Implementation of K-Means Cluster for Districts or Cities in West Java Province Based on Unemployment Indicators

Alifa Nur Oktaviani [(1)], Atika Nurani Ambarwati [(2)]

Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang.

Jl. Prof. Dr. Hamka Km 01 No. 17 Tambakaji Ngaliyan
e-mail: alifaoktaviani28@gmail.com[(1)], atika.nurani@gmail.com[(2)]

ABSTRAK

Tingginya disparitas tingkat pengangguran terbuka antarwilayah di Provinsi Jawa Barat menjadi isu yang memerlukan pemetaan berdasarkan karakteristik sosial ekonomi daerah. Tujuan dari studi ini adalah untuk mengelompokkan kabupaten/kota di Jawa Barat berdasarkan tingkat pengangguran terbuka dan faktor-faktor yang memengaruhinya dengan menggunakan metode K-Means Cluster. Data yang digunakan merupakan data sekunder tahun 2023 yang diperoleh dari Badan Pusat Statistik. Uji asumsi dilakukan menggunakan KMO dan Bartlett's Test untuk memastikan kecukupan sampel dan kelayakan struktur data. Hasil analisis menunjukkan bahwa wilayah di Jawa Barat terbagi menjadi tiga cluster. Cluster 1 terdiri dari daerah dengan pembangunan tinggi namun tingkat pengangguran juga tinggi, seperti Kota Bandung dan Kabupaten Bekasi. Cluster 2 mencakup wilayah dengan kondisi sosial ekonomi menengah, seperti Kota Depok dan Kabupaten Bandung. Cluster 3 terdiri dari wilayah dengan tingkat pembangunan rendah, IPM rendah, dan kemiskinan tinggi, seperti Ciamis, Garut, dan Pangandaran. Temuan ini menunjukkan adanya ketimpangan yang signifikan antarwilayah dan dapat menjadi dasar bagi penyusunan kebijakan penanggulangan pengangguran yang lebih terarah dan berbasis wilayah.
Kata kunci: *K-Means*; Pengangguran Terbuka; Jawa Barat

ABSTRACT

The high disparity of open unemployment rate among regions in West Java Province is an issue that requires mapping based on regional socioeconomic characteristics. The purpose of this study is to group districts/cities in West Java based on the open unemployment rate and its influencing factors using the K-Means Cluster method. The data used is secondary data for the year 2023 obtained from the Central Bureau of Statistics. Assumption test was conducted using KMO and Bartlett's Test to ensure sample adequacy and feasibility of data structure. The results of the analysis show that the regions in West Java are divided into three clusters. Cluster 1 consists of regions with high development but also high unemployment rates, such as Bandung City and Bekasi Regency. Cluster 2 includes regions with medium socioeconomic conditions, such as Depok City and Bandung Regency. Cluster 3 consists of regions with low development, low HDI, and high poverty, such as Ciamis, Garut, and Pangandaran. These findings indicate the existence of significant inequality among regions and can serve as a basis for the formulation of more targeted and region-based unemployment reduction policies.
Keywords: *K-Means; Open Unemployment; West Java*

INTRODUCTION

Indonesia is among the nations that possess a significant labor market potential, which could serve as a catalyst for economic growth. Nevertheless, the challenges related to population and workforce in developing nations pose a hindrance to national development, as elevated unemployment levels impact the success of economic advancement in Indonesia [1]. Based on information from the Badan Pusat Statistika (2011), open unemployment refers to those in the workforce who are either without a job or actively seeking employment. This category comprises individuals who are searching for jobs, starting a business, those who have given up on finding work due to a belief that job opportunities are nonexistent, and people who have secured employment but have yet to begin their roles (unemployed). Consequently, the open unemployment rate serves as a measure that reflects the count of individuals who are not employed [2]. Unemployment is not only a social problem but also an economic problem, because unemployment can cause problems in the form of changes in the economic growth of developing countries such as Indonesia [3]. According to the BPS report from 2023, Indonesia's open unemployment rate stands at 5. 32%, translating to 7. 86 million individuals. While this percentage might seem low, it still warrants concern due to its significant impact on various facets of life in Indonesia, including social and economic dimensions. The root cause of unemployment can be traced to wage structures that do not align with the balance between labor demand and supply, leading to a decline in the mobility of the workforce [4].

Most of the population classified as open unemployment comes from West Java Province. In addition, the unemployment rate in West Java Province has not met the regional target stated in 2019-2023 Regional Medium-Term Development Plan. The success of a region's economic development can be hampered by a high unemployment rate, because unemployment is closely related to other economic variables, so it is a very important parameter [2].

In this research, we aim to group regencies or cities within West Java Province by utilizing the unemployment rate as a key indicator through K-Means method. A previous study by Nurul Nurjanah et al., entitled "Implementation of K-Means Clustering to Group Unemployment Rates," which resulted in the formation of three clusters, it was proven that the K-Means method can be a useful tool for identifying trends in data related to unemployment issues [5]. K-Means is a straightforward algorithm for grouping data that operates in an unsupervised manner, meaning it does not require labeled inputs [6]. We anticipate that this approach will lead to more effective outcomes in decreasing the unemployment figures in West Java. The K-Means algorithm is widely recognized for its utility in clustering analysis in studies, primarily due to its ability to deliver optimal clusters with rapid convergence [7]. Clustering is a method of analyzing data aimed at categorizing or forecasting the value of a target variable. It seeks to partition the dataset in a way that items within the same cluster are more similar to one another than to those in separate clusters [8], [9]. In overcoming the problem of unemployment, local governments can determine districts or cities that require special attention through grouping or clustering so as to create appropriate program adjustments in the future [10].

METHOD

In this research, the approach used K-Means Clustering. The information utilized in this research consists of secondary data related to unemployment metrics in the West Java Province. This data has been sourced from the BPS publication for West Java from the year 2023.

Research Variabels

The factors examined in this research consist of predictor variables (X). These predictor variables represent elements that may influence Open Unemployment in Central Java. Here are the variables identified:

Table 1. Research Variables

Variables	Descriptions
X1	Open Unemployment Rate
X2	Human Development Index
X3	Percentage of Poor Population
X4	GRDP
X5	Labor Force Participation Rate
X6	Minimum Wage

Data Processing

The process for categorizing regencies or cities in West Java Province according to the Unemployment Indicator includes these actions:

1. Descriptive analysis related to unemployment in West Java Province in 2023
2. Perform Assumption Test for cluster analysis, specifically:
   a. Evaluation of sample representativeness

   The Kaiser-Meyer-Olkin (KMO) test is commonly employed to determine whether a sample accurately reflects the population.This test shows how suitable the sample is, with values ranging from 0 to 1. A KMO score between 0. 5 and 1 means the sample effectively represents the overall population [11].

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2} \quad (1)$$

Description:

r_{ij} = correlation coefficient between variable i and variable j

a_{ij} = partial correlation coefficient between variable i and variable j

p = total number of variables.

   b. Multicollinearity Examination

   Multicollinearity shows a strong connection among independent variables. In cluster analysis, it's important that variables are free from multicollinearity. To find multicollinearity, one can look at the correlation matrix. If the correlation coefficient between the variables is between 0. 8 and 1. 0, it indicates multicollinearity exists [12]

$$r_{(x,r)} = \frac{n(\sum_{i=1}^{n} X_1 Y_i) - (\sum_{i=1}^{n} Y_i)}{\sqrt{n(\sum_{i=1}^{n} X_i^2) - (\sum_{i=1}^{n} X_i)^2}} \quad (2)$$

Description:

r = the correlation coefficient between variables X and Y

n = number of research observations

3. Determination of k Value through the Elbow Method

The Elbow method is used to find the best clustering results by looking at the percentage of cluster members that form a right angle at a certain point. This point shows the best SSE value. The text includes the formula to calculate the SSE value [13]:

$$SSE = \sum_{i=1}^{n} \sum_{x_i \epsilon S_k} ||x_i - C_k||^2 \quad (3)$$

Description:

n : cluster

x_i : i-th data

S_k : k-th cluster element

C_k : average center in the k-th cluster/centroid

4. Clustering using K-Means Cluster

K-Means is a method for grouping data based on distance, using only numerical traits [14]. Its goal is to sort data into clusters based on a chosen k value, which decides how many groups there will be. Data points with similar features are placed in the same cluster. The main steps to follow when using the K-Means algorithm include:

a. Determining the cluster centroid value

b. Determining the Euclidean distance for every object in relation to the central point [15]. In calculating the Euclidean distance, the following equation can be used:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (4)$$

Descpription:

$d(x,y)$ = Euclidean distance

x = cluster center data

y = data on the attribute

x_i = data at the cluster center to i

y_i = data on every i-th

c. Classify objects based on their proximity to the center.

5. Interpretation of clustering results and cluster mapping

6. Make conclusions and recommendations

RESULT AND DISCUSSION

Descriptive Analytics

Before conducting a cluster analysis, it is necessary to examine the state of unemployment in West Java in 2023 to understand its characteristics.

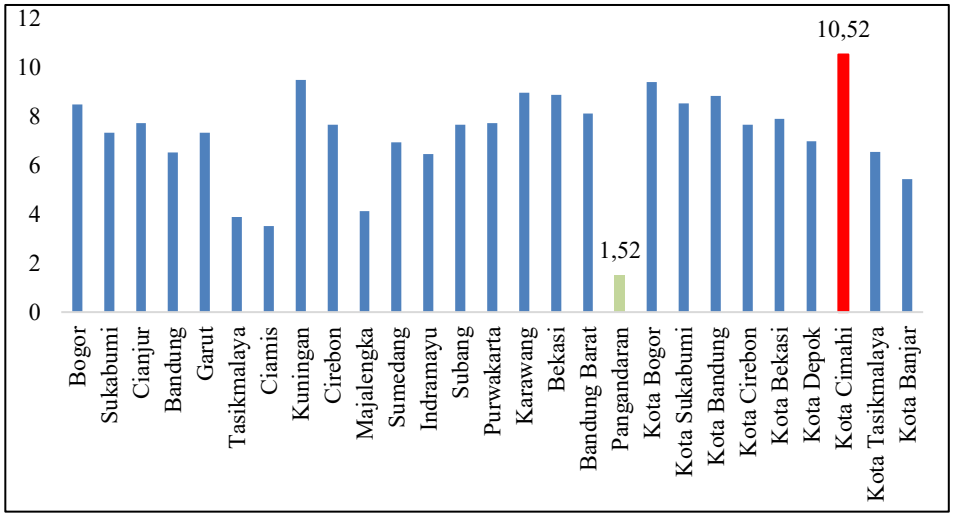

Figure 1. Open Unemployment Rate in West Java Province, 2023

The graph of the open unemployment rate in West Java Province in 2023 shows clear variations between districts and cities. Bekasi City recorded the highest open unemployment rate, possibly due to the high rate of urbanization and the mismatch between the available workforce and vacancies. Meanwhile, Pangandaran Regency shows the lowest unemployment rate, reflecting more stable labor market conditions. Other regions are at an intermediate level, illustrating the imbalance in labor distribution. This suggests the need for more focused employment policies in accordance with the character of each region, particularly to increase employment opportunities in urban areas.

K-Means Cluster Analysis Assumption

1. The sample represent the population

Table 2. KMO Value

KMO and Bartlett's Test	
Kaiser-Meyer-Olkin	0.701

Based on the table, the KMO value is greater than 0.5 and close to 1. Therefore, it can be concluded that the sample represents the population and can be analyzed using cluster analysis.

2. Multicollinearity test

Table 3. Matrix Correlation

Variabel	X1	X2	X3	X4	X5	X6
X1	1	0.408	-0.25	0.365	-0.553	0.547
X2	0.408	1	-0.774	0.29	-0.383	0.55
X3	-0.25	-0.774	1	-0.433	0.177	-0.634
X4	0.365	0.29	-0.433	1	-0.283	0.662

X5	-0.553	-0.383	0.177	-0.283	1	-0.404
X6	0.547	0.55	-0.634	0.663	-0.404	1

Based on the results of the correlation matrix above, it can be concluded that the data is free of multicollinearity. The correlation coefficient value between the variables in the matrix is less than 0.8, so the analysis can proceed to the next stage.

Using Elbow Method to Determine the Number of Clusters

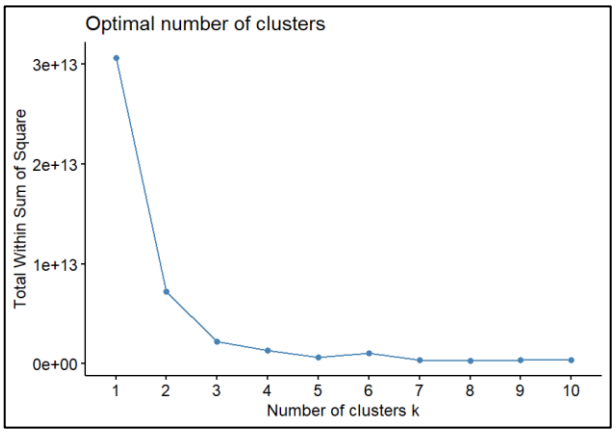

Figure 2. Elbow Method Plot

Based on Figure 2, it can be seen that the most optimal number of clusters is 3 as seen from the graph that shows a right-angled line. After the value of k is known, then the clustering stage can be carried out using the K-Means cluster method.

Clustering with the K-means Cluster

After figuring out how many clusters there are, the subsequent action is to find the centroid value for every cluster. The centroid value is derived from averaging the values in each cluster, beginning with a random selection for the initial cluster center value. The clustering procedure ends when the new initial cluster center value matches the previous one. Below is the initial center value that was calculated:

Table 4. Initial Cluster Centers

Initial Cluster Centers			
	Cluster		
	1	2	3
TPT	8.87	9.39	1.52
IPM	75.76	77.85	69.38
PPM	4.93	6.67	8.98
PDRB	99.00	20.78	11.68
TPAK	65.00	64.81	80.15
UM	97.92	83.97	10.57

Table 4 represents the initial phase of the data grouping procedure. Prior to conducting iterative tests, this information will be utilized to create three separate clusters.

Table 5. Iteration

Iteration History[a]			
	Change in Cluster Centers		
Iteration	1	2	3
1	23.483	17.898	18.243
2	0.000	2.305	1.855
3	0.000	0.000	0.000

Table 5 illustrates the iteration process in cluster grouping based on the initial table, where there are two iterations performed. In the first and second iterations, the resulting centeroids did not show significant differences. However, in the third iteration, the centeroids that appear show significant results. Therefore, all clusters have been formed, and the iteration process is stopped at the third iteration with a minimum distance of 71.451.

Table 6. Final Cluster Centers

Final Cluster Centers			
	Cluster		
	1	2	3
X1	8.78	7.90	6.14
X2	75.73	75.05	71.09
X3	6.01	6.91	9.82
X4	78.66	24.53	17.31
X5	64.90	66.67	67.83
X6	86.24	64.78	19.95

Table 5 shows the final results of the clustering process which consists of three clusters for each variable. The following is the interpretation of the K-Means cluster results above:

Cluster 1 : Cluster 1 shows the profile of a region with a relatively high open unemployment rate (8.78%) and a high Human Development Index (75.73). Regions in this cluster have the lowest percentage of people living in poverty (6.01%) and the highest Gross Regional Domestic Product (78.66) among the three clusters. Cluster 1's labor force participation rate is intermediate (64.9%), but this region has the highest minimum wage (86.24%). These characteristics suggest that Cluster 1 may represent an area with a strong economy and employment challenges, or an area with a dominant formal economy sector that offers high wages but has a certain level of unemployment.

Cluster 2 : Cluster 2 has a slightly lower open unemployment rate of 7. 90% and a Human Development Index of 75. 05 compared to Cluster 1. However, poverty levels rose above Cluster 1's 6. 91%, and the gross regional domestic product dropped by 24. 53%. The labor force participation rate in Cluster 2 is slightly higher than the 66. 67% of Cluster 1, and the minimum wage is moderate at 64. 78%. Overall,

Cluster 2 shows average socioeconomic conditions with moderate economic growth and increased poverty issues.

Cluster 3 : Cluster 3 has the lowest open unemployment rate (6.14%) and the lowest Human Development Index (71.09) of the three clusters. This cluster has the highest percentage of poor people (9.82%) and the lowest Gross Regional Domestic Product (17.31%). Cluster 3 has the highest labor force participation rate (67.83%), but the minimum wage is very low (19.95%). Overall, Cluster 3 indicates areas with high labor participation rates but low job quality, low wages, and high poverty rates.

Table 7. Districts/Cities in Each Cluster

Cluster 1	Cluster 2	Cluster 3
Kab. Bekasi	Bandung Barat	Ciamis
Kab.Bogor	Kab. Bandung	Cianjur
Karawang	Kab. Sukabumi	Kab. Cirebon
Kota Bandung	Kota Bekasi	Garut
	Kota Bogor	Indramayu
	Kota Cimahi	Kab. Tasikmalaya
	Kota Depok	Kota Banjar
	Purwakarta	Kota Cirebon
	Subang	Kota Sukabumi
	Sumedang	Kota Tasikmalaya
		Kuningan
		Majalengka
		Pangandaran

Table 7 shows the results of the K-means clustering analysis of the open unemployment rate in West Java. The regions are divided into three groups based on influencing factors. The first group consists of regions in the "good" category. These regions have high development indicators, such as HDI, GRDP, and minimum wage. However, the unemployment rate is still high due to urbanization pressure. Regions in this group include Bekasi Regency, Bogor Regency, Karawang, and Bandung City. The second group is in the medium category. These regions have fairly balanced socioeconomic conditions and are close to urban areas. Examples include Kota Depok, Kota Cimahi, and Kab. Bandung. The third group is in the low category and covers rural areas, such as Ciamis, Garut, and Pangandaran. These regions have low HDI and GRDP, and their level of development lags far behind the other two groups. This division illustrates regional inequality and is important for formulating targeted policies.

Table 8. Distance Between Clusters

Distance between final Cluster			
Cluster	1	2	3
1		58.273	90.606
2	58.273		45.714
3	90.606	45.714	

The space separating centroids, commonly known as centroid distance, measures how far apart the centroids of different clusters are from each other. Clustering is considered effective when this distance measurement is sufficiently large. A greater value indicates a larger gap between the clusters, which makes the distinctions between one cluster and another more apparent.

Table 9. Total Number Of Cluster Members

Number of Cases in each Cluster		
Cluster	1	4.000
	2	10.000
	3	13.000
Valid		27.000
Missing		0.000

This table shows the number of cases in each cluster from the clustering analysis. Cluster 1 has four cases, Cluster 2 has ten cases, and Cluster 3 has thirteen cases, making it the largest. There are 27 valid cases in total with no missing data. These numbers demonstrate that Cluster 3 is the largest, followed by Cluster 2; Cluster 1 is the smallest. This data is important for understanding the representation of each cluster in the sample and the relative significance of each group.

CONCLUSION

Analysis using the K-Means Cluster method divides West Java Province into three groups based on open unemployment rates and socioeconomic indicators. The first group consists of developed regions such as Bekasi Regency, Bogor Regency, Karawang, and Bandung City, which show high development despite having high open unemployment due to urbanization and lack of employment. The second group is medium-sized regions such as West Bandung, Bandung Regency, and other cities, which have stable socioeconomic conditions with unemployment and labor force participation in the medium category. The third group includes lagging regions such as Ciamis, Cianjur, and others, which show low development, high poverty rates, and low unemployment that may be influenced by the informal sector. These results show significant regional differences, so different policies are needed for each group to make unemployment reduction more effective and in line with regional characteristics.

REFERENCE

[1] A. A. Putrie and R. Sanjaya, "Pengelompokan Kabupaten/Kota Berdasarkan Indikator Tingkat Pengangguran Menggunakan Algoritma K-Means Clustering (Studi Kasus: Provinsi Jawa Barat)," vol. 2, no. 2, 2021, [Online]. Available: https://jabar.bps.go.id

[2] R. Ardian, U. Sultan, A. Tirtayasa, M. Syahputra, and D. Dermawan, "Pengaruh Pertumbuhan Ekonomi Terhadap Tingkat Pengangguran Terbuka Di Indonesia," vol. 1, no. 3, 2022.

[3] L. Luk, A. Mufida, and M. S. Nasir, "Analisis Dinamis Tingkat Pengangguran di Indonesia," 2021. [Online]. Available: https://economics.pubmedia.id/index.php/jmsd

[4] S. Pasuria and N. Triwahyuningtyas, "Pengaruh Angkatan Kerja, Pendidikan, Upah Minimum, Dan Produk Domestik Bruto Terhadap Pengangguran Di Indonesia," *Sibatik*

Journal: Jurnal Ilmiah Bidang Sosial, Ekonomi, Budaya, Teknologi, dan Pendidikan, vol. 1, no. 6, pp. 795–808, Apr. 2022, doi: 10.54443/sibatik.v1i6.94.

[5] N. Nurjanah, N. Suarna, W. Prihartono, and M. Kec Kesambi Kota Cirebon, "Implementasi K-Means Clustering Untuk Mengelompokan Tingkat Pengangguran," 2024.

[6] D. R. Ningsih, "Pengelompokan Produksi Daging Sapi Menurut Provinsi di Indonesia Tahun 2017-2022 dengan Menggunakan Metode K-Means," *ESTIMASI: Journal of Statistics and Its Application*, pp. 113–125, Jan. 2024, doi: 10.20956/ejsa.v5i1.26988.

[7] Muharni Sita dan Sigit Andriyanto, "Penerapan Metode K-Means Clustering Pada Data Tingkat Pengangguran Terbuka Tahun 2016-2018 Dan 2019-2021," *Penerapan Metode K-Means Clustering Pada Data Tingkat Pengangguran Terbuka Tahun 2016-2018 Dan 2019-2021*, vol. 22, pp. 90–99, 2022.

[8] F. A. Tanjung, A. P. Windarto, M. Fauzan, M. P. Studi, S. Informasi, and S. Tunas Bangsa, "Penerapan Metode K-Means Pada Pengelompokkan Pengangguran Di Indonesia," vol. 6, pp. 61–74, [Online]. Available: https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

[9] M. Athoillah, I. Irawan, M., and M. Imah, Elly, "Study Comparison of SVM-, K-NN- and Backpropagation-Based Classifier for Image Retrieval," *Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)*, 2015.

[10] J. Veronika, "Analisis Tingkat Pengangguran Di Kota Palopo Menggunakan Metode K-Means," 2024.

[11] D. R. Ningrat, D. Asih, I. Maruddani, and T. Wuryandari, "Analisis Cluster Dengan Algoritma K-Means Dan Fuzzy C-Means Clustering Untuk Pengelompokan Data Obligasi Korporasi," *Jurnal Gaussian*, vol. 5, no. 4, pp. 641–650, 2016, [Online]. Available: http://ejournal-s1.undip.ac.id/index.php/gaussian

[12] M. W. Talakua, Y. A. Lesnussa, and M. Y. Matdoan, "Analisis Klaster untuk Pengelompokkan Kabupaten/Kota di Provinsi Maluku Berdasarkan Indikator Pendidikan dengan Menggunakan Metode Ward," *Jurnal Statistika dan Aplikasinya*, vol. 5, no. 1, 2021.

[13] M. Desdianti, N. N. Debataraja, and E. Sulistianingsih, "Analisis K-Means Clustering Dengan Metode Elbow Pada Pengelompokan Tingkat Pengangguran Di Kalimantan Barat," 2024.

[14] M. Rais, R. Goejantoro, and S. Prangga, "Optimalisasi K-Means Cluster dengan Principal Component Analysis Optimization of K-Means Cluster with Principal Component Analysis on the Grouping of Districts/Cities on the Island of Kalimantan Based on Unemployment Rate Indicator," 2021.

[15] C. Auditiyah, "Pengelompokkan Daerah Rawan Demam Berdarah (DBD) di Jawa Timur Menggunakan Metode K-Means," *ESTIMASI: Journal of Statistics and Its Application*, pp. 205–215, Jul. 2024, doi: 10.20956/ejsa.v5i2.27091.