Journal of Student Research Exploration

https://shmpublisher.com/index.php/josre
p-ISSN 2964-1691 | e-ISSN 2964-8246

Analysis of k-means clustering algorithm in advanced country clustering using rapid miner

Ireneus Prabaswara[1], Dwika Ananda Agustina Pertiwi[2], Jumanto[3]
[1, 3]Informatics Engineering Study Program, Universitas Negeri Semarang, Indonesia
[2]Faculty of Technology Management, Universiti Tun Hussein Onn Malaysia, Malaysia

Article Info

Article history:

Received April 18, 2024
Revised July 4, 2024
Accepted July 5, 2024

Keywords:

Data mining
Clustering
K-means
Developed countries
Machine learning

ABSTRACT

In the era of globalization, the understanding of developed countries is no longer limited to the level of per capita income alone. As part of the analysis of developed countries based on aspects of government revenue, income balance, national savings, and domestic output based on sales. This research aims to cluster and to find out how these economic indicators are interrelated and affect the status of a country as a developed country. The K-means algorithm is used to identify patterns of countries with similar economic characteristics. From the research conducted, there are 4 clusters generated based on the characteristics of developed countries.

1. Introduction

In the 21st century, the understanding of developed countries is no longer limited to per capita income levels alone. Aspects such as general government revenue, income balance, national savings, and sales of domestic products become important focuses when analyzing a country's development. Economic theory argues that the movement of capital flows between countries benefits all parties and results in efficient resource allocation, increased productivity, and economic growth [1]. Capital inflows tend to be positively correlated with macroeconomic cycles, a phenomenon that Kaminsky et al. labeled "when it rains it pours" [2].

[1] ***Corresponding Author:***
Ireneus Prabaswara,
1Informatics Engineering Study Program,
Universitas Negeri Semarang,
Sekaran, Gunung Pati, Semarang City, Central Java, Indonesia.
Email: dina.lusianti@umk.ac.id
DOI: https://doi.org/10.52465/josre.v2i2.337

An in-depth understanding of how domestic sales reflect the productivity of the economy is important for developing sustainable development strategies. Examining the relationship between these indicators is expected to provide more comprehensive insights into economic trends in developed countries and their potential to address complex global challenges. By strengthening macroeconomic fundamentals, domestic factors that increase the return on domestic investment can prevent fire sales and sudden stoppages of capital flows and deter foreign investors from exiting when a country is in financial distress [3], [4].

There are many types of data mining. Data mining training methods fall into two broad classes: supervised learning and unsupervised learning [5]. Supervised learning focuses on tasks that target prediction, classification, and detection data. In addition, human experts may also provide labels for certain defects (rather than just signaling the presence of defects). This information is then used to train a predictive model [6]. Unsupervised learning is more focused on performing tasks that do not involve target data such as clustering [7]. Both have evolved into several smaller techniques as well. One type of unsupervised learning is K-Means. This algorithm is used when we have data that does not yet have a label, in other words, data that cannot be clustered.

K-Means algorithm is one of the popular machine learning methods to reduce the objective function that has been set in the clustering process. The objective here is to minimize variation by maximizing other cluster data. K-Means is a well-known partitioning method for clustering [8]. The K-Means algorithm is a data analysis method that groups data into similar categories or clusters. As part of the analysis of developed countries based on aspects of government revenue, balance of income, national savings, and domestic output based on sales, the K-means algorithm is used to identify patterns or patterns of countries with similar economic characteristics.

The K-Means algorithm allows for the analysis of developed countries to be divided into more homogeneous categories based on relevant economic indicators. The K-means algorithm requires specifying the number of clusters "K" at the beginning of the clustering process. Although users can usually determine the number of clusters, they are not always able to identify the most appropriate and accurate number [9]. Determining the optimal number of clusters can be done through various methods, including statistical indices, variance-based methods, and information theory approaches [10]. This provides greater insight into the diversity and similarities of the economies of these developed countries.

This research aims to find out how these economic indicators are interrelated and affect a country's status as a developed country by clustering. Clustering analysis aims to group unlabeled data sets into several clusters so that similar data are grouped into the same cluster and different data are separated into different clusters [11]. In this context, how government revenue reflects economic stability,

how income balance affects long-term economic growth, how national savings reflect social welfare.

2. Method

There are several stages carried out in this research. The flow of the research can be seen in Figure 1.

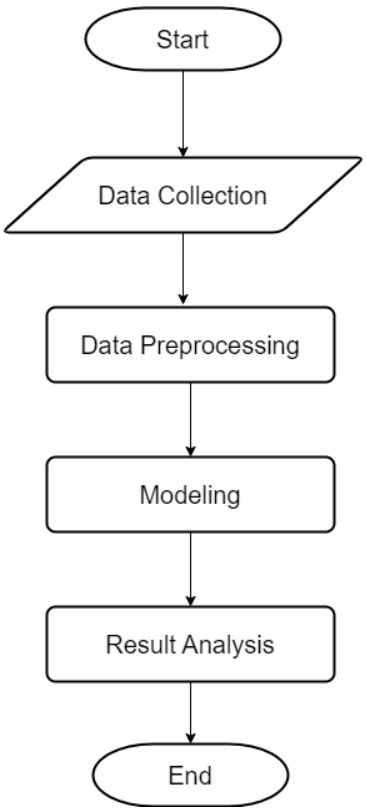

Figure 1. Research flowchat

In order to make the clustering of developed countries in this analysis more focused, I divided this research into four stages, namely; the first stage is data collection by creating a dataset obtained from GitHub which is still relevant to me with the data title is IMF. The second stage is data processing. In this analysis, I will perform imputation to eliminate missing values contained in the dataset.

The third stage will be done if the analysis is done using RapidMiner with the K-Means algorithm. This method begins by importing the data into the job page, then by selecting the attribute that we want to analyze in numeric form. After we get the desired attribute, we replace the Missing Value contained in the data table after that the data normalization process is carried out. The last stage is done by analyzing the cluster results generated by the K-Means model.

Data Collection

The dataset used in this analysis is a dummy dataset generated using the GitHub application with the data name "IMF.csv". In the dataset, there are already missing values that meet the criteria to be processed in RapidMiner using the K-Means algorithm method.

Data Preprocessing

Data pre-processing is done by checking and resolving missing value issues in the dataset, if any. Datasets often have missing values due to measurement errors, incomplete data collection, or due to the nature of the measurement itself [12]. The different patterns that can be adopted by missing values in a dataset can affect the performance of the algorithms used differently and significantly [13]. Therefore, there is a need to handle missing values. In this research, missing values are handled by deleting data rows that have missing values. In addition, this research also applies normalization techniques to make it easier for the model to understand the dataset.

Modeling

The clustering model is created using the K-Measn algorithm. The K-means clustering algorithm generates clusters using the average value of cluster objects [14], [15]. The cluster number is required as a user-defined parameter and is used in randomly selecting cluster centers from the dataset [16]. The workings of K-Means can be seen in Figure 2.

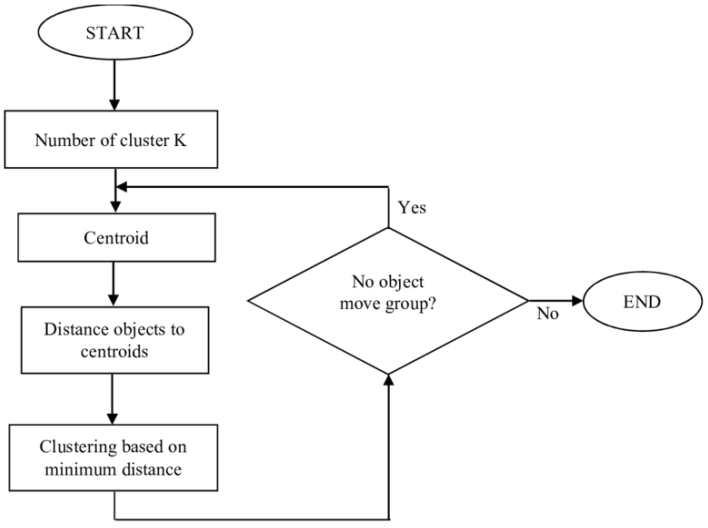

Figure 2. K-Means worfklow [17]

Result Analysis

At this stage, analysis is carried out by analyzing the characteristics of the clusters formed. This process produces information and characteristics of each cluster generated based on the dataset of the dataset used in the modeling process.

3. Results and Discussion

Data Collection

The dummy dataset obtained from the IMF's GitHub contains 186 data with several criteria used as the basis for assessing developed countries, including aspects of government revenue, income balance, national savings, and domestic product of sales. In accordance with the dataset prerequisites, there are several elements of missing values used in data processing.

Data Preprocessing

In this analysis Rapid Miner is used to help the analysis process. The following is a table column that shows the nominal value of several attributes before processing. There is a Missing Value marked with a question mark which will be processed and eliminated.

Figure 3. The result of checking for missing values

Any missing values in the data that has been collected are addressed using the Replace Missing Values Series method. This method is the simplest technique to handle missing values. The missing data will be replaced with the average value of the other values.

The data that appears is very varied, so data normalization is carried out so that the data can become normal before clustering. This normalization method uses range transformation so that the value becomes 0 to 1.

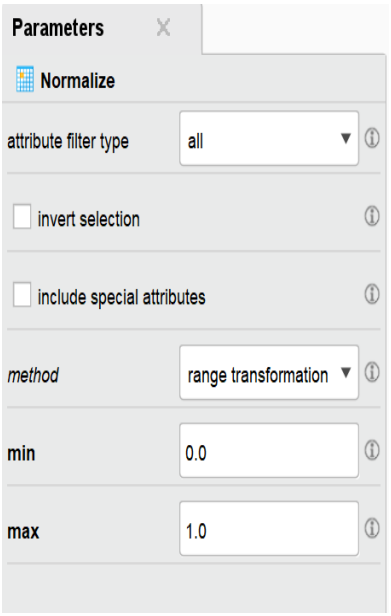

Figure 3. Normalize parameters

The results of normalization can be seen in Figure 4.

Current acc...	General gov...	Gross dome...	Gross natio...	Total invest...
0.456	0.221	0.002	0.581	0.385
0.268	0.277	0.002	0.356	0.371
0.500	0.433	0.017	0.843	0.712
0.519	0.535	0.008	0.458	0.081
0.247	0.227	0.000	0.380	0.446
0.416	0.444	0.044	0.471	0.340
0.228	0.206	0.001	0.387	0.490
0.373	0.365	0.061	0.489	0.391
0.445	0.603	0.023	0.499	0.277
0.758	0.567	0.006	0.815	0.210
0.452	0.297	0.002	0.640	0.483
0.429	0.068	0.018	0.566	0.348
0.307	0.417	0.000	0.235	0.114
0.224	0.508	0.009	0.538	0.732

attributes)

Figure 4. Normalization results on the dataset

Modeling

The results of the clustering process can be seen in Figure 5.

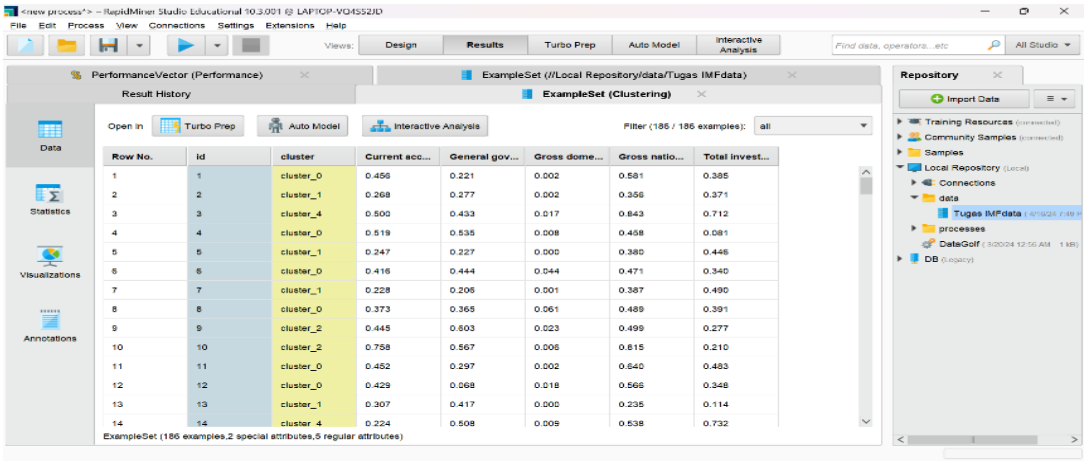

Figure 5. Clustering results

Furthermore, to compare with the best K value, a looping process is performed. Loop parameters used is by repeating the parameters option used previously to compare if we use another K value. Here with the looping parameters method, we use a value range between 10 with a step of 10 as a previous comparison. The results of the looping process can be seen in Figure 5 and Figure 6.

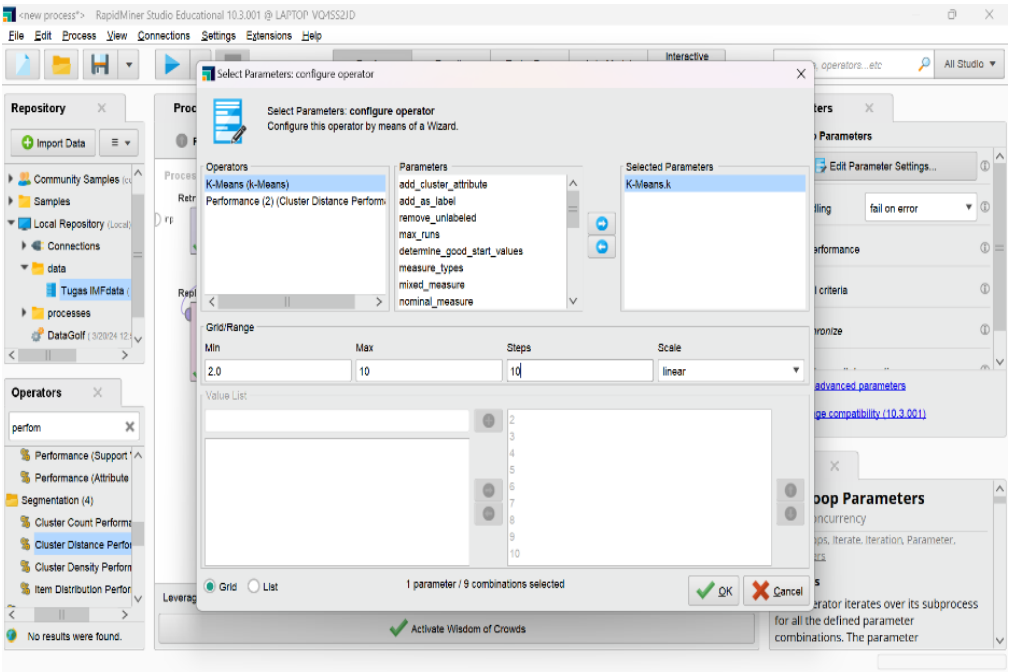

Figure 5. Looping range 10

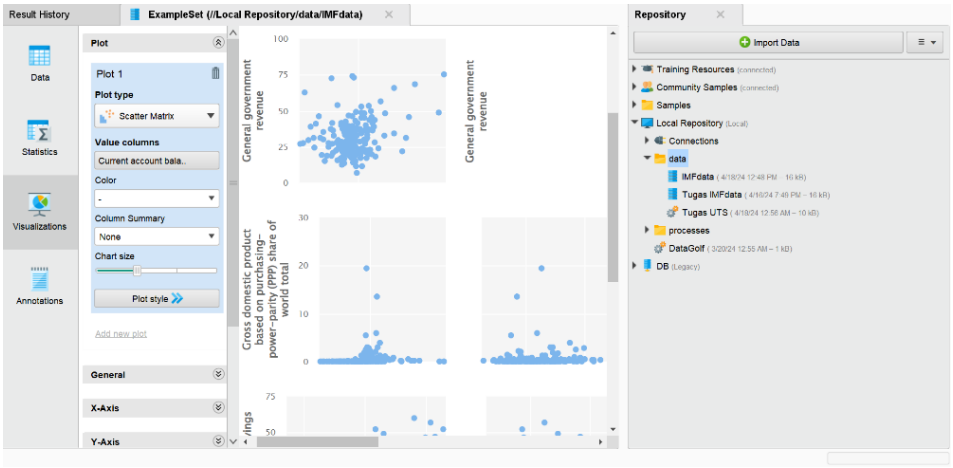

Figure 6. Visualization of k-means clustering

System Testing

The following is the arrangement in processing data starting from importing data to the work page, then selecting the attributes to be analyzed. Before clustering there is an error due to missing values, so the solution is to replace missing values. After the missing value is resolved, data normalization is carried out so that the data form is simpler. The result of normalization can be seen in Figure 7.

Row No.	Current acc...	General gov...	Gross dome...	Gross natio...	Total invest...	Country
19	4.890	33.165	0.064	24.987	17.007	Bolivia
20	-5.733	46.725	0.041	13.655	19.388	Bosnia and H
21	-1.992	29.416	0.037	27.939	29.481	Botswana
22	-2.206	35.424	2.928	18.033	20.239	Brazil
23	45.453	48.609	0.027	?	15.878	Brunei Darus.
24	-1.043	32.694	0.130	21.844	22.887	Bulgaria
25	-2.327	20.138	0.028	16.018	18.345	Burkina Faso
26	-9.387	37.304	0.007	10.111	19.920	Burundi
27	-3.916	17.034	0.041	13.431	17.348	Cambodia
28	-3.012	17.467	0.060	13.108	16.120	Cameroon
29	-3.131	38.425	1.786	19.074	22.204	Canada
30	-12.467	28.013	0.003	25.337	37.804	Cape Verde
31	-10.195	17.150	0.005	4.079	14.275	Central Africa
32	-3.511	25.271	0.025	38.905	42.416	Chad

Figure 7. Dataset after normalization process

Analysis of Cluster Results

Next is the result stage of the algorithm analysis using K-Means. With visualization using the Scatter Matrix type, it can be seen that the attributes have been clustered. The clusters produced by the model are as follows:

1. Cluster 1: Cluster 1 are countries that have more government revenue than other countries.
2. Cluster 2: Cluster 2 are countries that have higher government revenues than countries in cluster 1.
3. Cluster 3: Cluster 3 are countries that have medium government revenues, not high but not low either.
4. Cluster 4: Cluster 4 are countries that have the highest government revenues compared to countries in other clusters.

4. Conclusion

The analysis of the K-Means algorithm shows that there are four groupings of countries based on total government revenue. Cluster 1 consists of countries with the lowest government revenue, and Cluster 4 consists of countries with the highest government revenue. The results of this analysis can be used to understand the differences in government revenue around the world. The results can also be used to identify countries that have similar characteristics in terms of government revenue. For future research, it is recommended to optimize the clustering model by using feature selection techniques or stitching in the modeling process.

REFERENCES

[1] O. Ogrokhina and C. M. Rodriguez, "Inflation targeting and capital flows: A tale of two cycles in developing countries," *J. Int. Money Financ.*, vol. 146, p. 103121, Aug. 2024, doi: 10.1016/j.jimonfin.2024.103121.

[2] G. L. Kaminsky, C. M. Reinhart, and C. A. Végh, "When It Rains, It Pours: Procyclical Capital Flows and Macroeconomic Policies," *NBER Macroecon. Annu.*, vol. 19, pp. 11–53, Jan. 2004, doi: 10.1086/ma.19.3585327.

[3] R. J. Caballero and A. Simsek, "A Model of Fickle Capital Flows and Retrenchment," *J. Polit. Econ.*, vol. 128, no. 6, pp. 2288–2328, Aug. 2019, doi: 10.1086/705719.

[4] E. Cavallo, A. Izquierdo, and J. J. León-Díaz, "Preventing Sudden Stops in Net Capital Flows," Washington, D.C., Aug. 2020. doi: 10.18235/0002561.

[5] H. M. Ferreira, D. R. Carneiro, M. Â. Guimarães, and F. V. Oliveira, "Supervised and unsupervised techniques in textile quality inspections," *Procedia Comput. Sci.*, vol. 232, pp. 426–435, 2024, doi: 10.1016/j.procs.2024.01.042.

[6] B. Mohammadi, M. Fathy, and M. Sabokrou, "Image/Video Deep Anomaly Detection: A Survey," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.01739

[7] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009, doi: 10.1145/1541880.1541882.

[8] J. Heidari, N. Daneshpour, and A. Zangeneh, "A novel K-means and K-medoids algorithms for clustering non-spherical-shape clusters non-sensitive to outliers," *Pattern Recognit.*, vol. 155, p. 110639, Nov. 2024, doi: 10.1016/j.patcog.2024.110639.

[9] Yee Leung, Jiang-She Zhang, and Zong-Ben Xu, "Clustering by scale-space filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1396–1410, 2000, doi: 10.1109/34.895974.

[10] C. C. Aggarwal and C. Zhai, "Text classification," in *Data Classification: Algorithms and Applications*, 2014, pp. 287–336. doi: 10.1201/b17320.

[11] W. Wu, W. Wang, X. Jia, and X. Feng, "Transformer Autoencoder for K-means Efficient clustering," *Eng. Appl. Artif. Intell.*, vol. 133, p. 108612, Jul. 2024, doi: 10.1016/j.engappai.2024.108612.

[12] S. Piqueras *et al.*, "Handling Different Spatial Resolutions in Image Fusion by Multivariate Curve Resolution-Alternating Least Squares for Incomplete Image Multisets," *Anal. Chem.*, vol. 90, no. 11, pp. 6757–6765, Jun. 2018, doi: 10.1021/acs.analchem.8b00630.

[13] A. Gómez-Sánchez, R. Vitale, C. Ruckebusch, and A. de Juan, "Solving the missing value problem in PCA by Orthogonalized-Alternating Least Squares (O-ALS)," *Chemom. Intell. Lab. Syst.*, vol. 250, p. 105153, Jul. 2024, doi: 10.1016/j.chemolab.2024.105153.

[14] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[15] M. Capo, A. Perez, and J. A. A. Lozano, "An efficient Split-Merge re-start for the K-means algorithm," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2020, doi: 10.1109/TKDE.2020.3002926.

[16] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (Ny).*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.

[17] A. Bustamam, H. Tasman, N. Yuniarti, Frisca, and I. Mursidah, "Application of k-means clustering algorithm in grouping the DNA sequences of hepatitis B virus (HBV)," 2017, p. 030134. doi: 10.1063/1.4991238.