Strategic Allocation of Aid

PRML Minor Project | PROJECT 5

Kamuju Aashish | B21AI001
Poojita Oppangi | B21CS055
Kovidh Pothireddy | B21AI026

I. INTRODUCTION

In today's world, many countries face severe economic and health issues. Non-Governmental Organizations (NGOs) like HELP International work to help these countries in dire need of aid. However, it is crucial for HELP International to prioritize its aid and allocate its resources effectively. With a budget of 10 million dollars, HELP International needs to determine which countries are in the direst need of aid. Our task is to categorize the countries based on socio-economic and health factors that determine their overall development. This categorization will enable HELP International to identify the countries that require immediate assistance and allocate resources accordingly.Our approach involves collecting and analyzing data on various socio-economic and health factors, such as GDP, life expectancy, infant mortality rate, literacy rate, and poverty rate. We will then use machine learning algorithms to cluster the countries based on these factors.

II. DATA VISUALIZATION AND PREPROCESSING

To begin our analysis, we collected data on various socio-economic and health factors for 167 countries. After obtaining the data, we performed a preliminary data exploration to identify missing values and outliers. We found that there are no missing values.We also identified some outliers which has very good socio-health factors , but we won't remove them as every country has to be considered while distributing aid. Next, we performed data visualisation to gain insights into the data and identify any patterns.

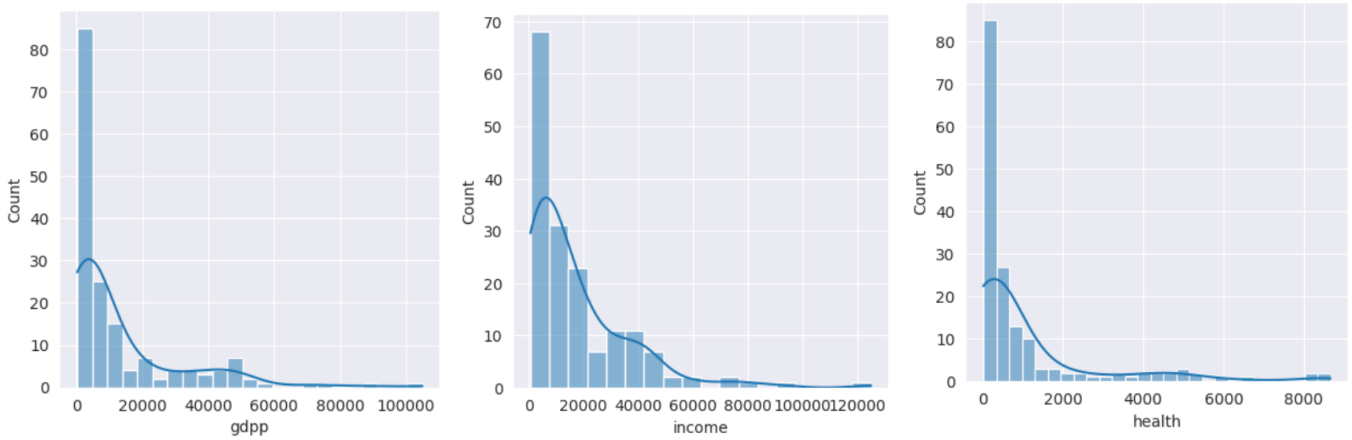

Figure 1: Plots of histogram

Preprocessing

A. Converting columns from percentages to actual values:

The 'exports', 'health', and 'imports' columns were expressed as percentages of GDP. We converted these columns to their actual values using the formula: actual value = percentage value * GDP.

B. Normalizing the data:

We used StandardScaler from the sklearn library to normalize the data. Normalization involves scaling the features to have a mean of 0 and a standard deviation of 1. Normalization of the data is crucial for machine learning models as it ensures that the range of values for each feature is similar. This eliminates the possibility of any one feature dominating the model's output.

C. Creating a new dataframe:

We created a new dataframe, 'scaled df,' from the normalized data. This new dataframe contains the same columns as the original dataset, but the values have been standardized.

III. MACHINE LEARNING MODELS

Principal Component Analysis (PCA):

It is a popular technique for dimensionality reduction that works by identifying the principal components of the dataset. It helps us reduce the number of features or variables in the dataset while retaining as much information as possible. It also helps improve the performance of machine learning models by reducing the noise and computational complexity of the dataset.

K Means clustering:

It is a partition-based clustering technique that aims to divide data into k different clusters based on the similarity of observations to each other. K Means assigns each observation to a cluster by minimizing the sum of squared distances between the centroid of the cluster and all the observations in that cluster.

Hierarchical clustering:

It is a hierarchical method of clustering where the data is successively divided into smaller clusters, which can be either merged or divided further. The result is a tree-like structure, called a dendrogram, which can be cut at a particular level to form a certain number of clusters.

DBSCAN clustering:

It is a density-based clustering technique that groups together points that are close to each other in high-density regions and separates points in low-density regions. DBSCAN assigns each point to a cluster based on its density and the density of its neighboring points.

Ensemble clustering:

It is a technique that combines the results of multiple clustering algorithms to improve the overall clustering performance. Ensemble clustering can be achieved using multiple clustering algorithms and averaging their results, combining different clustering results using a voting mechanism, or combining different clustering algorithms in a hierarchical or sequential manner.

Scoring methods to evaluate the performance of clustering algorithms and determine the quality of the clustering results:

Silhouette Score: Silhouette score is a measure of how well each data point fits into its assigned cluster, based on both the distance to the other points in the cluster and the distance to the nearest points in other clusters.
Calinski-Harabasz Index: The Calinski-Harabasz index measures the ratio of between-cluster variance to within-cluster variance for all clusters. A higher score indicates a good clustering result.
Davies-Bouldin Index: The Davies-Bouldin index measures the average similarity between each cluster and its most similar cluster, compared to the average dissimilarity between each cluster and its least similar cluster. A higher score for Silhouette Score, Calinski-Harabasz Index indicates better clustering performance, while a lower score for Davies-Bouldin Index indicates better clustering.

IV. MODEL BUILDING

Pipeline:

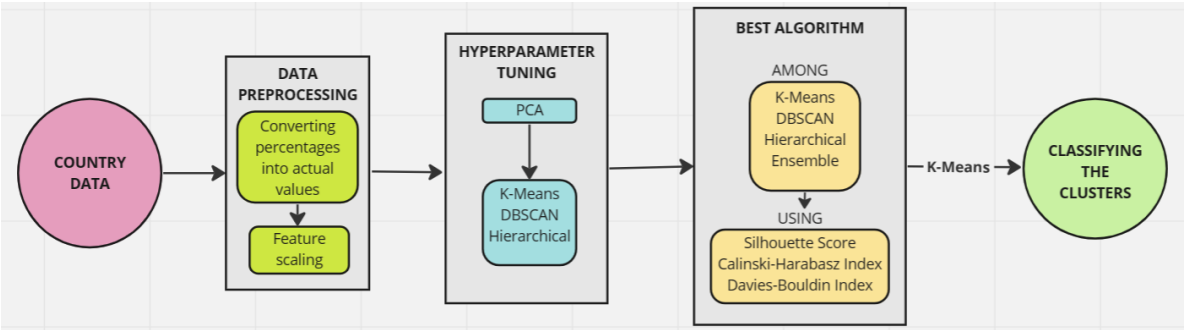

Figure 2: Pipeline

Hyperparameter tuning of PCA:

We performed PCA on our preprocessed dataset to determine the number of components required to retain (90-95) percentage of the variance in the dataset. We instantiated a PCA object and fitted it to the preprocessed data. We then calculated the explained variance ratio and the cumulative explained variance ratio to determine the number of components required to retain the desired percentage of variance and specified the number of components to be 2. These principal components are linear combinations of the original features of the dataset, which are now standardized and scaled.

Hyperparameter tuning of clusters:

Then , we perform hyperparameter tuning for three different clustering algorithms: K-means, Hierarchical, and DBSCAN. For K-means and Hierarchical clustering, we iterate over different numbers of clusters and calculate the Silhouette score for each cluster to determine the optimal number of clusters for each algorithm. For DBSCAN clustering, we iterate over different values of eps and min samples hyperparameters and If there are enough clusters, we print the Silhouette score for that set of hyperparameters. Overall, hyperparameter tuning allows us to identify the optimal hyperparameters for each clustering algorithm, which can improve the quality of the clustering results.

Figure 3: Plots of Silhouette Scores

Training the model

Then, the KMeans ,DBSCAN and Hierarchical clustering algorithms are trained on the data, and the linkage matrix is computed to visualize the dendrogram and determine the optimal number of clusters.The Ensemble Clustering algorithm is used to combine the results of the three clustering algorithms, by assigning each data point to the cluster that appears most frequently across the three algorithms.

V. CLUSTERING MODEL COMPARISION

The silhouette score, Calinski-Harabasz index, and Davies-Bouldin index are used to evaluate the performance of each clustering algorithm, as well as the Ensemble Clustering algorithm.

Clustering Model	Silhouette score	Calinski - Harabasz Index	Davies - Bouldin Index
K-Means	0.561	273.464	0.436
Hierarchical	0.551	259.034	0.455
DBSCAN	0.615	17.401	1.721
Ensemble	0.475	109.173	0.493

Best Clustering Algorithm:

As we can see , KMeans Clustering algorithm have the best scores inturn producing best clusters. Now, we can print countries in each cluster. Then , we creates a scatter plot of the data after applying the KMeans clustering algorithm. Each point in the scatter plot represents a country, and the color of the point corresponds to the cluster that the country belongs to, as determined by the KMeans algorithm.

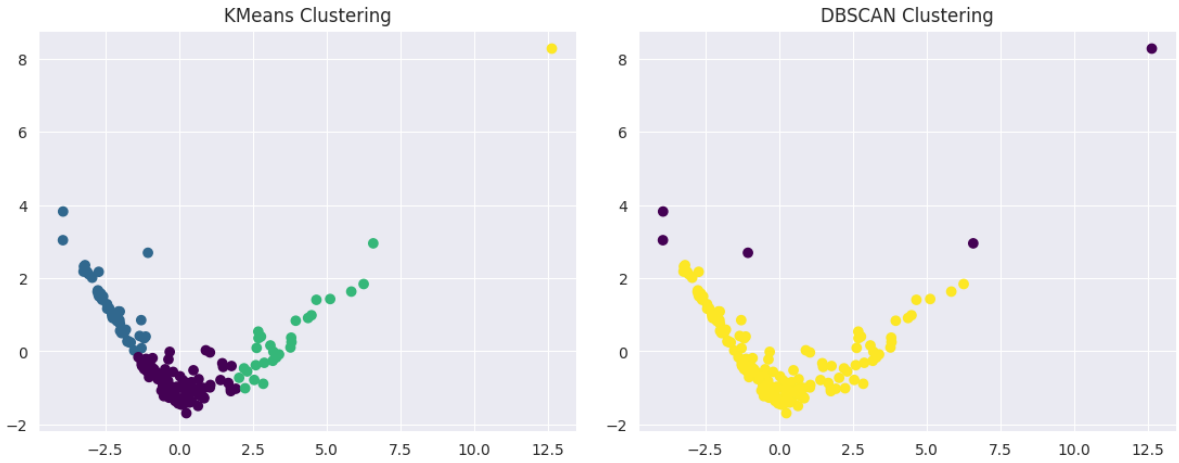

Figure 4: Plot of K-Means Clustering

VI. DISTRIBUTION OF MONEY AMONG CLUSTERS

Then, the total need of each cluster is calculated by summing up the values of GDP per capita (gdpp), health, and income for all countries in that cluster. The proportion of total aid needed for each cluster is then calculated by dividing the cluster need by the total need of all clusters. Based on the available funding of 10 million dollars, the aid amount for each cluster is distributed. Finally, for each cluster, the aid amount for each country in that cluster is distributed based on their proportion of total need.
Cluster 0 : 3808628.4 , Cluster 1 : 566283.08 , Cluster 2 : 5231481.95 , Cluster 3 : 393606.57

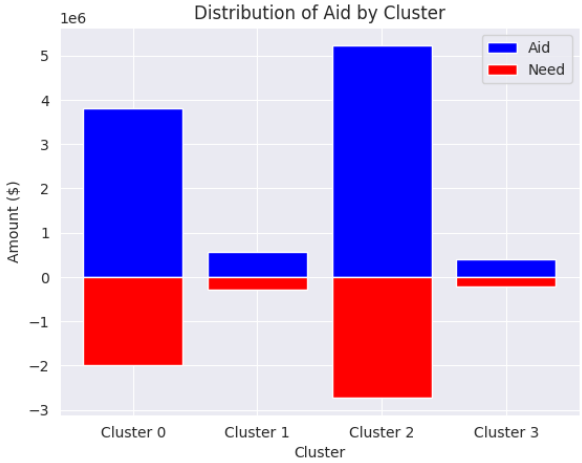

Figure 5: Distribution of money among Clusters

VII. CONCLUSION

We have categorized the countries using some socio-economic and health factors that determine the overall development of the country and finally we have categorized countries into 4 clusters .
Cluster 0 : **'Developing Countries'**.
Cluster 1 : **'Underdeveloped Countries'**.
Cluster 2 : **'Developed Countries'**
Cluster 3 : **'Well Developed Countries'**.
And no.of countries in each Cluster respectively are **89, 49, 28, 1**.

VIII. REFERENCES

https://www.researchgate.net/publication/314700681_Hierarchical_Clustering
https://github.com/kenanarslanbay/K-Means-Clustering
https://github.com/ugis22/clustering_analysis
https://towardsdatascience.com/performance-metrics-in-machine-learning-part-3-clustering-d69550662dc6