Journal of Innovative and Creativity
Journal Homepage: https://joecy.org/index.php/joecy
e-mail: journaljoecy@gmail.com

Faculty of Education
University of Pahlawan Tuanku Tambusai
ISSN 2775-771X (Online)
ISSN 2962-570X (Printed)
5 (3) 2025, Page: 30663-30669

Clustering Analysis Of Productive Age Unemployment Rates Using The K-Means Algorithm In Bekasi City

Rahayu Ningsih, Dhiya Firyal Suharsono, Sri Muryani, Sinta Rukiastiandari, Herlina Ferliyanti
[12345]*Teknologi Informasi Universitas Bina Sarana Informatika* Jakarta, Indonesia
Email: rahayu.ryh@bsi.ac.id

Published: November , 2025

ABSTRACT

Unemployment among the productive-age population remains a significant issue in urban areas, particularly in Bekasi City, which recorded an open unemployment rate of 7.82% in August 2024—much higher than the national average of 4.91%. This study aims to classify productive-age unemployment using a machine learning approach with the **K-Means Clustering** algorithm to provide a more comprehensive understanding of unemployment patterns. This research adopts a quantitative approach with the use of **SPSS version 25** and **RapidMiner** software. SPSS is used for validity, reliability, multicollinearity tests, optimal cluster determination, and ANOVA. Primary data were collected through a questionnaire distributed to 100 respondents, with 75 valid responses meeting the criteria (productive age, unemployed, and actively seeking employment). The research variables include age, education level, unemployment duration, perception of job opportunities, self-perception, and factors causing unemployment. The K-Means analysis resulted in three main clusters: the pessimistic cluster (low motivation), the neutral cluster (moderate perception), and the optimistic cluster (high motivation). Evaluation using ANOVA showed that the variables significantly differentiate between the clusters. These findings emphasize that productive-age unemployment is heterogeneous and requires cluster-based specific policies.
Keywords: Machine Learning, K-Means, Clustering, Productive-Age Unemployment

© The Author(s). 2021 Open Access This article is licensed under a Creative Commons Atibution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0

INTRODUCTION

Unemployment among the productive-age population is a critical socio-economic issue in Indonesia. In Bekasi City, the open unemployment rate reached 7.82% in August 2024, higher than the national average of 4.91%. This situation reflects an imbalance between labor force growth and job availability, influenced by skill mismatches, limited access to job information, and delayed adaptation to technological advancements.

Previous studies have widely applied data mining and clustering methods in various fields, but few have focused on socio-economic issues such as productive-age unemployment. Studies using the K-Means Clustering algorithm have shown effectiveness in identifying patterns in unemployment and poverty data. For example, several researchers used K-Means to cluster regions based on unemployment rates, showing its capability to reveal socio-economic distinctions. However, no study has specifically applied clustering algorithms to classify productive-age unemployment in urban areas.

This research aims to fill that gap by employing a quantitative approach based on data mining using the K-Means Clustering algorithm, supported by SPSS version 25 and RapidMiner for validity, reliability, multicollinearity tests, optimal cluster determination, and ANOVA analysis.

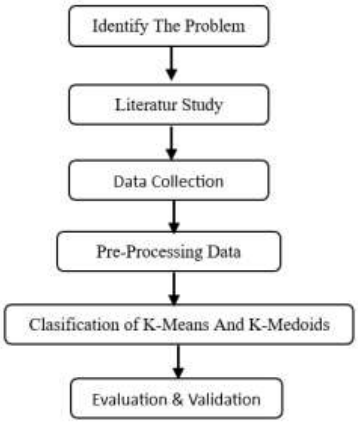


Figure 1. Research Steps

1. Identifying the Problem

The main issue in this study is the high open unemployment rate (TPT) in Bekasi City, which reached 7.82% in August 2024, higher than the national average of 4.91%. This condition emphasizes the need for more detailed mapping of the characteristics of productive-age unemployment.

2. Literature Study

Previous studies have focused on methodological/business aspects, but there has been a lack of attention to productive-age unemployment with clustering techniques.

3. Data Collection

Primary data was collected from 100 respondents, then filtered to 75 valid respondents using Slovin's formula (with a 10% margin of error). The variables include age, education, unemployment duration, perception of job opportunities, self-perception, and factors causing unemployment. Secondary data was obtained from BPS and the Bekasi Manpower Office (Disnaker).

4. Data Pre-Processing

Data cleaning, normalization, and transformation were carried out to ensure compatibility with the algorithms.

5. Clustering with K-Means and K-Medoids

The K-Means Clustering method produced three clusters. Validation was performed using ANOVA.

6. Evaluations & Validations

K-Means was tested using ANOVA ($p < 0.05$).

A. Populasi dan Sampel

The author used Slovin's formula (9) to calculate the sample size.

$$n = \frac{N}{1 + N(e)^2}$$

Figure 2. Slovin's Formula

n = sample size
N = population size
e = margin of error (0.1)

$$n = \frac{10.000}{1+10.000.(0,1)^2} \quad \frac{10.000}{1+100} \quad \frac{10.000}{101} \approx 99$$

Figure 3. Slovin's Calculation

The sample size for this study is 99 people, based on the number of unemployed individuals in Bekasi City. Therefore, Slovin's formula was used to determine the sample size, and for ease of distribution, the questionnaire was rounded to 100 respondents.

B. Validity Test

The validity test was conducted to ensure that the research instrument, including the questionnaire, accurately measures the constructs it intends to assess. The validity decision is based on the correlation coefficient (r) and statistical significance.

C. Reliability Test

The Cronbach's Alpha method was used to evaluate the reliability of this study. A good Cronbach's Alpha value (above 0.60) indicates that the items are more reliable as the value approaches 1.

D. Descriptive Statistics Test

Since K-Means cannot process missing values, any missing data can affect the analysis results. Therefore, a check of valid and missing data was performed using the frequency or descriptive feature of SPSS to determine whether the variables used in the analysis had missing (null) data.

E. Multicollinearity Assumption Test

High multicollinearity (VIF > 10) indicates that two or more variables have a very strong relationship, which can distort the clustering results.

F. Determination of Optimal K (Number of Clusters)

Elbow Method: The optimal **K** is chosen when the error reduction begins to slow down (elbow point). In SPSS, this is done by running K-Means for several values of K and recording the Error Sum of Squares (ESS).

G. One-Way ANOVA Test and P-Value

A one-way ANOVA test was conducted on each variable with the cluster as the factor. The F value and p-value were obtained. If $p < 0.05$, then the variable significantly differentiates the clusters.

.

RESULT AND DISCUSSION

The data used in this study includes all young individuals residing in Bekasi City who are unemployed but actively seeking employment, falling within the productive age range of 15 to 64 years. The variables used include age, education level, unemployment duration, perception of job opportunities, self-perception, and factors contributing to unemployment. The data used in the clustering process comes from 75 respondents who fully completed the questionnaire with a "Yes" response, while 25 respondents answered "No" because they did not meet the study criteria and were unable to continue filling out the questionnaire.

1. Data Processing with the K-Means Algorithm

a. Determination of Optimal K (Number of Clusters**)**

The determination of the number of clusters (K) in the K-Means analysis was done by setting the value of **K = 3**, based on conceptual considerations and the structure of the data.

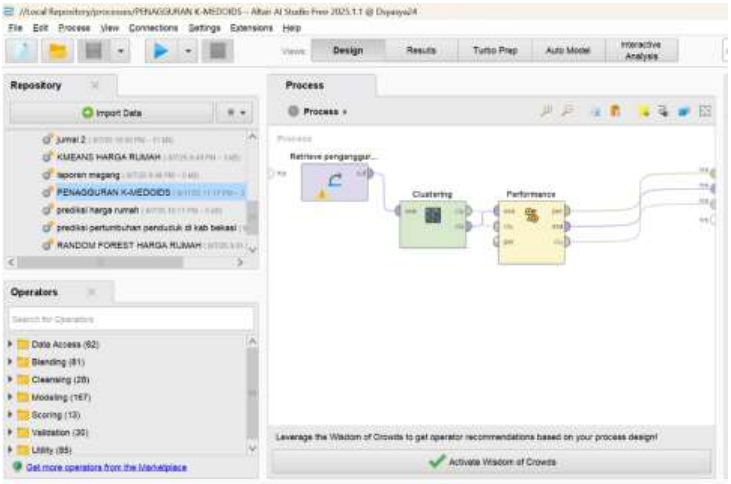
Final Cluster Centers

	Cluster		
	1	2	3
P1	4.58	4.54	3.75
P2	4.26	3.04	3.21
P3	4.21	4.39	4.36
P4	4.79	3.64	3.82
P5	4.21	3.21	3.61
P6	2.26	3.54	2.82
P7	1.26	4.18	3.04
P8	3.79	4.25	3.46
P9	2.11	4.29	3.39
P10	4.26	4.00	4.04
P11	4.53	4.82	4.04
P12	4.26	4.61	3.86
P13	4.21	3.29	2.75
P14	4.68	4.18	3.54
P15	4.53	4.25	3.75

Figure 4. Number of Clusters

b. One-Way ANOVA Test and P-Value

Based on the K-Means classification results, the data was successfully grouped into three clusters, with 19, 28, and 28 respondents in each cluster. Evaluation was performed using a One-Way ANOVA test, which showed that most variables had significance values below 0.05, indicating that these variables were effective in distinguishing between the clusters. Additionally, the size of each cluster is relatively balanced, and there are no clusters with very small or dominant numbers, suggesting that the selection of the number of clusters was appropriate.



ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
P1	5.670	2	.595	72	9.529	.000
P2	9.412	2	1.158	72	8.129	.001
P3	.201	2	.531	72	.378	.687
P4	8.153	2	1.024	72	7.966	.001
P5	5.618	2	.924	72	6.078	.004
P6	9.516	2	.955	72	9.965	.000
P7	48.116	2	.594	72	81.026	.000
P8	4.354	2	1.214	72	3.588	.033
P9	26.909	2	.753	72	35.758	.000
P10	.436	2	.731	72	.596	.554
P11	4.389	2	.442	72	9.935	.000
P12	3.944	2	.525	72	7.515	.001
P13	12.099	2	1.113	72	10.872	.000
P14	7.745	2	.600	72	12.915	.000
P15	3.715	2	.601	72	6.186	.003

Figure 5. One-Way ANOVA Test Results and P-Value

c. Evaluasi Hasil Klaster

75 respondents were effectively divided into three main clusters based on the clustering results obtained using the K-Means Clustering approach.

Figure 6. Cluster Evaluation Results

Number of Cases in each Cluster

Cluster	1	19.000
	2	28.000
	3	28.000
Valid		75.000
Missing		.000

Each cluster consists of the following number of respondents:

1. **Cluster 1** has 19 respondents, with variables P4, P11, and P14 showing the highest values in the first cluster, while P6 and P7 have the lowest values. This indicates that although respondents in this group have negative perceptions or are less engaged with indicators representing job opportunities or job search, they exhibit high satisfaction or responsiveness to certain indicators (P7).
2. **Cluster 2** has 28 respondents, representing a group with a neutral or moderate attitude toward the measured variables, with average scores across almost all indicators. This group may serve as a bridge between the two extreme perspectives.
3. **Cluster 3** has 28 respondents, with high scores on P3, P7, and P9, particularly on P7, which suggests they are actively seeking employment or view the process positively. This group likely has a strong work orientation and high motivation.

These findings show that the distribution of members in each cluster is relatively balanced, although Cluster 1 has a slightly lower number of members compared to the other two clusters. This suggests that a representative cluster was successfully formed using the K-Means method, with no cluster being too small or too dominant. Additionally, the clustering technique can be considered optimal and comprehensive since all respondent data (valid = 75) was distributed into clusters, and no data was missing (missing = 0).

2. Data Processing

a.Data preprocessing included data cleaning, normalization, and transformation to ensure compatibility with the clustering algorithm. Missing or inconsistent data were removed, and numerical scaling was applied to achieve uniformity.

Figure 7. RapidMiner Design

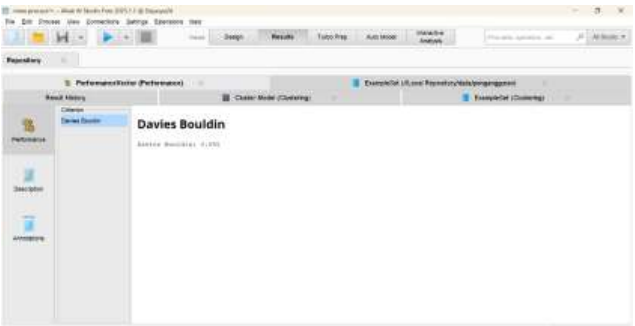
The data was processed with the parameter **k = 3**, measure type = **numerical measures**, and numerical measure = **Euclidean distance**. The result was 3 cluster models containing Cluster 0 = 42 items, Cluster 1 = 33 items, and Cluster 2 = 25 items.



Figure 8. Clustering Results

Next, the evaluation stage was conducted using the Davies Bouldin Index (DBI) by selecting the Cluster Distance Performance operator. The parameter used was the main criterion = Davies Bouldin, then select normalize and maximize to measure the closeness of data within a cluster by assessing its standard deviation.

Figure 9. Davies Bouldin Results



The results are presented in a visual display of the clustering data in a scatter plot. The blue color represents Cluster 0, green represents Cluster 1, and orange represents Cluster 3, with the x-axis representing (ID) and the y-axis representing (Cluster).

Figure 10. Visualization Display



This visualization display is used to provide a comparison of the clusters. With this option, it can be seen that the best cluster from the processed data is Cluster 2, which consists of 25 people, based on the highest cluster in the visualization display. The cluster with the worst condition is Cluster 0, which consists of 42 people, based on the visualization showing that this cluster has the lowest value compared to the others.

Cluster 0 consists of 42 respondents, with 18 males and 24 females. Twelve respondents are from outside Bekasi City, while the remaining 30 respondents are from Bekasi City. The majority of respondents are aged between 15–24 years. In terms of education, there are 21 high school/vocational school graduates, 8 D1/D3 graduates, and 13 S1 graduates. Regarding unemployment duration, 20 people have been unemployed for less than 3 months, 6 people for 3–6 months, 2 people for 7–12 months, and 14 people for over 1 year. The marital status shows that 6 people are married, and 36 people are unmarried. The main reasons for unemployment are the lack of job vacancy information and the mismatch with available job opportunities.

Cluster 1 consists of 33 respondents, with 19 males and 14 females. Twelve people are from outside Bekasi City, while 21 others are from Bekasi City. In terms of education, there is 1 person with a master's degree, 6 with a bachelor's degree, 6 with D1/D3 degrees, and 20 with high school/vocational school diplomas. Regarding unemployment duration, 8 people have been unemployed for less than 3 months, 12 people for 3–6 months, 6 people for 7–12 months, and 7 people for more than 1 year. In terms of marital status, 11 people are married, while 22 are unmarried. The main reasons for unemployment are the lack of job vacancy information, skill mismatches, and slow industrial and job growth around them. Regarding age, 21 people are aged 15–24 years, 2 are aged 25–34 years, and 10 are aged under 65.

Cluster 2 consists of individuals who are all employed, with a total of 25 people, and therefore does not belong to the unemployed category.

Based on the data analysis using the K-Means Clustering method with k = 3 and evaluation using the Davies Bouldin Index (DBI), which resulted in a value of 0.095, three main groups were identified.

Cluster 0 consists of 42 respondents, representing the most critical group, dominated by individuals aged 15–24 years with a high school or vocational school background, most of whom have been unemployed for more than three months. The main causes of unemployment in this cluster are lack of job information and job mismatches.

Cluster 1 consists of 33 respondents with a broader range of educational backgrounds and varying unemployment durations. The main unemployment factors in this cluster are skill mismatches and slow job market growth.

Cluster 2 consists of 25 respondents, all of whom are already employed, and therefore are not categorized as unemployed.

The Davies Bouldin Index (DBI) score of 0.095 indicates that the cluster separation and compactness achieved by the K-Means algorithm are very good. These findings demonstrate that K-Means

Clustering is effective in identifying groups with similar characteristics and can reveal significant differences in unemployment patterns among the productive-age population.
The differences in characteristics across the three clusters can serve as a foundation for designing more targeted and effective strategies to address unemployment issues, particularly for the younger age group with secondary education levels.

CONCLUSION

This study successfully classified productive-age unemployment in Bekasi City using the K-Means Clustering algorithm to identify clusters based on unemployment characteristics. The analysis resulted in three main clusters: the pessimistic cluster (low motivation), the neutral cluster (moderate perception), and the optimistic cluster (high motivation). Evaluation using the Davies Bouldin Index (0.095) indicated good cluster separation and quality, demonstrating that K-Means is effective in grouping respondents with similar characteristics. The method also proved efficient in handling homogeneous data and provided meaningful insights into the behavioral patterns of unemployed individuals. Overall, the results suggest that K-Means Clustering is a reliable and efficient method for analyzing unemployment data and can assist policymakers in developing targeted strategies to reduce unemployment rates.

REFERENCES

1. bekasikota.bps.go.id. bekasikota.bps.go.id. 2024. Tingkat Pengangguran Terbuka (TPT) sebesar 7,82 persen. Available from: https://bekasikota.bps.go.id/id/pressrelease/2024/11/11/270/tingkat-pengangguran-terbuka--tpt--sebesar-7-82-persen.html
2. Putrie AA, Sanjaya R. Pengelompokan Kabupaten/Kota Berdasarkan Indikator Tingkat Pengangguran Menggunakan Algoritma K-Means Clustering (Studi Kasus: Provinsi Jawa Barat). eProsiding Sist Inf. 2021;2(2):111–21.
3. Hendrastuty N. Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa. J Ilm Inform dan Ilmu Komput. 2024;3(1):46–56.
4. Jarupunphol P, Kuptabut S, Sudjarid W. Evaluating K-Means and K-Medoids clustering for household poverty analysis using random forests. Multidiscip Sci J. 2025;7(11).
5. Wafiq Azizah, Ade Irma Purnamasari, Agus Bahtiar, Kaslani. Clustering Analysis of Administrative Service Types Using K-Means (Study Case: Village bojongsalam). J Artif Intell Eng Appl. 2025;4(2):1303–10.
6. Soesmono S, Pertiwi R, Saputri B, Putri N, Widodo E. Pengelompokan Provinsi di Indonesia Berdasarkan Tingkat Pengangguran Tahun 2023 Menggunakan K-Medoids. Emerg Stat Data Sci J. 2025;3(1):498–515.
7. Surapati U, Jannah M. Penerapan Data Mining Menggunakan Metode K-Means Untuk Mengetahui Minat Customer Dalam Pembelian Merchandise Kpop. J Sains dan Teknol. 2024;5(3):875–84.
8. Maharani A, Juni Yanris G, Aini Nasution F. Implementation of the K-Means Clustering Method in Clustering Poor Population in Bandar Kumbul Village, Labuhanbatu Regency. Int J Sci Technol Manag. 2025;6(1):248–56.
9. Purnawansa sena wahyu., Wardhana A, Renggo yuniarti reny., Hudang adrianus kabubu., Darwin., Sayekti siskha putri., et al. Metodologi Penelitian Kuantitatif, Kualitatif, dan Kombinasi. Media Sains Indonesia. 2022. 1–225 p.
10. Waluyo edy, Septian Ahmad, Jerilian Ega AIS. Analisis data sampel menggunakan uji hipotesis penelitian perbandingan pendapatan menggunakan uji anova dan uji t. Ekon Dan Bisnis. 2024;2(30218365):775–85.