Inside | Jurnal Sistem Informatika Cerdas

Volume 2 Nomor 2 | bulan Desember tahun 2024 | halaman 5-9
e-ISSN: 3031 - 4860

Implementation of K-Means Algorithm for Poverty Clustering (Case Study: East Java Province)

Agung Muliawan[1*], Difari Afreyna Fauziah[2], Mas'ud Hermansyah[3], Nur Andita Prasetyo[4], M. Faiz Firdausi[5]

Institute of Technology and Science Mandala, Jember, Indonesian[1,2,4,5]

State Polytechnic of Jember, Jember, Indonesian[3]

Corresponding Author:
Agung Muliawan, Institute Technology and Science Mandala, Jember, 68121, Indonesia
Email: agung.muliawan@itsm.ac.id

Abstract

Poverty is a crucial issue that requires attention and effective handling to achieve equitable development. This study aims to utilise the K-Means algorithm in clustering poverty in East Java based on the classification of available data. This study uses poverty data from the Central Bureau of Statistics from 2013-2024, including indicators of the number of poor people by district/city in East Java. The K-Means algorithm is used to cluster the regions in East Java into homogeneous groups in terms of poverty. The process begins with data preprocessing, normalisation, and selection of the optimal K parameter for the algorithm. The clustering results are expected to provide a clearer picture of poverty distribution and enable more targeted and effective policy development. The results of this study obtained 3 main clusters of poverty in the province of East Java, including cluster 0 small poverty category 23 regions, cluster 1 medium poverty category 15 regions and cluster 2 high poverty category. The findings of this study are expected to contribute to the planning of more targeted social and economic interventions in East Java. With the use of the K-Means algorithm, it is expected that there will be an improvement in the effectiveness of poverty alleviation programmes through a data-driven and analytical approach.

Keywords : Clustering, K-Means, Poverty, Province, East Java

1 INTRODUCTION

Poverty is one of the most pressing social and economic issues in many developing countries, including Indonesia. At the provincial level, such as East Java, the challenge of overcoming poverty becomes even more complex due to the differences in socio-economic conditions and development levels between regions[1]. To effectively face this challenge, an in-depth understanding of the distribution of poverty in the region is necessary. East Java, as one of the most populous and economically diverse provinces in Indonesia, faces significant poverty disparities. The complex and diverse poverty data demands an analytical approach that is capable of revealing patterns that may not be directly visible [2]. One analytical method that can be used is clustering techniques, which allow to group regions based on their poverty characteristics.

K-Means algorithm is one of the clustering methods that has proven effective in big data analysis [3]. This algorithm works by dividing data into K groups based on similar features, making it easier to understand patterns and relationships between data. In this context, K-Means can be used to cluster regions in East Java covering indicators of the number of poor people by district/city in East Java and using the Central Bureau of Statistics poverty data from 2013-2024. With this mapping, a clearer insight into the distribution of poverty in the province is expected.

This research aims to utilise the K-Means algorithm in clustering poverty in East Java. Through this approach, it is hoped that a better understanding of the pattern of poverty distribution can be generated, as well as identifying clusters of areas that require special attention [4]. The results of this study can serve as a basis for policy makers and related institutions to design more effective interventions in addressing poverty and optimising the use of resources.

Through this research, it is expected that specific patterns of poverty in various regions of East Java can be identified, so that it can help in designing intervention strategies that are more effective and in

accordance with the needs of each regional group. By using the K-Means algorithm, it is hoped that this research can make a significant contribution to the planning and implementation of poverty alleviation policies in East Java, as well as encourage the development of more focused and data-based strategies [5].

2 RESEARCH METHOD

The research method carried out in this study has several stages starting from data collection, making K-Means problems and evaluating and validating to see the accuracy value. The following is a detailed explanation of each stage:

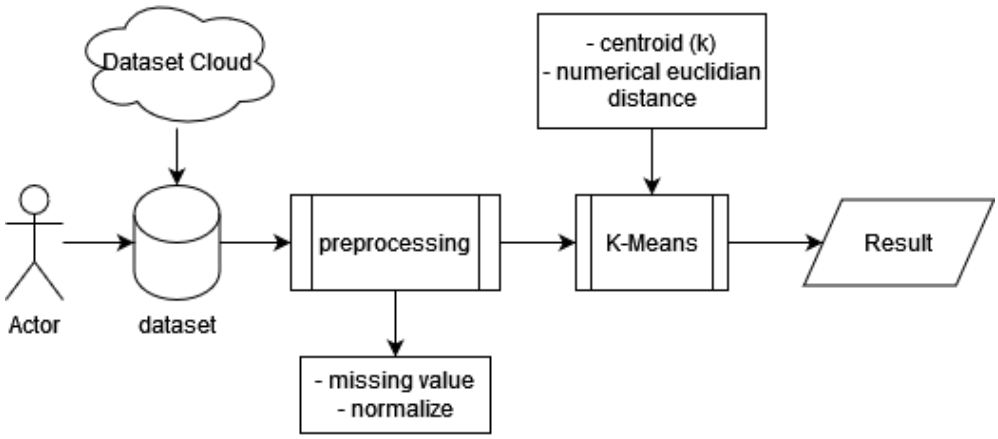

Fig 1. Research Methodology

In the picture above, the research methodology is explained, starting from taking open datasets on the jatim.bps.go.id site. Furthermore, the data obtained will be pre-processed before being processed in the k-means algorithm, the process carried out is missing value and data normalisation. The data that has been pre-processed will continue in making the k-means model to get the clustering of each region in East Java. The results obtained from the clustering process produce clusters according to the similarity of existing regions according to the resulting centroid value.

2.1 Dataset

In this study, researchers used poverty data from "Badan Pusat Statistik" from 2013-2024 with 39 districts / cities in East Java. The attributes used include occupation data per thousand of each existing region. The following is a table of sample data used in this study:

Table 1. Data Sample

No	Province	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
1	Pacitan	91,7	88,9	92,08	85,53	85,26	78,64	75,86	80,82	84,19	76,93	76,2	73,03
12	Situbondo	90,3	87,7	91,17	89,68	88,23	80,27	76,44	83,74	86,95	81,46	82,62	80,17
20	Ngawi	127,5	123,2	129,32	126,65	123,76	123,09	119,43	128,19	130,81	119,02	121,3	116,47
32	Tuban	196,9	191,1	196,59	198,35	196,1	178,64	170,8	187,13	192,58	178,05	177,25	171,24
39	Kota Surabaya	169,4	164,4	165,72	161,01	154,71	140,81	130,55	145,67	152,49	138,21	136,37	116,62

2.2 Pre-Processing

The data taken is still raw data so it needs to be preprocessed. The purpose of the preprocessing process is also to improve the results and accuracy of data mining to be better. Before being processed, the preprocessing process is carried out first. The steps of the preprocessing process start from data cleaning and normalizing [6]. Data cleaning is done by cleaning empty and inconsistent values and the normalization process is the process of normalizing values by reducing the value to be smaller to facilitate the calculation process [7] [8]. The data cleansing process is essential to ensure the quality and accuracy of data that will be used for analysis, decision-making, or predictive modelling [9]. Poor data can lead to inaccurate analysis results and risk to the organisation [10]. Where n is the number of valid data, and 'Valid Data' is data that is not empty. The following is the Data Cleansing process formula:

$$\text{Imputed Value} = \frac{\sum_{i=1}^{n} Valid\ Data}{n} \quad (1)$$

For the normalization process, using Z-Transformation or z-score normalization which is used to change the value of a dataset into a form based on the standard normal distribution (z-distribution) [11]. The following is the Z-Transformation process formula:

$$Z = \frac{X - \mu}{\sigma} \quad (2)$$

Description :
Z = skor-z (value after transformationi),
X = data real,
μ = population or sample mean,
σ = population standard deviation

2.3 K-Means

K-means clustering is a method used in machine learning and statistics to group data into different groups (or "clusters") [12]. The main goal of this algorithm is to divide a set of data into K-means different clusters, so that objects in one cluster are more similar to each other than to objects in other clusters [13]. The K-means algorithm is easy to understand and implement. The following is an overview of the k-means algorithm :

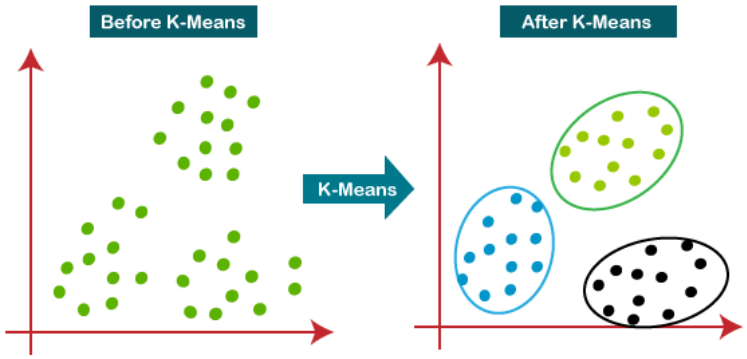

Fig 2. K-means algorithm

The basic concept of dividing data into clusters based on the distance to the centroid makes it intuitive for many applications. The following is the K-means formula [14] :

$$\sqrt{\sum_{j}^{n} = 1(Xij - \mu kj)^2} \quad (3)$$

Description :
xij = value j to data xi
μkj = value j to centroid μk

.

3 RESULTS AND ANALYSIS

From the results of the analysis carried out in this study, 3 clusters were obtained. This cluster includes 23 areas that show low or insignificant poverty levels. The areas in this cluster are relatively economically stable and do not face serious poverty problems. These results indicate that there are large areas in East Java that do not experience deep poverty problems. Cluster 1, there is only 1 area included in this cluster, which shows a very high level of poverty. This area may be experiencing very severe poverty conditions and requires immediate and intensive intervention from various parties, including the government and humanitarian organizations. Cluster 2, this cluster consists of 15 areas that are considered vulnerable to poverty. Although not as severe as the cluster that is very vulnerable to poverty, the areas in this cluster still show quite significant poverty indicators. This indicates that these areas require greater attention and intervention to overcome the existing poverty challenges. The following is a visualisation of k-means clustering and the value of each cluster in the regions of East Java :

Fig 3. Visualisation K-means

Attribute	cluster_0	cluster_1	cluster_2
2013.0	77.574	4771.260	207.253
2014.0	75.287	4786.790	201.113
2015.0	75.876	4789.120	202.932
2016.0	74.856	4703.300	198.774
2017.0	73.403	4617.010	195.247
2018.0	69.850	4332.590	181.738
2019.0	66.132	4112.250	172.749
2020.0	70.629	4419.100	186.309
2021.0	73.076	4572.730	192.800
2022.0	66.768	4181.290	176.374
2023.0	66.685	4188.810	177.006
2024.0	63.490	3982.690	168.161

Fig 4. Value cluster in regions East Java

4 CONCLUSION

Based on the clustering results, it is recommended that poverty alleviation policies and programs in East Java be adjusted to the characteristics of each cluster. For Cluster 0, efforts can be focused on sustainable development and improving the quality of life, health, and infrastructure development. For Cluster 1, the main focus should be on improving access and quality of education. While for Cluster 2, strategies should include strengthening local economic programs and improving the quality of social services. By using the K-Means algorithm, this study provides valuable insights in designing and implementing more effective and targeted poverty alleviation strategies in East Java. This analysis can also be the basis for evaluating future policies and programs to ensure that poverty alleviation efforts can be more targeted and have a positive impact on the community.

REFERENCES

[1] R. K. Dinata, S. Safwandi, N. Hasdyna, and N. Azizah, "Analisis K-Means Clustering pada Data Sepeda Motor," *INFORMAL Inform. J.*, vol. 5, no. 1, Art. no. 1, Apr. 2020, doi: 10.19184/isj.v5i1.17071.

[2] I. Sabilirrasyad, M. Hermansyah, A. Muliawan, A. Wahid, and R. Rakhmawati, "Strengthening the Economy and Competitiveness of MSMEs in Jelbuk Village, Jelbuk Sub-District Through MSME Digitization," *Prog. Conf.*, vol. 6, no. 1, Art. no. 1, Nov. 2023.

[3] M. Hermansyah, R. Hamdan, F. Sidik, and A. Wibowo, "Klasterisasi Data Travel Umroh di Marketplace Umroh.com Menggunakan Metode K-Means," *J. Ilmu Komput.*, vol. 13, p. 8, Sep. 2020, doi: 10.24843/JIK.2020.v13.i02.p06.

[4] H. Sulastri, H. Mubarok, and S. S. Iasha, "Implementasi Algoritma Machine Learning Untuk Penentuan Cluster Status Gizi Balita," *J. Rekayasa Teknol. Inf. JURTI*, vol. 5, no. 2, Art. no. 2, Dec. 2021, doi: 10.30872/jurti.v5i2.6779.

[5] D. I. Ramadhani, O. Damayanti, O. Thaushiyah, and A. R. Kadafi, "Penerapan Metode K-Means Untuk Clustering Desa Rawan Bencana Berdasarkan Data Kejadian Terjadinya Bencana Alam," *JURIKOM J. Ris. Komput.*, vol. 9, no. 3, Art. no. 3, Jun. 2022, doi: 10.30865/jurikom.v9i3.4326.

[6] A. Muliawan, A. Rizal, and S. Hadiyoso, "Heart Disease Prediction based on Physiological Parameters Using Ensemble Classifier and Parameter Optimization," *J. Appl. Eng. Technol. Sci. JAETS*, vol. 5, no. 1, Art. no. 1, Dec. 2023, doi: 10.37385/jaets.v5i1.2169.

[7] D. A. Fauziah, A. Muliawan, and M. A. Rohim, "Stock Price Prediction of PT. Jasa Marga (Persero) Tbk Using Linear Regression Algorithm," *PROCEEDING Int. Conf. Econ. Bus. Inf. Technol. ICEBIT*, vol. 4, pp. 810–815, Jul. 2023.

[8] A. Muliawan, T. Badriyah, and I. Syarif, "Membangun Sistem Rekomendasi Hotel dengan Content Based Filtering Menggunakan K-Nearest Neighbor dan Haversine Formula," *Technomedia J.*, vol. 7, pp. 231–247, Sep. 2022, doi: 10.33050/tmj.v7i2.1893.

[9] J. Badriyah, N. Ramadhani, A. Muliawan, K. R. Ummah, and A. Amrullah, "Penerapan Dimensi Reduksi Pada Machine Learning Dalam Klasifikasi Kanker Payudara Berdasarkan Parameter Medis," *J. RESTIKOM Ris. Tek. Inform. Dan Komput.*, vol. 6, no. 3, Art. no. 3, Dec. 2024, doi: 10.52005/restikom.v6i3.379.

[10] E. Afrianto, F. Wiranto, A. Muliawan, and M. Muhdar, "DATA MINING ANALYST FOR CLASSIFYING PLANT GROWTH DATA USING THE NAIVE BAYES METHOD," *PROCEEDING Int. Conf. Econ. Bus. Inf. Technol. ICEBIT*, vol. 5, pp. 233–239, Sep. 2024, doi: 10.31967/prmandala.v5i0.1190.

[11] F. Wiranto, A. Muliawan, and E. Afrianto, "Analysis of LQ45 Index Stock Movements using the ARIMA Method during Uncertainty in Global Economic Conditions in 2023," *PROCEEDING Int. Conf. Econ. Bus. Inf. Technol. ICEBIT*, vol. 4, pp. 816–827, Jul. 2023.

[12] Y. Yahya and M. Mahpuz, "Penggunaan Algoritma K-Means Untuk Menganalisis Pelanggan Potensial Pada Dealer SPS Motor Honda Lombok Timur Nusa Tenggara Barat," *Infotek J. Inform. Dan Teknol.*, vol. 2, no. 2, Art. no. 2, Aug. 2019, doi: 10.29408/jit.v2i2.1447.

[13] "Clustering in Machine Learning," GeeksforGeeks. Accessed: May 02, 2023. [Online]. Available: https://www.geeksforgeeks.org/clustering-in-machine-learning/

[14] M. F. A. Halik and L. Septiana, "Analisa Data Untuk Prediksi Daerah Rawan Bencana Alam Di Jawa Barat Menggunakan Algoritma K-Means Clustering," *JISAMAR J. Inf. Syst. Appl. Manag. Account. Res.*, vol. 6, no. 4, Art. no. 4, Nov. 2022, doi: 10.52362/jisamar.v6i4.939.