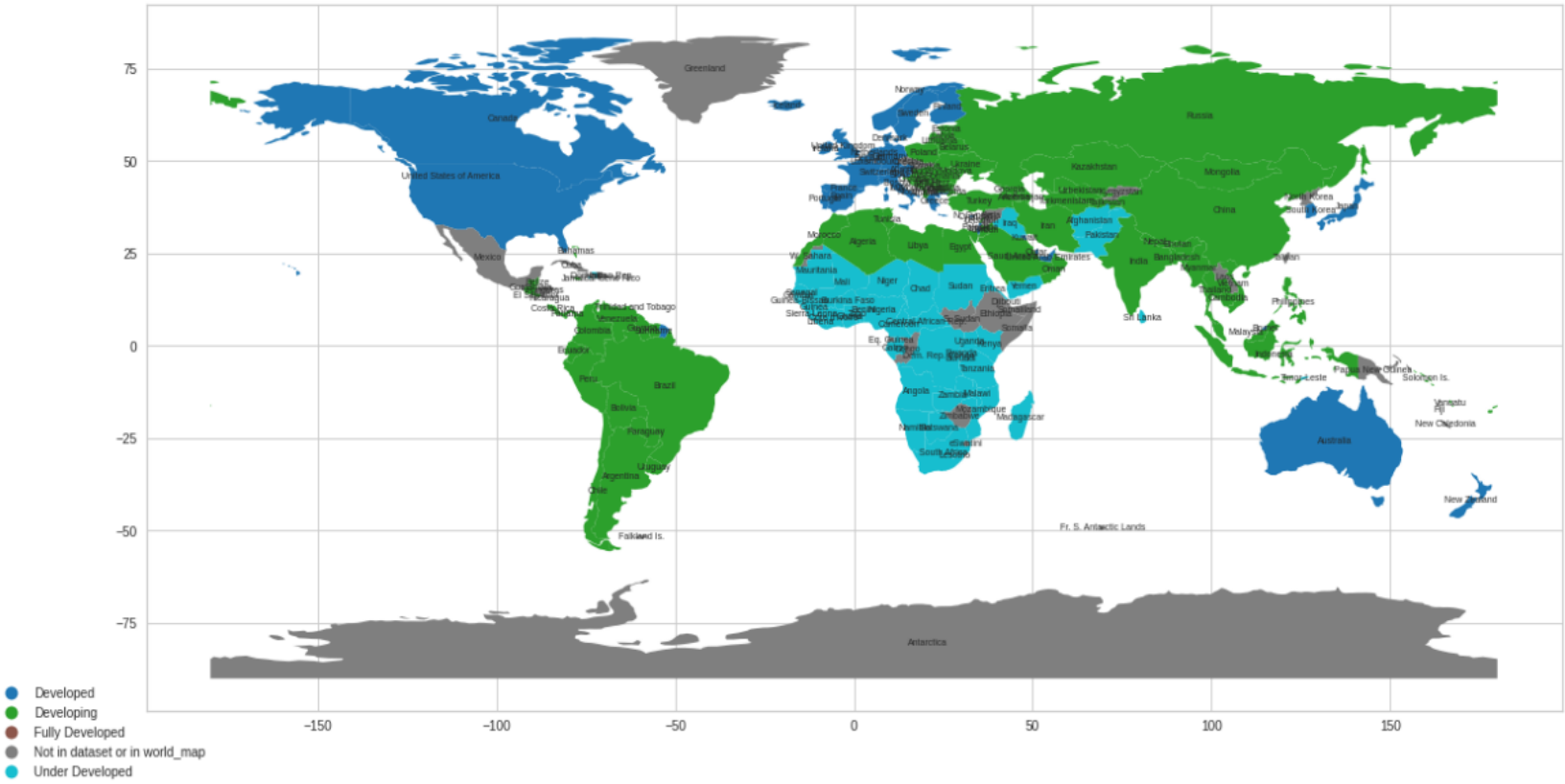
1. KMeans clustering: 0.32718347402877207
2. Gaussian mixture clustering: 0.1743791369752946
3. Hierarchical (Agglomerative) clustering: 0.30427188416905565
4. DBSCAN clustering: 0.1929253329808842

From the silhouette score of different clustering algorithms, we can observe that KMeans has highest Silhouette score and thus performing best on the dataset. Therefore, we will choose KMeans to cluster the data points and predict the countries that are in need of funding from HELP.
By observing the centroids of the clusters formed by KMeans, we can assign different country ranks to each cluster which are as follows:

1 -------> Under Developed
0 -------> Developing
2 -------> Developed
3 -------> Fully Developed

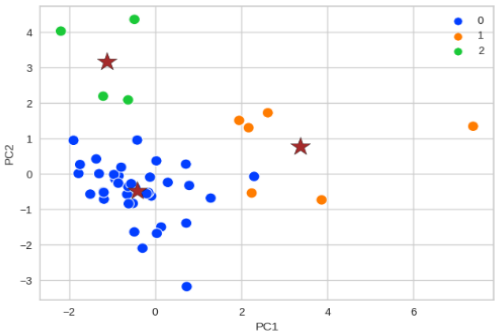


The obtained labels can be plotted on the world map as follows: -



The countries that are colored in cyan are Under Developed nations and are in need of funding from the HELP organization.

Now, after repeating the clustering process for the dataset which contains only data of under developed countries, we got the following silhouette scores
Gaussian 0.21882388381167545
KMeans 0.15455494211167753
alggomerative 0.3601544027088434
Dbscan 0.22315215063140725



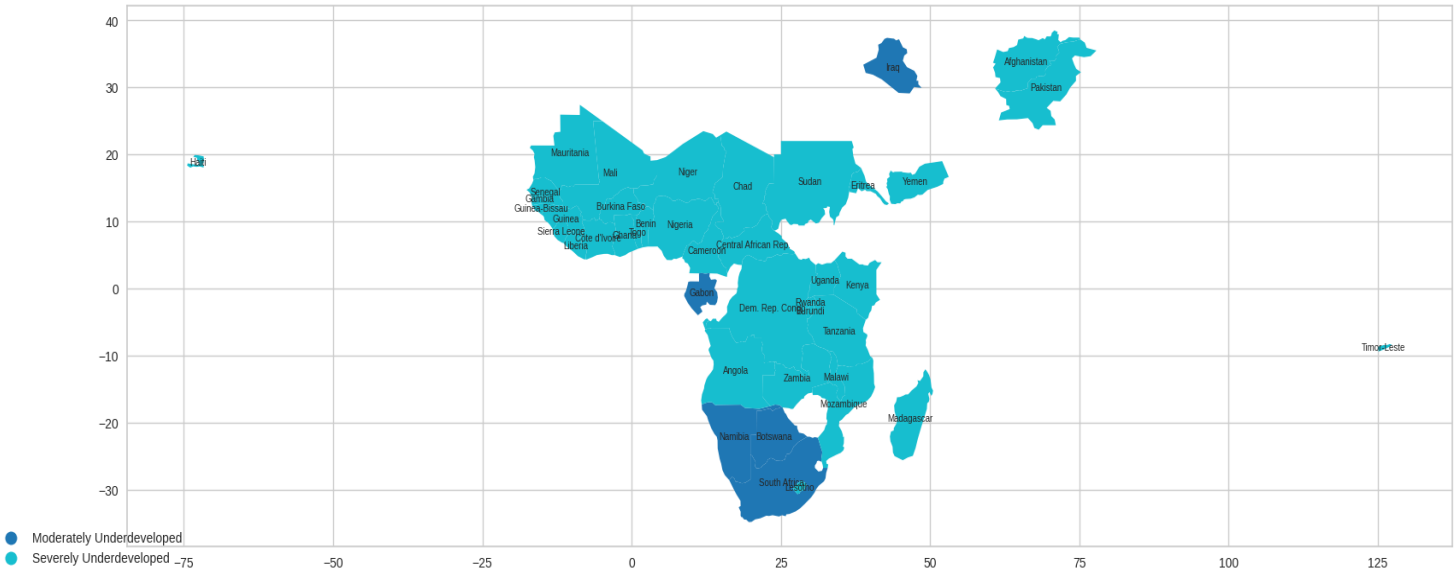
So, hierarchical clustering (alggomerative) performed best. So, we will use hierarchical clustering to cluster the datapoints.

Now, on the basis of centroids of clusters, we can see that countries in cluster 1 has highest mean gdpp of around 8000 which is even higher than that of developing nations (around 6000). These are the countries which are on the verge of becoming developing nations but due to some factors like high inflation and child mortality they are still under developed.Countries in the other two clusters have nearly same stats and thus we can merge them to form a single cluster.

Now we have only two clusters. Let's say one to be 'Moderately under developed' (includes cluster 1) and other 'severely under developed' (includes cluster 0 and 2)

Now HELP organisation can apply its own policy to divide the money among moderately and severely under developed countries. Let us consider they adopt a policy to divide the money such that a severely under developed country will get three times the money that a moderately under developed country will receive. In that case, we can divide the money as follows:
232557.0 $ can be given to the severely under developed countries individually
77519.0 $ can be given to the moderately under developed countries individually

The obtained labels for the underdeveloped countries are plotted on the world map as follows



Links

The Project can also be viewed from the given link
https://github.com/SameerSharma-57/Country_data_project