Clustering to Investigate about Countries – Task 2

1. Introduction

The analysis was tried to cluster the countries based on their information. Then each group has their similar characteristics and easy to identify their patterns. These groups were grouped based on countries' child morality, exports, health spending etc. When we receive the information from a new country which is not in the given dataset, these models can easily group it into relevant cluster and we can recognize and get an idea about its patterns.

2. Methodology

The analysis was used one of the best three clustering algorithms which are Logistic K Means Clustering (KM), Mean Shift Clustering (MS) and Hierarchical Clustering (HR) to build the best model. The main of the KM is to reduce the total distances between the data point and their corresponding clusters and also KM builds k clusters and they have n data points. Mean Shift algorithm also commonly used algorithm and it is a density based clustering algorithm. Hierarchical algorithm arranges the data into a hierarchical or tree structure.

The performance of these algorithms was checked using 'Silhouette Score' and it's a good method to calculate the quality of the clusters.

3. Results and Discussion

First and foremost, the analysis was tried to capture the patterns of the dataset using exploratory data analysis and necessary data preprocessing methods. Initially, the analysis was removed 'country' variable because it doesn't give any information for the analysis. Then it was captured the missing values, data distribution using density plots and correlation between each variable using heatmap and correlation with the response variable using bar chart of correlation matrix. Basically, it was selected only two independent variables based on the correlation which have the highest relationship between the variables with target variable. The initial model was built using 'gdpp' and 'income' variables [Task 2.1]. When the analysis was performed KM model, it was found the optimal clusters for KM using elbow method and also HR model was performed going through complete linkage method and single linkage method which measure the distance of two clusters using maximum and minimum distance between any two points of that two clusters respectively.

Figure 1, Figure 2 and Figure 3 represents the optimal clusters with 71%, 65.5% and 70.4% of the quality/ strong of the clusters from Silhouette score after using KM, MS and HR algorithms respectively.

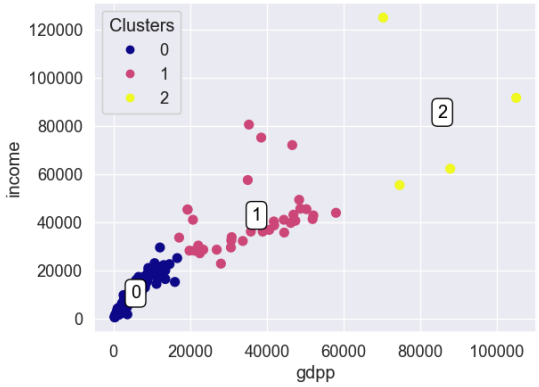

Figure 1: Clusters of KM algorithm

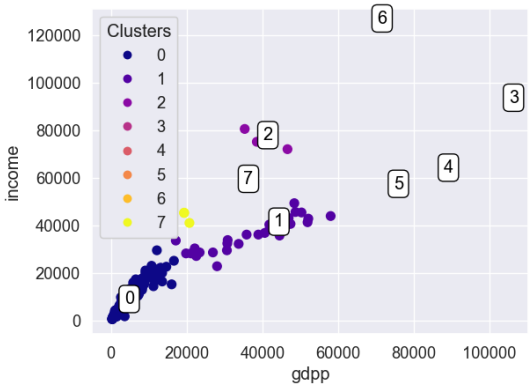

Figure 2: Clusters of MS algorithm

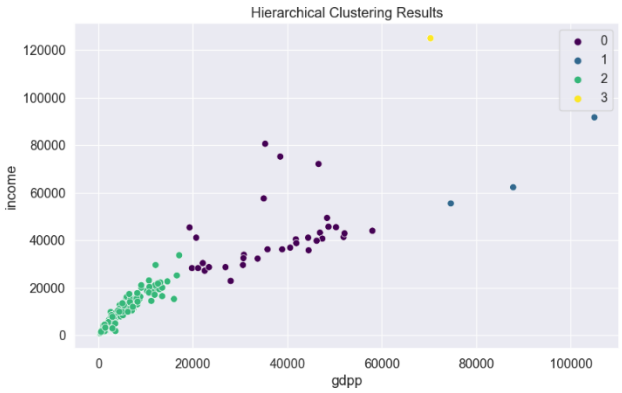

Figure 3: Clusters of HR algorithm

Figure 1 and Figure 3 grouped the data points better than Figure 2.

Secondly, the model was performed with all the independent variables under the same preprocessing techniques that were used in initial model and second initial model was given 71%, 65.5% and 70.3% of Silhouette score respectively for KM, MS and HR algorithms [Task 2.1].

Next, more data preprocessing techniques which are duplicate value checking, checking the data distribution, checking the correlation using heatmap, scaling the data using standardize (Std) and normalize (Nor) methods and experimenting the results with outlier removing and outlier imputation from median because data distribution of all the variables don't follow the normal distribution and not changing the outliers were added rather than initial model [Task 2.2].

TABLE 1 represents how the Silhouette values of the models vary and we can get better understanding about the quality of the clusters that was built by three algorithms with data preprocessing techniques.

TABLE 1: Overall Results After Adding Data Preprocessing Techniques

	Outlier Imputation ; Median			Outlier not Handling			Outlier Removal		
	Std	Nor	Unscaled	Std	Nor	Unscaled	Std	Nor	Unscaled
KM	20.9%	24%	61.6%	28.3%	33.9%	71%	27.6%	34.3%	61.9%
MS	22.5%	25.2%	60.8%	23.2%	18.8%	65.5%	N/A	N/A	58.6%
HR	16.4%	15%	47.9%	28.6%	29.9%	69.4%	24.6%	28%	51.7%

4. Conclusion and Recommendations

TABLE 2: Overall Results without using Data Preprocessing Techniques

	Using only gdpp and income variables	Using all the variables
KM	71%	71%
MS	65.5%	65.5%
HR	70.4%	70.3%

According to the final results we can conclude that the best Silhouette score which is the score for quality of the clusters is 71% and it was given by few models. Figure 1 visualizes one best model and it was clearly grouped the data points into three clusters and it was grouped using 'gdpp' and 'income' variables which have the highest correlation pair using KMeans clustering algorithm. Other best model was also given from the KMeans clustering algorithm without doing any data preprocessing analysis and there are also three clusters according to lbow method. Therefore, we can conclude that KMeans algorithm is the best approach for this dataset and data preprocessing techniques that were used and adding all the variables in the analysis can't give much effect to improve the model.

Finally, we can group the countries based on these models and experts can get their decisions after considering the patterns of variables in each cluster.