COMPARISONS OF K-MEANS AND FUZZY C-MEANS ALGORITHM BY CLUSTERING COUNTRIES

Authors
Talaat AL RAHHAL
Assoc. Prof. Ömer Faruk RENÇBER

Editor
Prof. Dr. Sadettin PAKSOY

İKSAD
Publishing House

COMPARISONS OF K-MEANS AND FUZZY C-MEANS ALGORITHM BY CLUSTERING COUNTRIES[1]

Authors

Talaat AL RAHHAL[2]

Assoc. Prof. Ömer Faruk RENÇBER[3]

Editor

Prof. Dr. Sadettin PAKSOY

İKSAD Publishing House

[1] This book has been produced from the master's thesis study at Gaziantep University, Institute of Social Sciences, Department of Business Administration.

[2] Gaziantep University Gaziantep, Türkiye

[3] Gaziantep University Gaziantep, Türkiye

Institution of Economic Development and Social Researches Publications®
(The Licence Number of Publicator: 2014/31220)
TURKEY TR: +90 342 606 06 75
USA: +1 631 685 0 853
E mail: iksadyayinevi@gmail.com
www.iksadyayinevi.com

ISBN: 978-625-8323-67-2
Cover Design: İbrahim KAYA
August / 2022
Ankara / Turkey
Size = 16x24 cm

PREFACE

In a rapidly developing and changing world, millions of data are produced every second. It is of great importance to transform these data into understandable and interpretable information. The main purpose of data mining techniques used for this purpose is to learn data and understand the past or predict the future.

Many applications in this field, also called machine learning techniques, affect human life directly or indirectly. For example, image processing in the engineering field or the creation of computer recommendations in the diagnosis and treatment of disease in the health field are some of these applications. In the social sciences; It can be said that machine learning techniques have many application areas for regression, classification, or clustering.

In this book study, theoretical and application examples of frequently used types of clustering algorithms are given. The main purpose of the book is to cluster the countries in terms of macroeconomic variables with K-Means and Fuzzy C-Means techniques.

Cluster analysis; hierarchical or non-hierarchical clustering methods are examined in two groups. Accordingly, clustering algorithms work with the logic of maximum similarity within the cluster and maximum difference between clusters. In hierarchical cluster analysis; The number of decision units is not determined, each of them is handled separately at the beginning, or they are clustered by the merging method or by the separation method, all of which is considered a cluster.

In this book, K-Means and Fuzzy C-Means techniques are used to cluster the countries. I hope that the book will lead international literature, data scientists and innovative approaches.

Finally, I would like to thank to Talaat AL RAHHAL and Assoc. Prof. Ömer Faruk RENÇBER, the authors of this book. In addition I would like to thank to Sefa Salih BİLDİRİCİ, President of İKSAD Publishing Group, who contributed to the publication of the book, to Dr. Mustafa Latif EMEK, to İbrahim KAYA and to all other publishing house staff.

Prof. Dr. Sadettin PAKSOY
Editör

CONTENTS

CHAPTER 1

INTRODUCTION

INTRODUCTION

Clustering is a concept that expresses the process of understanding, analyzing, interpreting, and transforming large amounts of data into information. Knowledge can be discovered through the use of data. Methods in this field include purposes such as pattern recognition, model creation, or making predictions based on the created model. However, the increase in digital data today brings with it problems such as not being able to be easily analyzed and interpreted.

This paper analyzes how to cluster countries according to multiple variables and argues that using the suitable clustering techniques can be determined according to the type of data set. Most studies have classified countries according to their income or financial position, and this idea is right and useful because that research can be useful for both firms and countries. See, for example, the study by Dirsehan (2015), where he clustered countries according to their Export Competitiveness, and another study by Nielsen (2011), where he clustered countries based on their Level of Development.

Moreover, a lot of classifications of countries are based on income level, such as New World Bank country classifications by income level. The purpose of firms when using countries' classification is to decide where to operate and which countries, they may focus on to use them for their development.

Even if the classification is in terms of development, the categories are named according to economies which are low- and middle-income as developing economies, and to economies which are upper middle-income and high-income as developed countries.

"The World Bank clusters the world's economies according to four income groups: high, upper-middle, lower-middle and low by considering Gross National Income (GNI) per capita (current US$)" (Cheung, 2020).

However, here again the classification is made according to the income, and that is one variable that you measure and according to it you decide where to categorize the country and which cluster to put it in.

For example: Big and noticeable differences can be seen in the income level distributed among citizens of different countries. An example can be given of an Indian citizen in 2010 who earns on average US $1,357.56 compared to US $ 7,761.65 for a Lebanese citizen.

In 2013 inequality in education in India was 42.1 % as compared to 24.1% in Lebanon. However, in 2013 inequality in income in India was 16.1% while in Lebanon it was 30%. Moreover, in 2013 the total unemployment rate (% of labour force) in Lebanon was 6.4% while it was 2.8 % in India. So how is it possible to understand which country is better than the other?

Moreover, a country which is better in one aspect than another may be worse in another aspect than the same country. It is understood here that not only the financial position refers to the position of the country or to define that it is a developing or developed country according to only one variable.

"The word pair developing/ developed countries became in the 1960s the more common way to characterize countries, especially in the context of policy discussions on transferring real resources from richer (developed) to poorer (developing) countries" (Lester B & Pearson , 1969).

To enable better understanding of the position of each country regarding the variations in social and economics result among others, it would be better to place them into groups/clusters. That would be possible by using one of the cluster analysis methods.

While a lot of studies would easily accept the fact that the USA is a better country than Algeria, they would be more undecided to locate Lebanon or India in the clusters classified. Where to put each country properly in which cluster is not obvious, and that requires using the right clustering method.

Therefore, this study is based on a group of interesting indicators selected from the Human Development Data Center where this resource of data is sourced from global agencies and experts with delegation to collect domestic data on particular indicators.

Firstly, eight interesting indicators are collected as follows: Gross domestic product (GDP) per capita (PPP $), Inequality in education, Inequality in income, Life expectancy Index, Mean years of schooling, Unemployment rate, Exports and imports and Foreign direct investment (net inflows).

Regarding the above-mentioned indicators, you may note that the data for all the variables have been collected for the same years which are from 2013,

2014 ... to 2018). Regarding data availability, missing values have been replaced by using SPSS software (the explanation of the process applied will be mentioned below).

In total in this research, there are 189 countries included. In this research, countries will be classified according to many variables that can give a clear answer regarding where a country is located in the table of classifications.

So, to start clustering countries or any other topic, first you need to understand what clustering analysis is, and the most important thing is to specify which clustering method/type you need to use to get the right results, or in other words, the most accurate ones.

In this study, each type of cluster analysis will be defined, but the focus will be on using two clustering methods, which are the k-means method and fuzzy c-means method, and make the comparison between them, which can be done by comparing the results of each cluster method with the other. It is to be hoped that this article will help to understand how to choose the right analysis method for the right type of variables.

CHAPTER 2

THEORITICAL REVIEW

2. THEORITICAL REVIEW

2.1. Cluster Analysis and Data Mining

2.1.1. Definitions and Usage

To understand the concept of cluster analysis in a better way is to understand that the word of cluster refers to a combination of data points gathered together due to particular similarities. A type of statistical method for processing data is cluster analysis. It processes in a way that organizes data and split them into groups, or classifies them based on the way how they match each other.

Cluster analysis cares in data matrices whose variables have not been classified priorly to standard against predictable sub-clusters. The aim of applying cluster analysis is to discover identical groups of items, where similarities between some items mean some universal metrics over a full set of characteristics.

So, the main idea of using cluster analysis, in a nutshell, is to find the similarities and differences between groups. *"Understanding the world requires conceptualizing the similarities and differences between the entities that compose it" (Tryon & Bailey, 1970).*

A Data mining Algorithm is an important algorithm that is helpful in dataset analyzing and improving techniques to determine significant patterns. It is one part of machine learning algorithms. This algorithm is applied to different programming like R language, Python.

There are many data mining algorithms that are based on statistical and mathematical formulas applied to the data set. "The process of grouping the data into classes or clusters, so that the objects within the same cluster have higher degree of similarity in comparison to one another but are very much dissimilar to the objects in different clusters" (Kaufmann, Han J, & Kamber M, 2000).

The reason why cluster analysis is defined under the algorithm of unsupervised learning is that because it is not known the number of clusters exists in the data set before running the proper cluster type. Unlike much of

other statistical methods, when there are no presumptions made about the probable relationships within the data, cluster analysis is used. It supplies information about where similarities and dissimilarities exist in data.

Clustering is an unsupervised machine learning method of determining and gathering homogeneous data points in bigger datasets. Clustering is typically applied in order to distribute data in a shape that can be more distinguished and comprehended.

While it is known that clustering is a common strategy but, it isn't a monolithic term, as there are multi algorithms that use cluster analysis with various techniques. In general, it might be said that the best way to deal with any data is to cluster them in groups where each group expresses what is the feature or the characteristic of itself.

"One of the most important means to deal with data is classifying or grouping it into clusters or categories". "Classification has played an important and an indispensable role throughout human history" (Yannis Goulermas, Ananiadou, Mu, & Brockmeier, 2018) (Wu et al, 2008).

2.1.2. Dealing With Large Unorganized Dataset

The value of using cluster analysis lies in the amount of effort that can reduced and taken off the hands. Comparing with other unsupervised learning machines, clustering can handle with big datasets and arrange them into something more understood and applicable. If you are not aiming to get completed inclusive results, clustering can provide quick answers about any dataset.

"Clustering techniques or methods are proposed based on some form of documents presentation, similarity and machine learning algorithms" (Tarau & Mihalcea, 2005).

2.1.3. Understanding How Many Clusters The Dataset To Be Divided

Despite having more organized or well-labelled datasets, it might not satisfy you or give you what you were looking for. Clustering is a good initial start in preparing a dataset and get the answers to queries. For example, it might be discovered that what was thought are two essential subsets are in fact three.

2.1.4 Why Use Clustering In Data Mining?

Data mining can be used by many organizations when they want to analyze their data. Running data mining model help business to collect valuable information that can raise revenue or reduces costs.

The function of data mining is finding the connections between data. There are a decent number of applicable data mining software obtainable to be used. *"Clustering and classification are both fundamental tasks in Data Mining. Classification is used mostly as a supervised learning method, clustering for unsupervised learning (some clustering models are for both). The goal of clustering is descriptive, that of classification is predictive" (Vayssieres & Plant, 1998).*

2.1.5. What Are The Benefits Of Clustering in Data Mining?

Companies are usually using data mining to focus on their clients. Many fields such as Financing or marketing benefit from data mining, the reason why data mining is so useful for these kinds of businesses is that it gives the opportunity to the companies to get the most benefits of their data, and using clustering techniques helps them again to have additional advantages from the data they get.

And this is how they can inspect the relationship between inner factors such as prices, skills of staff, product quality, and outer factors like competitions with other companies or demographic of clients.

Using clustering and analyzing data can have a massive effect on sales and clients' satisfaction, having precious appliances to raise revenues and decrease expenses.

It ought to not only confine the application of clustering on business but also there are a lot of applications that can be mentioned.

2.1.6. The Implementation of Cluster Analysis in Data Mining

The benefit of using cluster analysis is classification, items are classified into clusters so each item is more to be identical in its characteristics to other item in its cluster than items in other clusters.

In marketing domain, this benefit can be used to help in specifying categories such as ages group, revenue, urban or suburban areas or other categories. Cluster analysis can be used in separating clients into groups, so the goal can be various clients' groups.

Cluster analysis can be used by healthcare researchers to reach out the reason for any health problem by linking some variables that might be a possible factors contributing to these particular diseases.

Regardless of the type the study you are making, or the algorithm you are using, there will be a cross-point with using the clustering technique. Clustering can make marketers understand their database better and can distinguish their clients' groups according to their buying types.

Cluster analysis in the scope of biology might be used to derive plant and animal taxonomies and more. Cluster analysis in the scope of geography might be used to identify the group of houses in an area regarding the type, value, and location of the house. Cluster analysis in the scope of finding information helps in classifying documents.

Another application has been applied by (Tiwari & Misra, 2011) Where they mentioned in their article that "Agricultural expert systems are being used extensively almost in every walk of life. Various tools also have been evolved for evaluating, justifying, upgrading, and modifying the existing agricultural expert systems thus making them more useful in their intended purposes.

And to do so cluster analysis has been applied as the tool of improving agricultural management, predicting and suggesting solutions for its problems, and briefly discusses few such efforts" (Tiwari & Misra, 2011).

Another application has been applied by Albert S. Kyle and Anna A. Obizhaeva and Nitish Ranjan Sinha and Tugkan Tuzun Where they mentioned in their article that "News articles clustering is a wide area of research that has been on for a very long time in history which includes several tasks that range from segmenting events of news streams to tracking and detecting events" (Kyle, Obizhaeva, Sinha, & Tuzun, 2012).

Clustering of objects are used in many fields in life and in different periods of time as Rokach, L., & Maimon mentioned in their book that "Clustering of objects is as ancient as the human need for describing the salient characteristics of men and objects and identifying them with a type." (Rokach & Maimon, 2005).

CHAPTER 3

LITERATURE REVIEW

3. LITERATURE REVIEW

3.1. Types of Clustering

In general there are two types of clustering:

A. Hard Clustering: in this clustering, every data item is supposed to be related to one cluster entirely. In other words, each data item can be assigned only to a single cluster.

B. Soft Clustering: in this clustering, every data item has a probability of being a member in each cluster, in other words, each data item can be assigned to multi clusters.

"One can distinguish between so-called hard clustering methods that assign each feature vector x to exactly one cluster represented by the CV wj and soft clustering methods implying the concept of fuzzy membership of feature vectors in several clusters" (S. Theodoridis, 2009).

So, what explained earlier, what mentioned above in other words. There are two types, hard cluster, and soft cluster:

Hard clustering refers to a binary kind of grouping - either a data item related to a certain group, or it does not. So, when sorting through the dataset of cars and animals, a data point will either be:

A car (1) or not a car (0) or an animal (1) or not an animal (0). Using this type of clustering, there is no ambiguity in the grouping. On the other hand, there is soft clustering, however, is slightly more ambiguous. Instead of belonging to a certain group, and a certain group only, soft clustering groups together a set of data points based on its likelihood of belonging to a certain group.

For example, given a dataset of cats and dogs, an algorithm may evaluate a specific datapoint as: A cat (0.3) or a dog (0.7). Rather than grouping them in a yes-or-no fashion, there is much more ambiguity and lee-way in this method of clustering.

In general a lot of clustering methods and techniques to use, but there are (6) main techniques of cluster analysis that can be used in data science that will be focused on in this study and define each one of them and on which kind of

data that it deals with or on other words let's say which kind of results are possible to get by applying each of them.

3.2 Cluster Analysis Techniques:

3.2.1 Centroid Clustering (Partitioning Methods)

Centroid clustering is a common methodology used in clustering. In this methodology, the number of clusters that need to classify should be chosen. For instance, dividing the market into different geographic units, or dividing the customers by the number of products they buy.

Centroid Clustering (Partitioning methods) are clustering techniques that subdivide the data sets into a set of k groups, where k is the number of groups pre-specified by the analyst.

"There are different types of Centroid Clustering (Partitioning methods) methods. The most popular is the K-means clustering in which, each cluster is represented by the center or means of the data points belonging to the cluster" (MacQueen, 1967).

"An alternative to k-means clustering is the K-medoids clustering or PAM (Partitioning Around Medoids, which is less sensitive to outliers compared to k-means" (Kaufman & Rousseeuw, 1990).

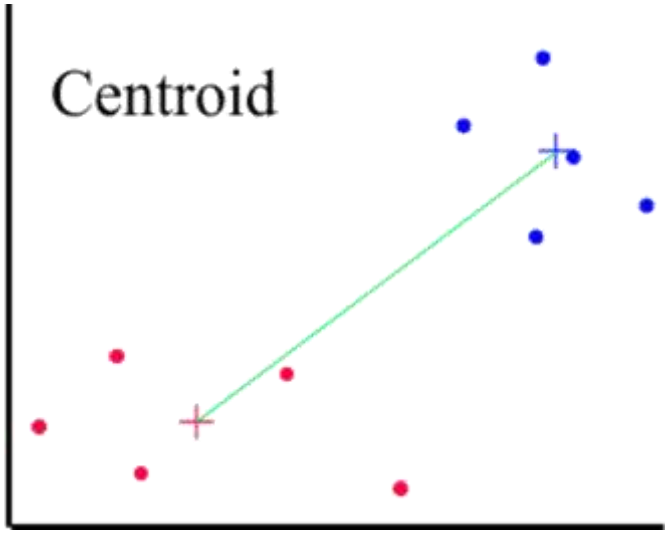

Figure 1: Centroid clustering, (Bansal,2020)

To be able to apply Centroid Clustering Methodology, some requirements need to be satisfied:

A. Each data item should be related to only one cluster.

B. Each cluster should have a specific purpose or criteria

3.2.2 Density Clustering (Model-based methods)

The way density clustering works is by grouping data items according to how they are intensely populated. In order to collect closely linked data items, the Density clustering algorithm works in the way of understanding that the more closely the data items are to each other the more related they are. To do that, the algorithm chooses a point randomly, then begins measuring the range between each point that surrounds it. After that, the algorithm identifies the rest data points that are within the allowed distance of relevance. The procedure continues repeating by choosing different data points randomly until getting the best cluster.

Density-based clustering is one of the partitioning methods that was introduced by Ester et al (1996). *"Density-based clustering can find out clusters of different shapes and sizes from data containing noise and outliers" (Ester, Martin, Xiaowei Xu., Sander, & Kriegel, 1996).*

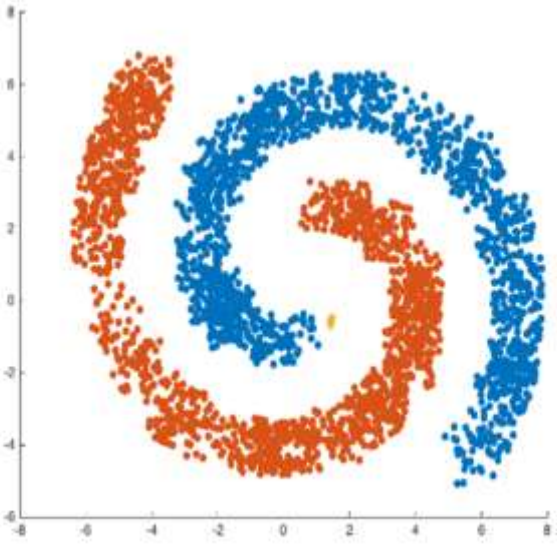

Figure 2: Density Clustering (Bansal,2020)

3.2.3. Distribution Clustering

"Distribution clustering identifies the probability that a point belongs to a cluster. Around each possible centroid, The algorithm defines the density distributions for each cluster, quantifying the probability of belonging based on those distributions The algorithm optimizes the characteristics of the distributions to best represent the data" (Gervelis, 2018).

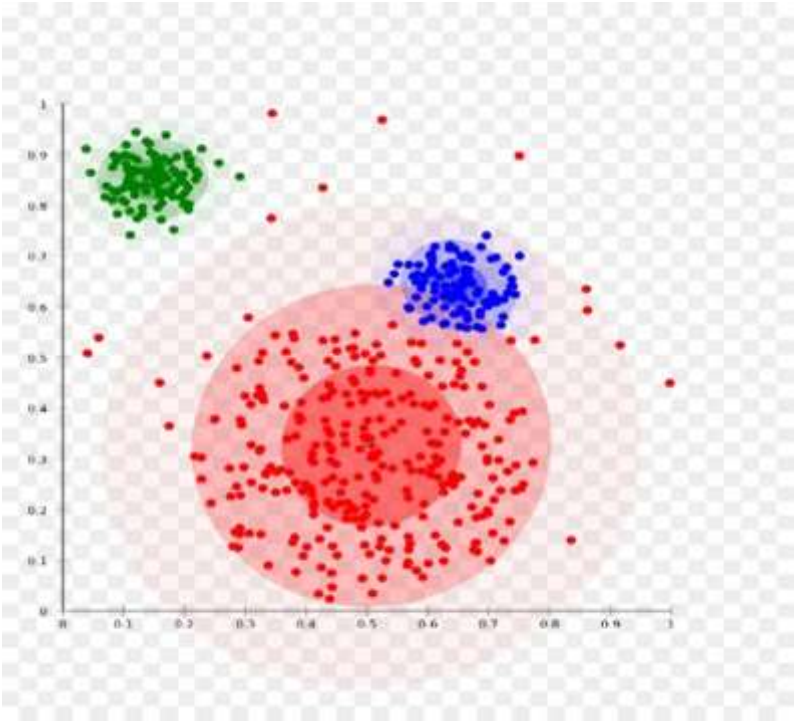

Figure 3: Distribution Clustering (Bansal,2020)

Looking at the clarification picture above, it is seen that data points can be identified to which cluster it belongs regarding where it locates among these bull's eyes.

The distribution clustering method is unlike the density clustering method, where Distribution clustering helps in assigning an outlier to one cluster, as counter to density clustering that does not. It is remarkable that the distribution clustering method only works with synthetic data. In other words data items that belong to predefined clusters.

3.2.4 Connectivity Clustering (Hierarchical Clustering)

Hierarchical clustering is the classification technique to group in “trees” with branches. Topic models are models in which the differentiation features for grouping are “topics” for elements, usually words, it is usually used in natural language processing. When the technique used to group topics is hierarchical, then you have the second concept.

Agglomerative hierarchical clustering, instead, builds clusters incrementally, producing a dendrogram. Based on the below picture, “the algorithm begins by assigning each sample to its own cluster (top-level). At each step, the two clusters that are the most similar are merged; the algorithm continues until all of the clusters have been merged” (Kumar, 2020).

There are two types of approaches for the creation of hierarchical decomposition, which are:

- Divisive: here the entire dataset is taken as a cluster and according to some criteria then trying to split the cluster and make new clusters.
- Agglomerative: here are the points that are taken as clusters and according to some criteria then trying to group the points and make clusters.

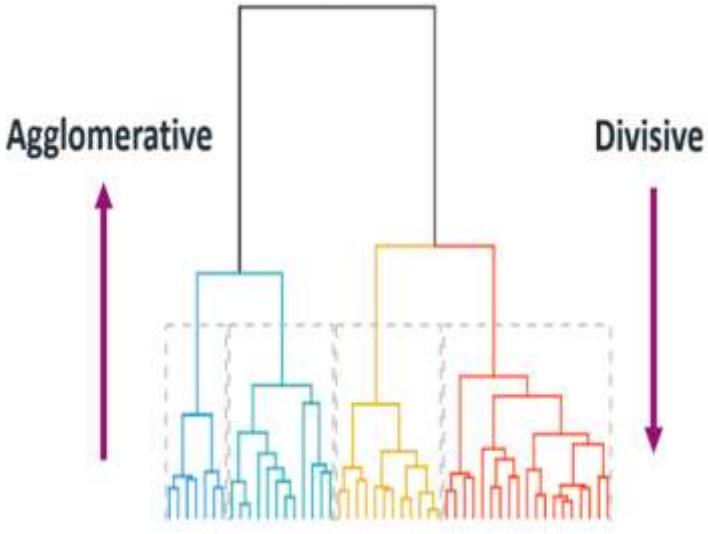

Figure 4: Connectivity Clustering (Bansal,2020)

"The result of the hierarchical methods is a dendrogram, representing the nested grouping of objects and similarity levels at which groupings change". "A clustering of the data objects is obtained by cutting the dendrogram at the desired similarity level". (Rokach & Maimon, 2005).

"The merging or division of clusters is performed according to some similarity measure, chosen so as to optimize some criterion (such as a sum of squares)". "The hierarchical clustering methods could be further divided according to the manner that the similarity measure is calculated" (Flynn, Murty & Jain , 1999).

3.2.5. Fuzzy Clustering

Fuzzy clustering clusters in such a way that each data point can belong to more than one cluster. The data point has a degree of membership in each cluster.

Fuzzy clustering is probabilistic clustering; individuals can belong to more than one cluster for example: (can be 40% cluster 1 and 60% cluster 2).

"In a fuzzy clustering, every object belongs to every cluster with a membership weight that is between 0 (absolutely doesn't belong) and 1 (absolutely belong)" (P.N. Tan , Kumar, & Steinbach, 2005).

There are multiple ways to evaluate membership values obtained from the fuzzy clustering algorithm. It can be either created hard-cuts based on the most probable membership or can look at the individual distributions across clusters to visualize patterns of membership and potential outliers.

3.2.6. Constraint-Based Clustering Method

"A constraint is defined as the desired properties of the clustering results, or a user's expectation on the clusters so formed – this can be in terms of a fixed number of clusters, or, the cluster size, or, important dimensions (variables) that are required for the clustering process" (Prasad, 2020).

Constrained clustering methodology is to restricts each cluster to have a minimum number of points. It might end up in a scenario where the optimal cluster ends up having very few or no points, which is not accepted, so the constrained version can be used to avoid this scenario.

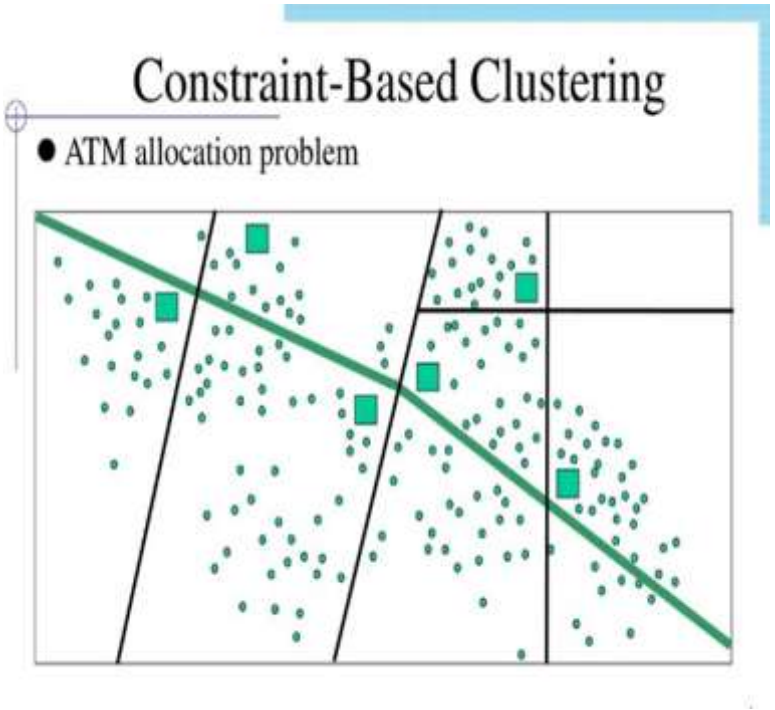

Figure 5: Constraint-Based Clustering, (Bansal,2020)

As it is seen that there are minimum number of ATM points, each point represents an unique cluster that has its own affiliated.

3.3. Choosing the right cluster type/method

Generally speaking, how to choose the best clustering type/method?

There is no definitive answer to this question or maybe it is not right to choose any clustering method and apply it without understanding the cluster models and which one is better to use according to the type of data or the results wanted to get.

Therefore, to understand how to select among various clustering algorithms, it ought to understand these cluster models. Otherwise, it can't be known beforehand which clustering algorithm would be more suitable to be applied. But, there are some guidelines that can be followed and proceed accordingly.

According to the variables (that will be defined later), it should be guaranteed that chosen cluster method has the characteristic which allows the following:

- The number of variables should be 1 at least.
- Input variables type should be quantitative (numerical).
- The number of classes should be chosen prior to computations.
- Results: class membership should be deterministic.

•Density model: is probabilistic model that uses the soft clustering approaches for spreading the points in different clusters.
•Distribution models: offer a "model-based clustering" that derives clusters using probabilistic models that describe the distribution of data. It is a model that describes the distribution of data then assesses probabilities that specific cases are members of specific classes.
•Hierarchical clustering: “if the number of clusters is determined at the beginning of the research, non-hierarchical clustering methods are preferred, and if the number of clusters is not decided, the hierarchical clustering method is preferred” (Burmaoğlu, 2011)
•Soft clustering (fuzzy clustering): each object belongs to each cluster to a specific degree where this led to end up with probabilistic results/class membership.

Finally, the only cluster method which has the characteristics that match what has been mentioned earlier is the partitioning cluster and precisely it is k-means clustering which is the most popular type of partitioning method.

However, in this research, k means and fuzzy c-means are going to be applied will be compared between the results gotten from both methods (hard clustering vs soft clustering).

CHAPTER 4

K-MEANS & FUZZY C-MEANS

4. K-MEANS & FUZZY C-MEANS

4.1. K-Means

4.1.1. K-Means Definition

The term "k-means" was first developed by James MacQueen in 1967 where he states that *"K-means clustering is one of the most commonly used unsupervised machine learning algorithm for partitioning a given data set into a set of k groups" (MacQueen, 1967).* K-Means is an unsupervised algorithm used for clustering. Saying unsupervised it is meant that only given data is unlabeled to run the model.

"The algorithm just depends on the dynamics of the independent features to make inferences on unseen data". "The common way to figure out between two points is the Euclidean distance". (Forina & Lanteri, 1984).

"This k-means approach is a centroid based partitional clustering method, where the centroids are the arithmetically calculated centers of the clusters and where of k is the number of clusters. The initial centroids for each cluster can either be randomly selected or pre-assigned from the data" (P.N. Tan , Kumar, & Steinbach, 2005).

4.1.2. K-Means Algorithm

The algorithm is composed of the following steps:

1-Initially, in a random way select k centroids/cluster centers. Where it is better to make them near the data but in the same spot next to each other.
2-Then allocate each data point to the nearest centroid.
3-Move the centroids to the average position of the data points allocated to them.
4-Repeat the preceding two steps until the allocations don't change.

"The simplest and most commonly used algorithm, employing a squared error criterion is the K-means algorithm. This algorithm partitions the data into K clusters

(c1, c2, . . . , ck), represented by their centers or means. The center of each cluster is calculated as the mean of all the instances belonging to that cluster" (Rokach & Maimon, 2005).

"The optimum number of clusters is the value of k that maximizes the strength" (Granville, 2019).

Accurate proof of the limited convergence of the K-means type algorithms has been given by Selim et al where they state that "The K-means algorithm may be viewed as a gradient- descent procedure, which begins with an initial set of K cluster-centers and iteratively updates it so as to decrease the error function". (Selim & Ismail, 1984).

4.1.3. How do you interpret k-means clustering results?

In K-Means method, each point gets allocated to only one centroid, where points allocated to the same centroid belong to the same cluster. To cluster faces the centroids will look like faces, because if average several faces another will be gotten. To cluster the dataset then centroids will show the average of all the different variants for a digit.

Centroids after clustering the faces dataset, each face is a centroid and looks like a face because it is the result of averaging the faces in its cluster. Kmeans is a form of vector quantization, you can represent each datapoint as its closest centroid so that gives you a lossy compression of the dataset to the number of centroids you have used for K-means.

This trick is used in several algorithms to speed up computation, first, you apply K-means then you apply some algorithm to the centroids and extend the results to all the points associated with that centroid.

You can also use K-means for dimensionality reduction representing each point in k-dimensions using the distance from the point to each centroid. Sometimes the "features" generated by K-Means will make a classifier work better than with the raw points.

K-Means also gives a Voronoi tessellation of the used dataset, each centroid represents a region of points that are closer to that centroid than to any other. This can be useful in several ways. For example, if you are approaching the Near-Neighbors problem you can compare the query point to the k-centroids and then to the points in the cluster for the closets centroid, this avoids having to compare the query against all points in the dataset, of course, if the query

point is near the border between two regions the result can be missing some close neighbors and that's why it is an approximation. So K-Means can give you a clustering result, dimensionality reduction, lossy data compression.

4.1.4. Implementation of K-Means Clustering

The Application of K-Means method is so can be used in many domains such as customer segmentation. Customer segmentation allows businesses to customize market programs that will be suitable for each of their customer segments.

Or it can be used in anomaly or fraud detection such as Separate valid activity groups from bots or detect fraudulent claims.

Also, other applications like inventory categorization based on sales or other manufacturing metrics, creating newsfeeds, cloud computing environment, environmental risks, and pattern recognition in images.

"The use of the K-means algorithm is often limited to numeric attributes" (Huang, 1998). There are many studies in the literature using k-means clustering algorithm for customer segmentation (Kim, K., & Ahn, H. , 2008) (Hruschka & Natter , 1999).

Grouping students according to different characteristics (Oyelade & Oladipupo, 2010), (Baradwaj, B. & Pal, S, 2012); image segmentation (Ng, H., Goh, P.,, Nowinski, W., Ong, s, & Foong, k, 2006), Dhanachandra et al., 2015), (Dhanachandra, N., Chanu, Y., & Manglem, K., 2015).

4.1.5. Reasons For The Algorithm's Popularity

Even if the number of cases is remarkably large, this algorithm is computationally suitable. "*The K-means algorithm has an advantage in comparison to other clustering methods (e.g., Hierarchical clustering methods), which have non-linear complexity". (Rokach & Maimon, 2005).*

Other reasons for the algorithm's popularity are *"its ease of interpretation, simplicity of implementation, speed of convergence and adaptability to sparse data" (Modha & Dhillon, 2001).*

4.1.6. K-means clustering advantages and disadvantages

4.1.6.1. Assumption:

- Assume balanced cluster size within the dataset.
- Assume the joint distribution of features within each cluster in a spherical shape.
- Clusters have similar density.

4.1.6.2 Disadvantages:

- Oftentimes produces clusters with relatively uniform size even if the input data have different cluster sizes.
- Does not work properly with clusters that have different densities but spherical shapes.
- K value is not known, thus running k means for a k value lots of times (20-100 times).
- Does not work properly with outliers.

4.1.6.3. Advantages:

- Normally work properly even some assumptions are broken.
- The simplicity of its implementation.
- Its easiness of interpreting the results of clustering.
- Quick and effective in terms of computational cost.

4.2. Fuzzy C-Means

4.2.1. Fuzzy C-Means Definition

As a counter method of the method of traditional k-means where each point belongs to only one cluster, in fuzzy clustering, every point has a probability of belonging to each cluster Separately and differently from another cluster.

Fuzzy c-means solve the problem where points are located in a way that can not be definitely defined which center is better to be belonged to, and this problem appears when the distances between one point and the surrounded centers are very similar to each other.

Fuzzy c-means specifies the centroids based on these probabilities. Applied processes of initialization, iteration, and termination are the same as the ones used in k-means.

As it is realized that fuzzy c-means and k-means are different from each other in a way of giving the probability to each point, where k-means simply gives 1 if the data point is closest to a centroid and 0 otherwise.

A fuzzy version of K-means, called Fuzzy C-means (FCM) (sometimes called fuzzy K-means), was proposed by (J. Bezdek, 1980), (Bezdek, 1981). Fuzzy C-mean (derived from fuzzy logic) is a clustering technique, which calculates the measure of similarity of each observation to each cluster. Indirectly it means that each observation belongs to one or more clusters at the same time, unlike the traditional clustering also in which each data point is a part of a single cluster only.

Fuzzy c-means is the most popular fuzzy clustering algorithm. Even though as Rokach mentioned that "Fuzzy c-means is better than the hard K-means algorithm at avoiding local minima. FCM can still converge to the local minima of the squared error criterion" (Rokach & Maimon, 2005)

" FCM has a soft membership function and a constant weight function. In general, FCM performs better than K-means" (G. Hamerly, 2003) and "it is less affected by the presence of uncertainty in the data" (A. Liew, S. Leung & W. Lau, , 2000). However, as in K-means it requires the user to specify the number of clusters in the data set.

Fuzzy Clustering method generates partitions; in one partition, each instance belongs to only one cluster. Therefore, the clusters in a hard clustering are disassembled. For instance as has been suggested by B. Höppner to extends this notion and suggests a soft clustering schema (B. Höppner, M. Bartelheım, R. Krauss, & M. Huıjsmans, 2005).

4.2.2. Algorithm of Fuzzy Clustering

- Initially, presume the fixed number of clusters c.
- Then, initialize the c-means by calculating the probabilities of the data points xi to the given clusters c.
- Recalculating the centroids of the clusters.
- Iterating until convergence or until being able to specify the number of iterations that have been reached.

Also, Fuzzy-c means clustering was used for image segmentation (Zhang & Chen, 2004); (Cai, W, Zhang, D, & Chen, S, 2007). Moreover, there are studies comparing the k-means and fuzzy-c means algorithms with various data sets (Mingoti, S & Lima, J, 2006) (Rong, 2011) (Ghosh & Dubey, 2013).

CHAPTER 5

VARIABLES DEFINITIONS

5. VARIABLES DEFINITIONS

After has defining above both methods K means and FCM and as has been mentioned earlier that this study intends to provide measures and criteria that are used for determining whether the results of K-means and FCM clustering are similar or dissimilar.

The variables are going to be defined that are collected for 189 countries and the way of applying both K means and FCM methods will be described and will be compared between both results accuracy or suitability to such data set collected.

5.1. Gross Domestic Product Per Capita (GDP Per Capita)

- Gross = Sum total of all (including depreciation, that is, including old and worn out).
- Domestic = A geographic region demarcated from another region by a boundary or a frontier.
- Product = Something that is ready for consumption. (No more value additions, the product is final)

So, Gross Domestic Product is the sum total of all final goods and services within a geographic region termed as domestic or home country. Products differ from each other, Shape, size, density, properties, aesthetics, time is taken for production, etc. Are sometimes used to determine the value of the product. This value is found by comparing one product with the other.

Example: A block of Gold of a particular size is heavier than a block of Silver of the same size. Here silver is kept as the standard, against which the Gold is compared. But here only one parameter is taken into consideration, which is the weight (or density). But, if the plant's seed is given and then it is been given a choice to choose a bucket of most fertile soil or a block of gold, what will be valuable at that point in time?

Thus, the value of a product is always determined by the demand for that product. This value is usually quoted by the domestic or home country's official currency, which is used as a standard to compare the value of different products. Hence, Gross Domestic Product is *"the sum total of the value of all final goods and services, quoted in the standard currency, produced within a geographic region termed as domestic or home country".*

How are the gross domestic product and the gross domestic product per capita different? GDP is the cumulative income of the country. It is decided by either how much industry produced or how much end-user consumed it along with some other adjustments to include missing income (like Indian companies income on foreign land ex TCS, L&T, etc) and adjust price-sensitive elements (inflation).

Per capita refers to the contribution of a country's citizens to the index, Let's suppose a country has a GDP of $100 billion with a total population of 1 billion. It would give per capita GDP as $100 ($100 billion / 1 billion). (The higher rate of GDP per capita is the better).

5.2. Educational Inequality

One huge inequality is that between those who can go to school and those who can't. Millions of children do not go to school as their governments do not have the money to support public education. Then there are the girls who don't get to go to school for cultural reasons. One even got shot in the head for doing so and won half the Nobel Peace Prize.

The inequality subsets, though including high-quality education, are far too weight-carrying and the education may not change the overall trend. Other subsets include the caste system, placement of people/groups/individuals above or below the poverty line, religion, race, family history, etc.

There are tremendous pressures, in general, working against the younger lot seeking jobs in desired departments and setups. In Pakistan, for example, what has happened is that successive governments are too busy holding on to a sham democracy and in the process, politicians make money even when the Sun is not shining. Money is gone. Jobs are not there. These lads vie to move out of the country and line up in neighboring countries. (The lower percentage is the better).

5.3. Income Inequality

It seems to be inevitable once people move beyond the hunter-gatherer stage. Individual humans do have significant differences, and these differences do tend to get magnified as they interact with each other and their environment. The larger and more complex the society, the greater the degree of differentiation and differing economic outcomes.

When inputs are unequal, outputs must be unequal, as well. If not, the total output falls spectacularly. Inequality in output (income, in this case) is one of the reasons the country's economy is so strong.

Many people confuse inequality with poverty. Inequality is generally the antidote to poverty. Relieving poverty is a good thing, but it doesn't do much to reduce quantitative measures of inequality. Inequality in rich modern societies, as measured by things like the Gini coefficient, is determined by how rich the richest people are, not how comfortable the poorest people are.

Many people worry about inequality because they fear the richest people obtain their wealth corruptly, and use it to acquire illegitimate power. While those are certainly bad things, and they do happen, the solution is to fight corruption and illegitimate power. Attacking inequality is too indirect and ineffective, and has too many bad consequences.

Another stand of view says that the mere fact of inequality is a feature and not a bug.

"Income inequality" is kind of a silly thing to worry about. It places too much importance on money itself and ignores the reality that one's income is merely the side effect of the value of their work. It's unreasonable to act as though all jobs would have the same value. (The lower percentage is the better).

5.4. Life Expectancy

The definition of “life expectancy” is how many years are expected for a person to live. It is based on the estimation of the average years that a person of a specific population group will live.

The most common measurement of life expectancy is life expectancy at birth. Life expectancy is a presumptive measurement. (The higher percentage is the better).

5.5. Mean Years of Schooling

An average number of completed years of education of a country's population aged 25 years and older, excluding years spent repeating individual grades. It is a statistical term and would require a group of people in order to calculate the MEAN (or average) number of years of schooling attained by the group in question. You’d take each individual’s years of schooling, add them all together, then divide by the number of people in the group in order to determine the mean years of schooling.

GDP per capita and mean years of schooling are correlated. Because, in essence, GDP per capita is a value-added created per person. More years of schooling usually mean a higher value-adding capacity of the person. (The higher rate is the better).

5.6. Unemployment

Unemployment definition simply is when a citizen from a population group is searching for a job but, unable to find out one.

Unemployment might at some points represents the economy's health.

The calculation of the unemployment rate is the number of people is unemployed divides the total labor force. (The lower percentage is the better).

5.7. Exports and Imports (% of GDP)

A country exports its locally produced goods to a foreign market in another country, which brings money into the country, which increases the exporting nation's GDP. A country imports goods from another country, that spend money out of the economy, which decreases the importing nation's GDP.

Net exports can be either positive or negative. When exports are greater than imports, net exports are positive. When exports are lower than imports, net exports are negative.

Net Exports = Value of Exports – Value of Imports/GDP

For more clarification it ought to mention that there are tariffs on both imports and exports. Import tariffs are used, on paper, at least, to protect the local industry.

Export tariffs are used to ensure that the local market has access to the goods being produced locally.

It's not unheard of, especially in third-world countries with very cheap labor, that it's more profitable for companies to sell their goods abroad, which lives the local market under-supplied.

Import duties are charged on goods entering a country from another. They are charged at the time of entry, the goods can not be received until duty is paid.

Export duties are charged on goods produced in a country that is about to leave. They are paid before the goods actually leave as you cannot load goods onto a vessel without proof of duties paid. (The higher percentage is the better).

5.8. Foreign Direct Investment Net Inflows (of GDP)

When a foreign investor invests in business in another country and has control over the company purchased, it is called foreign direct investment. There are some advantages for both the company owned by the foreigner and the foreign country which is often a developing country such as accessing to markets, accessing to resources, or reducing the cost of production. (The higher percentage is the better).

CHAPTER 6

APPLICATIONS

6. APPLICATIONS

6.1. Aim and Scope

The aim of this study is to understand the difference between KM and FCM methods, and in order to understand how to select the right analysis method to apply it to the data that fit it. KM and FCM methods are applied in this study, to do a comparison between their results later on and, to understand which method is more suitable according to the data here. Therefore this research will study the effect of the eight variables on a country's clustering and how to pick up the right clustering method.

And to get the right conclusion so it could be able to find effective solutions.

6.2. Importance of The Study

The importance of using the appropriate analysis to reach the correct result, through which depends to access to the most appropriate solutions. The importance of the research is in the use of the statistical clustering method, in order to facilitate the process of comparison between countries through their classification and understanding their convergence.

And to understand the criteria that are used for determining whether the method of either K-means or FCM clustering is better fits the data here. Moreover, this study facilitates the way to better understanding the most important indicators that led to heterogeneity and convergence between countries, makes them work to develop this aspect or indicator to provide the best rates or values they could.

6.3. Implementation

To summarize the application section, first to run both K-means algorithm then FCM algorithm with the number of clusters fixed at five, thus, K=5 five clusters are gotten and named according to their values as follows:

(Lowest - Lower Midde - Middle - Upper Middle – Highest)

Using the data set consisting of 189 cases.

According to the results, it could be derived what cluster technique is more suitable for this dataset.

Moreover, there are some measures that indicate which clustering method is better than the other. Firstly the reported R2 is the ratio of the between sums of squares and total sums of squares, where A higher the R-squared, the better the model fits the data.

The Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC) also measure the goodness-of-fit of a model, but on top of that penalize for the number of free parameters of the model, where for both mentioned measures a lower score is better.

Lastly, the Silhouette score displays the average internal consistency of the clustering and for Silhouette a higher value is better.

Results of K-Means Method on Year (2013)

Table 1: K-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
5	189	0.528	789.840	919.510	0.220

Table 2: Cluster Centroids of Year (2013) - K-Means

Cluster	1	2	3	4	5
Number of members	53	18	38	73	7
GDPpercapita2011PPP$_1	-0.676	-0.570	1.286	-0.203	1.721
Inequalityineducation_1	0.000	0.568	1.957	1.554	1.487
Inequalityinincome_1	2.225	0.000	2.911	2.234	2.892
LifeexpectancyIndex_1	-1.007	-0.753	0.919	0.347	0.952
Meanyearsofschoolingyears_1	-1.241	-0.462	1.048	0.398	0.737
Unemploymenttotaloflabourforce_1	0.976	0.000	0.716	0.073	0.813

ExportsandimportsofGDP_1	-0.353	0.230	0.155	-0.090	3.356
Foreigndirectinvestmentnetinflow GDP_1	0.026	-0.246	-0.501	-0.001	3.166
Mean of each group =	-0.006	-0.212	1.061	0.539	1.891

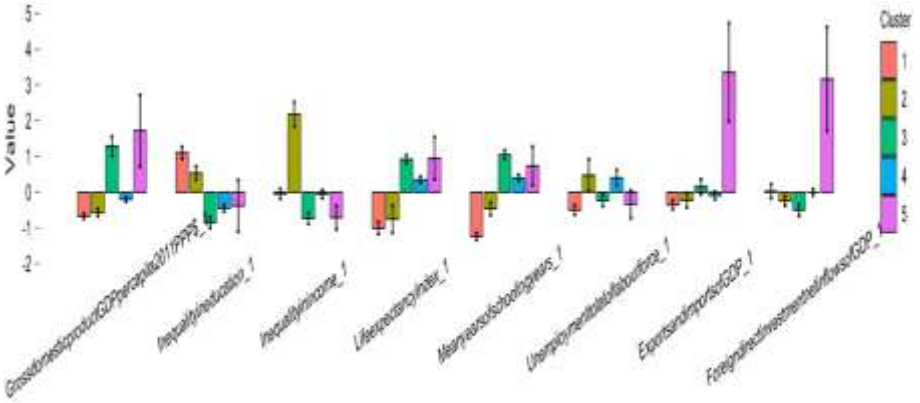

Figure 6: K-Means Cluster Mean Plots All predictors 2013

It can be seen from the chart of cluster mean plots of k-means 2013 (Figure 6) that:

Regarding the variable "Gross domestic product GDP per capita" the highest value/score is in cluster5, and the lowest value/score is in cluster1.

And regarding the variable "Inequality in education" the highest value/score is in cluster1, and the lowest value/score is in cluster3.

And regarding the variable "Inequality in income," the highest value/score is in cluster2, and the lowest value/score is in cluster5.

And regarding the variable "Life expectancy Index" the highest value/score is in cluster5, and the lowest value/score is in cluster1.

And regarding the variable "Mean years of schooling years," the highest value/score is in cluster3, and the lowest value/score is in cluster1.

And regarding the variable “Unemployment total of the labor force” the highest value/score is in cluster2, and the lowest value/score is in cluster1.

And regarding the variable “Exports and imports of GDP” the highest value/score is in cluster5, and the lowest value/score is in cluster1.

And regarding the variable “FDINI” the highest value/score is in cluster5, and the lowest value/score is in cluster3.

Table 3: Fuzzy C-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
5	189	0.348	1004.090	1133.760	0.070

Table 4: Cluster Centroids of Year (2013) – Fuzzy C-Means

Cluster	1	2	3	4	5
Number of members	3	68	11	43	64
GDPpercapita2011PPP$_1	1.747	-0.342	2.033	1.100	-0.267
Inequalityineducation_1	1.396	0.000	1.473	1.700	0.134
Inequalityinincome_1	0.000	0.524	2.138	2.309	0.597
LifeexpectancyIndex_1	0.902	0.261	1.367	1.099	0.632
Meanyearsofschoolingyears_1	1.530	-0.011	1.570	1.213	-0.075
Unemploymenttotaloflaborforce_1	0.135	0.703	0.550	0.000	0.245
ExportsandimportsofGDP_1	-1.109	-0.668	0.691	-0.228	-0.075
ForeigndirectinvestmentnetinflowsofGDP_1	-0.340	0.012	-0.960	-0.699	0.116
Mean of each group =	0.533	0.060	1.108	0.812	0.163

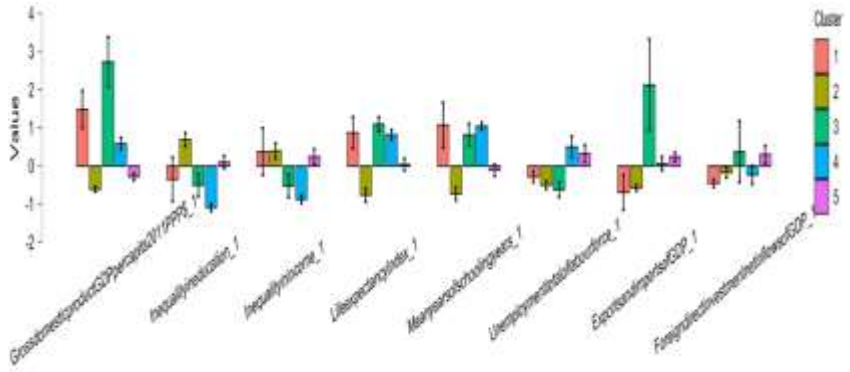

Figure 7: Fuzzy C -Means Cluster Mean Plots All predictors 2013

It can bee seen from the chart of cluster mean plots of fuzzy c -means 2013 (Figure 7):

That regarding the variable “Gross domestic product GDP per capita” the highest value/score is in cluster3, and the lowest value/score is in cluster2.

And regarding the variable “Inequality in education” the highest value/score is in cluster3, and the lowest value/score is in cluster4.

And regarding the variable “Inequality in income” the highest value/score is in cluster1, and the lowest value/score is in cluster4.

And regarding the variable “Life expectancy Index” the highest value/score is in cluster3, and the lowest value/score is in cluster2.

And regarding the variable “Mean years of schooling years” the highest value/score is in cluster1, and the lowest value/score is in cluster2.

And regarding the variable “Unemployment total of labour force” the highest value/score is in cluster4, and the lowest value/score is in cluster3.

And regarding the variable “Exports and imports of GDP” the highest value/score is in cluster3, and the lowest value/score is in cluster1.

And regarding the variable “FDINI” the highest value/score is in cluster3, and the lowest value/score is in cluster1.

Results of Fuzzy C-Means Method on Year (2014)

Table 7: Fuzzy C-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
5	189	0.467	921.930	1051.600	0.130

Table 8: Cluster Centroids of Year (2014) – Fuzzy C-Means

Cluster	1	2	3	4	5
Number of members	23	32	12	57	65
GDP Percapita2011ppp$_1	-0.819	-0.700	-0.260	0.699	-0.287
ınequalityineducation_1	0.000	0.311	2.435	2.677	1.759
ınequalityinincome_1	1.761	0.780	0.000	0.848	0.523
lifeexpectancyındex_1	-1.047	-0.949	0.472	0.281	0.643
meanyearsofschoolingyears_1	-1.516	-1.243	0.353	0.813	-0.056
unemploymenttotaloflabourforce_1	4.262	2.953	0.000	4.039	3.281
exportsandimportsofgdp_1	-0.471	0.175	0.326	0.023	-0.190
foreigndirectinvestmentnetinflowsofgdp_1	-0.626	0.722	-0.589	-0.281	-0.217
Mean of each group =	0.193	0.256	0.342	1.137	0.682

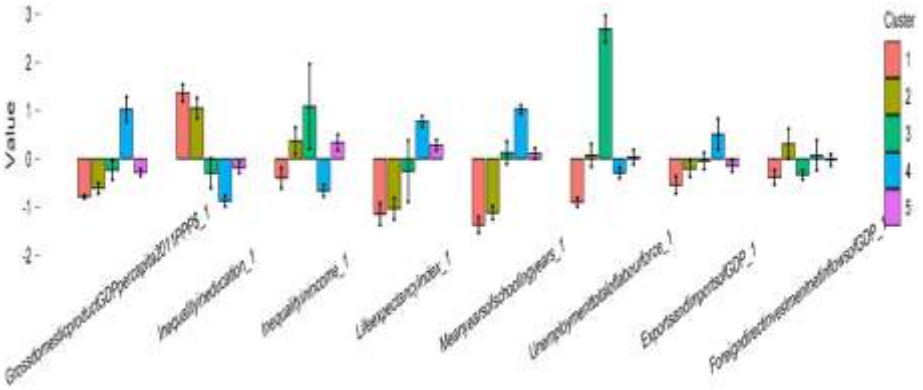

Figure 9: Fuzzy C -Means Cluster Mean Plots All predictors 2014

It can be seen from the chart of cluster mean plots of fuzzy c -means 2014 (Figure 9) that:

Regarding the variable “Gross domestic product GDP per capita” the highest value/score is in cluster4, and the lowest value/score is in cluster1.

And regarding the variable “Inequality in education” the highest value/score is in cluster1, and the lowest value/score is in cluster4.

And regarding the variable “Inequality in income” the highest value/score is in cluster3, and the lowest value/score is in cluster4.

And regarding the variable “Life expectancy Index” the highest value/score is in cluster4, and the lowest value/score is in cluster1.

And regarding the variable “Mean years of schooling years” the highest value/score is in cluster4, and the lowest value/score is in cluster1.

And regarding the variable “Unemployment total of labour force” the highest value/score is in cluster3, and the lowest value/score is in cluster1.

And regarding the variable “Exports and imports of GDP” the highest value/score is in cluster4, and the lowest value/score is in cluster1.

And regarding the variable “FDINI” the highest value/score is in cluster2, and the lowest value/score is in cluster1.

Results of K-Means Method on Year (2015)

Table 9: K-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
5	189	0.578	715.050	844.720	0.240

Table 10: Cluster Centroids of Year (2015) – K-Means

Cluster	1	2	3	4	5
Number of members	15	76	7	39	52
GDP percapita2011ppp$_1	-0.606	-0.234	1.800	1.283	-0.688
ınequalityineducation_1	0.598	1.628	1.756	2.078	0.000
ınequalityinincome_1	0.000	2.327	2.916	3.021	2.431
lifeexpectancyındex_1	-1.075	0.332	0.968	0.920	-0.996
meanyearsofschoolingyears_1	-0.456	0.329	0.824	1.071	-1.264
unemploymenttotaloflabourforce_1	0.000	0.555	0.932	1.237	1.432
exportsandimportsofgdp_1	-0.272	-0.113	3.490	0.170	-0.353
foreigndirectinvestmentnetinflowsofgdp_1	-0.270	-0.075	4.076	-0.302	-0.135
Mean of each group =	-0.260	0.594	2.095	1.185	0.053

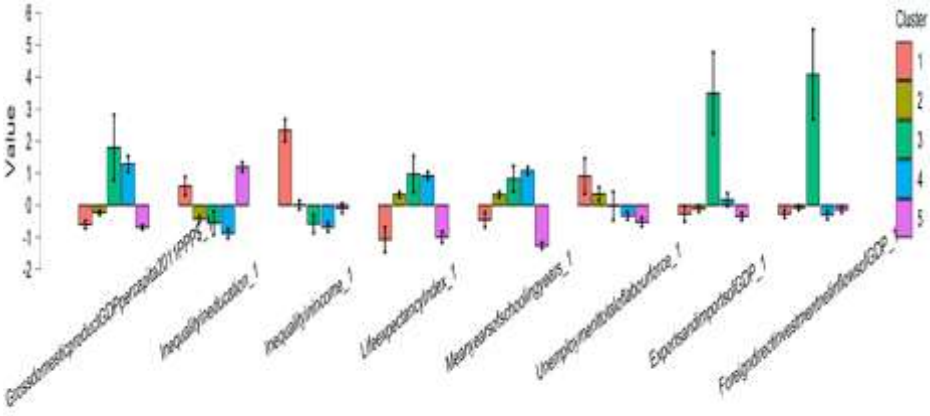

Figure 10: K-Means Cluster Mean Plots All predictors 2015

It can be seen from the chart of cluster mean plots of k-means 2015 (Figure 10) that:

Regarding the variable "Gross domestic product GDP per capita" the highest value/score is in cluster3, and the lowest value/score is in cluster5.

And regarding the variable "Inequality in education" the highest value/score is in cluster5, and the lowest value/score is in cluster4.

And regarding the variable "Inequality in income" the highest value/score is in cluster1, and the lowest value/score is in cluster4.

And regarding the variable "Life expectancy Index" the highest value/score is in cluster3, and the lowest value/score is in cluster1.

And regarding the variable "Mean years of schooling years" the highest value/score is in cluster4, and the lowest value/score is in cluster5.

And regarding the variable "Unemployment total of labour force" the highest value/score is in cluster1, and the lowest value/score is in cluster5.

And regarding the variable "Exports and imports of GDP" the highest value/score is in cluster3, and the lowest value/score is in cluster5.

And regarding the variable "FDINI" the highest value/score is in cluster3, and the lowest value/score is in cluster1.

Results of Fuzzy C-Means Method on Year (2015)

Table 11: Fuzzy C-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
5	189	0.487	876.580	1006.250	0.190

Table 12: Cluster Centroids of Year (2015) – Fuzzy C-Means

Cluster	1	2	3	4	5
Number of members	48	48	62	14	17
GDP percapita2011ppp$_1	-0.609	0.077	-0.245	0.728	-0.392
ınequalityineducation_1	0.000	2.230	0.881	2.155	0.030
ınequalityinincome_1	0.803	2.559	0.692	1.470	0.000
lifeexpectancyındex_1	-1.481	0.310	0.739	1.365	-0.090
meanyearsofschoolingyears_1	-1.112	1.133	0.034	0.410	-0.433
unemploymenttotaloflabourforce_1	2.541	2.660	2.644	0.000	1.528
exportsandimportsofgdp_1	-0.456	0.445	-0.246	-0.439	-0.931
foreigndirectinvestmentnetinflowsof gdp_1	0.259	-0.202	-0.083	-0.288	-0.295
Mean of each group =	-0.007	1.152	0.552	0.675	-0.073

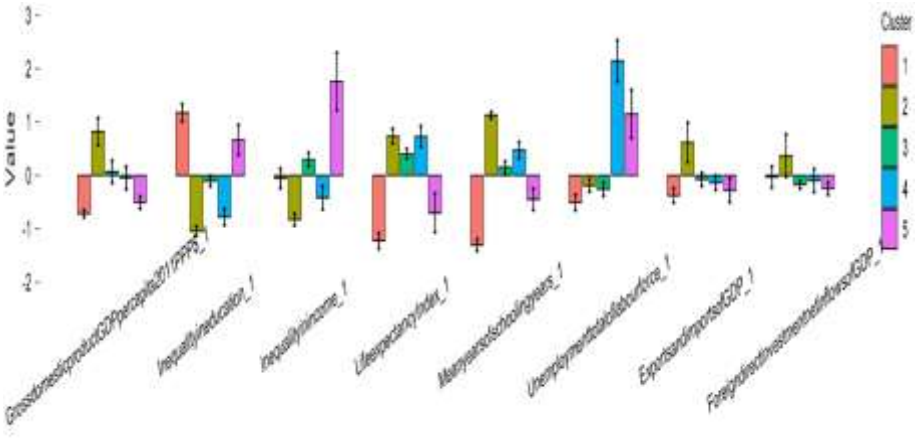

Figure 11: Fuzzy C -Means Cluster Mean Plots All predictors 2015

It can be seen from the chart of cluster mean plots of fuzzy c -means 2015 (Figure 11) that:

Regarding the variable "Gross domestic product GDP per capita" the highest value/score is in cluster2, and the lowest value/score is in cluster1.

And regarding the variable "Inequality in education" the highest value/score is in cluster1, and the lowest value/score is in cluster2.

And regarding the variable "Inequality in income" the highest value/score is in cluster5, and the lowest value/score is in cluster2.

And regarding the variable "Life expectancy Index" the highest value/score is in cluster4, and the lowest value/score is in cluster1.

And regarding the variable "Mean years of schooling years" the highest value/score is in cluster2, and the lowest value/score is in cluster1.

And regarding the variable "Unemployment total of labour force" the highest value/score is in cluster4, and the lowest value/score is in cluster1.

And regarding the variable "Exports and imports of GDP" the highest value/score is in cluster2, and the lowest value/score is in cluster1.

And regarding the variable "FDINI" the highest value/score is in cluster2, and the lowest value/score is in cluster5.

Results of K-Means Method on Year (2016)

Table 13: K-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
5	189	0.582	708.450	838.120	0.240

Table 14: Cluster Centroids of Year (2016) – K-Means

Cluster	1	2	3	4	5
Number of members	51	7	13	79	39
GDP percapita2011ppp$_1	-0.710	1.988	-0.571	-0.250	1.270
ınequalityineducation_1	0.000	2.080	0.471	1.528	2.084
ınequalityinincome_1	2.655	3.038	0.000	2.521	3.172
lifeexpectancyındex_1	-1.096	1.220	-1.131	0.335	0.913
meanyearsofschoolingyears_1	-1.240	1.078	-0.533	0.269	1.062
unemploymenttotaloflabourforce_1	1.802	1.498	0.000	0.947	1.646
exportsandimportsofgdp_1	-0.360	3.502	-0.319	-0.096	0.143
foreigndirectinvestmentnetinflowsofgdp_1	-0.150	3.860	-0.347	-0.100	-0.178
The mean of each group =	0.113	2.283	-0.304	0.644	1.264

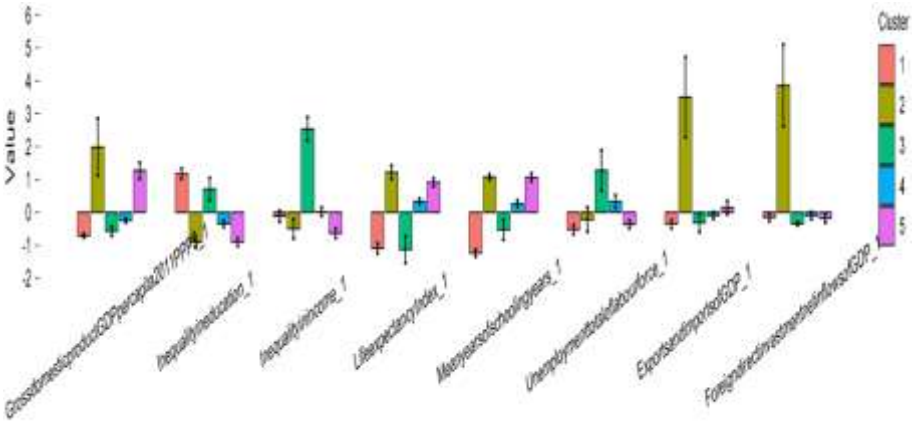

Figure 12: K-Means Cluster Mean Plots All predictors 2016

It can be seen from the chart of cluster mean plots of k-means 2016 (Figure 12) that:

Regarding the variable “Gross domestic product GDP per capita” the highest value/score is in cluster2, and the lowest value/score is in cluster1.

And regarding the variable “Inequality in education” the highest value/score is in cluster1, and the lowest value/score is in cluster5.

And regarding the variable “Inequality in income” the highest value/score is in cluster3, and the lowest value/score is in cluster5.

And regarding the variable “Life expectancy Index” the highest value/score is in cluster2, and the lowest value/score is in cluster3.

And regarding the variable “Mean years of schooling years” the highest value/score is in cluster2, and the lowest value/score is in cluster1.

And regarding the variable “Unemployment total of labour force” the highest value/score is in cluster3, and the lowest value/score is in cluster1.

And regarding the variable “Exports and imports of GDP” the highest value/score is in cluster2, and the lowest value/score is in cluster1.

And regarding the variable “FDINI” the highest value/score is in cluster2, and the lowest value/score is in cluster3.

Results of Fuzzy C-Means Method on Year (2016)

Table 15: Fuzzy C-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
5	189	0.516	923.060	1052.730	0.110

Table 16: Cluster Centroids of Year (2016) – K-Means

Cluster	1	2	3	4	5
Number of members	31	26	46	54	32
GDP percapita2011ppp$_1	0.862	-0.825	0.503	0.390	-0.669
ınequalityineducation_1	1.865	0.127	2.580	1.899	0.000
ınequalityinincome_1	0.899	1.276	0.718	0.808	0.000
lifeexpectancyındex_1	1.245	-1.717	0.180	1.045	-2.245
meanyearsofschoolingyears_1	1.101	-1.926	0.729	0.229	-0.756
unemploymenttotaloflabourforce_1	1.209	0.297	1.277	0.000	0.845
exportsandimportsofgdp_1	-0.144	-0.420	0.023	-0.123	-1.155
foreigndirectinvestmentnetinflowsofgdp_1	-0.346	-0.219	-0.412	-0.007	-0.363
Mean of each group =	0.836	0.426	0.700	0.530	-0.543

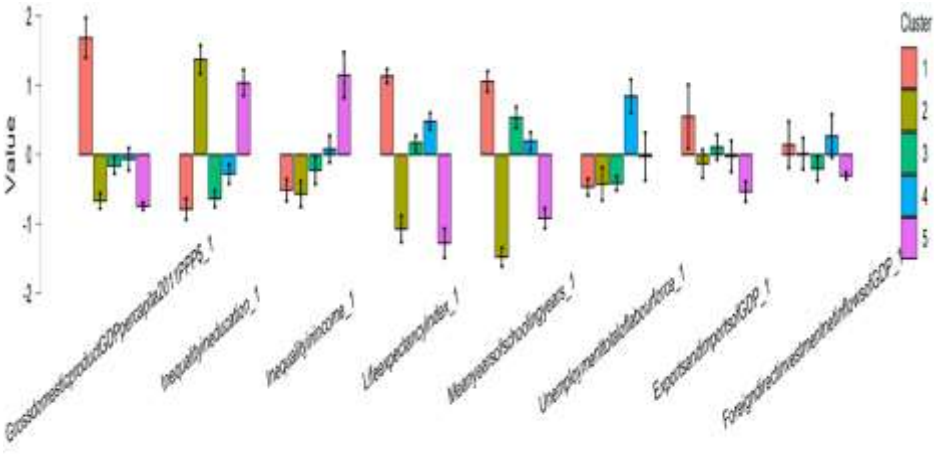

Figure 13: Fuzzy C -Means Cluster Mean Plots All predictors 2016

It can be seen from the chart of cluster mean plots of fuzzy c -means 2016 (Figure 13) that:

Regarding the variable “Gross domestic product GDP per capita” the highest value/score is in cluster1, and the lowest value/score is in cluster5.

And regarding the variable “Inequality in education” the highest value/score is in cluster2, and the lowest value/score is in cluster1.

And regarding the variable “Inequality in income” the highest value/score is in cluster5, and the lowest value/score is in cluster2.

And regarding the variable “Life expectancy Index” the highest value/score is in cluster1, and the lowest value/score is in cluster5.

And regarding the variable “Mean years of schooling years” the highest value/score is in cluster1, and the lowest value/score is in cluster2.

And regarding the variable “Unemployment total of labour force” the highest value/score is in cluster4, and the lowest value/score is in cluster1.

And regarding the variable “Exports and imports of GDP” the highest value/score is in cluster1, and the lowest value/score is in cluster5.

And regarding the variable “FDINI” the highest value/score is in cluster4, and the lowest value/score is in cluster5.

Results of K-Means Method on Year (2017)

Table 17: K-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
5	189	0.571	725.000	854.670	0.230

Table 18: Cluster Centroids of Year (2017) – K-Means

Cluster	1	2	3	4	5
Number of members	12	41	6	84	46
GDP percapita2011ppp$_1	-0.691	1.259	2.156	-0.263	-0.744
ınequalityineducation_1	0.404	2.105	2.031	1.603	0.000
ınequalityinincome_1	0.000	2.953	2.838	2.311	2.388
lifeexpectancyındex_1	-1.400	0.907	1.344	0.275	-1.120
meanyearsofschoolingyears_1	-0.783	1.053	1.062	0.225	-1.283
unemploymenttotaloflabourforce_1	0.000	1.724	1.660	0.960	1.925
exportsandimportsofgdp_1	0.491	0.220	3.586	-0.086	-0.379
foreigndirectinvestmentnetinflowsofgdp_1	0.331	-0.414	3.998	0.014	-0.091
Mean of each group =	0.412	1.226	2.334	0.630	0.087

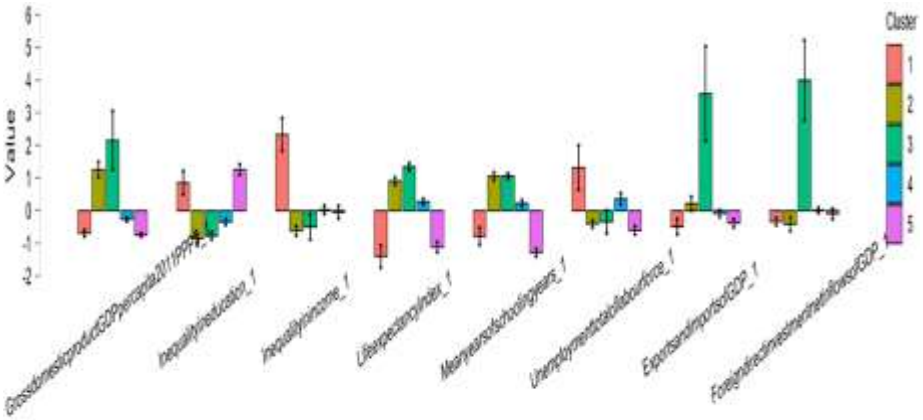

Figure 14: K -Means Cluster Mean Plots All predictors 2017

It can be seen from the chart of cluster mean plots of k -means 2017 (Figure 14) that:

Regarding the variable “Gross domestic product GDP per capita” the highest value/score is in cluster3, and the lowest value/score is in cluster5.

And regarding the variable “Inequality in education” the highest value/score is in cluster5, and the lowest value/score is in cluster2.

And regarding the variable “Inequality in income” the highest value/score is in cluster1, and the lowest value/score is in cluster2.

And regarding the variable “Life expectancy Index” the highest value/score is in cluster3, and the lowest value/score is in cluster1.

And regarding the variable “Mean years of schooling years” the highest value/score is in cluster3, and the lowest value/score is in cluster5.

And regarding the variable “Unemployment total of labour force” the highest value/score is in cluster1, and the lowest value/score is in cluster5.

And regarding the variable “Exports and imports of GDP” the highest value/score is in cluster3, and the lowest value/score is in cluster1.

And regarding the variable “FDINI” the highest value/score is in cluster3, and the lowest value/score is in cluster2.

Results of Fuzzy C-Means Method on Year (2017)

Table 19: Fuzzy C-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
5	189	0.205	1109.640	1239.310	-0.030

Table 20: Cluster Centroids of Year (2017) – Fuzzy C-Means

Cluster	1	2	3	4	5
Number of members	44	11	17	92	25
GDP percapita2011ppp$_1	0.415	0.285	0.261	-0.334	2.346
ınequalityineducation_1	0.000	1.148	1.107	0.397	0.973
ınequalityinincome_1	0.000	0.460	1.269	0.056	0.820
lifeexpectancyındex_1	0.071	-0.017	0.053	0.442	1.255
meanyearsofschoolingyears_1	0.589	0.966	0.932	0.543	1.242
unemploymenttotaloflabourforce_1	0.385	0.000	0.063	0.133	0.219
exportsandimportsofgdp_1	0.005	-0.712	-0.477	-0.623	-0.334
foreigndirectinvestmentnetinflowsofgdp_1	-0.747	-0.298	-0.160	-0.299	-0.352
Mean of each group =	0.090	0.229	0.381	0.039	0.771

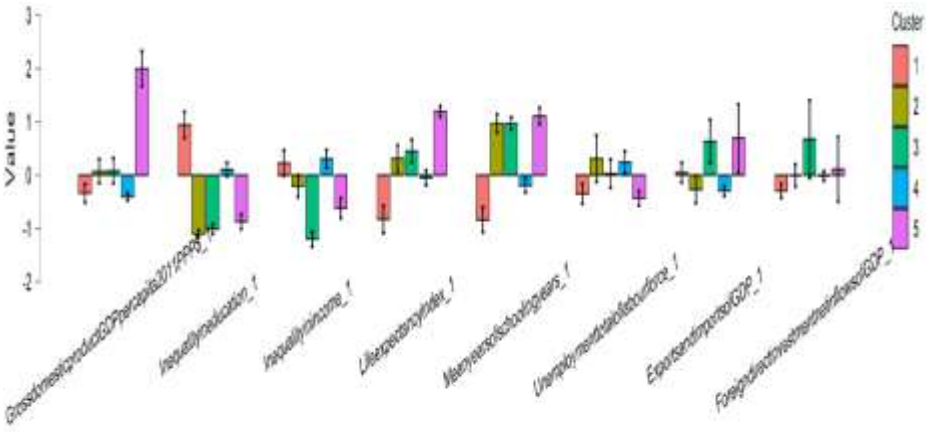

Figure 15: Fuzzy C -Means Cluster Mean Plots All predictors 2017

It can be seen from the chart of cluster mean plots of fuzzy c -means 2017 (Figure 15) that:

Regarding the variable “Gross domestic product GDP per capita” the highest value/score is in cluster5, and the lowest value/score is in cluster4.

And regarding the variable “Inequality in education” the highest value/score is in cluster1, and the lowest value/score is in cluster2.

And regarding the variable “Inequality in income” the highest value/score is in cluster4, and the lowest value/score is in cluster3.

And regarding the variable “Life expectancy Index” the highest value/score is in cluster5, and the lowest value/score is in cluster1.

And regarding the variable “Mean years of schooling years” the highest value/score is in cluster5, and the lowest value/score is in cluster1.

And regarding the variable “Unemployment total of labour force” the highest value/score is in cluster2, and the lowest value/score is in cluster5.

And regarding the variable “Exports and imports of GDP” the highest value/score is in cluster5, and the lowest value/score is in cluster3.

And regarding the variable “FDINI” the highest value/score is in cluster3, and the lowest value/score is in cluster1.

Results of K-Means Method on Year (2018)

Table 21: K-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
5	189	0.562	738.630	868.300	0.240

Table 22: Cluster Centroids of Year (2018) – K-Means

Cluster	1	2	3	4	5
Number of members	52	4	26	67	40
GDP percapita2011ppp$_1	-0.758	2.667	-0.314	-0.237	1.319
ınequalityineducation_1	0.000	1.988	1.381	1.634	2.148
ınequalityinincome_1	0.000	0.781	0.235	0.297	1.184
lifeexpectancyındex_1	-1.197	1.427	0.000	0.255	0.987
meanyearsofschoolingyears_1	-1.259	1.017	0.114	0.227	1.081
unemploymenttotaloflabourforce_1	2.370	2.543	0.000	2.190	2.384
exportsandimportsofgdp_1	-0.304	4.991	-0.094	-0.211	0.310
foreigndirectinvestmentnetinflowsofgdp_1	-0.002	3.160	0.208	0.082	-0.586
Mean of each group =	-0.144	2.322	0.191	0.530	1.103

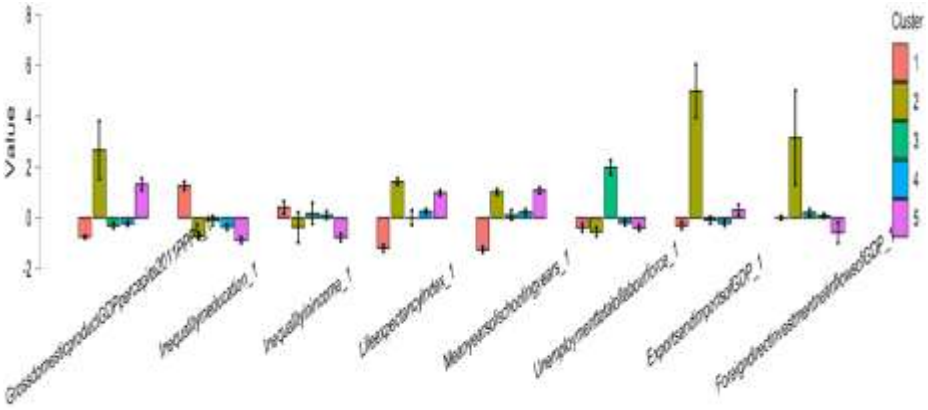

Figure 16: K-Means Cluster Mean Plots All predictors 2018

It can be seen from the chart of cluster mean plots of k-means 2018 (Figure 16) that:

Regarding the variable "Gross domestic product GDP per capita" the highest value/score is in cluster2, and the lowest value/score is in cluster1.

And regarding the variable "Inequality in education" the highest value/score is in cluster1, and the lowest value/score is in cluster5.

And regarding the variable "Inequality in income" the highest value/score is in cluster1, and the lowest value/score is in cluster5.

And regarding the variable "Life expectancy Index" the highest value/score is in cluster2, and the lowest value/score is in cluster1.

And regarding the variable "Mean years of schooling years" the highest value/score is in cluster5, and the lowest value/score is in cluster1.

And regarding the variable "Unemployment total of labour force" the highest value/score is in cluster3, and the lowest value/score is in cluster2.

And regarding the variable "Exports and imports of GDP" the highest value/score is in cluster2, and the lowest value/score is in cluster1.

And regarding the variable "FDINI" the highest value/score is in cluster2, and the lowest value/score is in cluster5.

Results of Fuzzy C-Means Method on Year (2018)

Table 23: Fuzzy C-Means Clustering

Clusters	N	R^2	AIC	BIC	Silhouette
5	189	0.321	920.790	1050.460	0.120

Table 24: Cluster Centroids of Year (2018) – Fuzzy C-Means

Cluster	1	2	3	4	5
Number of members	31	50	5	34	69
GDP percapita2011ppp$_1	-0.480	0.889	-0.321	-0.290	-0.282
ınequalityineducation_1	0.099	0.543	0.351	0.637	0.000
ınequalityinincome_1	1.965	1.799	2.249	1.665	0.000
lifeexpectancyındex_1	0.212	0.657	-0.374	0.500	-0.177
meanyearsofschoolingyears_1	0.505	0.389	0.470	0.700	0.047
unemploymenttotaloflabourforce_1	0.000	2.019	1.514	1.809	1.288
exportsandimportsofgdp_1	-0.131	0.213	0.525	-0.771	-0.034
foreigndirectinvestmentnetinflowsofgdp_1	0.006	-0.031	2.396	-0.015	0.503
Mean of each group =	0.272	0.810	0.851	0.529	0.168

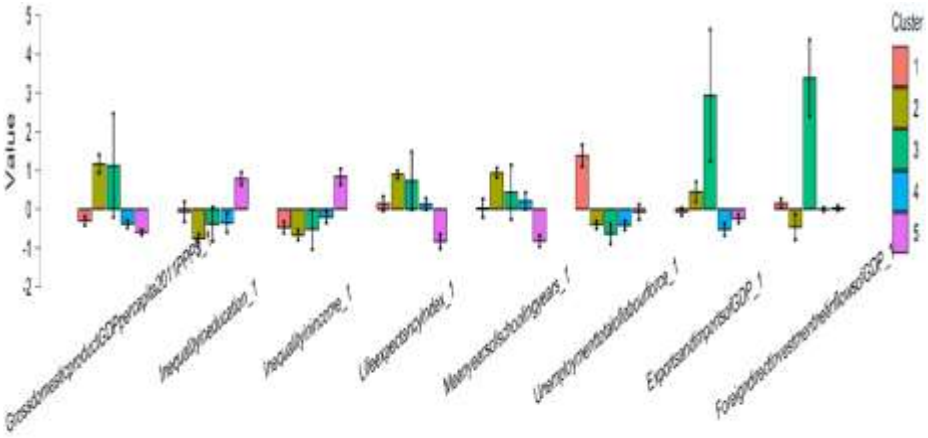

Figure 17: Fuzzy C -Means Cluster Mean Plots All predictors 2018

It can be seen from the chart of cluster mean plots of fuzzy c -means 2018 (Figure 17) that:

Regarding the variable “Gross domestic product GDP per capita” the highest value/score is in cluster2, and the lowest value/score is in cluster5.

And regarding the variable “Inequality in education” the highest value/score is in cluster5, and the lowest value/score is in cluster2.

And regarding the variable “Inequality in income” the highest value/score is in cluster5, and the lowest value/score is in cluster2.

And regarding the variable “Life expectancy Index” the highest value/score is in cluster2, and the lowest value/score is in cluster5.

And regarding the variable “Mean years of schooling years” the highest value/score is in cluster2, and the lowest value/score is in cluster5.

And regarding the variable “Unemployment total of labour force” the highest value/score is in cluster1, and the lowest value/score is in cluster3.

And regarding the variable “Exports and imports of GDP” the highest value/score is in cluster3, and the lowest value/score is in cluster4.

And regarding the variable “FDINI” the highest value/score is in cluster3, and the lowest value/score is in cluster2.

CHAPTER 7

FINDINGS AND DISCUSSION

7. FINDINGS AND DISCUSSION

7.1. FINDINGS

The aim of this study is to understand the difference between KM and FCM methods by comparing them, and in order to understand how to select the right analysis method to apply it to the data that fit it. In order to understand the results at the beginning, it ought to define the names of the results gotten from each clustering technique is applied as follows: (R-squared, AIC, BIC, and Silhouette) and what does each one represents?

Then, it ought to understand what rate or value of these results is better than the other then it could be understood which cluster method is more suitable or accurate regarding to the data used.

The results' outputs and the better value or percentage that fit for R-squared where A higher the R-squared, the better the model fits the data, for AIC a lower score is better, for BIC a lower score is better and for Silhouette a higher value is better.

To simplify matters, first running both K-means algorithm then FCM algorithm with the number of clusters fixed at five, thus, K=5 to be later elaborated on the algorithm based on the output. Both tables of the Results of K-means method or Fuzzy C-means method on a year show the fit scores for the model with K=5 clusters using the data set consisting of 189 cases.

The reported R2 is the ratio of the between sums of squares and total sums of squares. A model with an R2 close to the upper bound of one is perceived as a good fitting model, whereas an R2 close to the lower bound of zero indicates that the model fits poorly. The R2 however does not make a distinction between a well-fitted model, and a model that overfits. The Akaike Information Criterion (AIC), (Akaike, 1987) and the Bayesian Information Criterion (BIC) (Schwarz, 1978) also measure the goodness-of-fit of a model, but on top of that penalize for the number of free parameters of the model. By penalizing the number of parameters these information criteria try to safeguard against overfitting.

Models that have a lower information criterion are perceived as models that generalize better. Lastly, the silhouette score displays the average internal consistency of the clustering by assessing how similar each case is with respect to its own cluster compared to other clusters. For silhouette scores, the general rule is that the closer it is to the upper bound of 1, the more consistent the clustering is, whereas Silhouette scores close to the lower bound of -1 indicate a bad match. Years (2013,2014,2015,2016,2017 and 2018) show the fit scores for the model with K=5 clusters using the data set consisting of 189 cases.

Table 25: Interpretation of KM and FCM methods' results for year (2013)

Measures	KM results	FCM results	Conclusion
R^2	0.528	0.348	K-means method has better coefficient of determination.
AIC	789.840	1004.090	K-means method is better in terms of fit for the data.
BIC	919.510	1133.760	K-means method is better in terms of criterion for model selection among a finite set of models.
Silhouette	0.220	0.070	K-means method is more consistent.

Table 26: Interpretation of KM and FCM methods' results for year (2014)

Measures	KM results	FCM results	Conclusion
R^2	0.569	0.467	K-means method has better coefficient of determination.
AIC	728.090	921.930	K-means method is better in terms of fit for the data.
BIC	857.760	1051.600	K-means method is better in terms of criterion for model selection among a finite set of models.
Silhouette	0.240	0.130	K-means method is more consistent.

Table 27: Interpretation of KM and FCM methods' results for year (2015)

Measures	KM results	FCM results	Conclusion
R^2	0.578	0.487	K-means method has better coefficient of determination.
AIC	715.050	876.580	K-means method is better in terms of fit for the data.
BIC	844.720	1006.250	K-means method is better in terms of criterion for model selection among a finite set of models.
Silhouette	0.240	0.190	K-means method is more consistent.

Table 28: Interpretation of KM and FCM methods' results for year (2016)

Measures	KM results	FCM results	Conclusion
R^2	0.582	0.516	K-means method has better coefficient of determination.
AIC	708.450	923.060	K-means method is better in terms of fit for the data.
BIC	838.120	1052.730	K-means method is better in terms of criterion for model selection among a finite set of models.
Silhouette	0.240	0.110	K-means method is more consistent.

Table 29: Interpretation of KM and FCM methods' results for year (2017)

Measures	KM results	FCM results	Conclusion
R^2	0.571	0.205	K-means method has better coefficient of determination.
AIC	725.000	1109.640	K-means method is better in terms of fit for the data.

BIC	854.670	1239.310	K-means method is better in terms of criterion for model selection among a finite set of models.
Silhouette	0.230	-0.030	K-means method is more consistent.

Table 30: Interpretation of KM and FCM methods' results for year (2018)

Measures	KM results	FCM results	Conclusion
R^2	0.562	0.321	K-means method has better coefficient of determination.
AIC	738.630	920.790	K-means method is better in terms of fit for the data.
BIC	868.300	1050.460	K-means method is better in terms of criterion for model selection among a finite set of models.
Silhouette	0.240	0.120	K-means method is more consistent.

Regarding all years compared, it is concluded that K-means method is better than Fuzzy C-means method in clustering according to the data set applied on.

7.2. DISCUSSION

As this study aim to compare between both k-means and fuzzy c-means methods to understand which method is more suitable according to the data here. It is concluded that these results get affected by the applied cluster technique, as an example of this:

In year (2013)

Albania gets the cluster Middle when applying K means method, but it gets the cluster Lower Middle when applying FCM method. Algeria gets the cluster Middle when applying K means method, but it gets the cluster Lower Middle when applying FCM method.

Belgium gets the cluster Upper Middle when applying K means method but, it gets the cluster Highest when applying FCM method. Bosnia and Herzegovina gets the cluster Middle when applying K means method but, it gets the cluster Upper Middle when applying FCM method.

However, on the other hand, some cases have the same result of cluster classification after applying both methods on them, for example: Brazil gets the cluster Lowest when applying both methods (K-means & FCM), Austria gets the cluster Lowest when applying both methods (K-means & FCM), Cambodia gets the cluster Lowest when applying both methods (K-means & FCM).

In year (2014)

Albania gets the cluster Middle when applying K means method, but it gets the cluster Upper Middle when applying FCM method. Algeria gets the cluster Middle when applying K means method, but it gets the cluster Upper Middle when applying FCM method.

Belgium gets the cluster Upper Middle when applying K means method but, it gets the cluster Highest when applying FCM method. Brazil gets the cluster Middle when applying K means method but, it gets the cluster Upper Middle when applying FCM method. Austria gets the cluster Upper Middle when applying K means method but, it gets the cluster Highest when applying FCM method.

However, on the other hand, some cases have the same result of cluster classification after applying both methods on them, for example: Bosnia and Herzegovina gets the cluster Middle when applying both methods (K-means & FCM), Cambodia gets the cluster Lower Middle when applying both methods (K-means & FCM).

In year (2015)

Albania gets the cluster Middle when applying K means method, but it gets the cluster Upper Middle when applying FCM method. Bosnia and Herzegovina gets the cluster Middle when applying K means method but, it gets the cluster Upper Middle when applying FCM method. Belgium gets the cluster Upper Middle when applying K means method but, it gets the cluster Highest when applying FCM method.

Brazil gets the cluster Middle when applying K means method but, it gets the cluster Lowest when applying FCM method. Austria gets the cluster Upper Middle when applying K means method but, it gets the cluster Highest when applying FCM method.

However, on the other hand, some cases have the same result of cluster classification after applying both methods on them, for example: Cambodia gets the cluster Lower Middle when applying both methods (K-means & FCM), Algeria gets the cluster Middle when applying both methods (K-means & FCM).

In year (2016)

Belgium gets the cluster Upper Middle when applying K means method but, it gets the cluster Highest when applying FCM method. Austria gets the cluster Upper Middle when applying K means method but, it gets the cluster Highest when applying FCM method.

However, on the other hand, some cases have the same result of cluster classification after applying both methods on them, for example: Cambodia gets the cluster Lower Middle when applying both methods (K-means & FCM), Algeria gets the cluster Middle when applying both methods (K-means & FCM). Albania gets the cluster Middle when applying both methods (K-means & FCM). Bosnia and Herzegovina gets the cluster Middle when applying both methods (K-means & FCM). Brazil gets the cluster Middle when applying both methods (K-means & FCM).

In year (2017)

Albania gets the cluster Middle when applying K means method but, it gets the cluster Upper Middle when applying FCM method. Algeria gets the cluster Middle when applying K means method but, it gets the cluster Lowest when applying FCM method.

Belgium gets the cluster Upper Middle when applying K means method but, it gets the cluster Highest when applying FCM method. Austria gets the cluster Upper Middle when applying K means method but, it gets the cluster Highest when applying FCM method.

Cambodia gets the Lower Middle when applying K means method but, it gets the cluster Lowest when applying FCM method.

Bosnia and Herzegovina gets the Middle when applying K means method but, it gets the cluster Lowest when applying FCM method. Brazil gets the Middle when applying K means method but, it gets the cluster Lowest when applying FCM method.

However, on the other hand, some cases have the same result of cluster classification after applying both methods on them, for example: Armenia gets the cluster Middle when applying both methods (K-means & FCM).

In year (2018)

Cambodia gets the Lowest when applying K means method but, it gets the cluster Highest when applying FCM method. Brazil gets the Lower Middle when applying K means method but, it gets the cluster Lowest when applying FCM method.

However, on the other hand, some cases have the same result of cluster classification after applying both methods on them, for example: Albania gets the cluster Lower Middle when applying both methods (K-means & FCM). Algeria gets the cluster Lower Middle when applying both methods (K-means & FCM). Belgium gets the cluster Upper Middle when applying both methods (K-means & FCM). Austria gets the cluster Upper Middle when applying both methods (K-means & FCM). Bosnia and Herzegovina gets the cluster Lower Middle when applying both methods (K-means & FCM).

As has been seen above after making the comparison of both methods in terms of the results gotten, it is seen a clear difference in classifying the cases, However, comparing the results in order to get the best resulting clustering is not easy and that is because of a lot of several metrics used in the comparison that is not provided by every algorithm.

Classic metrics will be used within this comparison such as the results gotten from the tables of cluster method (R-squared, AIC, BIC and Silhouette). Those are the results when applying both methods on the data set applied that have which it was unlabeled data and not fuzzy.

It has been observed by Rokach that "both perform better than their classical counterparts, the K-means algorithm, and the fuzzy c-means algorithm" (Rokach & Maimon, 2005).

Noting the underlying differences of the fuzzy and crisp clustering (Hard clustering), some researchers have investigated the efficiency measures such as number of iterations and time complexity (Ghosh & Dubey, 2013), (Cebeci, Z. & Yildiz, F, 2015).

Regarding the comparison of both K means and FCM methods, there are few stands of views, see for instance in terms of iterations you can see in the paper of (Hakyemez, Bozanta, & Coşkun, 2018), the Mobile App dataset that contains features of 7196 available applications was clustered using two popular clustering algorithms (K means and FCM) they mentioned in terms of iterations that there are efficiency measures that can be used as a basis for comparing the computational performances of the algorithms. For the best models, k-means algorithms converged after only 4 iterations whereas it takes 25 iterations for fuzzy c- means algorithm to reach the optimal cluster centers.

And in terms of time, An empirical study was presented by alsultan and Khan (Al-Sultana & Khanb, 1996)the K-means method is the most efficient in terms of execution time comparing with other methods.

Soft clustering algorithms are slower than hard clustering algorithm as there are more values to compute and as a result, it takes longer for the algorithms to converge. (Malik, 2019).

Another stand of view of study results of Celebi et al indicate that "fuzzy c-means does not seem to offer any advantage over hard c-means. Furthermore, due to the intensive membership calculations involved, fuzzy c-means is significantly slower than hard c-means, which makes it unsuitable for time-critical applications" (Celebi , Quan Wen, & Schaefer, 2011).

In the conclusion of a study on the Featured Mobile Applications Benchmark that has been done by (Hakyemez, Bozanta, & Coşkun, 2018), that "The performance indicators suggest that fuzzy c-means prevailed over the k-means algorithm in terms of cluster quality measured by Xie and Beni index. But, from the efficiency point of view, k-means algorithm is far better than its counterpart. More clearly, it converges more rapidly with less iteration, which may cause huge efficiency differences in larger datasets, thus being more preferable in some cases" (Hakyemez, Bozanta, & Coşkun, 2018).

CHAPTER 8

CONCLUSION AND SUGGESTIONS

8. CONCLUSION AND SUGGESTIONS

In this paper, K-means (hard) and fuzzy c-means clustering algorithms were compared within the context of classifying countries. The results in all years compared 2013,2014,2015,2016,2017 and 2018 indicate that K-means method is better than Fuzzy C-means method in clustering according to the data set applied on.

Where in all years compared, it is concluded that in terms of R2, K-means method has a better coefficient of determination. And regarding the outputs of AIC, K-means method is better in terms of fit for the data, and in terms of BIC, K-means method is better in terms of criterion for model selection among a finite set of models. Finally, regarding the output Silhouette, it is concluded that K-means method is more consistent.

Soft clustering algorithms are the solution for the cases where a data point could be affiliated to multiple groups or when wanted to find how similar a data point is to multiple given centers. As Malik states “if you are attempting to forecast the rating changes for the counterparties who you trade with then you can use the soft clustering technique. The algorithm can create clusters for each rating and indicate the likelihood of a counterparty to belonging to a cluster” (Malik, 2019).

According to the large dataset, (Sheshasayee, A & Sharmila, P, 2014)claime that FCM does not run quicker than KM, in the opposite of what (Cebeci, Z. & Yildiz, F, 2015), where they found that FCM is slower in all datasets in a remarkable way.

“Clusters' number quickly increases the time complexity of FCM increases. So, it is understood that runtimes of FCM are mainly affected by the number of clusters rather than their sizes” (Cebeci, Z. & Yildiz, F, 2015). However as it is been reported also that KM and FCM were equally efficient to find clusters in all datasets having the clusters scattering with regular patterns, KM was superior to FCM when the computing cost was also concerned as a privilege factor in the choice of an appropriate algorithm.

It is been recommendedby Cebeci et al that "KM can be used for its lower computing time cost and higher number of correctly found clusters" (Cebeci, Z. & Yildiz, F, 2015),. Similar conclusions were reported by (Madhukumar, S & Santhiyakumari, N, 2015).

K-means was faster than Fuzzy c-means in all datasets that contain the clusters in irregular or regular patterns. Fuzzy c-means is an algorithm based on more iterative fuzzy calculations. That means soft clustering algorithm is slower than hard clustering algorithm but, on other hand, it does not get affected when there is uncertainty in the data.

Similar results were stated by Panda et al for Iris (Panda, S, Sahu, S, Jena, P, & Chattopadhyay, S , 2012), Wine and Lens datasets; by (Gohokar & Jipkate, 2012) for segmentation of images; by (Ghosh & Dubey, 2013) for Iris dataset; by (Bora & Gupta, 2014) for Iris dataset; by Sivarathri & Govardhan (2014) for diabetes data; and by (Madhukumar, S & Santhiyakumari, N, 2015) for brain MR images data.

From the efficiency point of view, k-means algorithm is far better than its counterpart. Where K-Means clustering was more efficient in terms of execution time comparing with FCM and the number of iterations and the accuracy of the results and the suitability of the method applied on the data used.

Come to a final conclusion, there is no algorithm that is the best for all cases. Hence, the given data sets should be accurately examined in order to determine which algorithm would be more suitable. To achieve this, it is ought to understand that an important factor in selecting a better and more suitable clustering algorithm is the type of data used.

So relatively to what mentioned above, it could be found in some studies that applying FCM clustering is better than KM clustering, for instance (Govardhan & Sivarathri, 2014)Sivarathri & Govardhan (2014) revealed that FCM is better than KM in term of the accuracy of clusters on the diabetes dataset obtained from the UCI repository. However, in this study, KM was more appropriate to use, or in other words, KM clustering is successful to find better clusters for each case appeared in the used data set in this study.

The writer would like to suggest other researchers conduct further studies on this topic. Another research may investigate the same topic, but with different Cluster Analysis Techniques, for example: another study might be on

the same data set but with applying other statistical Analysis Techniques, like comparing between Distribution Clustering and Fuzzy C-means. Or, between Connectivity Clustering (Hierarchical clustering) and K-means.

REFERENCES

A. Liew, S. Leung, & W. Lau, . (2000). Fuzzy Image Clustering Incorporating Spatial Continuity. In IEE Proceedings Vision, Image and Signal Processing, vol. 147, no. 2.

Akaike, H. (1987). Factor analysis and AIC. Psychometrika. s. 317–332.

Al-Sultana, K., & Khanb, M. (1996, March). Computational experience on four algorithms for the hard clustering problem. s. 295-308.

B. Höppner, M. Bartelheım, R. Krauss, & M. Huıjsmans. (2005, May). Prehistoric copper production in the Inn Valley (Austria), and the earliest copper in central Europe.

Baradwaj, B., & Pal, S. (2012). Mining educational data to analyze students' performance. arXiv.

Bezdek. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press.

Bora, D., & Gupta, A. (2014). A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm. Int. J. of Computer Trends and Technology, vol. 10, no. 2, pp. 108-113.

Burmaoğlu, S. (2011). Determining Purchasing Alternatives By Using Multivariate Statistical Techniques: An Application on Sniper Rifles.

Cai, W, Zhang, D, & Chen, S. (2007). Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. Pattern recognition,. s. 825-838.

Cebeci, Z., & Yildiz, F. (2015). . Comparison of K-means and Fuzzy C-means algorithms on different cluster structures. Agrarınformatıka /Journal Of Agrıcultural Informatıcs. s. 13-23.

Celebi , M., Quan Wen, & Schaefer, G. (2011, April). A Comparatıve Study Of K-Means And Fuzzy C-Means For Color Reductıon. Gerald Schaefer.

Centroid clustering reside in Machine Learning: Different Types of Clustering Methods and Applications by Sunit Prasad (July 5, 2020) India's Top Ranked Data Science Institute Web site retrieved June 16,2021, from https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/

Cheung, Z. (2020, August). Factor analysis & Cluster analysis on Countries Classification.

Connectivity Clustering (Hierarchical clustering) reside in Machine Learning: Different Types of Clustering Methods and Applications by Sunit Prasad (July 5, 2020) India's Top Ranked Data Science Institute Web site retrieved June 16,2021, from https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/

Constraint-Based Clustering Method reside in Machine Learning: Different Types of Clustering Methods and Applications by Sunit Prasad (July 5, 2020) India's Top Ranked Data Science Institute Web site retrieved June 16,2021, from https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/

Density Clustering (Model-based methods) reside in Machine Learning: Different Types of Clustering Methods and Applications by Sunit Prasad (July 5, 2020) India's Top Ranked Data Science Institute Web site retrieved June 16,2021, from https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/

Dhanachandra, N., Chanu, Y., & Manglem, K. (2015). Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. Procedia Computer Science. s. 764-771.

Dirsehan, T. (2015). Classifying Countries According To Their Export Competitiveness The Position Of Turkey As An Emerging Economy.

Distribution Clustering reside in Machine Learning: Different Types of Clustering Methods and Applications by Sunit Prasad (July 5, 2020) India's Top Ranked Data Science Institute Web site retrieved June 16,2021, from https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/

Dunn, J. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics. s. 32-57.

Ester, Martin, Xiaowei Xu., Sander, J., & Kriegel, H.-P. (1996). “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.”. Ester, Martin.

Forina, M., & Lanteri, S. (1984). Data Analysis in Food Chemistry.

Fuzzy Clustering reside in Machine Learning: Different Types of Clustering Methods and Applications by Sunit Prasad (July 5, 2020) India's Top

Ranked Data Science Institute Web site retrieved June 16,2021, from https://www.analytixlabs.co.in/blog/types-of-clustering-algorithms/

G. Hamerly. (2003). Learning Structure and Concepts in Data using Data Clustering, PhD Thesis. University of California, San Diego.

Gervelis, G. (2018). 4 Types of Cluster Analysis Techniques.

Ghosh, S., & Dubey, S. (2013). Comparative analysis of k-means and fuzzy c-means algorithms. International Journal of Advanced Computer Science and Applications. 35-38.

Ghosh,, S., & Dubey, , S. (2013). Comparative Analysis of K-Means and Fuzzy C-Means Algorithms. Int. J. Advanced Computer Science and Applications,, vol. 4, no.4, pp. 35-39.

Gohokar, V., & Jipkate, B. (2012). A Comparative Analysis of Fuzzy C-Means Clustering and K Means Clustering Algorithms. Int. J. of Computational Engineering, vol. 2, no. 3, pp. 737-739.

Govardhan, & Sivarathri. (2014). 'Experiments on Hypothesis Fuzzy K-Means is Better Than K-Means for Clustering' Int. J. Data Mining & Knowledge Management Process, vol. 4,.

Granville, V. (2019, March 13). How to Automatically Determine the Number of Clusters in your Data.

Hakyemez, T. C., Bozanta, A., & Coşkun, M. (2018, October). K-Means vs. Fuzzy C-Means: A Comparative Analysis of Two Popular Clustering Techniques on the Featured Mobile Applications Benchmark. s. 24-26.

Hruschka , H., & Natter , M. (1999). Comparing performance of feedforward neural nets and K-means for cluster-based market segmentation.

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values.

J. Bezdek. (1980). A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence. s. 1–8.

Kaufman , L., & Rousseeuw, P. (1990, March 08). Partitioning Around Medoids (Program PAM).

Kaufmann, M., Han, J., , & Kamber, M.:. (2000, September). Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann.

Kim, K., , & Ahn, H. . (2008). A recommender system using GA K-means clustering in an online.

Kumar, S. (2020). An overview of agglomerative hierarchical clustering, dendrogram and their implementation in python. Towards Data Science.

Kyle, A., Obizhaeva, A., Sinha, N., & Tuzun, T. (2012). News Articles and the Invariance Hypothesis.

Lester B , & Pearson . (1969, January 1). Partners in Development: Report of the Commission on International Development.

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. s. 281–97.

Madhukumar, S, & Santhiyakumari, N. (2015). Evaluation of K-Means and Fuzzy C-means Segmentation on MR Images of Brain. The Egyptian J. of Radiology and Nuclear Medicine,, ol. 46, no. 2, pp. 475–479.

Malik, F. (2019, Jun 7). Machine Learning Hard Vs Soft Clustering article.

Mingoti, S , & Lima, J. (2006). Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. European Journal of Operational Research.

Misra, D., & Tiwari, M. (2011, December). Application of Cluster Analysis in Agriculture – A Review Article.

Modha , D., & Dhillon , I. (2001). Concept Decompositions for Large Sparse Text Data Using Clustering Inderjit.

Ng, H., Goh, P.,, Nowinski, W., Ong, s, & Foong, k. (2006). Medical image segmentation using kmeans clustering and improved watershed algorithm. Image Analysis and Interpretation. s. 61-65.

Nielsen, L. (2011, February). Classifications of Countries Based on Their Level of Development.

Oyelade, & Oladipupo. (2010). Application of k Means Clustering algorithm for prediction of Students Academic Performance.

P.J. Flynn, Muraty, M., & Jain , A. (1999, September). Data Clustering: A Review.

P.N. Tan , Kumar, V., & Steinbach, M. (2005, January). cluster Analysis: Basic consepts and algorithms.

Panda, S, Sahu, S, Jena, P, & Chattopadhyay, S . (2012). Comparing Fuzzy-C Means and K-Means Clustering Techniques. A Comprehensive Study'. Advances in Intelligent and Soft Computing, vol. 166, pp. 451-460.

Prasad, S. (5th July 2020). Different Types of Clustering Methods and Applications.

Rokach , L., & Maimon, O. (2005). Data Mining and Knowledge Discovery Handbook.

Rokach, L., & Maimon, O. (2005). Data Mining and Knowledge Discovery Handbook, Clustering Methods. s. 321–352.

Rong, C. (2011). A comparison between k-means and fuzzy c-means in the cloud. s. 565-569.

S. Theodoridis. (2009, December). Clustering Algorithms III: Schemes Based on Function Optimization.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics. s. 461–464.

Selim, S.Z. , & Ismail, M.A. . (1984). K-Means Type Algorithms a generalized convergence theorem and characterization of local optimality.

Sheshasayee, A, & Sharmila, P. (2014). Comparative Study of Fuzzy C-means and K-means Algorithm for Requirements Clustering. Indian J. of Science and Technology, no 6, pp. 853–857.

Tarau , P., & Mihalcea , R. (2005). A Language Independent Algorithm for Single and Multiple Document Summarization.

Tryon , R., & Bailey, D. (1970). Cluster Analysis.

Vayssieres, M., & Plant, R. (1998). Identification of Vegetation State and-transition Domains in California's Hardwood Rangelands.

Wu et al,. (2008). Effects of 2000-2050 global change on ozone air quality.

Yannis Goulermas, J., Ananiadou, S., Mu, T., & Brockmeier, A. (2018). Self-Tuned Descriptive Document Clustering Using a Predictive Network.

Zhang, & Chen, C. (2004). A novel kernelized fuzzy c-means algorithm with application in medical image segmentation. Artificial intelligence in medicine. s. 37-50.

ISBN: 978-625-8323-67-2