Published online on the journal's webpage: **http://journal.ittelkom-pwt.ac.id/index.php/dinda**

Journal of Dinda

Data Science, Information Technology, and Data Analytics

Vol. 4 No. 1 (2024) 23 - 27 E-ISSN: 2809-8064

K-Means Clustering Algorithm: A Study on Unemployment Rates in Districts/Cities in Three Highest Provinces

Mohammad Dian Purnama[1*], Mutia Eva Mustafidah[2]

[1*]Department of Mathematics, Faculty of Mathematics and Natural Sciences, State University of Surabaya

[2]Department of Mathematics, Faculty of Mathematics and Natural Sciences, State University of Surabaya

[1*]mohammaddian.20053@mhs.unesa.ac.id, [2]mutia.20010@mhs.unesa.ac.id

Abstract

Unemployment is a recurring issue every year, particularly in provinces with high unemployment rates, posing economic and social challenges. West Java, Riau Islands, and Banten are identified as the three provinces with the highest unemployment rates, exceeding 8% in the year 2022. Hence, this study aims to delve into the unemployment scenario in these provinces, considering various influencing factors drawn from relevant previous research. The primary objective of this research is to obtain the classification results of regencies/cities in West Java, Riau Islands, and Banten based on unemployment indicators. The findings reveal four clusters: Cluster 1 comprises 13 regencies/cities with the lowest unemployment rates, Cluster 2 includes 4 regencies/cities with low unemployment rates, Cluster 3 consists of 13 regencies/cities with moderate unemployment rates, and Cluster 4 encompasses 12 regencies/cities with high unemployment rates.

Keywords: K-means, Clustering, Unemployment.

1. Introduction

As a developing country, Indonesia faces challenges related to unemployment, which is also a common issue in other developing nations. According to the employment indicators from the Central Statistics Agency (Badan Pusat Statistik or BPS), unemployment refers to individuals actively seeking employment, preparing for new ventures, or already accepted for work but not yet commenced [1]. Unemployment reflects the condition of individuals in the workforce who have not secured employment and are actively searching [1].

Based on BPS data for the year 2022, the unemployment rate in Indonesia is 5.86%, showing a decrease compared to 7.07% and 6.49% in 2020 and 2021, respectively. This decline can be attributed to the significant impact of the COVID-19 pandemic in increasing unemployment during those periods. However, when compared to the year before the COVID-19 pandemic, i.e., 2019, the unemployment rate was 5.23%, lower than that of 2022. This study will focus on the unemployment rates in three provinces with the highest unemployment rates exceeding 8%, namely West Java (8.31%), Riau Islands (8.23%), and Banten (8.09%). The high unemployment rates in these provinces, surpassing the national average, have the potential to create various economic and social problems. Therefore, an in-depth investigation is needed to explore the factors influencing changes in unemployment rates, particularly using clustering methods.

A commonly used clustering method in research is clustering analysis, with one of its techniques being the K-means algorithm. Clustering analysis is a grouping technique used in data mining [2]. K-means clustering analysis is frequently encountered in research due to its ability to produce optimal clusters with fast convergence [2].

Based on the above description, this study aims to cluster the Districts/Cities in West Java, Riau Islands, and Banten using K-means clustering analysis, considering the indicators that influence the unemployment rate. This study refers to findings from previous research highlighting factors such as population density, GDP, education level, health level, and minimum wage at the District/City level affecting the unemployment rate in Jambi [3]. The second study involves clustering the unemployment rate in Maluku using indicators such as the average population, average TPAK (labor force participation rate), labor force size, average population aged 15 and above, and TPT (open unemployment rate) [4]. The third study investigates the factors affecting unemployment, focusing on minimum wage at the District/City level and the Human Development Index

Received: 13-02-2024 | Revised: 22-02-2024 | Published: 22-02-2024

[5]. The fourth study explores factors such as economic growth rate, average years of schooling, and life expectancy affecting unemployment. The fifth study examines factors influencing unemployment including the poverty rate, Human Development Index, and minimum wage at the District/City level [6]. The sixth study investigates factors affecting unemployment such as population growth rate and minimum wage at the District/City level [7].

Prior to employing clustering in this study, Principal Component Analysis (PCA) was conducted on the variables based on factors influencing unemployment in previous research. During PCA, 11 factors were considered, including TPAK, economic growth rate, average years of schooling, life expectancy, Human Development Index, population density, minimum wage at the District/City level, population growth rate, poverty rate, GDP, and labor force size. After conducting PCA, six influential factors were identified for use in this study, namely economic growth rate, Human Development Index, population density, minimum wage at the District/City level, and poverty rate. The objective of this research is to cluster regions in three provinces, namely West Java, Riau Islands, and Banten, using open unemployment rate indicators while examining the influencing factors.

- **Research Methods**

This study utilizes secondary data obtained from the Central Statistics Agency (Badan Pusat Statistik or BPS) of West Java, Riau Islands, and Banten. The data includes the economic growth rate in percentage, average years of schooling in percentage, Human Development Index (HDI) in percentage, population density in number of people, minimum wage at the district/city level (UMK) in Indonesian Rupiah, and the percentage of the population living in poverty. All this data is sourced from the year 2022. These six indicators serve as variables in the clustering process for 42 districts/cities across three provinces, namely West Java, Riau Islands, and Banten.

2.1. Clustering

Clustering is a technique within the functionality of data mining, where clustering algorithms are responsible for grouping a set of data into specific clusters [2]. In the clustering process, determining or describing the quantitative values of the similarity or dissimilarity of data (proximity measure) plays a crucial role. Therefore, it is necessary to compare several commonly used methods, such as Euclidean, Manhattan, and Minkowski distances [5]. The determination of the number of clusters will use the Elbow method, which aims to interpret and test the consistency of the optimal number of clusters by examining the Sum of Squared Errors (SSE) value [8]. At a certain point, the graph will show a significant decrease with a bend referred to as the elbow criterion. This value is then considered the optimal number of clusters [8].

2.2. K-Means

K-means is an algorithmic technique used to group items or research subjects into K clusters by minimizing the sum of square distances with each cluster. This algorithm utilizes the Euclidean distance measure and iteratively assigns each record to the original cluster. The steps begin by selecting K with initial records as cluster centers (initial seed) and determining each record closest to the cluster. Each new record is then added to the cluster, and the cluster center is recalculated to reflect the new membership. This iterative process is repeated until convergence is achieved, and the migration of records into the cluster no longer produces significant changes in the solution [9]. As a heuristic algorithm, K-means separates a dataset into K clusters by minimizing the sum of squared distances within each cluster [8]. The main principle of this technique is to form K partitions or centroids, which represent the average of the dataset. The K-means algorithm begins with the formation of cluster partitions initially and iteratively refines these partitions until no significant changes occur in the cluster partitions [2]. One advantage of K-means lies in its ease of use, as the number of clusters to be formed is predetermined before conducting the analysis. In general, the K-means clustering analysis method uses the following algorithm [10].

- Selecting k as the number of clusters to be formed.
- Choosing k initial centroids (cluster center points) randomly.
- Calculating the distance between each object and each centroid in every cluster. The method used to measure the distance between variables and centroids employs the Euclidean Distance.
- Assigning members to each cluster based on the results of the nearest distance calculations.

$$d(u,v) = \sqrt{\sum_{i=0}^{p}(u_i - v_i)^2} \quad (1)$$

where:
u_i : Object at position i used in clustering.
v_i : Centroid at position i within the cluster.
p : Total number of objects.

- Determining new centroid locations (C_j) by calculating the average value of the data within the same centroid.

$$C_j = \left(\frac{1}{n_j}\right)\sum d_i \quad (2)$$

where:
n_j : number of members in the k cluster

d_i : members in k cluster

- Iterating again from step 3 to step 5 until there are no changes in the members of each formed cluster.

3. Results and Discussion

In the stage of determining the number of clusters formed, the Elbow method is employed. The results obtained from this method are derived from a graph illustrating the number of groupings based on the unemployment rate and its influencing factors. The factors or variables used encompass economic growth rate, average years of schooling, Human Development Index, population density, minimum wage at the District/City level, and the percentage of the population living in poverty. These variables are grouped into four clusters because in the fourth cluster, there is a noticeable decline in the graph, and it begins to flatten out with a bend, known as the elbow criterion. The four clusters are labeled as Cluster 1, Cluster 2, Cluster 3, and Cluster 4. The results of the Elbow method are depicted in Figure 1.

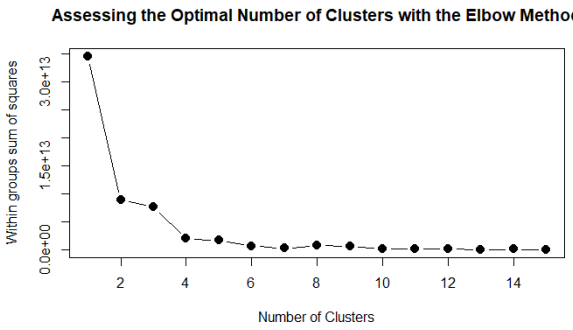

Figure 1. Elbow Method Results

Subsequently, using K-means clustering analysis, the following results are obtained:

- The randomly selected initial centroid points are C1 (5.2, 8, 71.2, 1861, 4217206, 7.73), C2 (5.24, 7.68, 71.56, 1036, 4173568, 8.7), C3 (4.38, 8.48, 72.65, 2448, 3348765, 6.87), and C4 (3.86, 6.59, 64.71, 433, 2773590, 9.24).
- After determining the initial centroids, clusters are defined using the Euclidean Distance formula to calculate the distance of each data point to the nearest centroid. The membership of each cluster is then determined. After the first iteration, the results are as follows: Cluster 1 with 11 Districts/Cities, Cluster 2 with 4 Districts/Cities, Cluster 3 with 10 Districts/Cities, and Cluster 4 with 17 Districts/Cities.
- From the results of the first iteration, new centroid points are obtained: C1 (5.41, 9.71, 75.92, 5857, 4426325, 5.24), C2 (5.55, 9.6, 77.17, 4937, 3996328, 6.05), C3 (3.99, 8.31, 72.19, 1983, 3285922, 7.61), and C4 (4.63, 7.87, 70.14, 1919, 2359714, 10.5).
- Perform Euclidean Distance calculations with the new centroids until the same pattern is achieved as in the previous iterations. The process stops at the fifth iteration, and the results from each iteration are shown in Table 1.

Table 1. Iteration Result

Iterations	C1	C2	C3	C4	Total
1	11	4	10	17	42
2	11	5	10	16	42
3	13	3	13	13	42
4	13	4	13	12	42
5	13	4	13	12	42

In this research, the iteration process halted at the 5th iteration, as explained in the K-means process, where an iteration is considered stable if the new iteration results are identical to the previous iteration. Therefore, the iteration process is deemed stable since the results of the 4th and 5th iterations are the same. In the 5th iteration, the clustering is obtained as follows: Cluster 1 (Bogor, Purwakarta, Karawang, Bekasi, Kota Bogor, Kota Bekasi, Kota Depok, Batam, Tangerang, Serang, Kota Tangerang, Kota Cilegon, Kota Tangerang Selatan), Cluster 2 (Kota Bandung, Kota Serang, Bintan, Kepulauan Anambas), Cluster 3 (Sukabumi, Bandung, Cianjur, Sumedang, Subang, Bandung Barat, Kota Cimahi, Karimun, Natuna, Lingga, Tanjung Pinang, Pandeglang, Lebak), Cluster 4 (Garut, Tasikmalaya, Ciamis, Kuningan, Cirebon, Majalengka, Indramayu, Pangandaran, Kota Sukabumi, Kota Cirebon, Kota Tasikmalaya, Kota Banjar). The clustering results using K-means with 4 clusters are depicted in Figure 2.

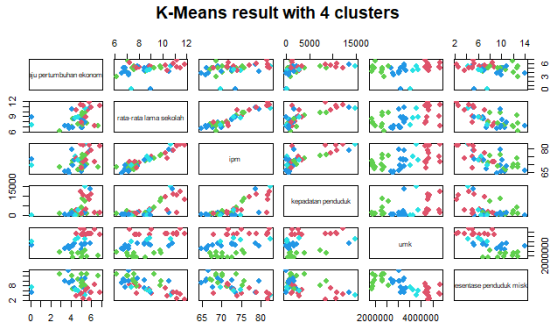

Figure 2. K-means Clustering Result with 4 Cluster

After analyzing the number of clusters formed, the results indicate that the distribution of open unemployment rates in the provinces of West Java, Riau Islands, and Banten is divided into 4 groups. This implies that there are four clusters of regions in the provinces of West Java, Riau Islands, and Banten with

different unemployment rates. To determine the characteristics of each cluster, the average variables for each cluster can be determined from the 5th iteration, as shown in Table 2.

Table 2. Cluster Average (Centroid) of Each Cluster

Iterations	C1	C2	C3	C4
Economic Growth Rate	5,5	3,66	4,28	4,88
Average Length of Schooling	9,66	8,91	8,08	7,98
Human Development Index (HDI)	76,03	75,02	70,77	70,81
Population Density	5130,38	4423	1759,76	2448,25
Minimum Wage of the District/City (UMK)	4388424	3698087	3082288	2144804
Poverty Rate	5,5	6,06	8,9	10,41
Average of All Variables	732275,105	617100,651	514023,301	357891,098

Based on the average values of variables for each cluster, the characteristics of the four clusters can be identified, and the clusters can be interpreted as follows.

- Cluster 1 consists of 13 districts/cities: Bogor, Purwakarta, Karawang, Bekasi, Bogor City, Bekasi City, Depok City, Batam, Tangerang, Serang, Tangerang City, Cilegon City, South Tangerang City. Cluster 1 has the highest average population growth rate, average length of schooling, Human Development Index (HDI), population density, and Minimum Wage of the district/city (UMK) compared to the other clusters. Additionally, it has the lowest rate of poverty among the clusters. Therefore, Cluster 1 can be categorized as the cluster with the lowest unemployment rate, with an average of all variables amounting to 732,275.105.
- Cluster 2 consists of 4 districts/cities: Bandung City, Serang City, Bintan, Anambas Islands. Cluster 2 has the lowest average economic growth rate compared to the other clusters, but it has higher averages for length of schooling, HDI, population density, and UMK compared to Clusters 3 and 4. Additionally, it has lower poverty rates than Clusters 3 and 4. Therefore, Cluster 2 can be categorized as a cluster with a low unemployment rate, with an average of all variables amounting to 617,100.651.
- Cluster 3 consists of 13 districts/cities: Sukabumi, Bandung, Cianjur, Sumedang, Subang, West Bandung, Cimahi City, Karimun, Natuna, Lingga, Tanjung Pinang, Pandeglang, Lebak. Cluster 3 has an average economic growth rate larger than Cluster 2 and smaller than Cluster 4. The average length of schooling and UMK is higher than Cluster 4, while poverty rates and population density are lower than Cluster 4. Therefore, Cluster 3 can be categorized as a cluster with a moderate unemployment rate, with an average of all variables amounting to 514,023.301.
- Cluster 4 consists of 12 districts/cities: Garut, Tasikmalaya, Ciamis, Kuningan, Cirebon, Majalengka, Indramayu, Pangandaran, Sukabumi City, Cirebon City, Tasikmalaya City, Banjar City. Cluster 4 has the second-highest average population growth rate after Cluster 1, and HDI is higher than Cluster 3. However, the average length of schooling and UMK are the lowest compared to other clusters. It has the highest poverty rates and higher population density than Cluster 3. Therefore, Cluster 4 can be categorized as a cluster with a high unemployment rate, with an average of all variables amounting to 357,891.098.

4. Conclusion

Based on the results and discussions presented, it can be concluded that there are 4 groups or clusters obtained from the conducted analysis. The members of each cluster were determined in the 5th iteration. The clusters are as follows:

- Cluster 1 consists of 13 districts/cities and is categorized as an area with the lowest unemployment rate.
- Cluster 2 consists of 4 districts/cities and is categorized as an area with a low unemployment rate.
- Cluster 3 consists of 13 districts/cities and is categorized as an area with a moderate unemployment rate.
- Cluster 4 consists of 12 districts/cities and is categorized as an area with a high unemployment rate. In the conclusion there should be no reference. Conclusions contain the facts obtained, simply answering the problem or purpose of the study (do not constitute any more discussion); State the possibilities of application, implications, and speculation accordingly. If needed, provide advice for future research.

References

[1] Sukirno, Sadono. 2008. *Makro Ekonomi Teori Pengantar*. Jakarta : PT. Raja Grafindo Persada.

[2] Sibuea, M. L., & Safta, A., 2017. Pemetaan Siswa Berprestasi Menggunakan Metode K-means Clustring, *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, 4(1), 85-92.

[3] Harlik, H., Amir, A., & Hardiani, H., 2013. Faktor-faktor yang mempengaruhi kemiskinan dan pengangguran di Kota Jambi, *Jurnal Perspektif Pembiayaan dan Pembangunan Daerah*, 1(2), 109-120.

[4] Rahakbauw, D. L., Sinay, L. J., & Enus, V., 2017. Aplikasi Metode Fuzzy C-Means Untuk Menentukan Tingkat Pengangguran, *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, 11(2), 95-100.

[5] Nishom, M., 2019. Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-means Clustering berbasis Chi-Square, *Jurnal Informatika*, 4(01), 20-24.

[6] Prastikasari, A. C., 2022. Analisis Pengaruh Indeks Pembangunan Manusia, Tingkat Kesempatan Kerja, Jumlah Penduduk Miskin, Dan Upah Minimum Terhadap Tingkat Pengangguran Terbuka Di Eks Karesidenan Semarang Periode 2017-2021, *PARETO: Jurnal Ekonomi dan Kebijakan Publik*, 5(2), 123-132.

[7] Syam, S., 2015. Pengaruh upah dan Pertumbuhan Penduduk terhadap tingkat pengangguran di kota Makassar, *Jurnal Iqtisaduna*, 1(1), pp.30-45.

[8] Savitri, A. D., Bachtiar, F. A., & Setyawan, N. Y, 2018. Segmentasi Pelanggan Menggunakan Metode K-means Clustering Berdasarkan Model RFM Pada Klinik Kecantikan (Studi Kasus: Belle Crown Malang), Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, 2(9), 2957-2966.

[9] Rumiarti, C. D., & Budi, I., 2017. Segmentasi pelanggan pada customer relationship management di perusahaan ritel: studi kasus PT Gramedia Asri Media, *Jurnal Sistem Informasi*, 13(1), 1-10.

[10] Fajriani, F., 2020. K-means clustering analysis pada Persebaran Tingkat pengangguran kabupaten/kota di Sulawesi selatan, *Jurnal Varian*, 3(2), pp.103-112.

[11] Badan Pusat Statistik (BPS) Provinsi Banten, 2022. *Banten Dalam Angka 2022*. Banten, 25 February 2022, Badan Pusat Statistik (BPS) Provinsi Banten: Banten.

[12] Badan Pusat Statistik (BPS) Provinsi Jawa Barat, 2022. *Jawa Barat Dalam Angka 2022*. Jawa Barat, 25 February 2022, Badan Pusat Statistik (BPS) Provinsi Jawa Barat: West Java.

[13] Badan Pusat Statistik (BPS) Provinsi Kepulauan Riau, 2022. *Kepulauan Riau Dalam Angka 2022*. Kepulauan Riau, 25 February 2022, Badan Pusat Statistik (BPS) Provinsi Kepulauan Riau: Riau Island.