Machine Learning for Regencies-Cities Clustering Based on Inflation and Poverty Rates in Indonesia

R Gustriansyah[*1], J Alie[2], A Sanmorino[3], R Heriansyah[4], M N M M Noor[5]

[1,3,4]Faculty of Computer Science, Universitas Indo Global Mandiri, Indonesia

[2]Faculty of Economics, Universitas Indo Global Mandiri, Indonesia

[5]Departement of Computer Engineering Technology, Malaysian Institute of Information Technology, Universiti Kuala Lumpur, Malaysia

E-mail: rendra@uigm.ac.id[1], juhaini@uigm.ac.id[2], sanmorino@uigm.ac.id[3], rudi@uigm.ac.id[4], megatnorulazmi@unikl.edu.my[5]

Submitted: 12 March 2022, revised: 28 August 2022, accepted: 31 August 2022

Abstract. The COVID-19 pandemic has increased inflation and poverty rates in many cities that requires considerable attention from the government as a policymaker. Therefore, this study aims to cluster regencies and cities that need mitigation priorities from the Indonesian government based on inflation and poverty rates in 2021. Four machine learning methods, namely k-Means (KM), Partitioning around medoids (PAM), Ward, and Divisive analysis (Diana) were utilized and compared to achieve that purpose. The clustering of 90 regencies and cities in Indonesia produced five optimal clusters. Furthermore, the clustering results were validated using the Silhouette width (SW) and Dunn index (DI). The results showed that the k-means method produced the most compact cluster. Hence, this study's results can be utilized as a reference for the government in determining the steps and priorities of economic policy in Indonesia.

Keywords: clustering, cluster validation, inflation rate, machine learning, poverty rate.

1. Introduction

Inflation is defined as an increase in the prices of goods and services in general in a certain period continuously [1]. Inflation is one of the economic indicators that have a very vast influence on the socio-economic life of the community. High and unstable inflation has resulted in higher poverty levels in an area, especially during the Covid-19 pandemic.

Therefore, the government must be able to determine actions and policies to control inflation and at the same time reduce the poverty rate. One of these policies is to stabilize the economic sector, including the distribution of social safety nets, subsidies, or other stimulus packages based on regional priorities. Regency/city clustering can be the basis for determining regional priorities.

Several studies summarized in Table 1 have clustered regencies/cities based on inflation or poverty rates. Two studies [2] and [3] applied Partitioning around medoids (PAM) method to cluster regencies/cities in Indonesia based on inflation and poverty, respectively. Whereas [4] used the k-Means (KM) method and fuzzy c-means to cluster 15 cities in Indonesia. Furthermore, [5] clustering was based on poverty rate using multivariate analysis, and [6] clustering was based on commodity inflation patterns using the ward method. These studies have inspired to involve inflation and poverty rates as two important interrelated variables, including validating the clusters that have been generated.

Table 1. Summary of related studies.

References	Year	No. of Cities	Variable	Clustering Method(s)	Cluster Validation	Clustering Results
[2]	2020	90	Inflation	k-Medoids (PAM)	-	5
[3]	2022	38	Poverty	k-Medoids (PAM)	-	2
[4]	2017	15	Inflation	k-means and fuzzy c-means	-	2
[5]	2018	37	Poverty	multivariate analysis	-	3
[6]	2021	24	Commodity Inflation Pattern	ward	-	5

Therefore, this study aims to cluster regencies/cities in Indonesia based on inflation and poverty rates, then validate the resulting cluster with *SW* and *DI* approaches. These results are expected to be a reference for the Indonesian government in determining effective economic policies and actions related to priorities for mitigating the impact of inflation and poverty during the Covid-19 pandemic. In addition, the clustering methods utilized in this study are the KM method, PAM (partition category), Ward (agglomeration hierarchy category), and Diana method (division hierarchy category) because of their convenience and representation of various categories of clustering methods. It is also an external cluster validation. The four methods are compared and measured based on compactness in clustering regencies/cities.

2. Method

2.1. Data

This study utilized data from Statistics Indonesia (BPS) in 2021. The data consisted of inflation and poverty rates in 90 regencies/cities in Indonesia [7], [8]. Then, the data were analyzed and clustered using four machine learning methods, namely KM, PAM, Ward, and Diana.

2.2. Research Workflow

The workflow of this research consists of five stages shown in Figure 1. Stage 1 is pre-processing of data using a boxplot. The next stage is to cluster regencies/cities using the KM, PAM, Ward, and Diana methods, after determining the optimal number of clusters using the thirty-two-cluster validity index [9]. Finally, the clustering results were validated using *SW* and *DI* and then interpreted.

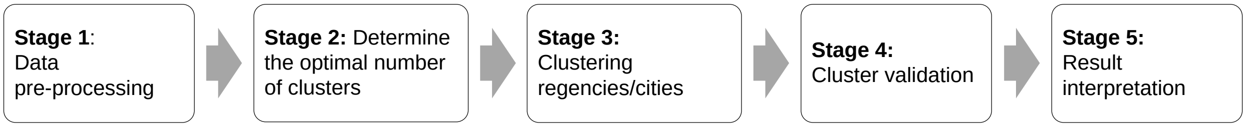

Stage 1: Data pre-processing
Stage 2: Determine the optimal number of clusters
Stage 3: Clustering regencies/cities
Stage 4: Cluster validation
Stage 5: Result interpretation

Figure 1. The five stages of the research workflow

2.3. The Optimal Number of Clusters (TONC) [10]

In general, there are many cluster validity indices (CVI) that can be utilized to determine TONC, including elbow, tracew, trcovw, marriot, scott, ccc, hartigan, ch, kl, ball, ratkowsky, beale, pseudot2, duda, db, cindex, rubin, friedman, silhouette, sdbw, dindex, sdindex, hubert, dunn, tau, gplus, gamma, mcclain, frey, gap, ptbiserial, and mixture. Each method has its advantages and limitations. This study used 32 CVI to determine TONC.

2.4. Clustering Method

2.4.1. K-means (KM) [11], [12]

KM is a clustering method based on partitioning that partitions *N* data into *k* clusters (N ≥ k). In this study, the KM method was initialized by determining the number of clusters as the cluster centre (centroid), then placing each data to join the nearest cluster using the equation (1). Next, the centroids were updated and re-clustered until the centroids did not change anymore.

$$KM = \sum_{j=1}^{k} \sum_{i=1}^{n} dist(i, s_j) \quad (1)$$

Where:

n = the number of objects in cluster *i*

dist(i, S_j) = the distance between objects *i* and centroid s_j.

2.4.2. Partitioning around medoids (PAM) [13]

In principle, PAM is similar to KM. However, PAM is considered more robust to outliers because it minimizes a sum of dissimilarities instead of a sum of the squared distances. The determination of the centroid uses the median value, not the average distance value like KM.

2.4.3. Ward [14]

Ward is a representative of the agglomerative hierarchical clustering method. In this study, Ward began with *n* clusters, each containing one data. Then the cluster pairs with minimum distance between clusters were merged sequentially to become a large cluster that contained all the data. It aimed to minimize the total variance in the cluster, which was calculated by the error sum of squares as in the equation (2).

$$Ward(c_1, c_2) = \delta^2(c_1, c_2) = \frac{n_1 n_2}{n_1 + n_2} dist(s_1, s_2) \quad (2)$$

Where:

Ward (c_1, c_2) = the merging of clusters c_1 and c_2

n_j = the number of objects in cluster *j*

s_j = the centroid of cluster *j*.

2.4.4. Divisive analysis (Diana) [15]

Diana represents a divisive hierarchical clustering method, the reverse of Agglomerative Hierarchical Clustering. In this study, Diana began with the largest cluster that contained all objects as in equation (3). Then the clusters were split sequentially until each cluster contained only one object.

$$diam(C) = \max_{c_1, c_2 \in C} dist(c_1, c_2) \quad (3)$$

Where:

diam(C) = the largest cluster *C*

dist(i, j) = the distance between objects *i* and *j*.

2.5. Cluster Validation

Essentially, a good cluster has minimal within-cluster variation and maximal between-cluster variation. Therefore, cluster validation is required to find out. Two cluster validation methods were utilized in this study.

2.5.1. The silhouette width (SW) [16]

SW is a cluster validation approach to measure how well-clustered objects are. *SW* estimates the mean of the intra-cluster distance and the closest cluster distance as in the equation (4).

$$SW(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \quad (4)$$

Where:

i = the object

a_i = the average distance between the *i*-th object to all other objects in the same cluster

b_i = is the average distance between the *i*-th object to its nearest neighbour cluster

$$a_i = \frac{1}{n_{C_i}} \sum_{j \in C_i, i \neq j} dist(i,j) \text{ and } b_i = \min_{k \neq i} \sum_{j \in C_k} \frac{dist(i,j)}{n_{C_k}} \quad (5)$$

Where:

C_i = the cluster containing data *i*,

n_C = the cardinality of cluster *C*

dist(i, j) = the distance between objects *i* and *j* (e.g. Euclidean [17]).

SW has the interval [-1, 1]. *SW* close to 1 is a good clustering result. A small *SW* (around 0) means the object is located between two clusters. Furthermore, a negative *SW* indicates that probably the object is located in the wrong cluster.

2.5.2. Dunn index (DI) [18]

DI is a cluster validation approach to calculate cluster compactness based on the comparison of the maximum intra-cluster distance to the minimum distance between data points that are not in the same cluster. It is calculated as in the equation (6).

$$DI(C) = \min_{C_k, C_i \in C, C_k \neq C_i} \left\{ \min_{i \in C_k, j \in C_i} \left\{ \frac{dist(i,j)}{\max_{C_m \in C} \Delta(C_o)} \right\} \right\} \quad (6)$$

Where $\Delta(C_o)$ is the distance between objects in cluster C_o.
DI has the interval [0, 1]. The higher *DI* indicates better clustering results.

3. Result and Discussion

Table 2 presents a descriptive statistical analysis of inflation and poverty rates in 2021 for 90 regencies/cities in Indonesia, and Figure 2 illustrates the inflation and poverty rates. The highest inflation rate was 4.62 in Sampit city (Central Kalimantan province), and the lowest inflation rate was 0.54 in Gunungsitoli city (North Sumatra province). Meanwhile, the highest poverty rate was 29.68 in Waingapu city (East Nusa Tenggara province), and the lowest poverty rate was 2.58 in Depok city (West Java province). The standard deviation shows that the data is homogeneous.

Table 2. Descriptive statistics of inflation and poverty rates in 2021 in Indonesia.

Variables	Observation	Minimum	Maximum	Mean	Std. Deviation
Inflation rate	90	0.54	4.62	2.16	0.78
Poverty rate	90	2.58	29.68	8.44	4.76

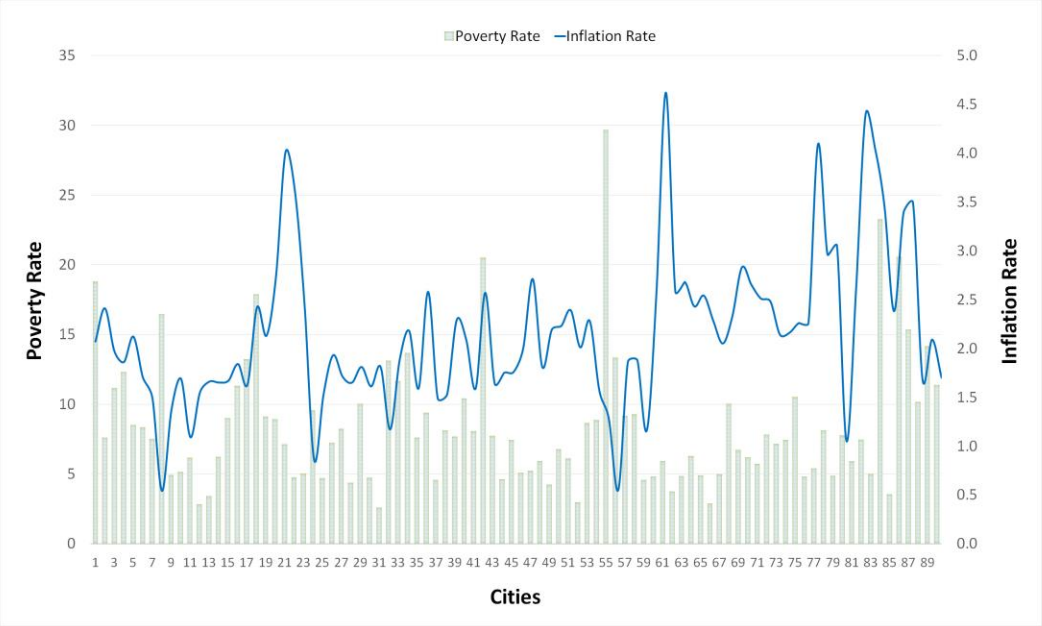


Figure 2. Inflation and poverty rates in 2021 for 90 regencies/cities in Indonesia [7], [8].

3.1. Data pre-processing

Outliers were removed from the dataset based on the boxplot chart illustrated in Figure 3 to optimize the clustering results. Five outliers in the inflation rate dataset (Tanjung Pandan, Sampit, Pare-Pare, Mamuju, and Ambon), and six outliers in the poverty rate dataset (Aceh Barat/ Meulaboh, Bengkulu, Sumenep, East Sumba/Waingapu, Tual, and Monokwari) were combined into one cluster. Therefore, they were not involved in calculating TONC.

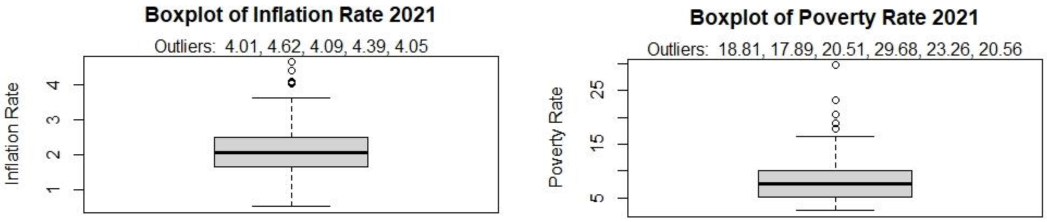


Figure 3. The boxplot of inflation and poverty rate in 2021 for 90 regencies/cities in Indonesia.

3.2. Determine TONC

TONC is determined with the assistance of R programming. Figure 4 shows that two and four are the highest number of clusters supported by 10 of the 32 CVI used (31.25%). However, the number four was selected to make it easier to interpret. Thus, TONC was 5 (including 1 cluster of outliers).

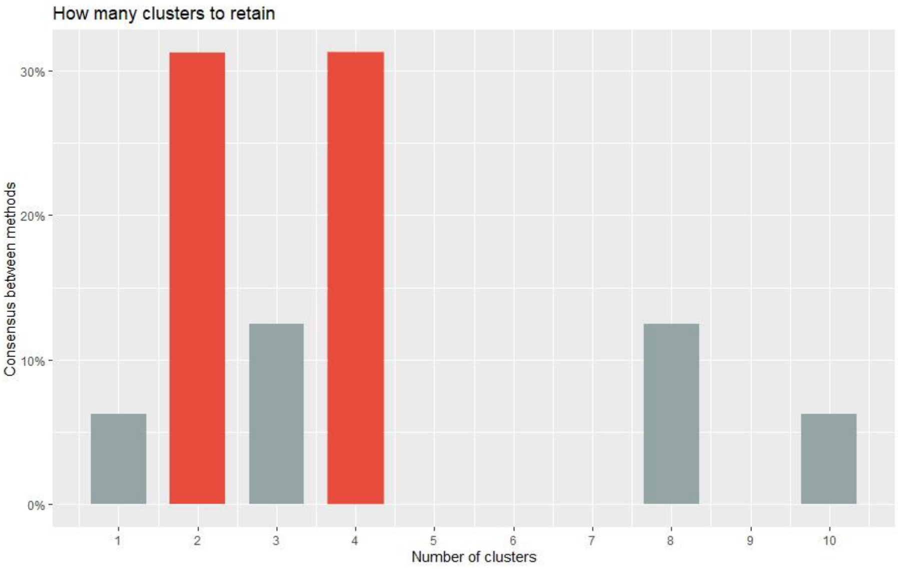


Figure 4. The highest number of clusters.

3.3. Clustering

This study used the KM, PAM, Ward, and Diana methods for clustering. The clustering results can be seen in Table 3.

Table 3. The number of regencies/cities for each cluster per method

	The number of regencies/cities			
Cluster	**KM**	**PAM**	**Ward**	**Diana**
1	27	7	7	6
2	6	26	35	43
3	7	32	28	32
4	20	19	14	8
5	30	6	6	1

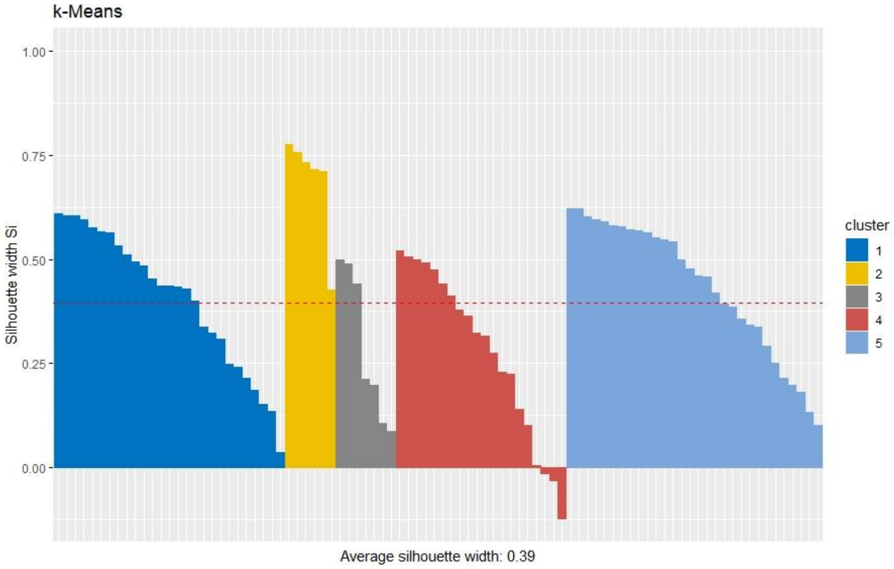
3.4. Cluster Validation

Table 4 shows that the *DI* value of Ward's method outperformed the other methods. However, KM was the best clustering method based on the *SW* value. Therefore, the further analysis was needed to obtain the most compact clustering method based on the comparison of clustering results and *SW* visualization.

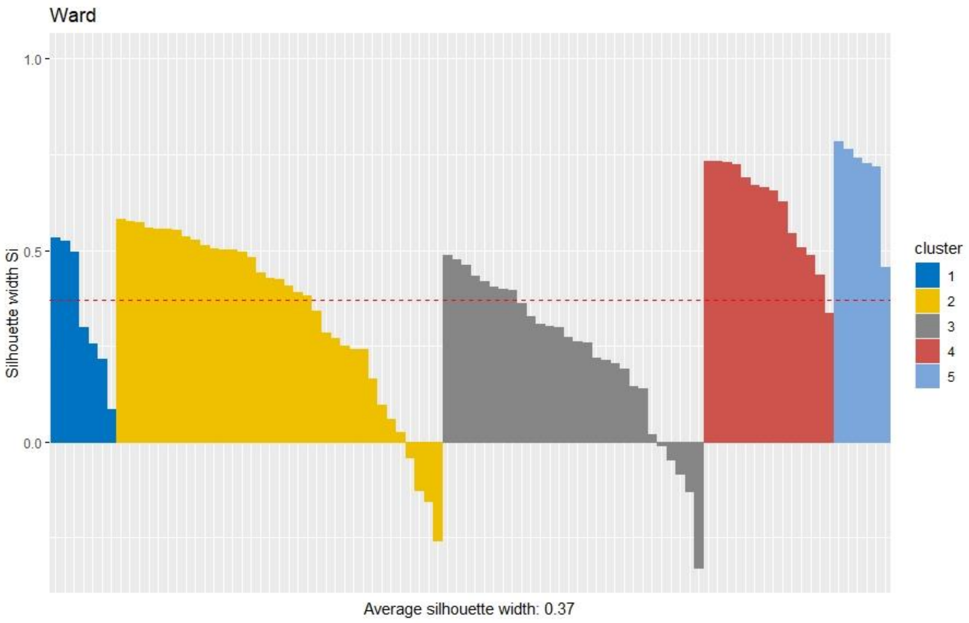
Table 4. The cluster validation using *DI* and *SW*.

Method	**KM**	**PAM**	**Ward**	**Diana**
DI	0.033028	0.031298	0.053273	0.046636
SW	0.395741	0.388256	0.371616	0.365340

The *SW* visualization in Figure 5(a) demonstrates that only three cities in cluster 4 of the KM method have negative values. It means that the cities (Tanjung Pinang, Jambi, and Bima) should probably be in another cluster. On the other hand, in Figure 5(b), the *SW* visualization from Ward's method contained nine cities with negative values, namely four cities in cluster 2 and five cities in cluster 3. Therefore, it can be concluded that KM is a more compact machine learning method for clustering based on inflation and poverty rates in this study.



(a)



(b)

Figure 5. The *SW* visualization for (a) KM and (b) Ward methods.

The clustering results using KM can be seen in Figure 6, and the details are presented in Table 5.

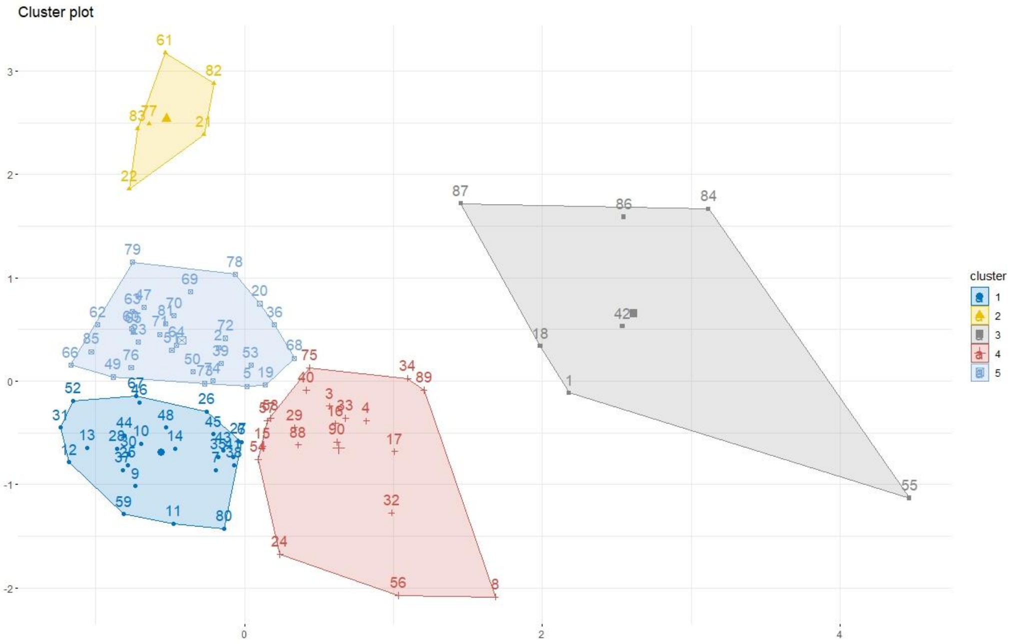


Figure 6. The clustering results using the KM method.

Table 5. Clustering regencies/cities using the KM method.

Regency/City	Cluster
(6) Medan, (7) Padangsidimpuan, (9) Padang, (10) Bukittinggi, (11) Tembilahan, (12) Pekanbaru, (13) Dumai, (14) Bungo, (25) DKI Jakarta, (26) Bogor, (27) Sukabumi, (28) Bandung, (30) Bekasi, (31) Depok, (35) Kudus, (37) Semarang, (38) Tegal, (41) Banyuwangi, (43) Kediri, (44) Malang, (45) Probolinggo, (46) Madiun, (48) Tangerang, (52) Denpasar, (59) Pontianak, (67) Samarinda, (80) Bua-Bau	1
(21) Tanjung Pandan, (22) Pangkal Pinang, (61) Sampit, (77) Pare-Pare, (82) Mamuju, (83) Ambon	2
(1) Meulaboh, (18) Bengkulu, (42) Sumenep, (55) Waingapu, (84) Tual, (86) Manokwari, (87) Sorong	3
(3) Lhokseumawe, (4) Sibolga, (8) Gunungsitoli, (15) Jambi, (16) Palembang, (17) Lubuklinggau, (24) Tanjung Pinang, (29) Cirebon, (32) Tasikmalaya, (33) Cilacap, (34) Purwokerto, (40) Jember, (54) Bima, (56) Maumere, (57) Kupang, (58) Sintang, (75) Watampone, (88) Merauke, (89) Timika, (90) Jayapura	4
(2) Banda Aceh, (5) Pematang Siantar, (19) Bandar Lampung, (20) Metro, (23) Batam, (36) Surakarta, (39) Yogyakarta, (47) Surabaya, (49) Cilegon, (50) Serang, (51) Singaraja, (53) Mataram, (60) Singkawang, (62) Palangka Raya, (63) Baru, (64) Tanjung, (65) Banjarmasin, (66) Balikpapan, (68) Tanjung Selor, (69) Tarakan, (70) Manado, (71) Kotamobagu, (72) Luwuk, (73) Palu, (74) Bulukumba, (76) Makassar, (78) Palopo, (79) Kendari, (81) Gorontalo, (85) Ternate	5

Therefore, based on the division of quadrants, cluster 2 contained six cities with the highest inflation rate, and the low poverty rate was the top priority. Cluster 3 included seven cities with a moderate inflation rate and high poverty rate being the second priority. Cluster 5 contained thirty regencies/cities with moderate inflation rates, and a low poverty rate was the third priority. Furthermore, cluster 4 was twenty regencies/cities with a low inflation rate and moderate poverty rate being the fourth priority. Finally, cluster 1 included twenty-seven regencies/cities with low inflation and poverty rates were the last priority. These results can be used as a reference for the Indonesian government in determining regency/city priorities to mitigate the impact of inflation and poverty during the Covid-19 pandemic, such as the distribution of social safety nets, subsidies, or other stimulus packages based on regency/city priorities.

4. Conclusion

This study discussed the application and comparison of four machine learning methods in regency/city clustering in Indonesia based on inflation and poverty rates. The study results indicated that the k-Means method was the most compact method for grouping 90 regencies/cities in Indonesia based on inflation and poverty rates in 2021. The five clusters produced in this study consisting of six cities were the top priority. Seven cities were the second priority. Thirty regencies/cities were the third priority. Twenty regencies/cities were the fourth priority, and twenty-seven regencies/cities were the last priority. This study results can be used as a reference for the government to determine priority actions and economic policies related to mitigating the impact of inflation and poverty during the Covid-19 pandemic.

The authors plan to use other methods such as deep learning and neural networks and their derivatives to optimize clustering results. The implementation in other cases will also be done as future work.

References

[1] Statistics-Indonesia, “Consumer Prices Indices,” *Statistics Indonesia*, 2022. [Online]. Available: https://www.bps.go.id/subject/3/inflasi.html. [Accessed: 25-Feb-2022].

[2] M. A. Hanafiah, “K-Medoids: Inflation Clustering of 90 Cities in Indonesia,” *Int. J. Inf. Syst. Technol.*, vol. 4, no. 1, pp. 1–9, 2020.

[3] F. Alfiah, A. Almadayani, D. Al Farizi, and E. Widodo, “Analisis Clustering K-Medoids Berdasarkan Indikator Kemiskinan di Jawa Timur Tahun 2020,” *J. Ilm. Sains*, vol. 22, no. 1, pp. 1–7, Dec. 2022.

[4] A. Setiawan, B. Susanto, and T. Mahatma, “Inflation data clustering of some cities in Indonesia,” *J. Phys. Conf. Ser.*, vol. 855, pp. 1–9, Jun. 2017.

[5] D. V. Ferezagia, “Analisis Tingkat Kemiskinan di Indonesia,” *J. Sos. Hum. Terap.*, vol. 1, no. 1, pp. 1–6, 2018.

[6] T. Hendrawati, A. H. Wigena, I. M. Sumertajaya, and B. Sartono, “Clustering of Commodity Inflation Pattern based on Estimated ARIMA Model,” *J. Phys. Conf. Ser.*, vol. 1863, no. 1, pp. 1–10, Mar. 2021.

[7] Statistics-Indonesia, “Inflation of 90 City (General) 2021,” *Statistics Indonesia*, 2022. [Online]. Available: https://www.bps.go.id/indicator/3/1708/2/inflasi-90-kota-umum-.html. [Accessed: 25-Feb-2022].

[8] Statistics-Indonesia, “Percentage of Poor Population (P0) by Regency/City,” *Statistics Indonesia*, 2022. [Online]. Available: https://www.bps.go.id/indicator/23/621/1/persentase-penduduk-miskin-menurut-kabupaten-kota.html. [Accessed: 25-Feb-2022].

[9] R. Gustriansyah, N. Suhandi, and F. Antony, “Clustering optimization in RFM analysis based on k-means,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 18, no. 1, pp. 470–477, 2020.

[10] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, “NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set,” *J. Stat. Softw.*, vol. 61, no. 6, pp. 1–36, 2014.

[11] J. A. M. Nugraha, “Medicine Inventory Grouping using Clustering Data Mining,” *Indones. J. Inf. Syst.*, vol. 2, no. 1, pp. 33–44, Aug. 2019.

[12] R. Gustriansyah, Ermatita, D. P. Rini, and R. F. Malik, “Integration of Decision-Making Method and Data-Mining Method as A Preliminary Study of Novel Sales Forecasting Method,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 5730–5735, Aug. 2020.

[13] E. Schubert and P. J. Rousseeuw, “Fast and Eager k-Medoids Clustering: O(k) Runtime Improvement of the PAM, CLARA, and CLARANS Algorithms,” *Inf. Syst.*, vol. 101, pp. 1–19, Aug. 2021.

[14] Y. Ogasawara and M. Kon, “Two clustering methods based on the Ward’s method and dendrograms with interval-valued dissimilarities for interval-valued data,” *Int. J. Approx. Reason.*, vol. 129, pp. 103–121, Feb. 2021.

[15] L. Bellanger, A. Coulon, and P. Husi, “Determination of cultural areas based on medieval pottery using an original divisive hierarchical clustering method with geographical constraint (MapClust),” *J. Archaeol. Sci.*, vol. 132, pp. 1–18, Aug. 2021.

[16] F. Batool and C. Hennig, “Clustering with the Average Silhouette Width,” *Comput. Stat. Data Anal.*, vol. 158, pp. 1–18, Jun. 2021.

[17] S. Cahyani, R. Wiryasaputra, and R. Gustriansyah, “Identifikasi Huruf Kapital Tulisan Tangan Menggunakan Linear Discriminant Analysis dan Euclidean Distance,” *J. Sist. Inf. Bisnis*, vol. 8, no. 1, pp. 57–67, Apr. 2018.

[18] S. Liang, D. Han, and Y. Yang, “Cluster validity index for irregular clustering results,” *Appl. Soft Comput.*, vol. 95, pp. 1–17, Oct. 2020.