Jurnal Pendidikan Matematika (Kudus)
P-ISSN 2615-3939 | E-ISSN 2723-1186
https://journal.iainkudus.ac.id/index.php/jmtk
DOI: http://dx.doi.org/10.21043/jpmk.v6i1.19529
Volume 6, Number 1, June 2023, pp. 49-58

Implementation of K-Means to Classify Poverty Based on Housing Characteristics in Central Java in 2021

Laras Indah Setyaningsih[1], Anjelina Rafika Wulandari[1], Prizka Rismawati Arum[1]*
[1]Department of Statistics, Universitas Muhammadiyah Semarang, Semarang, Indonesia
**Correspondence: prizka.rismawatiarum@unimus.ac.id*

Abstract

Poverty is a condition in which a person's inability to meet basic needs such as food, clothing, shelter, and education so that he is unable to guarantee his own survival. To support the successful implementation of development programs, especially those aimed at reducing poverty, grouping districts/cities using cluster analysis can be assisted. Cluster analysis can be carried out to identify how the poverty rate is based on housing characteristics in Central Java which can be taken into consideration so that development programs are more targeted. Cluster analysis is a grouping method in which a group has the same characteristics, while between groups have different characteristics. K-means is one of the algorithms in data mining that can be used for grouping/clustering. The purpose of this study was to determine the classification of poverty in Central Java districts/cities based on housing indicators which include the floor area of the house, building materials for the widest floor, sources of drinking water, main building materials for roofs, and main fuel for cooking. This study yielded three clusters, with cluster 1 consisting of 22 districts and cities, cluster 2 consisting of 5 districts and cities, and cluster 3 consisting of 8 districts and cities. Cluster 1 grouping indicators were based on the sources of drinking water and the type of fuel used for cooking, cluster 2 grouping indicators were based on the size of the house's floor plan, and cluster 3 grouping indicators were based on the materials used to construct the house's widest floor and its main roof.

Keywords: Clustering; Housing; K-means; Poverty

Abstrak

Kemiskinan merupakan kondisi dimana ketidakmampuan seseorang untuk memenuhi kebutuhan pokok seperti pangan, sandang, papan dan pendidikan sehingga tidak mampu menjamin kelangsungan hidupnya sendiri. Untuk menunjang keberhasilan pelaksanaan program-program pembangunan, khususnya yang ditujukan untuk mengurangi kemiskinan dapat dibantu dengan mengelompokkan kabupaten/kota dengan analisis cluster. Analisis cluster dapat dilakukan untuk mengenali bagaimana tingkat kemiskinan berdasarkan karakteristik perumahan di

Jawa Tengah yang dapat dijadikan pertimbangan agar program-program pembangunan lebih tepat sasaran. Analisis cluster merupakan suatu metode pengelompokan dimana dalam suatu kelompok mempunyai karakteristik yang sama, sedangkan antar kelompok mempunyai karakteristik yang berbeda. K-means merupakan salah satu algoritma dalam data mining yang dapat digunakan untuk mengelompokkan/clustering. Tujuan penelitian ini ialah untuk mengetahui pengelompokan kemiskinan Kabupaten/Kota Jawa Tengah berdasarkan indikator perumahan yang meliputi luas lantai rumah, bahan bangunan untuk lantai terluas, sumber air minum, bahan bangunan utama untuk atap rumah, dan bahan bakar utama untuk memasak. Penelitian ini menghasilkan 3 cluster dengan cluster 1 memiliki anggota 22 Kabupaten/Kota, cluster 2 memiliki anggota 5 Kabupaten/ Kota, cluster 3 memiliki anggota 8 Kabupaten/Kota. Indikator pengelompokkan cluster 1 didasarkan kepada sumber air minum dan bahan bakar untuk memasak yang digunakan, indikator pengelompokan cluster 2 didasarkan kepada luas lantai rumah, sedangkan indikator pengelompokkan cluster 3 didasarkan kepada bahan bangunan untuk lantai terluas dan bahan bangunan utama untuk atap rumah.

Kata Kunci: *Clustering*; K-means; Kemiskinan; Perumahan,

Introduction

Central Java Province is one of the provinces in Java which is located between two large provinces, namely West Java and East Java. Central Java Province consists of 29 regencies and 6 cities with an area of Central Java recorded at 3.28 million hectares or around 25.04 percent of the total area of Java Island (BPS, 2021). Central Java province is confronted with a singular challenge, specifically the presence of poverty. Poverty refers to a state wherein individuals lack the means to fulfil essential requirements like nourishment, attire, housing, and education, thereby impeding their ability to ensure their own sustenance and well-being. To support the successful implementation of development programs, especially those aimed at reducing poverty, grouping districts or cities can be assisted. Districts/cities with homogeneous indicators and characteristics of poverty can be included in a group and can be analyzed by grouping analysis or cluster analysis. The strength of the relationship between objects becomes the basis for forming clusters. Cluster analysis is a grouping method in which a group has the same characteristics, while between groups have different characteristics (Dhuhita, 2015). One of the methods used in clustering is k-means. K-means is one of the algorithms in data mining that can be used for grouping/clustering. Researchers use the k-means method because k-means is a simple clustering method that can handle numerical data with fast computation and has been used by many previous researchers. There are various ways to form a cluster, one of which is by setting rules to determine members of the same group based on their level of equality. Previous research (Rianda, 2022) regarding the Application of K-Means and K-Medoids Algorithms in Grouping

Provinces in Indonesia Based on Household Housing Indicators in 2020 resulted in the conclusion that the K-Means algorithm is better for grouping provinces in Indonesia based on household housing indicators compared to the K-Medoids algorithm. The purpose of this study was to determine the classification of poverty in Central Java districts or cities based on housing indicators which include the floor area of the house, building materials for the widest floor, sources of drinking water, main building materials for roofs, and main fuel for cooking.

Method

The K-means algorithm was employed in this study as a methodology. The K-means algorithm is an algorithm that functions for clustering. According to (Asroni et al, 2018) the clustering process with the K-Means algorithm is as follows:

1. Determine the desired number of clusters
2. Distribute data according to the number of clusters that have been determined
3. Determine the centroid value for each cluster
4. Calculate the shortest distance using the Euclidean formula
5. Display results based on the lowest distance from the Euclidean formula calculation results
6. If you haven't got the appropriate results, then continue the iteration again using step 3, the iteration will be stopped if the clustering results are the same as the previous iteration.

The centroid value can be determined based on the range value that is in the data source by selecting according to the selected centroid value. According to (Sani, 2018) distance is used formula Euclidean as follows:

$$d = \sqrt{\sum\nolimits_{k=1}^{n} (p_k - q_k)^2}$$

description:
d : object distance
p_k : coordinates of object p
q_k : coordinates of object q
k : the order of the coordinates
n : objects

Secondary data was the research's main information source. Secondary data is research data obtained indirectly through intermediary media obtained and recorded by other parties. This study utilized housing-based poverty data in Central Java for the year 2021, which was obtained from the Central Statistics Agency (BPS). The data was collected based on several indicators, including the floor area of the house, the building materials used for the widest floor (land), the sources of drinking water, the main building materials for roofs, and the primary fuel utilized for cooking. To better target development projects and combat poverty, these data are processed using the K-Means Algorithm to create groups and clusters.

Results

The research data used consisted of housing characteristics in the districts/cities of Central Java, including the floor area of the houses, the building materials used for the widest floor, the source of drinking water, the primary building material for the house's roof, and the main fuel used for cooking. There are 35 data for each variable.

Summary

This section will show the summary of the data which will show the statistics for each variable. Following are the results of the test:

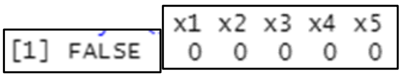
Table 2. Summary Data

		x1	x2	x3	x4	x5
N	Valid	35	35	35	35	35
	Missing	0	0	0	0	0
Mean		1.5183	10.1277	4.9451	.1680	12.7857
Std. Error of Mean		.39011	1.68507	.83839	.04016	1.83917
Median		.6900	7.7200	3.2900	.0900	10.2700
Mode		.10[a]	.38[a]	.00	.00	.26[a]
Std. Deviation		2.30795	9.96904	4.95999	.23758	10.88065
Variance		5.327	99.382	24.602	.056	118.389
Range		9.57	42.98	17.18	.86	40.76
Minimum		.00	.38	.00	.00	.26
Maximum		9.57	43.36	17.18	.86	41.02
Sum		53.14	354.47	173.08	5.88	447.50

From the table above, you can see the summary or characteristics of the data such as the amount of data in each variable totalling 35, there are no missing values in the data, as well as seeing the mean, median, modus, maximum value, minimum value, and many more.

Data Preparation

The initial step in data preparation involves data input, where the relevant data is collected and recorded. Subsequently, the focus shifts towards selecting the appropriate columns for analysis. Following this, it is essential to examine to identify and address any potential outliers in the data. These are the outcomes of checking for outliers:



```
[1] FALSE
```

```
x1 x2 x3 x4 x5
 0  0  0  0  0
```

Figure 1. Output Outliers

Based on Figure 1, it can be seen that output values x1 to x5 were all 0. Therefore, it could be interpreted that the data used had no outliers. So, It was not necessary to remove data. After that, knowing whether there was a correlation between variables. The Pearson correlation method is the one that is utilized to calculate correlations since it assesses the linear relationship between two numerical variables. After checking, the results were as follows:

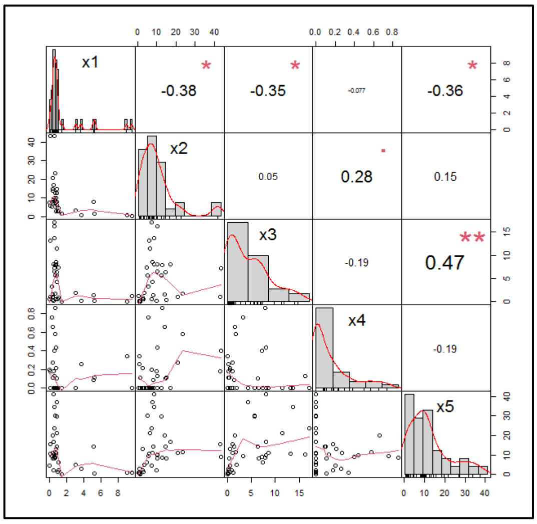


Figure 2. Correlation output

Based on Figure 2 above, it can be observed that the correlation values between x1 and x2 were -0.38, between x1 and x3, were -0.35, and so on, until the correlation value between x4 and x5 was -0.19. The figure indicated a correlation or relationship between x3 and x5, marked with **. Consequently, only one variable could be utilized. As a result, variable x5 was omitted. The subsequent step involved deleting variable 5 and standardizing the data. Data standardization was carried out to ensure consistent units for each variable.

Clustering Distance Measures

The choice of spacing size is an important step in grouping. It defines how the similarity of two elements (x, y) is calculated and it will affect the shape of the cluster. The distance matrix can be seen in the image below:

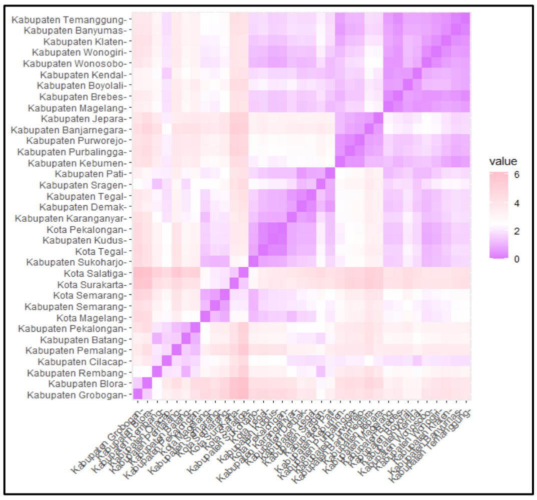


Figure 3. Distance Matrix Visualization

Referring to the presented Figure 3, the colour gradation provided valuable insights regarding distances. A deeper purple shade indicated a shorter distance, implying a higher likelihood of belonging to the same cluster. Conversely, a pinker hue signified a greater distance, suggesting a higher probability of belonging to different clusters.

Calculating K-Means Clustering

The first thing to do is determine the number of clusters used. Determination of the number of clusters using the Elbow method with the following results:

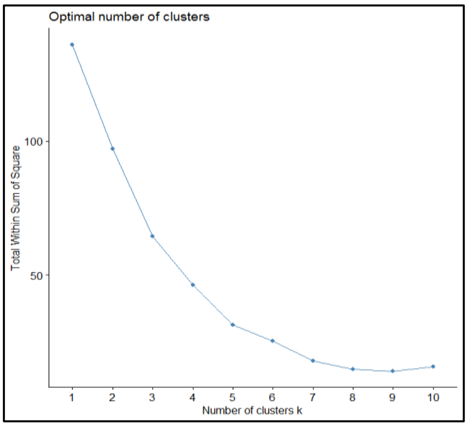


Figure 4. Elbow Outputs

From the picture above, it can be seen that when he ideal number of clusters is three groups, as evidenced by the slope movement at the cluster point at index number 3. Then determine the cluster members with k = 3, the following results are obtained:

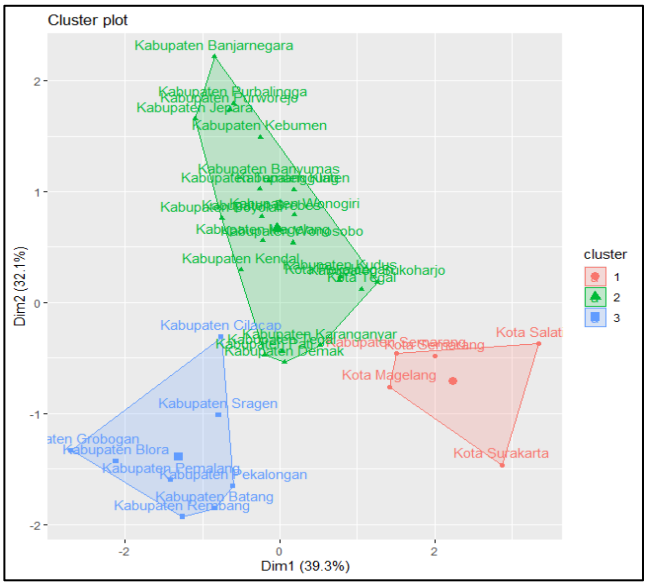


Figure 5. Cluster Visualization

From the picture above, it was found that cluster 1 comprised individuals from 22 districts and cities, cluster 2 was from 5 districts and cities, and cluster 3 was from 8 districts and cities. Then The features of each cluster can be determined by comparing the standardized data to the original data. The results were as follows:

Table 3. Result of Data Unstandardization

Clusters	x1	x2	x3	x4	x5
1	0.711	7.75	6.27	0.0423	15.1
2	6.51	2.26	0.61	0.16	3.95
3	0.454	21.6	4	0.519	11.9

The numbers in the table above show the result of unstandardized data or the result of returning the data to the original scale. From the table above, it can be concluded that Cluster 1's similarity indicators were x3 and x5, Cluster 2's similarity indicator was x1, and Cluster 3's similarity indicators were x2 and x4.

Discussion

Cluster analysis is a grouping method in which a group has the same characteristics, while between groups have different characteristics (Dhuhita, 2015). The information used in this study pertained to housing characteristics in the districts and cities of Central Java, including the size of the home's floor area, the materials used to construct the widest floor, the source of the home's drinking water, the main material used to construct the home's roof, and the main cooking fuel. The study revealed distinct clusters based on various factors. Cluster 1 comprised 22 regencies/cities, characterized by similar patterns in terms of sources of drinking water and fuel for cooking. Cluster 2 consisted of 5 regencies/cities, distinguished by similarities in house floor area. Lastly, cluster 3 encompassed 8 districts/cities, identified by commonalities in building materials used for the widest floor and the main materials employed for roofing houses. This result is to support the successful implementation of development programs, especially those aimed at reducing poverty in Central Java. These results are the same as the research conducted by Anggraini & Muharom (2017) regarding the grouping of sub-districts based on the education sector using the k-means cluster method. The variables used are students, the number of schools, the number of teachers, and the number of students.

Conclusion

Based on the above analysis, the following conclusions can be obtained: Regencies Banyumas, Purbalingga, Banjarnegara, Kebumen, Purworejo, Wonosobo, Magelang, Boyolali, Klaten, Sukoharjo, Wonogiri, Karanganyar, Pati, Kudus, Jepara, Demak, Temanggung, Kendal, and Brebes, as well as the cities of Pekalongan and Tegal, were part of Cluster 1. The Semarang Regency, the Cities of Magelang, Surakarta, Salatiga, and Semarang were part of Cluster 2. Cilacap, Sragen, Grobogan, Blora, Rembang, Batang, Pekalongan, and Pemalang Regencies are part of Cluster 3. Cluster 1 grouping indicators were based on the sources of drinking water and cooking fuel used, cluster 2 grouping indicators were based on the floor area of the house, while Cluster 3 grouping indicators were based on building materials for the widest floor and the main building material for the roof of the house.

References

Alfiansyah, D. N., Nastiti, V. R., & Hayatin, N. (2022). Penerapan Metode K-Means pada Data Penduduk Miskin Per Kecamatan Kabupaten Blitar. *J Repos*, 49-58. doi:10.22219/repository.v4i1.1416

Asroni, A., Fitri, H., & Prasetyo, E. (2018). Penerapan Metode Clustering dengan Algoritma K-Means pada Pengelompokkan Data Calon Mahasiswa Baru di Universitas Muhammadiyah Yogyakarta (Studi Kasus: Fakultas Kedokteran dan Ilmu Kesehatan, dan Fakultas Ilmu Sosial dan Ilmu Politik). *Semesta Tek*, 60-64. doi:10.18196/st.211211

BPS Jawa Tengah. (2021). *Provinsi Dalam Angka Jawa Tengah.* Semarang: BPS Jawa Tengah.

BPS Jawa Tengah. (2021). Distribusi Presentase Rumah Tangga menurut Kabupaten/Kota dan Jenis Lantai Terluas di Provinsi Jawa Tengah 2020-2021. https://jateng.bps.go.id/statictable/2021/04/12/2377/distribusi-persentase-rumah-tangga-menurut-kabupaten-kota-dan-jenis-lantai-terluas-di-provinsi-jawa-tengah-2020---2021.html

BPS Jawa Tengah. (2021). Presentase Rumah Tangga menurut Kabupaten/Kota dan Luas Lantai Rumah. https://jateng.bps.go.id/indicator/29/1515/1/persentase-rumah-tangga-menurut-kabupaten-kota-dan-luas-lantai-rumah-m2-.html

BPS Jawa Tengah. (2021). Distribusi Presentase Rumah Tangga menurut Kabupaten/Kota dan Sumber Air Minum di Provinsi Jawa Tengah. https://jateng.bps.go.id/indicator/29/1022/1/distribusi-persentase-rumah-tangga-menurut-kabupaten-kota-dan-sumber-air-minum-di-provinsi-jawa-tengah.html

BPS Jawa Tengah. (2021). Presentase Rumah Tangga menurut Kabupaten/Kota dan Bahan Bangunan Utama Atap Rumah Terluas. https://jateng.bps.go.id/indicator/29/1525/1/persentase-rumah-tangga-menurut-kabupaten-kota-dan-bahan-bangunan-utama-atap-rumah-terluas.html

BPS Jawa Tengah. (2021). Presentase Rumah Tangga menurut Kabupaten/Kota dan Bahan Bakar Utama yang Digunakan untuk Memasak. https://jateng.bps.go.id/indicator/29/1570/1/persentase-rumah-tangga-menurut-kabupaten-kota-dan-bahan-bakar-utama-yang-digunakan-untuk-memasak.html

Dhuhita, W. (2015). Clustering Menggunakan Metode K-Mean untuk Menentukan Status Gizi Balita. *J Inform Darmajaya*, 160-174.

Mustafidah, R., & Atok, R. M. (2017). Pengelompokan Kabupaten/Kota Di Provinsi Jawa Tengah Berdasarkan Indikator Kemiskinan Dengan C-Means Dan Fuzzy C-Means Clustering. *Published online*.

Rianda, F. (2022). Penerapan Algoritma K-Means dan K-Medoids dalam Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Perumahan Rumah Tangga Tahun 2020. *Estimasi: Journal of Statistics and Its Application*., 94-108. doi: 10.20956/ejsa.vi.18849

Sani, A. (2018). Penerapan Metode K-Means Clustering pada Perusahaan. *J I*(Agustina et al., 2012; Nasari & Darma, 2013; Poerwanto, 2016; Sumadikarta & Abeiza, 2014)*lm Progr Pascasarj Magister Ilmu Komput STMIK Nusa Mandiri.*, 1-7.

Agustina, S., Yhudo, D., Santoso, H., & ... (2012). Clustering Kualitas Beras Berdasarkan Ciri Fisik Menggunakan Metode K-Means. *Universitas Brawijaya ….* https://www.academia.edu/download/46692771/clustering-kualitas-beras-dengan-k-means.pdf

Dhuhita, W. (2015). Clustering Menggunakan Metode K-Mean Untuk Menentukan Status Gizi Balita. *Jurnal Informatika Darmajaya*, *15*(2), 160–174.

Dinata, R. K., Safwandi, S., Hasdyna, N., & Azizah, N. (2020). Analisis K-Means Clustering pada Data Sepeda Motor. *INFORMAL: Informatics Journal*, *5*(1), 10. https://doi.org/10.19184/isj.v5i1.17071

Harahap, B. (2019). Penerapan Algoritma K-Means Untuk Menentukan Bahan Bangunan Laris (Studi Kasus Pada UD. Toko Bangunan YD Indarung). *Regional Development Industry & Health Science, Technology and Art of Life*, 394–403. https://ptki.ac.id/jurnal/index.php/readystar/article/view/82

Hardiani, T. (2022). Analisis Clustering Kasus Covid 19 di Indonesia Menggunakan Algoritma K-Means. *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, *11*(2), 156–165. https://doi.org/10.23887/janapati.v11i2.45376

Nasari, F., & Darma, S. (2013). Penerapan K-Means Clustering Pada Data Penerimaan Mahasiswa Baru (Studi Kasus : Universitas Potensi Utama). *Semnasteknomedia Online*, *3*(1), 2-1–73. https://ojs.amikom.ac.id/index.php/semnasteknomedia/article/view/837

Poerwanto, B. (2016). *Analisis Cluster K-Means Dalam Pengelompokan Kemampuan Mahasiswa*. *December*.

Sumadikarta, I., & Abeiza, E. (2014). Penerapan Algoritma K-Means Pada Data Mining Untuk Memilih Produk Dan Pelanggan Potensial. *Jurnal Satya Informatika*, *1*, 12–22. https://lppm.usni.ac.id/jurnal/Istiqomah-Sumadikarta-Evan-Abeiza.pdf