CSL2050 : Pattern Recognition and Machine Learning
Minor Project Report

Project 5 : Analyzing Country Data for Classification/Clustering Task

---

Submitted in partial fulfillment of the requirements of the Minor Project for the Course CSL2050 : Pattern Recognition and Machine Learning

by

Manish	B21CS044
Pranav Chakravarthy	B21EE050
Krishna Gaurang	B21EE086

॥ त्वं ज्ञानमयो विज्ञानमयोऽसि ॥

Indian Institute of Technology Jodhpur
NH 62, Surpura Bypass Rd, Karwar, Rajasthan 342030

March 2023

ABSTRACT

The goal of this study was to develop an unsupervised learning model for classifying the Countries dataset into developed, underdeveloped, and developing countries. In our research, we have used Feature Selection and Dimensionality Reduction techniques such as Principal Components Analysis (PCA), Linear Discriminant Analysis (LDA), and Independent Component Analysis (ICA). For the main purpose of clustering, we used the methods of KMeans Clustering, DBSCAN, and Hierarchical Clustering. We have combined all these different approaches and presented the results to find the best model or technique for the current task at hand. We have presented the data and results in a pictorial format whenever and wherever appropriate in order to better understand the outcomes that these methods provide for us.

For this project, we adhered to the entire machine learning pipeline.

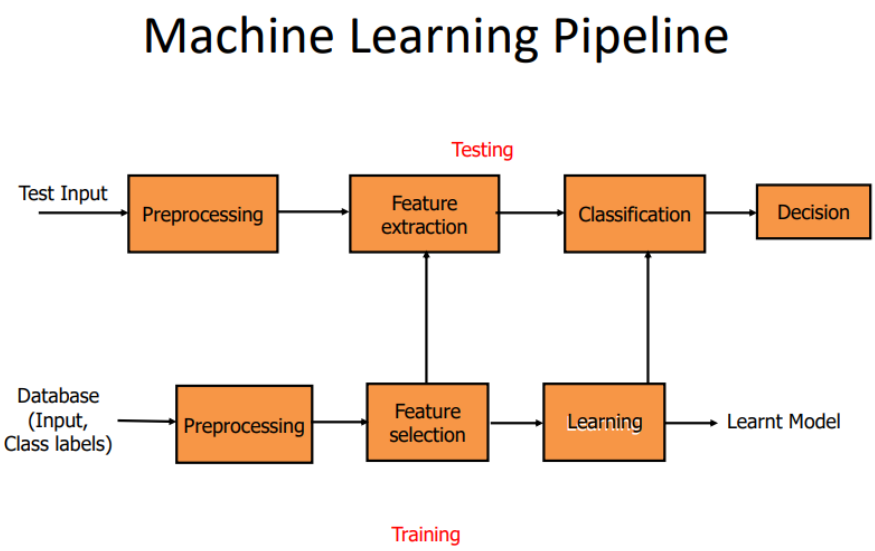


Figure: Machine Learning Pipeline

INTRODUCTION

In this study, we made an effort to group the countries' data points into three categories: developing, under-developed, and developed. For this, we have used the Principal Component Analysis, Linear Discriminant Analysis, and Independent Component Analysis 3 dimensionality reduction/feature selection techniques. Along with these, we also used the clustering algorithms KMeans, DBSCAN, and Hierarchical Clustering. Using dimensionality reduction/feature selection techniques and the best clustering model that applies to them going forward, we have tried to identify the best features that best represent the data.

The process of transforming data from a high-dimensional space into a low-dimensional space, ideally one that is close to the intrinsic dimension of the original data, is known as dimensionality reduction.

Principal Component Analysis (PCA): In PCA, we examine the data from the perspective of variance and seek to identify the dimensions that most accurately capture the variance in the data. We then transform the given feature space into that space.

In linear discriminant analysis, we seek a new transformation of the initial feature set that produces the greatest interclass distance and the smallest intraclass distance.

Independent Component Analysis (ICA) seeks to produce a statistically independent feature space with zeros in its off-diagonal covariance matrix.

We attempt to investigate the effects that each of these various methods may have on the dataset and the current clustering task.

When various dimensionality reduction techniques are applied, the clustering algorithms, such as K Means, DBSCAN, and Hierarchical Clustering, each perform differently and have their own flaws.

Every time an optimization issue arose, we computed the best value for the iterable using the Elbow Method and Silhouette Score.

We have attempted to investigate each of the nine combinations that can be made using the three dimensionality reduction techniques and the three clustering algorithms in order to determine which one is most effective for the countries data. Additionally, we made an effort to interpret the data trends and analyze the anticipated outcomes.

PRELIMINARY ANALYSIS OF THE DATASET

In this step, we have performed the pre-processing. We see that there are no null values in the dataset. We have used the additional dat given on Kaggle webpage to ascertain what each of the columns in the dataset represent:

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. 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.

Table: Column Name and their description in the dataset

We can see that all of the columns contain relevant information to the task at hand and we need not remove any columns.

Visualizing the dataset using histogram, we have:

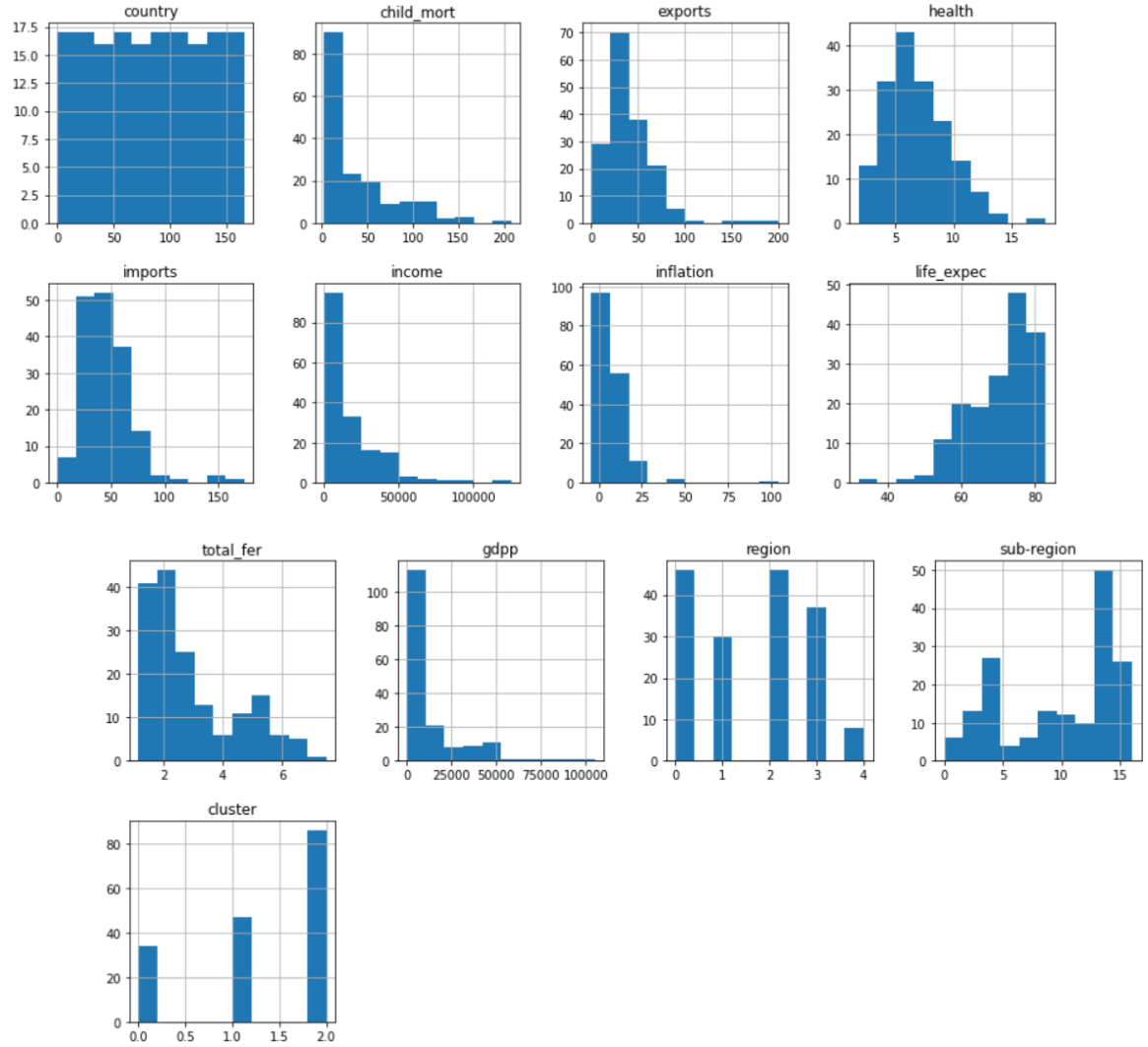


Figure: Histograms for visualizing the dataset

The data frame was then duplicated nine times to create a dataframe for every implementation. The sklearn library function StandardScaler() then scaled the data.

We can see from the boxplots that the data contains outliers, so we anticipate that algorithms like DBSCAN that are sensitive to outliers will perform poorly.

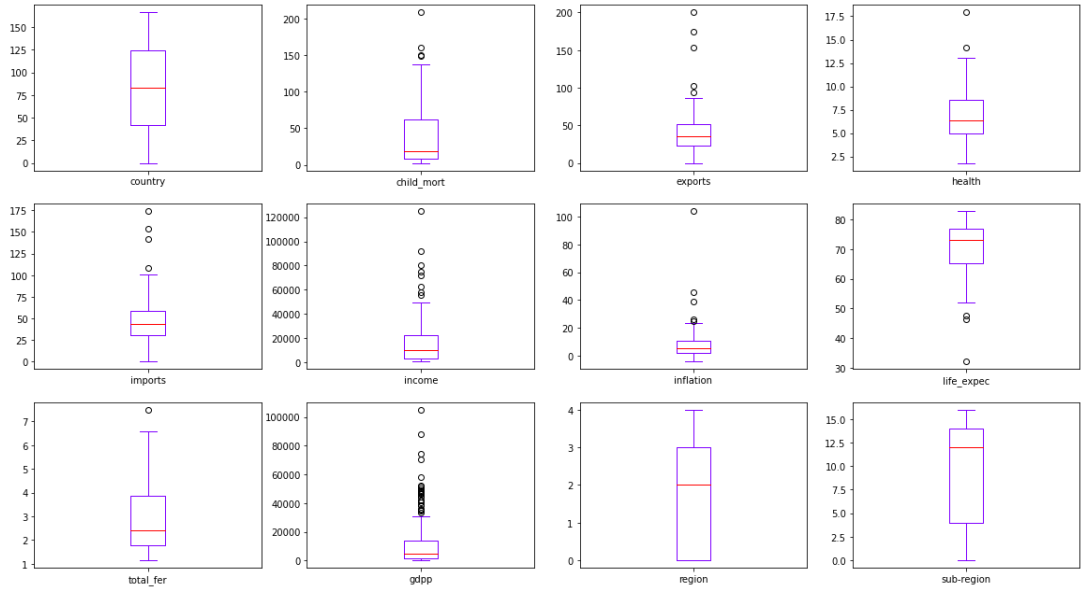


Figure: Boxplots for visualizing outliers in each feature

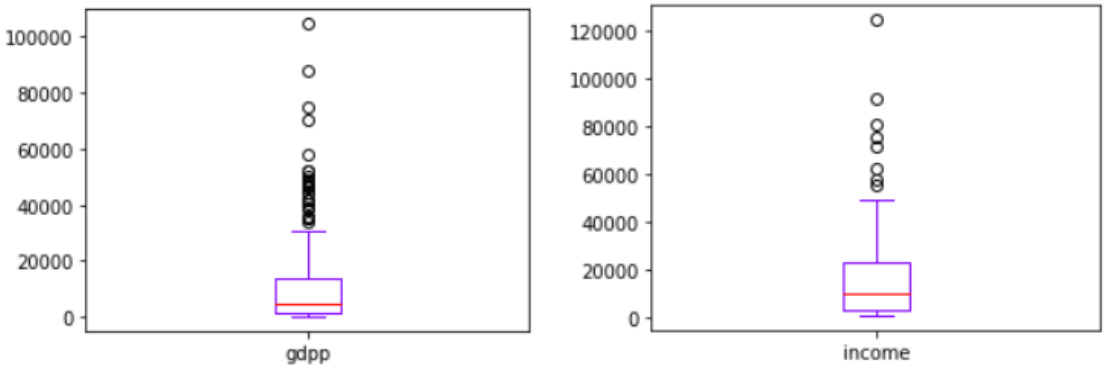


Figure: Box plot of gdpp and income for outliers

PRINCIPAL COMPONENT ANALYSIS AND K MEANS

We discover that we must maintain 7 principal components in PCA in order to maintain 95% of the explained variance ratio. The covariance matrix found using principal components for the new dataset has diagonal elements that are off and Therefore, we can say that in this case, the PCA makes the new features statistically independent. The covariance matrix's heat map demonstrates this matrix.

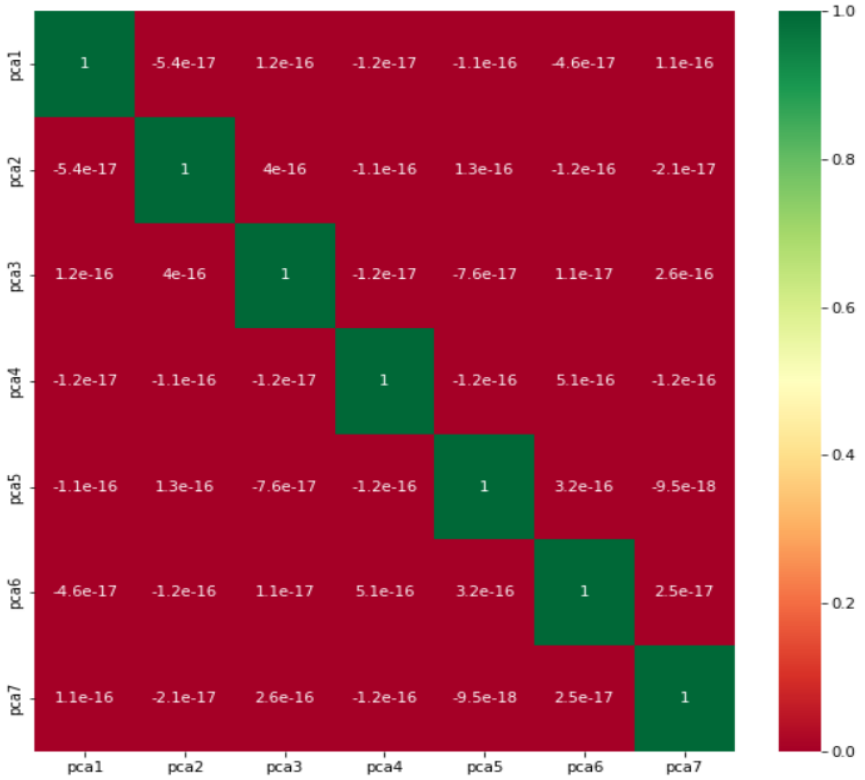


Figure: Heatmap of covariance matrix of PCA performed on the dataset

Measuring the optimum number of clusters has been done with the Elbow Method (with inertia as the parameter) and the Silhouette Score. Also considering the nature of the clustering task i.e. clustering countries into ones that need most help, may need help and do not need help, we have chosen the number of clusters as 3.

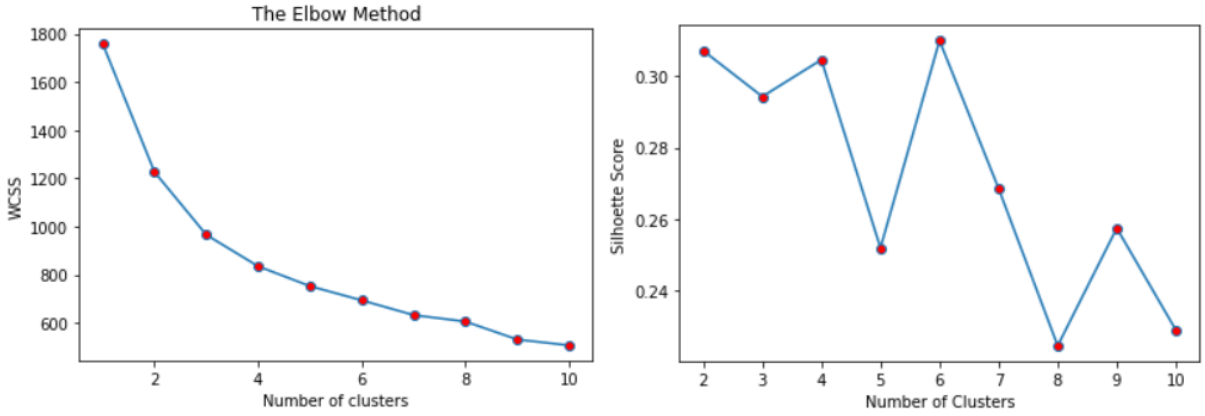


Figure: Elbow Method

Using K Means Clustering with $k = 3$ and randomly initialized centroids, we get the clustered points/countries as:

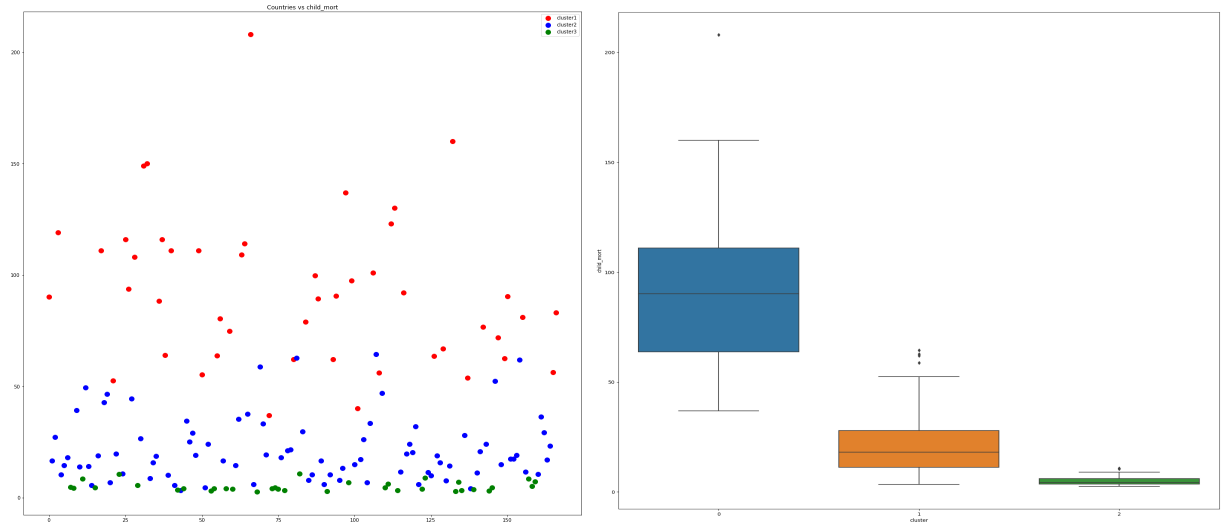


Figure: Plot of Child Mortality of countries with color indicating the cluster to which they have been assigned

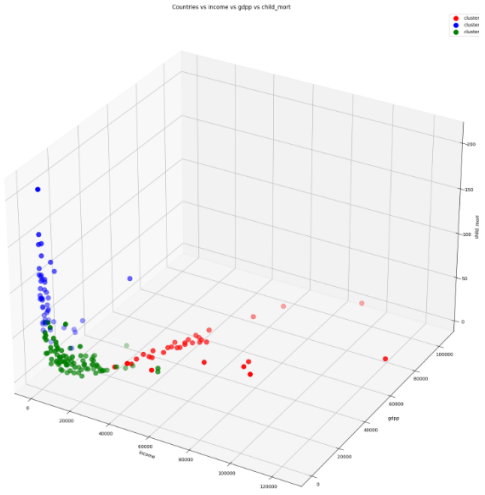


Figure: 3D plot of Clusters

Because we know that there exists high child mortality rates in countries that need help and very less Mortality Rate in countries that are developed, we can label the clusters based on these known universal facts. Thus, we get the following clusters on a world map:

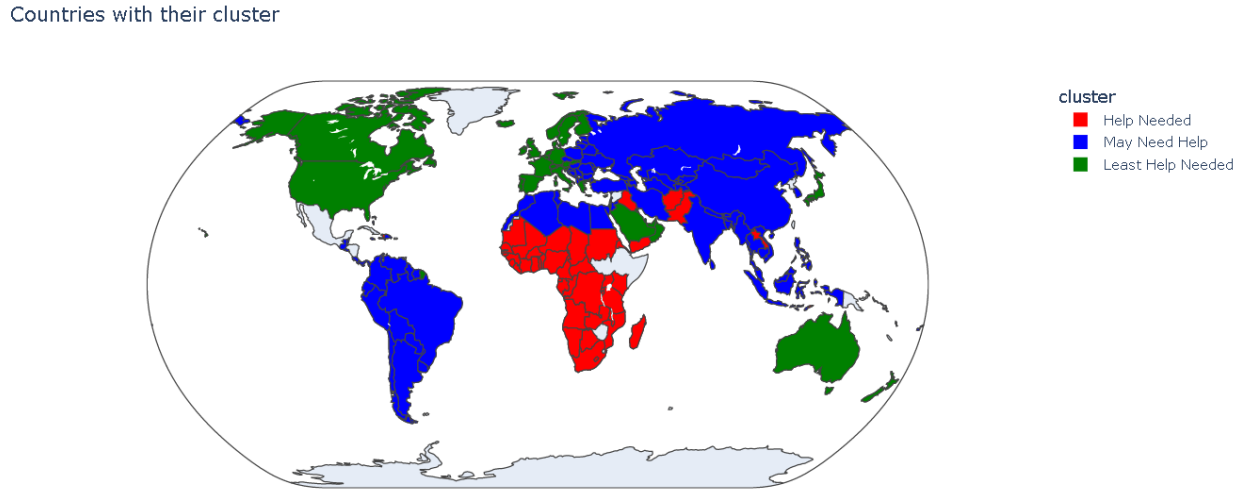


Figure: World map of Countries with their cluster

PCA with K Means has performed well and we can actually see in this visualization and from our general knowledge that the countries have been clustered correctly.

PCA WITH HIERARCHICAL CLUSTERING

On the same PCA dataset, we have tried out the Hierarchical Clustering (Agglomerative) method for clustering. The dendrogram that we obtained is as follows:

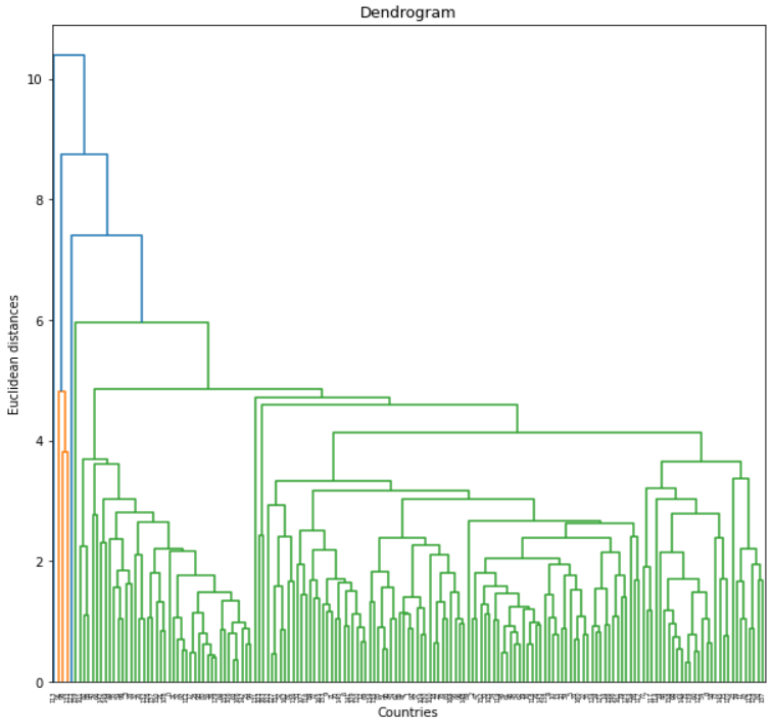


Figure: Dendrogram for PCA, Agglomerative Hierarchical Clustering

In this case, however, on putting the hyperparameter n_clusters=3, we have obtained 2 major clusters and an isolated cluster. Hence, we decided to use 2 clusters only here.

On comparing the life expectancy in these clusters, we have:

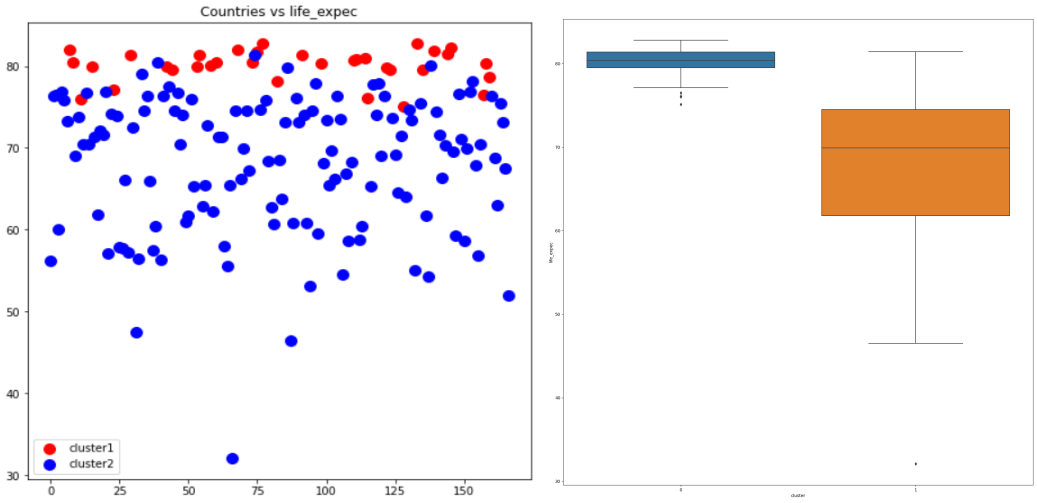


Figure: Plot of countries vs life expectancy with color indicating the cluster

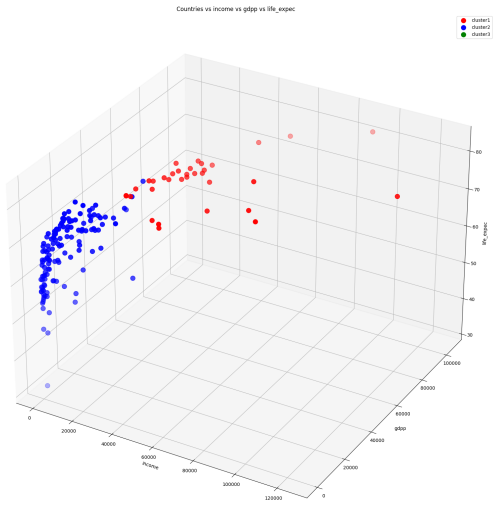


Figure: 3D plot of Clusters

On the world map, we can interpret these results as below.

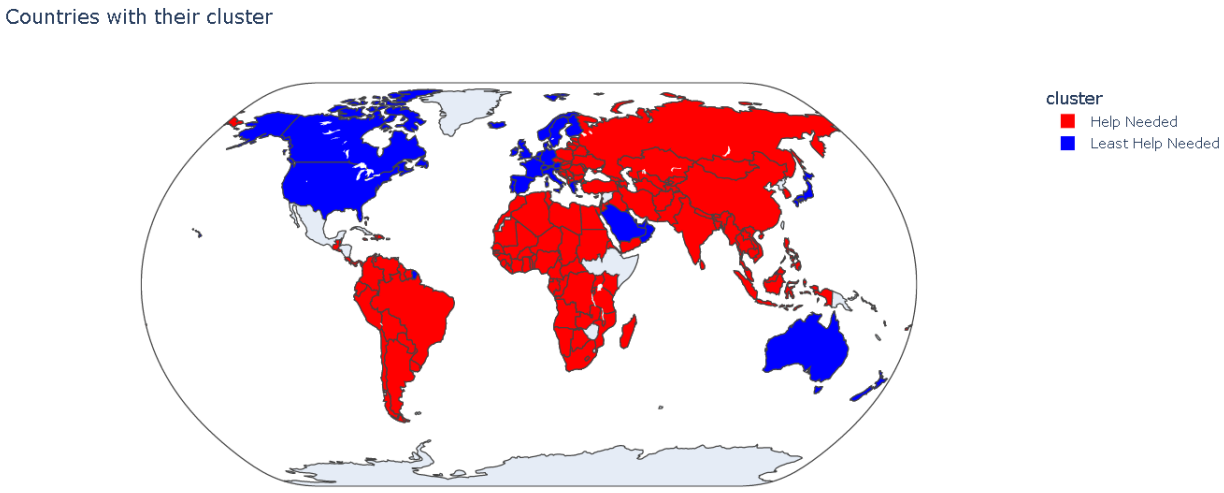


Figure: World map of Countries with their cluster

As can be seen in this map, the clustering algorithm works well in this case.

PCA WITH DBSCAN

In our visualization using the boxplots, we had seen that there were a few outliers in the dataset. Thus, we would expect DBSCAN to perform poorly on this data. With the hyperparameters **eps=1.5** and **min_samples=2**, we obtain 6 natural clusters and Noise.

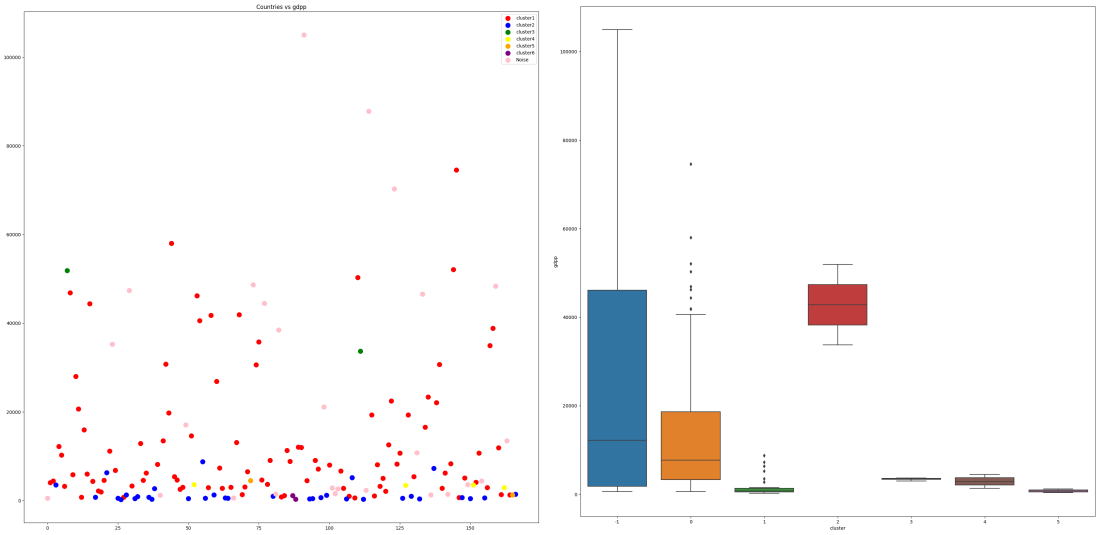
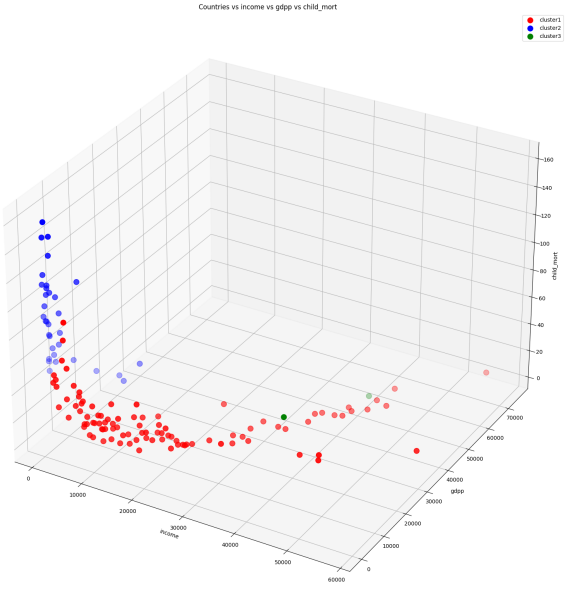


Figure: Countries vs gdpp with color representing the cluster



We have further manually grouped the 6 clusters into 2 clusters with 3 each and kept the noise separately.

On the world map, these can be shown as follows:

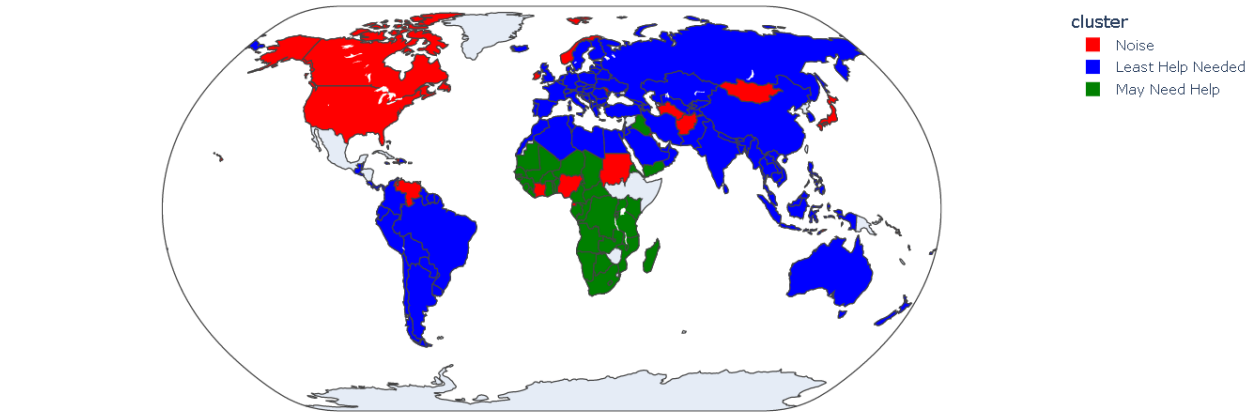


Figure: World map of Countries with their cluster

The countries that are in noise have either very high parameters in the feature space or very low and are mostly outliers. For example, the US has very high gdpp and very low child mortality whereas Mongolia has very low gdpp and a high child mortality but they are both included in noise.

LDA AND K MEANS

For retaining 95% of the explained variance ratio after performing LDA, we need to keep 3 LDA components. After the LDA transformation, we see that the new dataset has a covariance matrix with the off diagonal elements tending to 0. Thus, we can assume that the new transformed features are statistically independent.

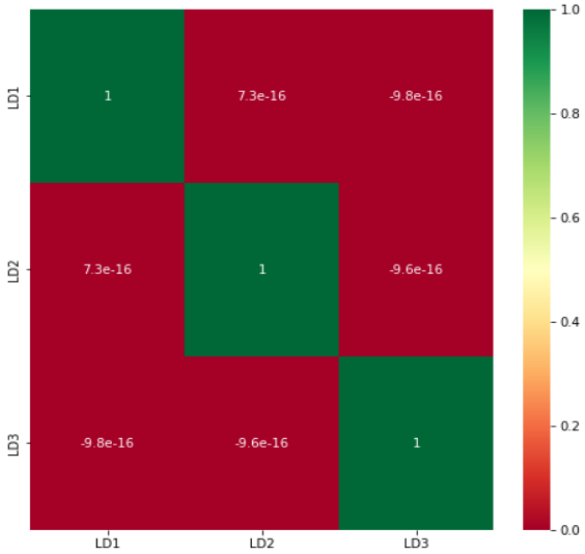


Figure: Heatmap of covariance matrix after LDA

Since we are using K-Means clustering, we have obtained the best number of clusters using the Elbow Method with parameters as inertia and Silhouette Score. With this, we obtained an elbow at $k = 4$ and also best silhouette score at $k = 4$. However, the task requires that we use $k = 3$ and thus we have done that.

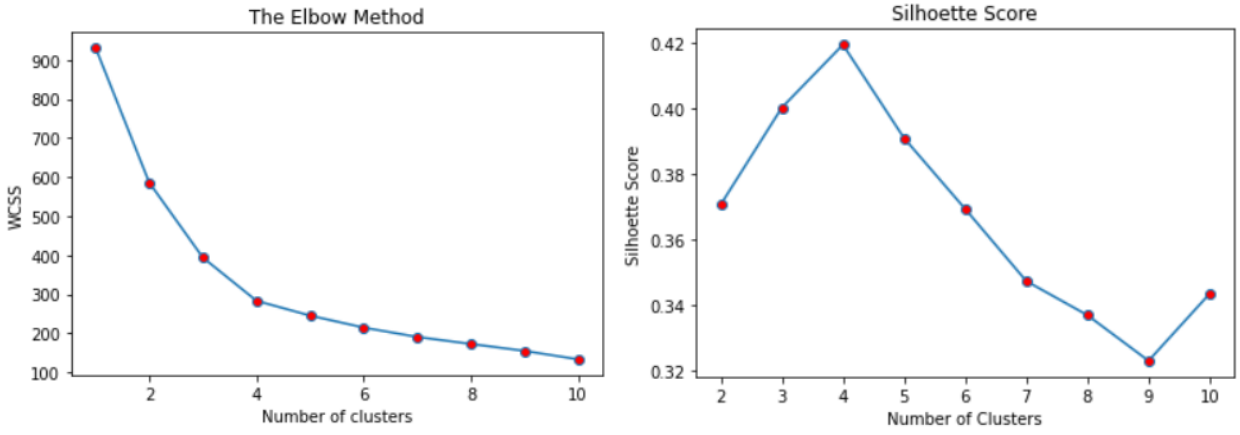


Figure: Elbow method and Silhouette Score for best k

Using LDA and K-Means we have obtained the 3 natural clusters that we wanted.

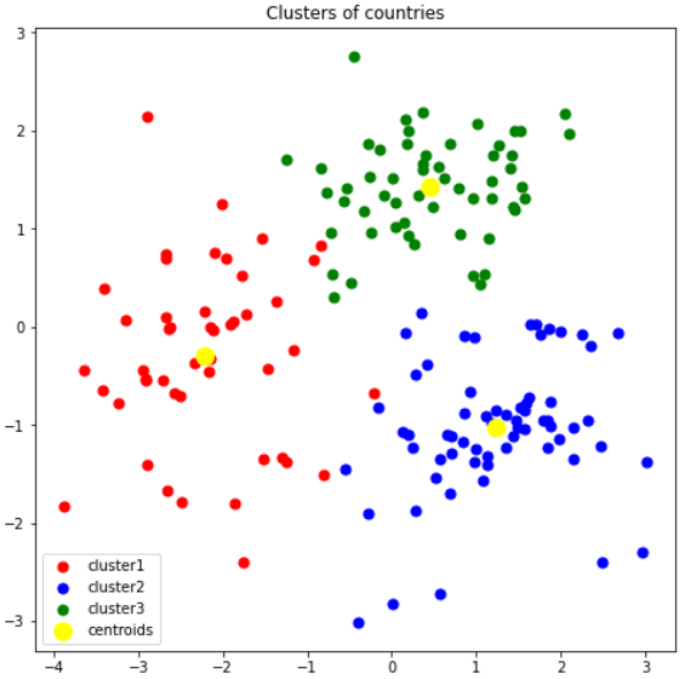


Figure: Clusters obtained using LDA and K Means Clustering

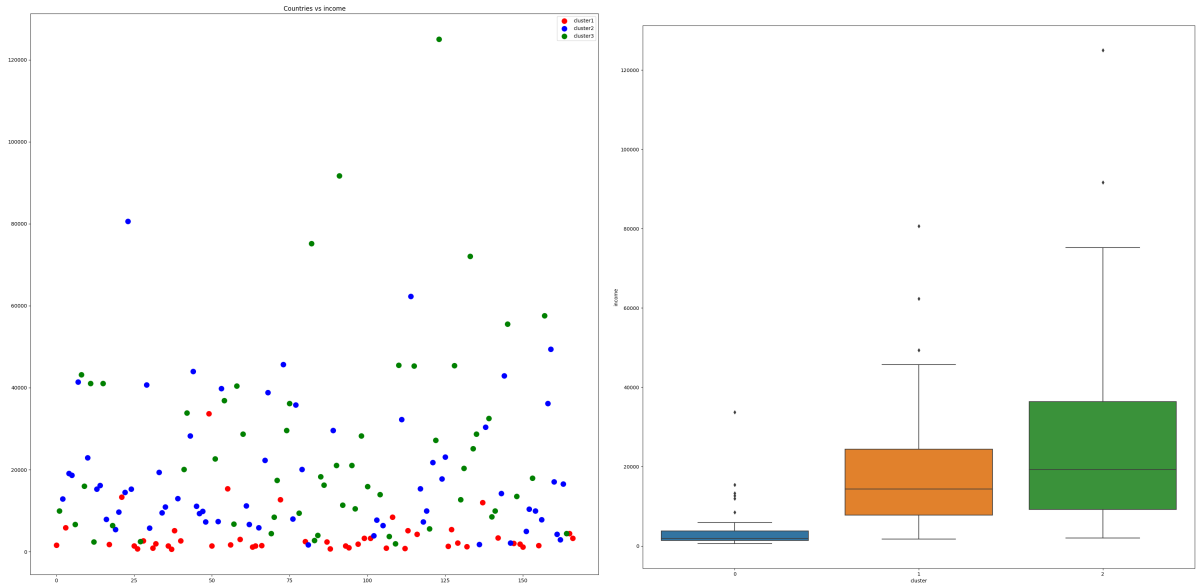


Figure: Plots for Countries vs Income with color representing cluster

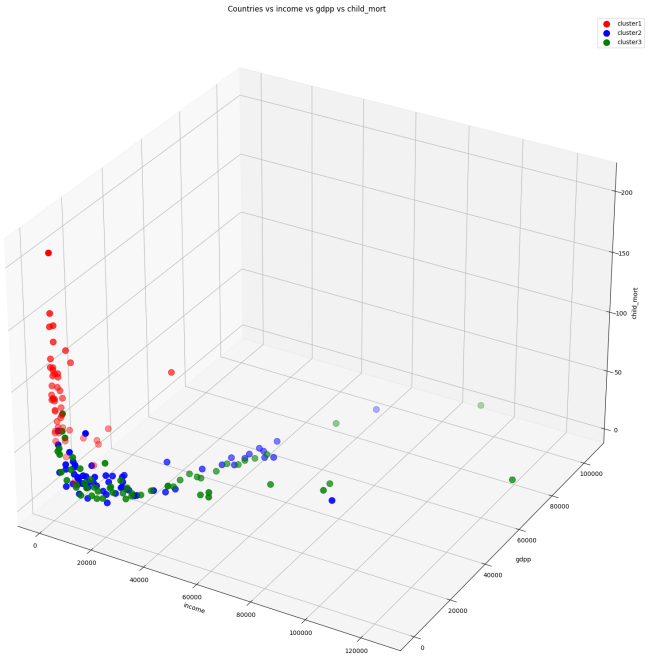


Figure: 3D plot of Clusters

On the world map, we can plot the results as follows:

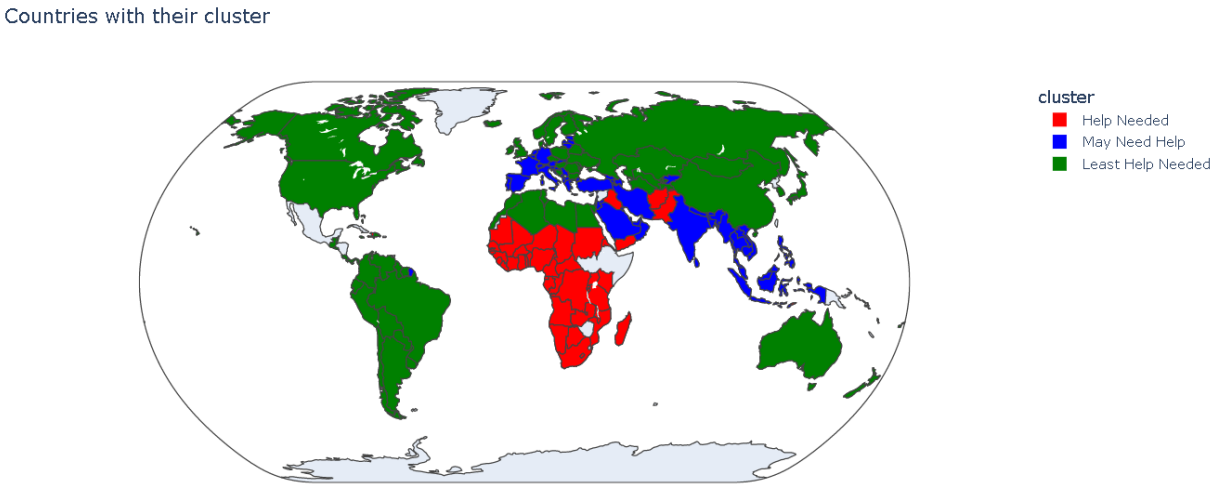


Figure: Countries with their Cluster and Labels

LDA WITH DBSCAN

Using DBSCAN on the LDA data, we get 5 natural clusters and noise.

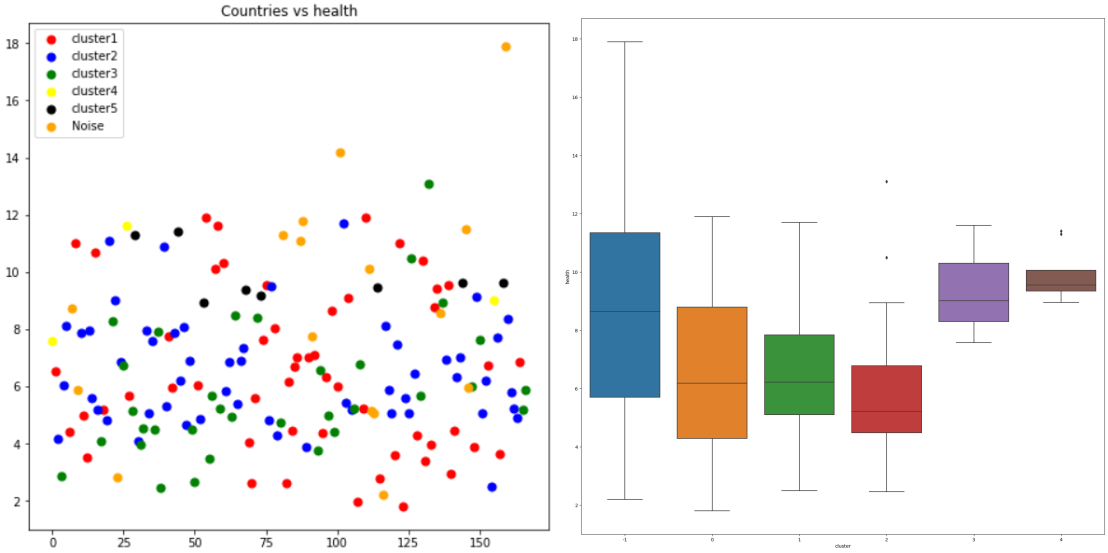


Figure: DBSCAN on LDA data

From above we can see that cluster 1,5 needs less help than cluster 2,3,4.
Hence, grouping them, we get on the world map:

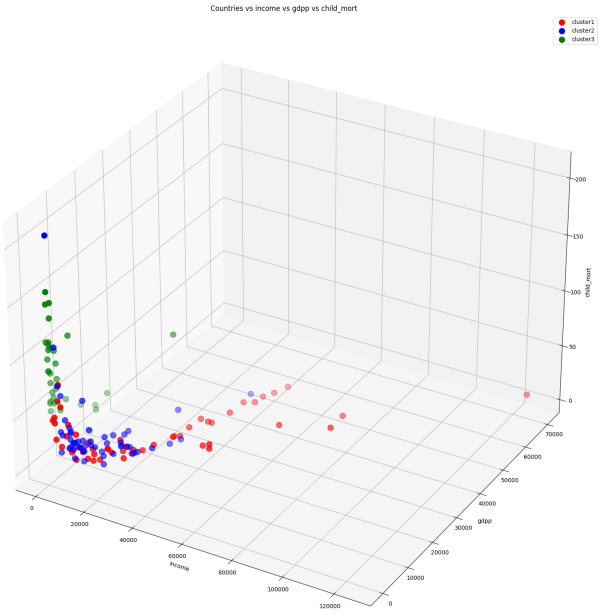


Figure: 3D plot of Clusters

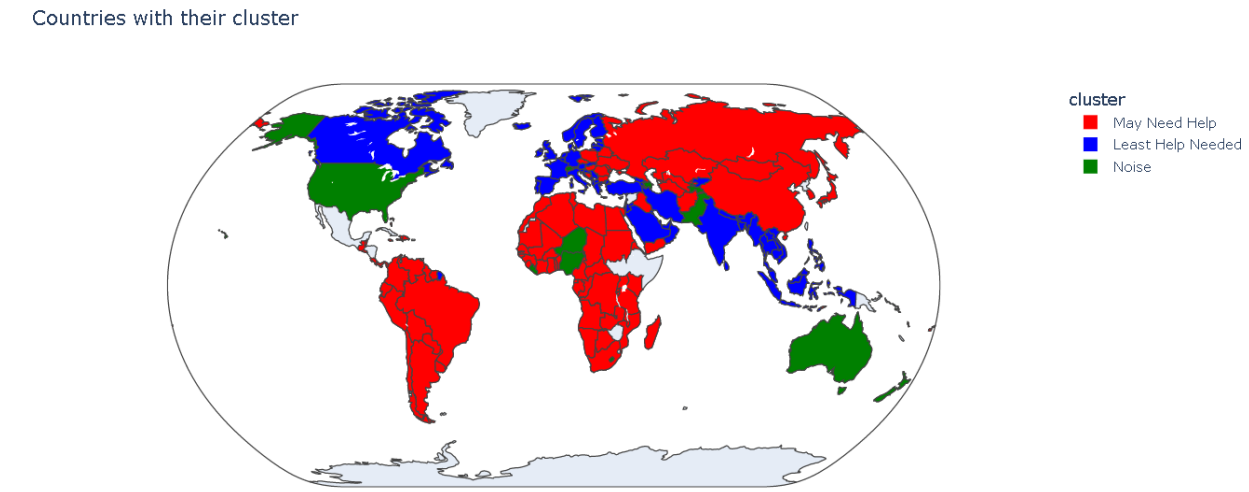


Figure: World map of Countries with their clusters

We see that DBSCAN doesn't work very well on this dataset as there are a lot of outliers and it is necessary to cluster even the outliers. The countries in the noise that we get from DBSCAN also belong to all 3 clusters and there is no general trend in them.

LDA WITH HIERARCHICAL CLUSTERING

On the same LDA dataset, we have tried out Agglomerative Clustering. The dendrogram that we have obtained is as follows:

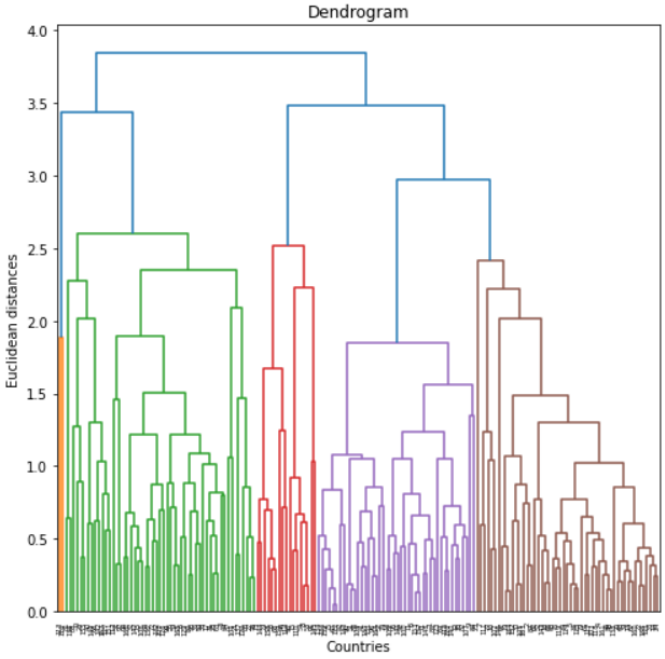


Figure: Dendrogram for LDA

Using hyperparameter n_cluters=3 and using Average linkage, we get 3 natural clusters.

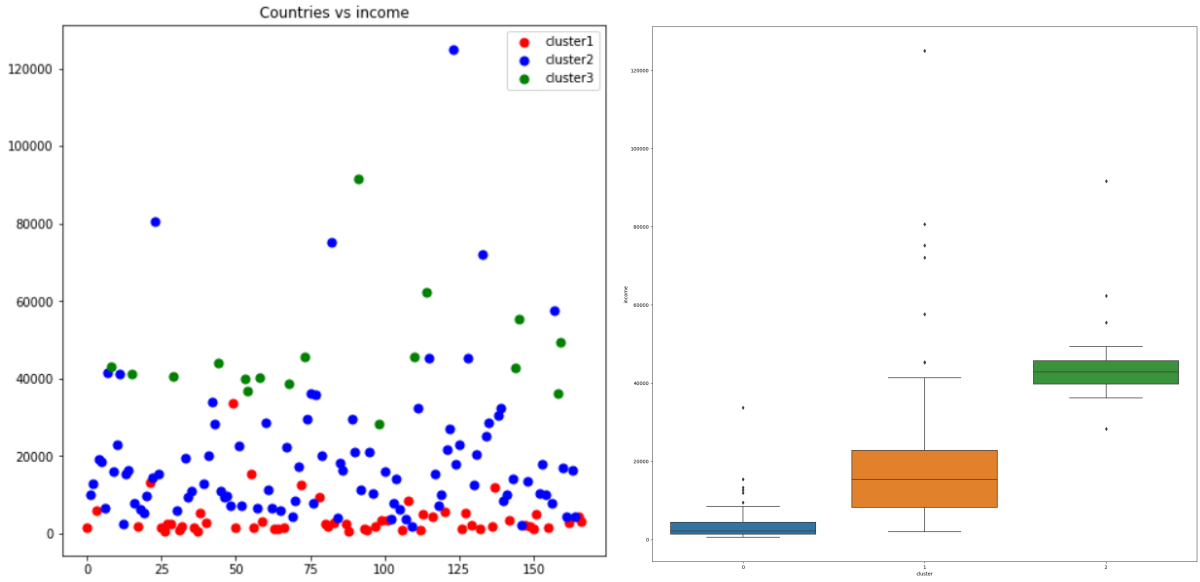


Figure: Clusters obtained using Agglomerative Clustering

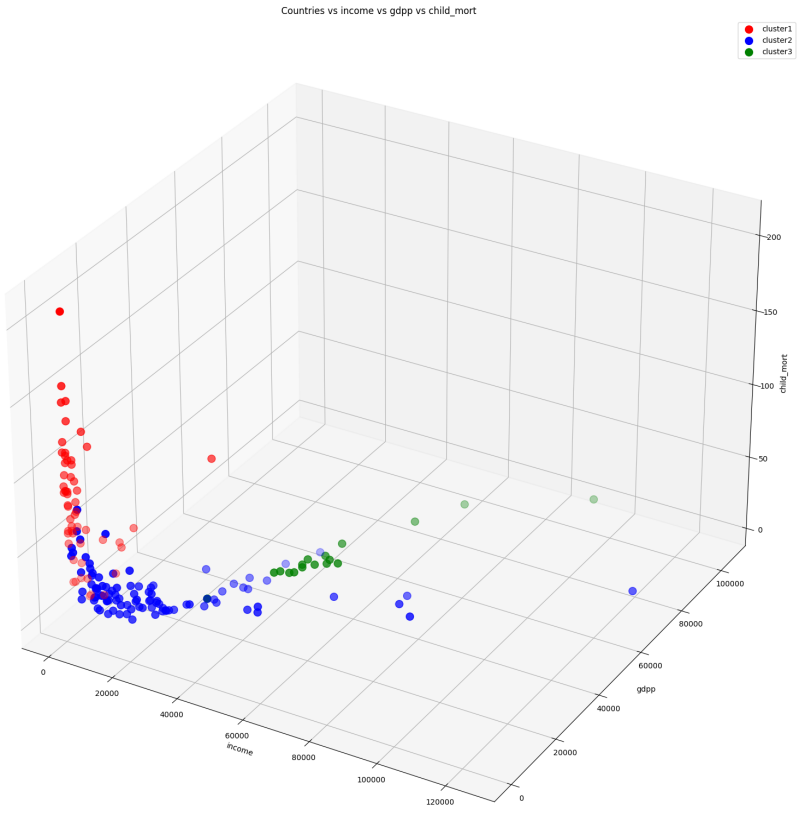


Figure: 3D plot of Clusters

From the above graph, we can ascertain the exact labels (developed, developing, under-developed) for the countries with income ranking from highest to lowest.

On the world map, we have this as:

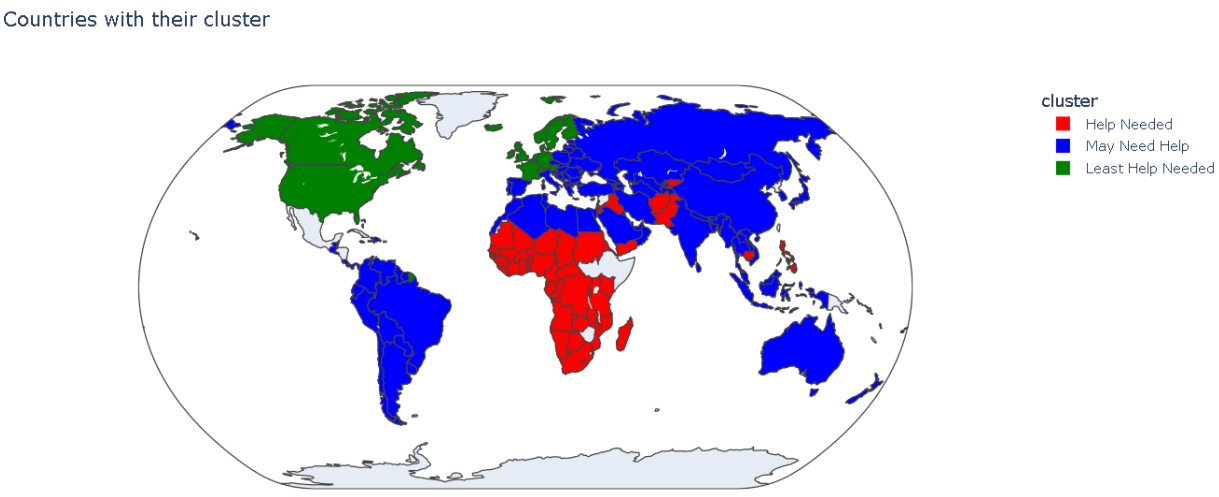


Figure: Final Clusters after Hierarchical Clustering after LDA

Thus, the Hierarchical Clustering works very well after performing LDA on the dataset and we get accurate results.

ICA WITH K MEANS

ICA (Independent Component Analysis) is a dimensionality reduction method which aims to make the data statistically independent such that the off-diagonal elements of the covariance matrix are 0 by using a linear transformation.

After performing ICA, we have taken 3 independent components and have plotted the heatmap of their covariance matrix.

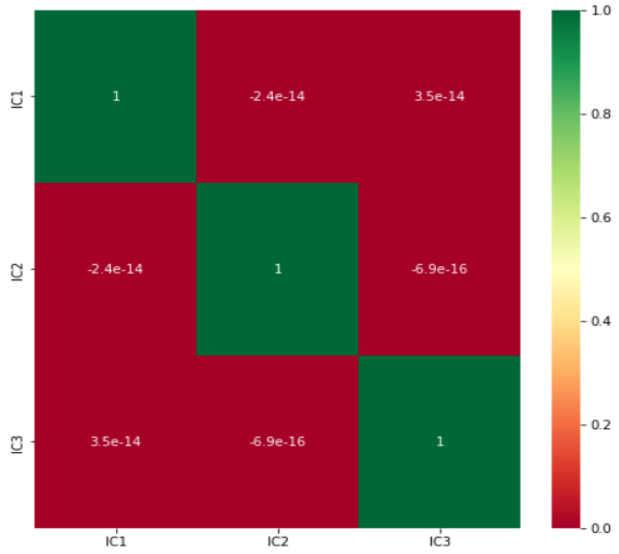


Figure: Heatmap of the Covariance Matrix after ICA

We see that the off-diagonal elements of the matrix tend to 0, thus the new features are statistically independent.

Since we are using K-Means for the clustering task, for finding the best number of clusters, we have used the Elbow method with inertia as a parameter and the Silhouette Score.

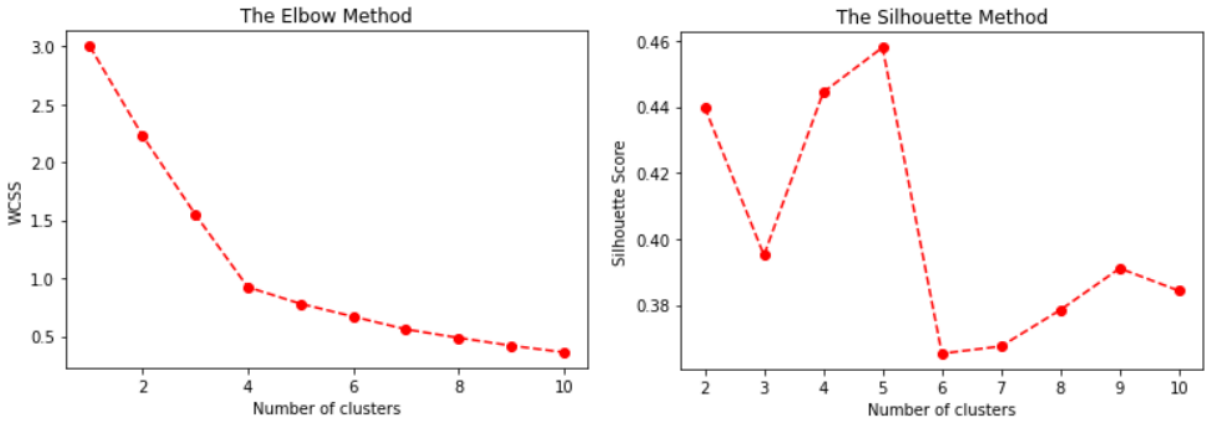


Figure: Elbow Method and Silhouette Score for K-Means after ICA

We find the best value of k to be 4 using the elbow method.

Thus, using k=4, we get the clusters as:

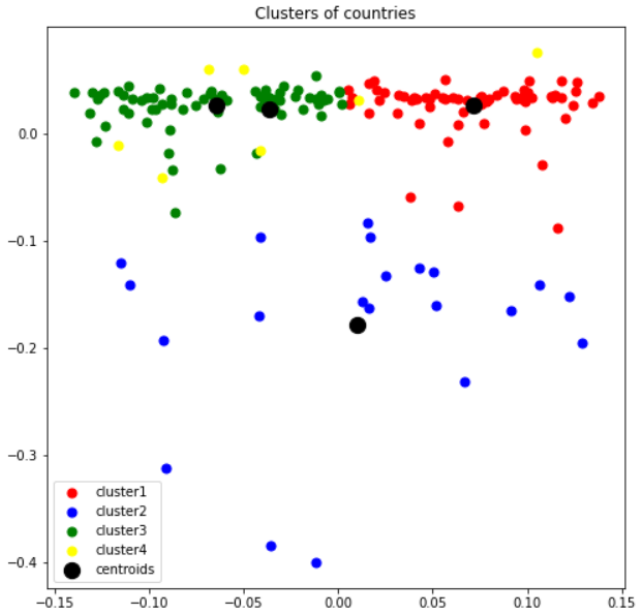


Figure: Clustering using K-Means after ICA

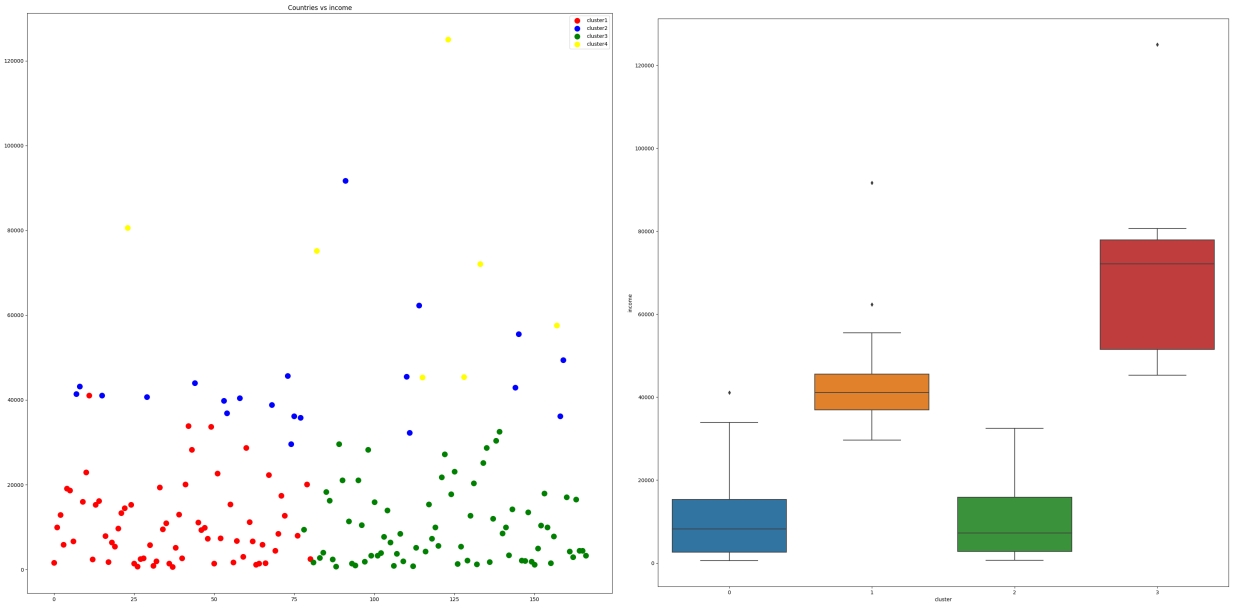


Figure: Plots for Countries vs Income with color representing cluster

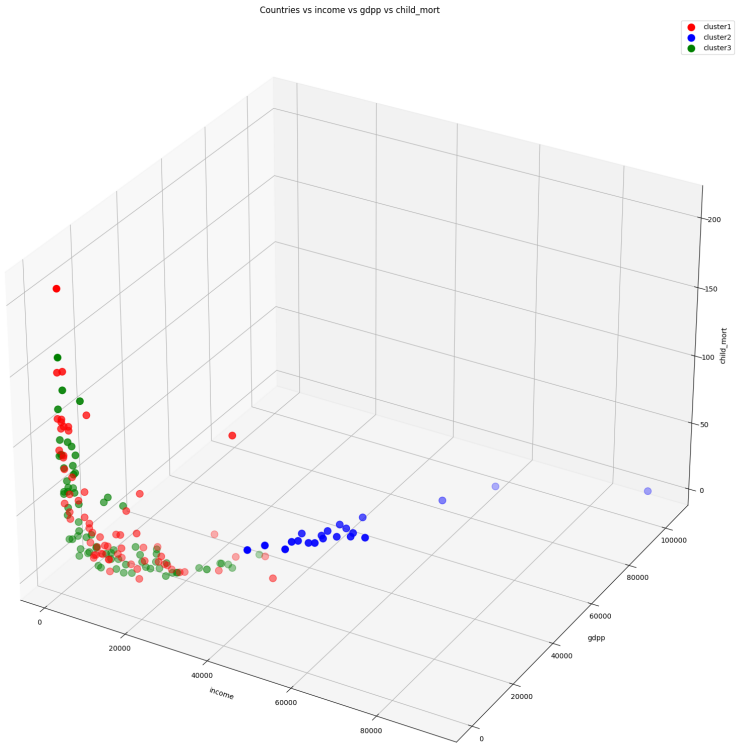


Figure: 3D plot of Clusters

We can see that cluster 1 and 3 need more help as compared to cluster 2 and 4. Thus, combining these clusters, we have on the world map:

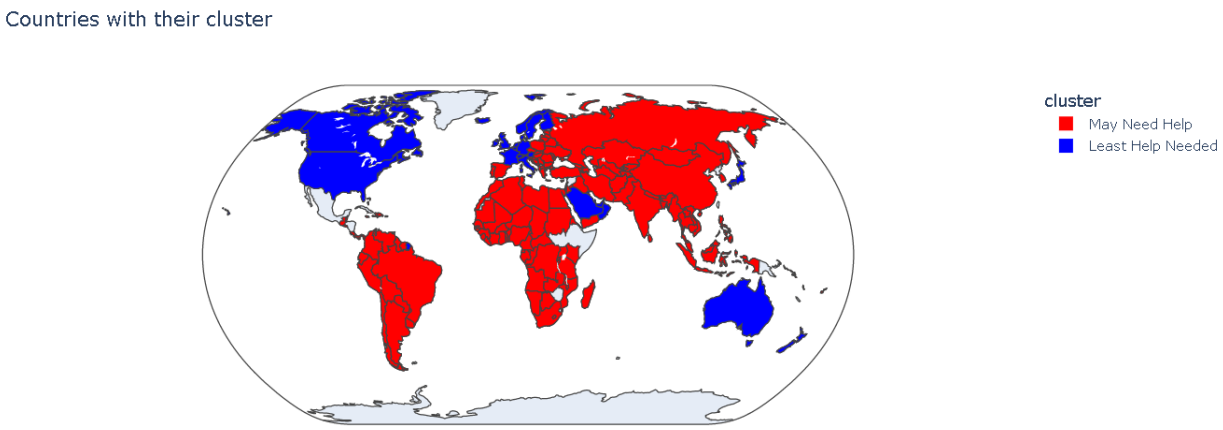


Figure: Countries with their Cluster

K-Means performs very well after performing ICA on the dataset.

ICA AND HIERARCHICAL CLUSTERING

On the ICA implemented dataset, using Agglomerative Hierarchical Clustering, we get the following Dendrogram:

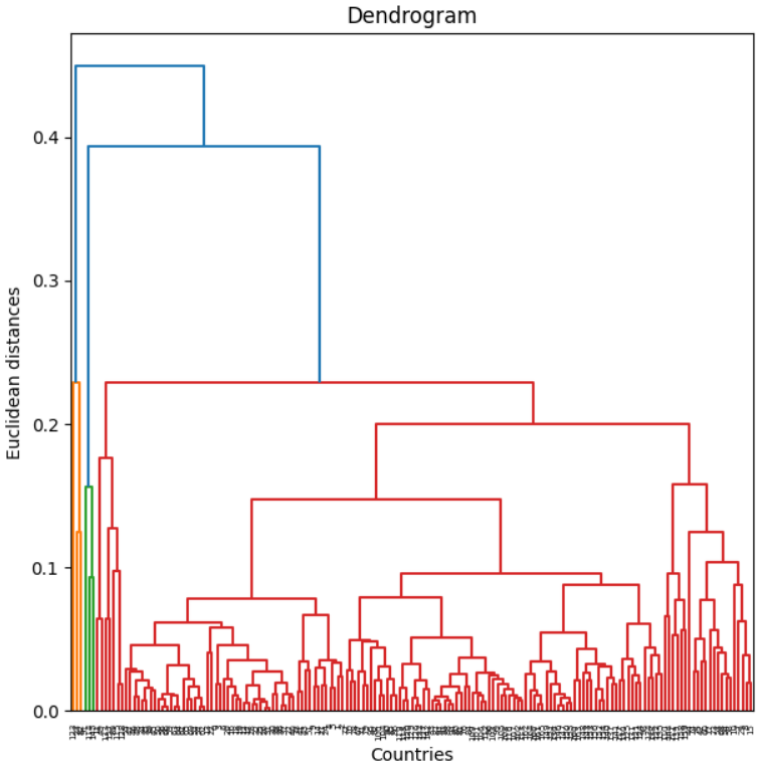


Figure: Dendrogram on ICA dataset

Taking the number of clusters=4, we get the following plot for countries vs income.

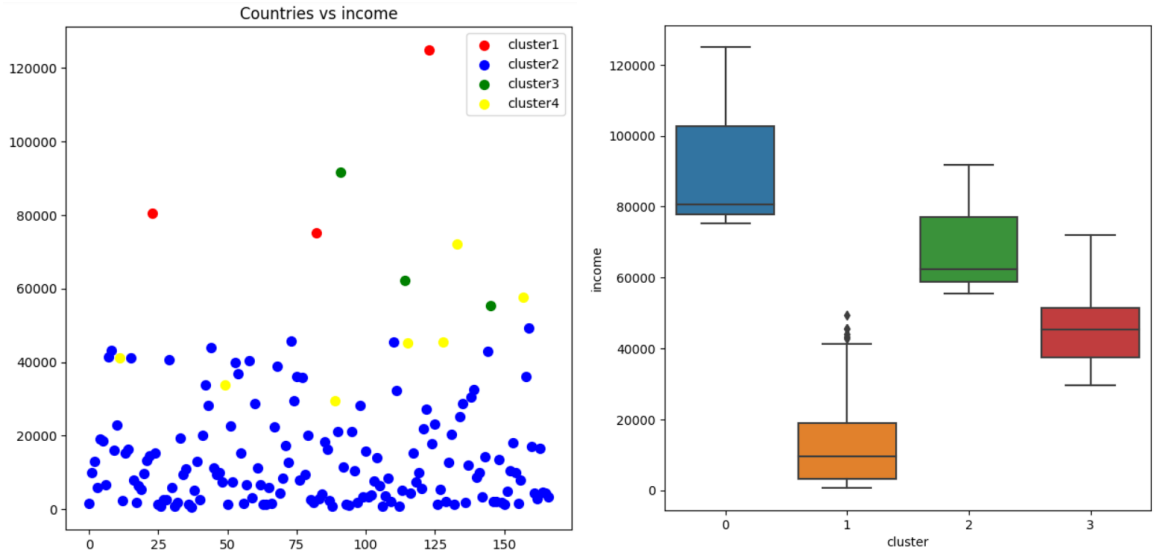


Figure: Plots for Countries vs Income with color representing cluster

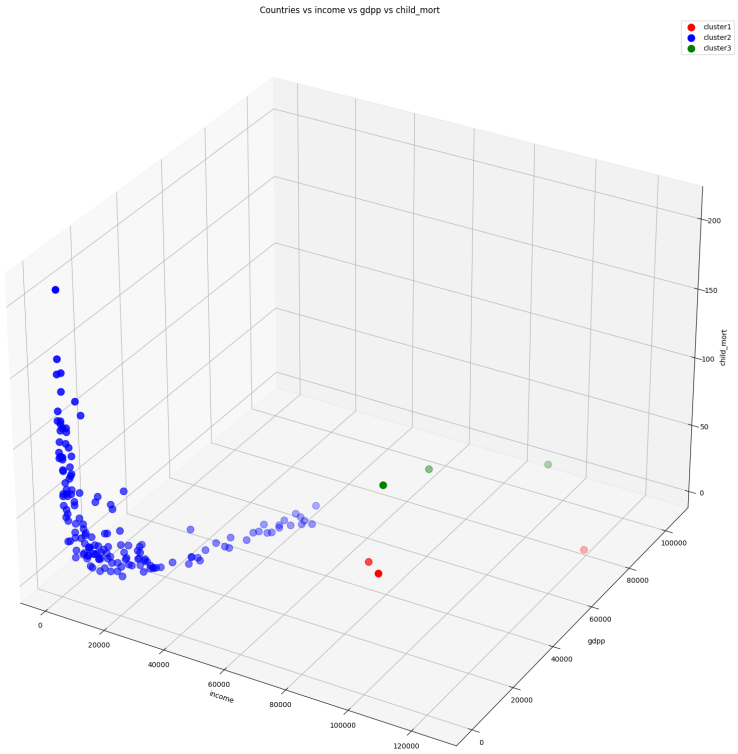


Figure: 3D plot of Clusters

Plotting the results obtained by this method on the world map, we have:

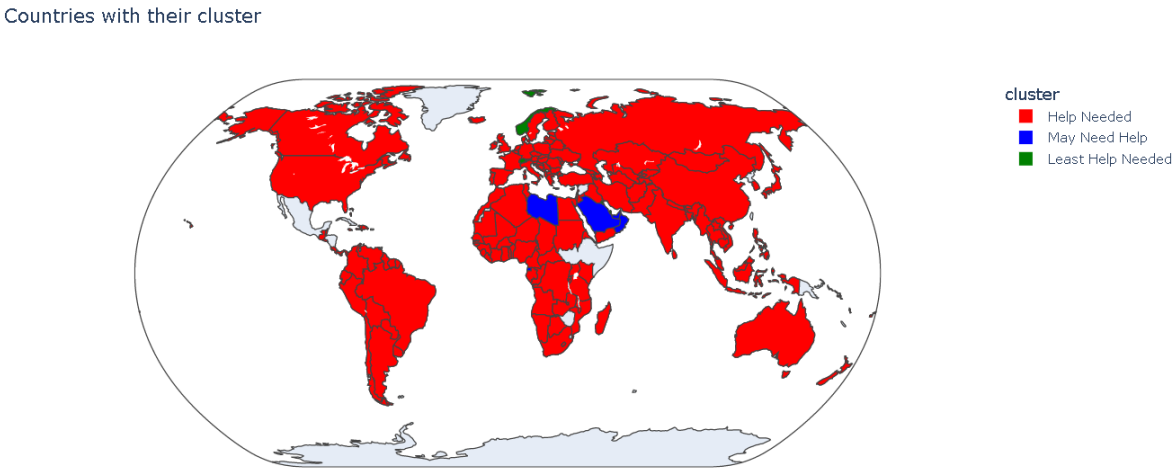


Figure: Clusters on World map for ICA and Agglomerative Clustering

From this image, we can directly see that this method performs extremely poorly on the dataset as most of the countries are grouped in a single cluster.

ICA WITH DBSCAN

DBSCAN is an algorithm that is very sensitive to outliers and our data contains many outliers, we expect a lot of noise in DBSCAN.

We obtain the following clusters using DBSCAN:

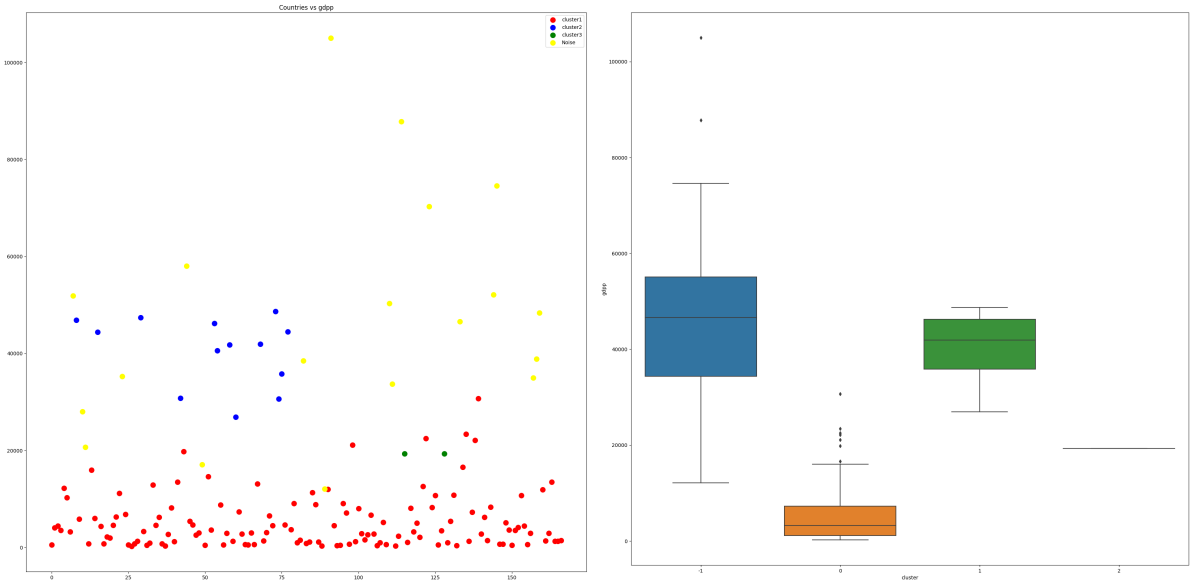


Figure: Plot of Countries vs gdpp

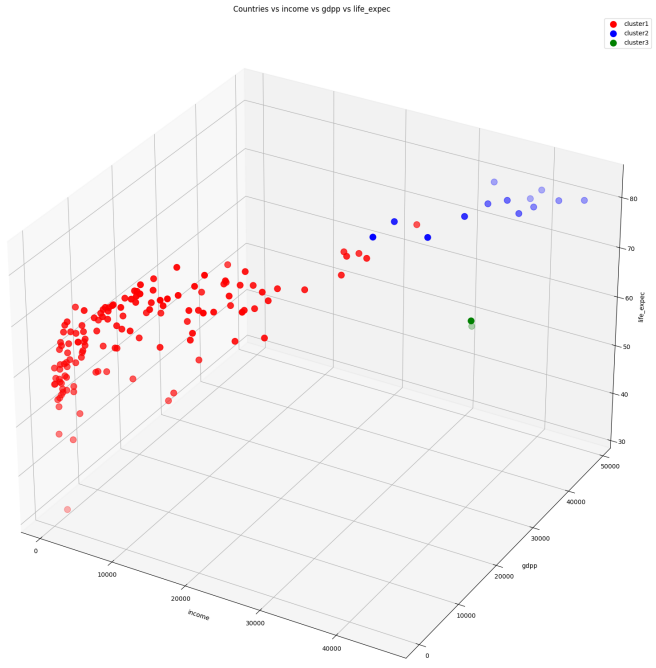


Figure: 3D plot of Clusters

On the world map, we can visualize this as follows:

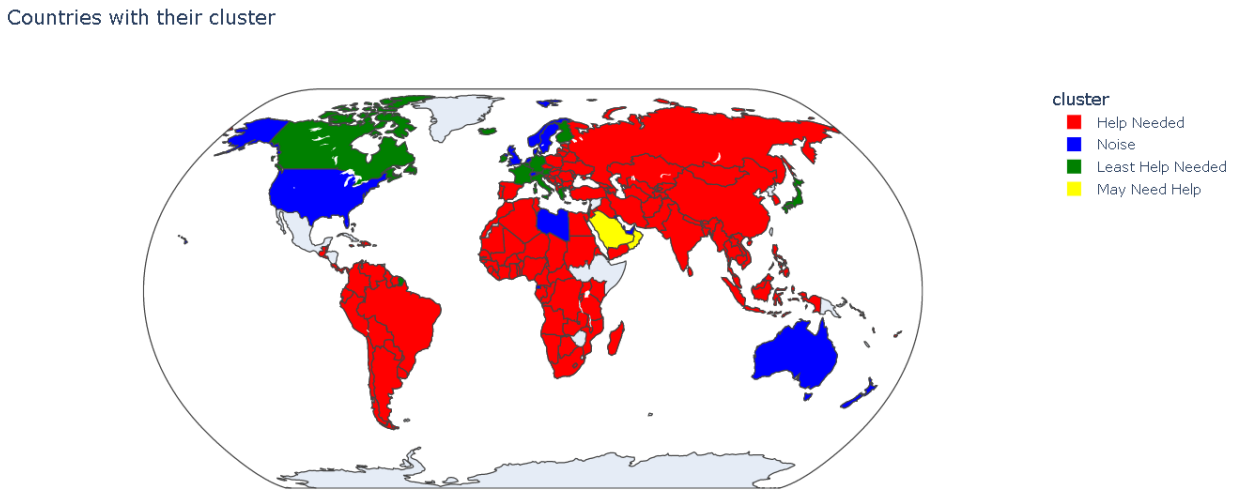
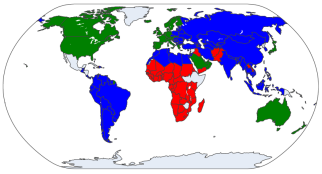
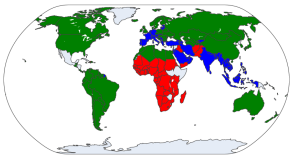
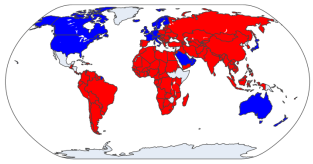
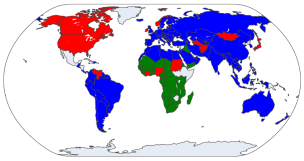
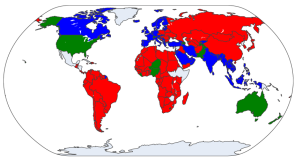
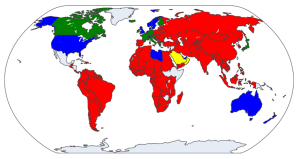
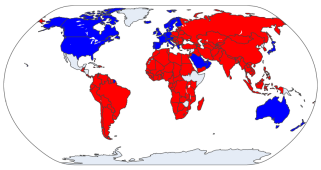
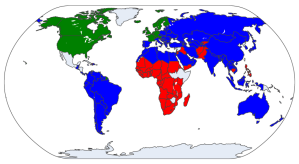
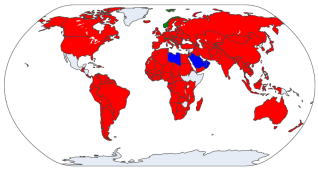


Figure: Counties with color representing their Cluster

We can clearly see that DBSCAN classifies a lot of the data points as noise. These points ideally belong to different clusters and thus nothing conclusive can be said about them.

CONCLUSION

World Maps Corresponding To All Models

	PCA	LDA	ICA
K-MEANS			
DBSCAN			
AGGLOMERATIVE			

Comparison Between All Models

	PCA	LDA	ICA
K-MEANS	Number of clusters can be specified in K-Means. Is sensitive to outliers. PCA captures dimensions along which there is maximum variation. Has the best performance	LDA is a supervised learning algorithm. We have used Continents and sub-region as the labels in this case. LDA along with K-Means has performed very poorly	ICA makes the new components statistically independent. However, it performs average in the given clustering task with K-Means.It has clubbed the May need help and Help Needed clusters.
DBSCAN	DBSCAN is very sensitive to outliers and classifies them as noise regardless of the closest cluster. In this dataset, there are a lot of outliers and DBSCAN performs very poorly.	DBSCAN in LDA has performed well.It has clustered the countries correctly. Only problem with this method is that it has clustered in a few countries as noise.	Same as in PCA, DBSCAN it hasn't performed well.
AGGLOMERATIVE	Agglomerative clustering has also performed fairly well and is second only to -Means in this clustering task.	Agglomerative clustering on the LDA dataset has given similar results to those by K-Means. Thus, it has also performed very well on the LDA dataset.	Agglomerative clustering has not performed very well on the ICA dataset and we get a lot of countries in a single cluster.

We find that the best clustering is achieved by K-Means with PCA, followed by Agglomerative Clustering. When compared to the other two clustering algorithms, ICA fares even worse. Due to the high number of outliers, DBSCAN generally has poor results.

Consequently, PCA with K-means emerges as the solution of choice at the end of the day. This model can be used to make fair decisions.

REFERENCES

1) www.wikipedia.com
2) Pattern Classification Second Edition by Duda et. al.
3) Lecture Slides, CSL:2050 Spring Term 2023, Dr. Richa Singh, IIT Jodhpur