Clustering Analysis using K means and Fuzzy C means algorithm

MTH514A: Multivariate Analysis Project Report

May 10, 2021

Ankit gupta(191016)
Rajat Agarwal (191104)
Vinay kumar sharma(191171)
Sourav Mandal (191149)

Department of Mathematics and Statistics , IIT KANPUR

Contents

1 Acknowledgement

We would like here to express our heartfelt gratitude to Dr. Minerva Mukhopadhyay for helping us in every difficulty which we face during this Project. It has been a great learning experience and has also provided us with a practical insight of the theoretical knowledge gathered during the course : MTH-514A(Multivariate analysis).

We also take this opportunity to thank the authors and publishers of the various books ,journals and blogs we have consulted. Without those this task would not have been completed.We would also like to thank our parents for their extensive support throughout the session. Their constant encouragement has enabled us to complete the project within the stipulated time-period.

2 Introduction

One of the popular machine learning technique is unsupervised learning . Using unsupervised machine learning techniques machine learns to assign a data in particular cluster or classes based on the feature or attribute of the observation or data .Classes or cluster are constructed such that there is high intra class similarity and low inter class similarity . One interesting thing in unsupervised learning is that ; here there is no response variable belonging to each data or observation thus in general we say in unsupervised learning there is no supervision of any response variable while assigning the observation to different classes .
In our report we will present the clustering analysis of country dataset using two popularily known clustering method namely :-

- K-Means Clustering (Hard clustering) :- Each data point belong to exactly one cluster
- Fuzzy C-means Clustering (Soft clustering):- Each data point belong to more than one cluster based on its belongingness or membership to clusters

3 K Means Clustering

K-means clustering is a unsupervised Machine learning algorithm. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the the nearest cluster centers or cluster centroid.

Let $C_1, C_2, ..., C_K$ be the sets containing ... respective observations in each cluster in such that - $C_i \cap C_j = \varnothing$ for all $i \neq j$ i.e. clusters are non overlapping.
Then we define Within Cluster variation by-
$W(C_k) = \sum_{i \epsilon C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$; $\bar{x}_{kj}$ mean for feature j in cluster C_k

Or main target is to minimise sum of all within cluster variablity i.e. $\min_{C_1, C_2, ..., C_k} \sum_{k=1}^{K} W(C_k)$ as small as possible.

Thus our overall optimization problem is $\min_{C_1, C_2, ..., C_k} \sum_{k=1}^{K} \sum_{i \epsilon C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$

3.1 Algorithm to implement the K-means algorithm

1. Randomly assign a number from 1 to K to data observation and consider this K observation as initial cluster center
2. Assign each observation to the nearest initial cluster
3. For each of the K cluster compute the cluster centroid
4. Repeat 2nd and 3rd step until the cluster centroid is not changing

3.2 Advantages of using K-means algorithm

1. Relatively simple to implement.

2. Scales to large data sets.
3. Guarantees convergence.
4. Can warm-start the positions of centroids.
5. Easily adapts to new examples.
6. Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

3.3 Dis advantages of using K-means algorithm

1. We have to choose the value of K manually.
2. K-means Clustering is dependent on initial values
3. K-means has trouble clustering data where clusters are of varying sizes and density.
4. K-means Clustering is very sensitive for outliers.the Cluster Centroids can be dragged into one sides. So ouliers should be removed before implementing K-means Clustering.
5. K-means Clustering is a distance-based similarity measure, so if the dimensionality increases, it converges to a constant value .

4 Fuzzy C-Means Clustering

It is an extension of K-mean algorithm developed by J.C. Dunn in 1973 and further improved by J.C. Bezdek in 1981 .It allows data point to assigned into more than one cluster .Basically , it assigns a degree of membership of each data point to several cluster , which sums to one for all cluster . Degree of membership of each data point to various cluster is calculated based on the distance between data point and cluster . more closer the data point is to cluster than higher will be the degree of association of the data point to the that or nearest cluster . Illustration of fuzzy c mean clustering are depicted below:-

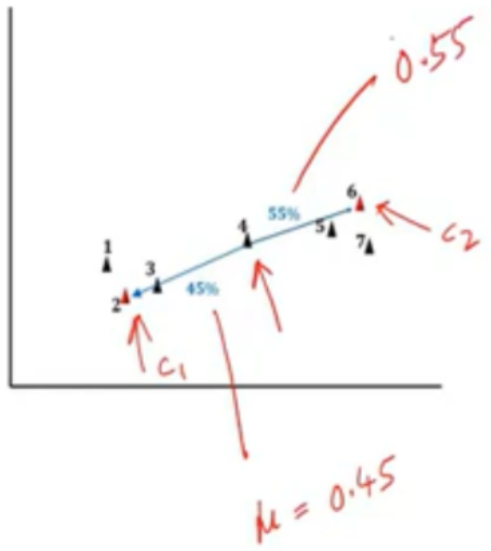


Figure 1: Fuzzy C-Means Clustering

In the above figure (1,2,3,4,5,6,7) are the data point and (2) and (6) are the cluster center C1 and C2 respectively . Based on the distance of data point (4) from centre C1 and C2 ,a degree of association of 4th data point to two center is given which is equal to 45% with center C1 and 55% with center C2 . this is what fuzzy c mean clustering algorithm do .

4.1 Steps to implement the Fuzzy C-means algorithm

Suppose we have dataset $(X_1, X_2, X_3, ..., X_n)$. Where each $X_i \epsilon R^d$. d is dimension of feature of each data point X_i. . Now to implement the fuzzy c means algorithm we will follow the below steps :-

1. Randomly select K cluster center from the data $(X_1, X_2, X_3, ..., X_n)$. here K < n . K represnt the total number of cluster center and n is total data points .

2. Calculate the association of each data point with respect to different cluster center

using the below formula :

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{K} \left(\frac{\left\|\mathbf{x_i}-\mathbf{c_j}\right\|}{\|\mathbf{x_i}-\mathbf{c_k}\|}\right)^{\left(\frac{2}{(m-1)}\right)}}$$

$\|\mathbf{x_i} - \mathbf{c_j}\| = [\sum_{d=1}^{D} (x_{id} - c_{id})^2]^{\frac{1}{2}}$; D is the dimension of feature space of the data point
where μ_{ij} represents the degree of association of X_i data point ; i = 1(1)n to C_j cluster center ; j = 1(1)K .Also , $\sum_{j=1}^{k=K} \mu_{ij} = 1$, i.e. sum of degree of association of each data point with K different cluster is equal to 1 . m$\epsilon[1, \infty]$ is weighting parameter controls the amount of fuzziness in the clustering process.

3. Now , after calculating the degree of association . we will recalculate our cluster center based on the above information. Each cluster coordinates for every cluster can be calculated as follows :

$$c_{jd} = \frac{\sum_{i=1}^{n} (\mu_{ij})^m x_{id}}{\sum_{i=1}^{n} (\mu_{ij})^m}$$

here j = 1(1)K and d is the dimension of the feature space of the data point .

4. Let $U^{(k)}$ be the matrix of association i.e. $U^{(k)} = [\mu_{ij}^{(k)}]; i = 1(1)n; j = 1(1)K$. So , we will repeat step 2 and step 3 until a stopping criteria is reached that :-

$$\left\|\mathbf{U^{(k+1)}} - \mathbf{U^{(k)}}\right\| < \epsilon$$

where ϵ is any small threshold or stopping value .$\epsilon > 0$

The objective function in the fuzzy c-means algorithm is:

$$\arg\min_{U,C} \sum_{i=1}^{n} \sum_{j=1}^{K} (\mu_{ij})^m \|\mathbf{x_i} - \mathbf{c_j}\|$$

$\|\mathbf{x_i} - \mathbf{c_j}\| = [\sum_{d=1}^{D} (x_{id} - c_{id})^2]^{\frac{1}{2}}$; D is the dimension of feature space of the data point .
U = $[\mu_{ij}]$is the association matrix and C = $[c_{jd}]$ is the matrix of cluster center .

4.2 Advantages of using Fuzzy c-means algorithm

1. In case of overlapped dataset fuzzy c means give comparatively better result than k-means algorithm

2. Here data point are assigned a degree of association as a result data point may belong to more than one cluster

4.3 Dis advantages of using Fuzzy c-means algorithm

1. We do not the number of cluster here

2. As we decrease the threshold("ϵ") we need more iteration to converge the algorithm

5 Clustering Analysis of Country Data using K-Means algorithm and Fuzzy C-Means algorithm

5.1 Objective

Our objective is to help the "HELP International"(an international humanitarian NGO, fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities) by categorising the countries using socio-economic and health factors that determine the overall development of the country.

AIM :- Find a group of country with similar socio-economic and health factor ,so that NGO can help the group of country which don't have a good socio-economic and health condition in there country

5.2 Description of the Dataset

We take Country dataset (from Kaggle) which have 167 distinct country socio-economic information . Dataset has 167 rows and 10 columns variable . The columns variable of the data are following:

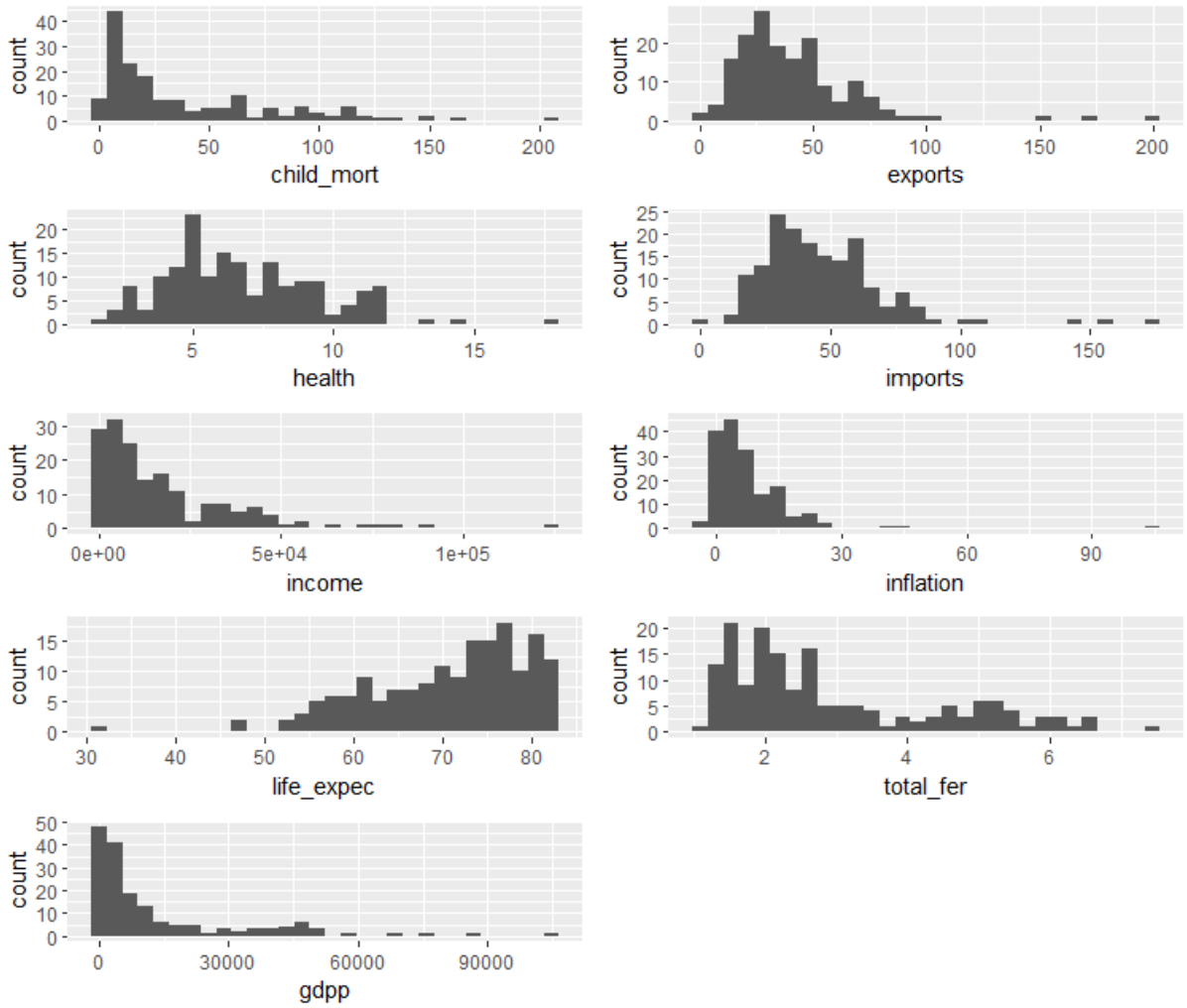
1. Country :- Name of the country
2. Child_mort :- Death of children under 5 years of age per 1000 live births
3. Exports :- Exports of goods and services per capita . given as % age of the GDP per capita.
4. Health :- total spending for health per capita . Given as %age of gdp per capita
5. Imports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

5.3 EDA (Exploratory Data Analysis)

As, there are 10 predictors column in our data set in which one is for identity of country and rest 9 are basically used for further study.All these variables are numeric in nature.On analysing data we found that there is no missing observation in data.The basic summary of different predictors is given as

```
   country           child_mort        exports           health          imports           income         inflation      
 Length:167         Min.   :  2.60   Min.   :  0.109   Min.   : 1.810   Min.   :  0.0659   Min.   :   609   Min.   : -4.210  
 Class :character   1st Qu.:  8.25   1st Qu.: 23.800   1st Qu.: 4.920   1st Qu.: 30.2000   1st Qu.:  3355   1st Qu.:  1.810  
 Mode  :character   Median : 19.30   Median : 35.000   Median : 6.320   Median : 43.3000   Median :  9960   Median :  5.390  
                    Mean   : 38.27   Mean   : 41.109   Mean   : 6.816   Mean   : 46.8902   Mean   : 17145   Mean   :  7.782  
                    3rd Qu.: 62.10   3rd Qu.: 51.350   3rd Qu.: 8.600   3rd Qu.: 58.7500   3rd Qu.: 22800   3rd Qu.: 10.750  
                    Max.   :208.00   Max.   :200.000   Max.   :17.900   Max.   :174.0000   Max.   :125000   Max.   :104.000  
  life_expec      total_fer          gdpp        
 Min.   :32.10   Min.   :1.150   Min.   :   231  
 1st Qu.:65.30   1st Qu.:1.795   1st Qu.:  1330  
 Median :73.10   Median :2.410   Median :  4660  
 Mean   :70.56   Mean   :2.948   Mean   : 12964  
 3rd Qu.:76.80   3rd Qu.:3.880   3rd Qu.: 14050  
 Max.   :82.80   Max.   :7.490   Max.   :105000  
```

Histograms of different predictors are given below.By seeing them we can say that "life expectancy" is left skewed and all the other predictors are right skewed.

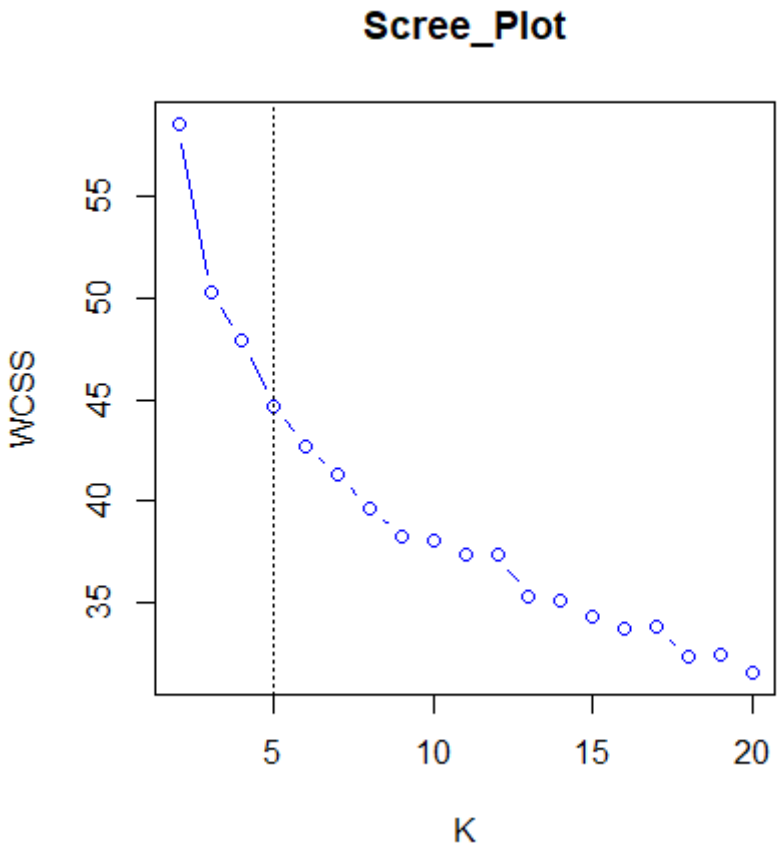


5.4 Methodology

We implement K-means Clustering and Fuzzy C-mean clustering algorithm for clustering the country Country dataset with the basis of their socio-economic and health factors .

5.5 Clustering Analysis using K-Means

Choosing K-value For choosing suitable no. of clusters K, we need such value for k so that within cluster variation is minimum . we are using scree plot to choose optimum number of cluster . Below is the scree plot :-

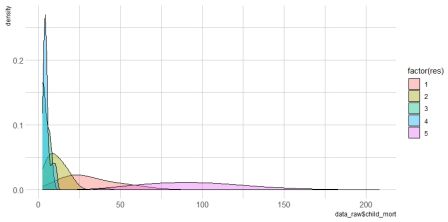


.5

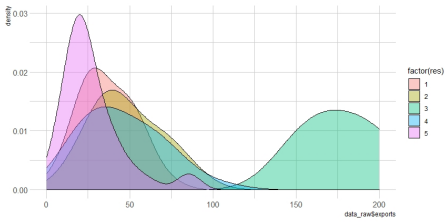
From the plot it is clear that after k = 5 the value WCSS shows a strict trail off . so , we can consider that optimum number of cluster that partition the data is 5 . Using k mean clustering with k = 5 , we have cluster the data point in 5 cluster and the results below shows the 10 country belonging to each cluster ordered in the ascending order of their income within cluster :-

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Nepal	Liberia	Greece	Maldova	Congo, Dem. Rep.
Tajikistan	Rwanda	Israel	Vietnam	Burundi
Bangladesh	Kiribati	New Zealand	Georgia	Niger
Cambodia	Solomon Island	Spain	Ukraine	Central Africa R.
Kyrgyz Republic	Lesotho	Japan	Bosnia and Herz.	Mozambique
Vanuatu	Micronesia, Fed. Sts.	Italy	Albania	Malawai
Myanmar	Namibia	United Kingdom	Tunisia	Guinea
Lao	South Africa	France	Maldives	Togo
Uzbekistan	Iraq	Iceland	Macedonia	Sierra Leone
India	Botswana	Finland	Serbia	Guinea-Bissau

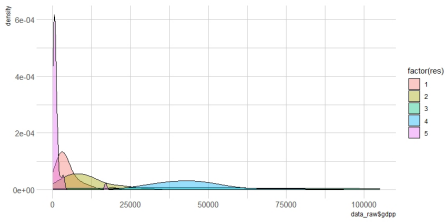
We have plotted density plot of socio-economic feature of observation belonging to each cluster .Below are the plots:-



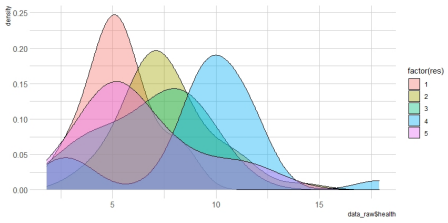
(a) Child mortality



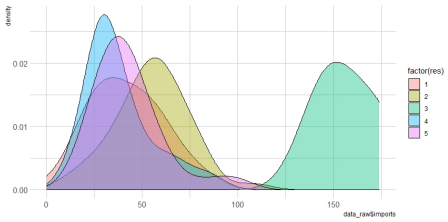
(b) Exports



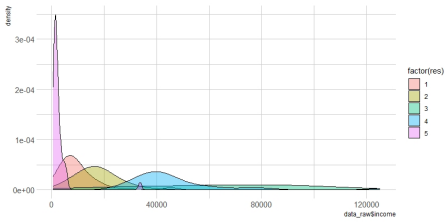
(c) GDP per Capita



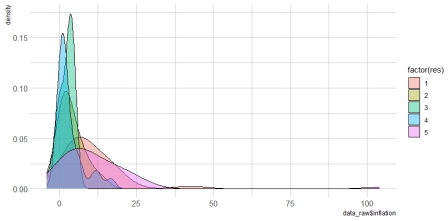
(d) Health



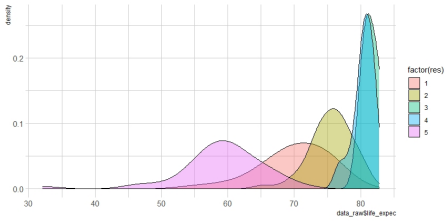
(e) Import



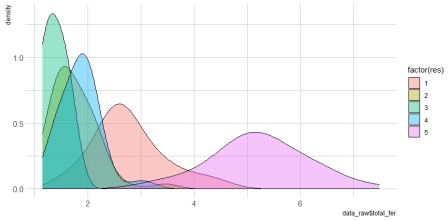
(f) Income



(g) Inflation



(h) Life Expectancy



(i) Total fertility

Figure 2: Density plot of variable belonging to each cluster

5.5.1 Cluster wise Characteristic

Here, we draw this conclusion from the above graphs and summary statistics. We compare the clusters on basis of different sectors like health, economy, gdpp etc.

1st Cluster

In 1st cluster Life expectancy is around 72. Total fertility is around near 3. It has high rate of child mortality around 25 per 1000 live births. 5% of total gdp is invested in health sectors.

GDPP in this cluster is very low but better than Cluster 5. Performance in both of the exporting and importing per capita are average overall for this Cluster.

Income very low but relatively better than cluster 5. Many countries in this cluster has higher inflation rate.

2nd Cluster

Life expectancy is around 76. Total fertility rate is around near 1. These countries achieved a descent success in child mortality rate, the min value of this cluster is around 8 per 1000 live birth. Around 7 % of gdp in invested in heath.

GDPP is below average. Average in overall exporting , but high import(58% of gdpp) for per capita is observed for this cluster.

These countries' Income is average and Inflation rate is moderate.

3rd Cluster

Life expectancy is around 82. Total fertility rate is around near 1(less than). It has very low in child moratality rate (5 per 1000 live births). High variability of investment in health sector is shown for different countries(from 5 to 10 % of gdpp).

GDPP is very high in these countries. The cluster has a tendency of very high export (near 175) and very high import (average above 150) per capita.

Average income is highly variable . Inflation rates are low.

4th Cluster

People have very high life expectancy (around 82). Total fertility rate is around near 2. Very Successful in controling child mortality rate i.e. around 6 per 1000 live birth. Heavy Investment in health sector(on an average of 10 % of gdp)

GDPP is very high (45000 usd). Overall total export for per capita is average and import per capita is low.

Average Income is very high (45000 usd). Inflation rates is low.

5th Cluster

Life expectancy is around 60, which is lowest among all other clusters. Total fertility is very high (around 5). It has very high rate of child mortality (around 80 per 1000 live births).Most of the countries in this cluster invested 5% gdp in health, while some invests around 10% of gdp.
GDPP is lowest among all the clusters. low export per capita.average import per capita. Income is extremely low. Most of the countries in this cluster has higher inflation rate.

Conclusion from K-Means Clustering

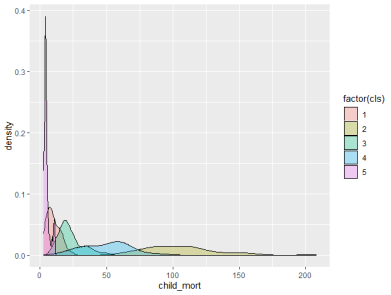
Finally we can say that Cluster 4 is the collection of best possible Countries. It has higher life expectancy, very low child mortality rate, highest income,low imports and average exports and Inflation rates are low. Cluster 3 immediately comes after cluster 4 for overall development beacuse of similarity of cluster 4, but the import and export are very high. Cluster 2 performs intermediate among all the cluster. While Cluster 5 consists of worst country of all aspects from health to income. Though after very bad report of cluster 1, it manages relatively better in situation than Cluster 5 in life expectancy, income, export and gdp per capita.

5.6 Clustering Analysis using Fuzzy C-means

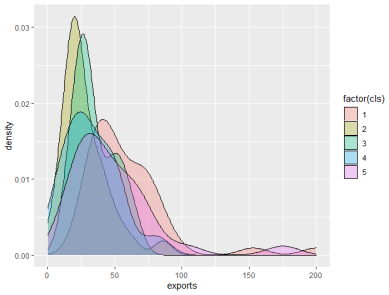
In this clustering technique, again whole data is divided into five cluster, and some members of these clusters are given in table below,

2

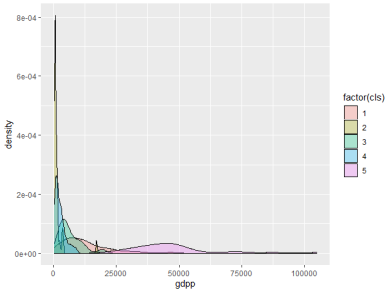
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Maldova	Portugal	Rwanda	Bangladesh	Congo, Dem. Rep.
Vietnam	Greece	Madagascar	Uzbekistan	Liberia
Georgia	Israel	Eritrea	Cape Verde	Burundi
Ukraine	New Zealand	Kiribati	Bhutan	Niger
Bosnia and Herzegovina	Spain	Solomon Islands	Morocco	Central African Republic
Albania	Japan	Nepal	Armenia	Mozambique
Tunisia	Italy	Tajikistan	Paraguay	Malawi
Maldives	UK	Kenya	El Salvador	Guinea
Thailand	France	Cambodia	Fiji	Togo
Panama	Iceland	Kyrgyz Republic	Mongolia	Sierra Leone
Malaysia	Finland	Vanuatu	Belize	Guinea-Bissau
Poland	Germany	Ghana	Jamaica	Comoros



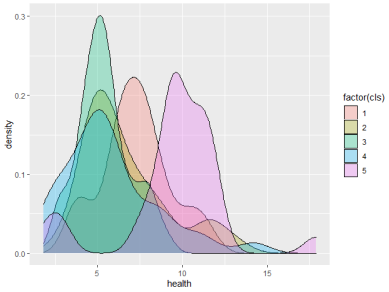
(a) Child mortality



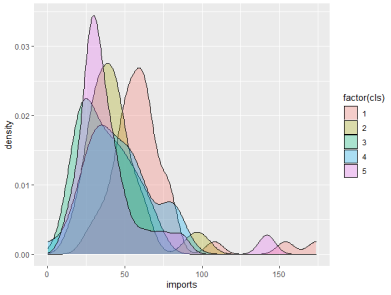
(b) Exports



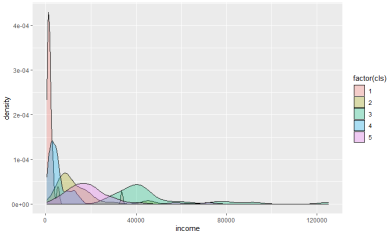
(c) GDP Per Capita



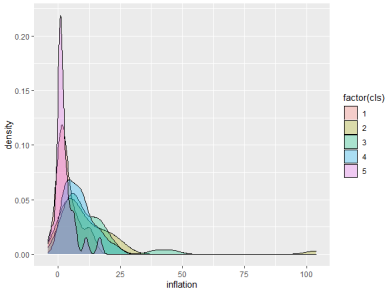
(d) Health



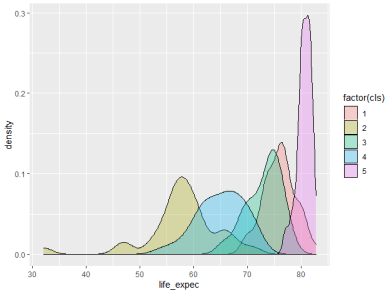
(e) Import



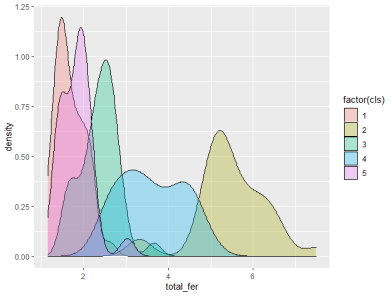
(f) Income



(g) Inflation



(h) Life Expectancy



(i) Total fertility

Figure 3: Density plot of variable belonging to each cluster

5.6.1 Cluster wise Characteristics

Now, we will visualise nature of different variables in different cluster and with the help of summary statistics and these graphs we draw some meaningful information from these cluster.

First cluster
There are 40 countries we categories in this cluster.The child mortality is around 10 children per 1000 on average, where as these country spend around 7.5 percentage of their GDP on health issues.Average net income per person is 20360 dollar. The GDP per capita is around 12700.

Second cluster
27 countries are in this cluster.Child mortality for these countries is minimum,also they spent largest portion of GDP on health sector.Net income per person and GDP are also highest for these countries in comparison to the countries of other cluster.

Third cluster
32 countries, we classified into this cluster. Mortality rate is comparatively high from first two cluster. On Average 5.5 percent of GDP is spent on health. Net income per person and GDP is fairly small.

Fourth cluster
There are 36 countries we categories in this cluster.The child mortality is around 22 children per 1000 on average, where as these country spend around 5.4 percentage of their GDP on health issues.Average net income per person is 13500 dollar. The GDP per capita is around 6500.

Fifth cluster
32 countries are classified into this cluster. Mortality rate for these countries is highest.Also percent share of GDP on health sector is small. Average net income per person is lowest compare to other clusters. GDP per capita is also very small.

Conclusion from fuzzy c means analysis
From the above discussion, we clearly conclude that countries of cluster five have high mortality rate, having low hospital facilities, GDP and income per person.So largest portion of fund should be allotted to these countries. On the basis of these criteria, the countries of third cluster must be get second highest portion of fund And clearly countries of second cluster are most prosper and rich.

6 Conclusion

In the report we have briefly discussed about the two popular unsupervised machine learning algorithm i.e. K Means and Fuzzy C Means clustering algorithm . We have discussed that for a overlapped data how fuzzy C Means algorithm gives comparatively better result than K Means by assigning a degree of belongingness to different cluster . One common disadvantage in Fuzzy C Means and K Means is that we don't know the optimum number of cluster which we have to choose wisely before doing any clustering . we have illustrated how scree plot helps to choose optimum number of cluster for a given dataset .

References

[1] James C Bezdek, Robert Ehrlich, and William Full. Fcm: The fuzzy c-means clustering algorithm. *Computers & geosciences*, 10(2-3):191–203, 1984.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[3] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

7 R Code and Data file link

https://drive.google.com/drive/folders/1fh0B70KbsTNV9lIb8DpcIH4XxpDQCP89?usp=sharing