Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2024. Vol. 3. No. 2; e-ISSN : 2808-4519

Analysis Of Poverty Level Mapping In Riau Province Using The K-Means Method

Erlisa Santri[1]*, Dede Brahma[2]

[1]*Riau University*
[2]*Indonesian Islamic University*
erlisa.santri2807@student.unri.ac.id , dede.brahma2@gmail.com

Abstract

This research aims to understand the pattern of poverty distribution in Riau Province, identify clusters that reflect similar characteristics and provide a basis for developing more targeted policies. This approach uses machine learning techniques, especially the K-Means algorithm, to form clusters based on poverty level data. The results of the analysis show cluster 0 (C0) with a high poverty level and cluster 1 (C1) with a low poverty level. K-Means proved effective in grouping areas with similar levels of poverty, and provided a strong foundation for further analysis. Evaluation results using the Adjusted Rand Index (ARI), Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index show that the quality of cluster formation is good. This analysis provides detailed insight into poverty patterns in Riau Province and provides an empirical basis for implementing more contextual policies.

Keywords: *Riau Province, Machine Learning, K-Means algorithm, Silhouette Score*

1. Introduction

Indonesia, as a country with a diverse population and heterogeneous economic landscape, continues to be committed to overcoming the challenge of poverty as an integral part of national development. The poverty level in Indonesia reflects complex dynamics that require a sophisticated and focused analytical approach. Therefore, poverty level mapping analysis is essential in developing effective and sustainable policy strategies. The level of poverty in Indonesia is a major focus in efforts to understand and overcome social and economic inequality. To respond to this challenge, mapping analysis provides a solid foundation for developing smarter and more contextual poverty reduction strategies.

The importance of discussing poverty level mapping analysis lies in the need for an in-depth understanding of the more local distribution of poverty. Without a good understanding of the patterns and characteristics of poverty in each region, poverty reduction efforts may not be optimal and less focused. Therefore, mapping analysis is the key to opening new perspectives in developing policies that can address poverty problems more effectively.

In facing complex challenges such as mapping poverty levels, conventional analysis methods are often unable to dig up hidden information well. The use of machine learning technology and the K-Means method is becoming increasingly important. This method is able to identify patterns and clusters automatically, providing a deeper understanding of the distribution of poverty at the local level.

In previous research, in 2023, clustering research was carried out on poverty level data in West Java Province using the K-Means algorithm. The research produced 3 clusters, namely cluster 1 was categorized as an area not prone to poverty, cluster 2 was categorized as an area prone to poverty, and cluster 3 was categorized as an area very prone to poverty [1]. In research [2] in conducting clustering on poverty information data in West Java using the K-Means algorithm, 3 clusters were obtained. And in research [3] in conducting poverty data for West Java Province using the K-Means algorithm, 2 clusters were obtained, namely cluster 0 with a high level of poverty and cluster 1 with a low level of poverty.

Based on the background above, this research will conduct a mapping analysis of poverty levels in Riau Province using the K-Means method. By conducting a mapping analysis of poverty-prone areas in Riau Province, it is hoped that the Riau regional government can make the right decisions in overcoming and preventing poverty problems.

2. Literature review

2.1. Clustering

Clustering is a known process of forming a data set from an unknown data set based on similarities. Clustering analysis is a multivariable analysis technique used to group objects (data or variables) to produce information that is useful for testing objects [4]. Apart from that, clustering can also be referred to as grouping data points into two or more groups so that data points belonging to the same group are more similar to each other than data points belonging to different groups, based only on the information provided. available at these data points [5]. Clustering is included in unsupervised classification, where the process of grouping or classifying objects is based on information

obtained from data that describes the relationship between objects, maximizing similarities between members of one class and similarities between classes/clusters [6].

2.2. K-Means clustering algorithm

The K-Means algorithm is a form of non-hierarchical clustering where data will later be divided into one or more clusters, a grouping technique for dividing data into clusters, grouping data with the same features together and grouping data with different characteristics into in different groups [7]. Apart from that, K-Means is also a data mining algorithm that can be used for data clustering. There are many approaches to creating clusters, including creating rules that determine membership in the same group based on similarities between members. Another approach is to create a group of functions that measure some clustering property as a function of some clustering parameters [8].

2.3. Silhouette scores

Silhouette Score is an evaluation metric for measuring how well objects match the cluster they belong to. This score is in the range -1 to 1, where high values indicate that the objects match the cluster they belong to, while low values indicate that the objects may be a better match to neighboring clusters. In the context of K-Means, Silhouette Score helps measure the extent to which objects in a cluster are the same as or different from other clusters. Silhouette analysis is used to study and understand the separation distance between the resulting clusters. This analysis is used to measure how close each object in one cluster is close to another objects in another cluster. Silhouette score values lie between -1 to +1. The value of +1 indicates correct clustering of objects while the value of -1 show that objects are not properly clustered [9]

3. Research methodology

A flow diagram illustrating the research can be seen in Figure 1. The research framework was carried out starting from data collection to evaluation of the methods used in the research.

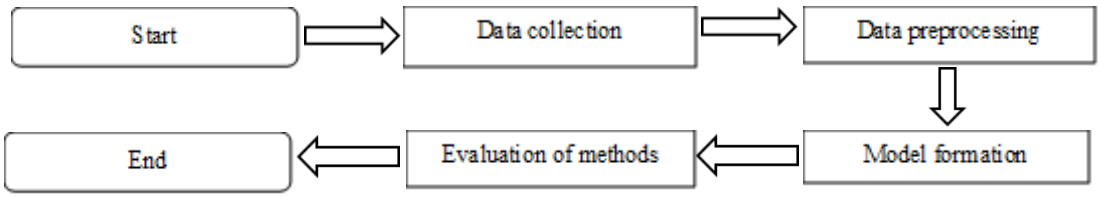

Fig. 1: Research framework

3.1. Data collection

Data collection was obtained from the publication of data at the Riau Province Central Statistics Agency (BPS) regarding data on the number of poor people in Riau Province by Regency/City from 2018 to 2022 [7].

3.2. Data preprocessing

Data preprocessing is done by data cleaning. Data cleaning uses missing values. Missing value will detect empty data or data. At this stage, we also ensure that the resulting data is of good quality.

3.3. Model formation

At this stage, a model or algorithm is formed based on the data that has been processed. This step involves selecting and applying analysis methods, such as establishing a machine learning model by selecting an algorithm, namely the K-Means algorithm. K-Means is an algorithm that determines K Clusters in a data set partition. As well as algorithms that are simple to implement and relatively easy to adapt in practice

3.4. Evaluation of methods

This evaluation stage focuses on the model that has been obtained, and it is hoped that the model will have good quality. Apart from that, at this stage effectiveness is carried out, using an evaluation matrix, namely the Adjusted Rand Index (ARI), Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index, where later the evaluation matrix will be used to test the quality of the clusters that have been obtained.

4. Results and Discussion

Based on the results of the research, grouping (clustering) with stages using the Silhouette Score method as well as using the K-Means algorithm and using the Python programming language for research assisted by Google Collab tools. With the following discussion:

4.1. Data collection

From data collection obtained through BPS publications in Riau Province, the data obtained is data on the number of poor people in Riau Province from 2018 to 2022. The amount of data obtained was 60 data. The data obtained came from 12 regencies/cities in Riau Province as can be seen in Table 1.

Table 1: Research data

Regency/City	Number of Poor Population (Thousand People)				
	2018	2019	2020	2021	2022
Kuantan Singingi	32.10	31.22	29.34	28.90	26.61
Indragiri Hulu	27.22	26.66	26.66	27.35	27.46
Indragiri Hilir	51.42	48.29	44.29	44.61	43.22
Pelalawan	44.29	45.98	45.88	49.30	47.84
Siak	25.81	24.49	25.38	25.77	25.71
Kampar	69.32	66.81	65.30	68.74	63.55
Rokan Hulu	72.28	72.21	73.35	74.73	73.81
Bengkalis	35.11	35.83	36.96	37.66	36.03
Rokan Hilir	48.92	49.80	48.85	51.97	49.59
Kepulauan Meranti	51.17	49.89	47.10	48.50	45.25
Pekanbaru	31.62	28.60	30.40	32.73	35.96
Dumai	11.19	10.95	9.88	10.57	10.00

4.2. Data preprocessing

After data collection is carried out, the data preprocessing stage is then carried out: cleaning. The data cleaning carried out is carrying out missing values. The results of the data preprocessing stage can be seen in Figure 2.

Fig. 2: Missing value

4.3. Model formation

The next stage is creating a model using the K-Means algorithm. It can be seen in Figure 3 showing IK-Means with the specified values. The specified value is the value 2 for the cluster.

Fig. 3: K-Means

It can be seen in Figure 4 that the results of clustering from "year" and "number of poor people" resulted in 2 clusters, namely cluster 0 and cluster 1. Cluster 0 is blue with a range of poor people of < 10 thousand people - < 40 thousand people, then from these results it can be concluded that cluster 0 is a cluster for low poverty levels. Meanwhile, cluster 1 is colored red with a range of poor population of > 40 thousand people - > 70 thousand people, so from these results it can be concluded that cluster 1 is a cluster with a high level of poverty.

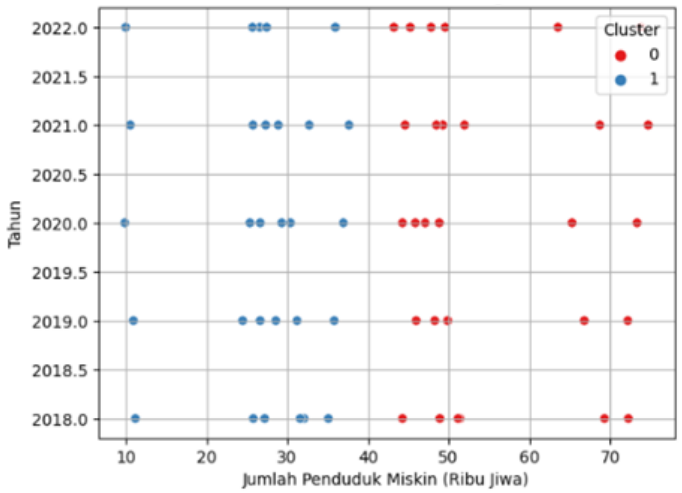

Fig. 4: Clustering 1

It can be seen in Figure 5 that the clustering results of "regency/city" and "number of poor people" resulted in 2 clusters, namely cluster 0 and cluster 1. In cluster 0, the color blue shows that there are 6 districts/cities in Riau Province that are included in cluster 0 or Low poverty levels, namely Kuantan Singingi, Indragiri Hulu, Siak, Bengkalis, Pekanbaru and Dumai. Meanwhile, cluster 1 in red shows that there are 6 districts/cities in Riau Province that are included in cluster 1 or high poverty levels, namely Indragiri Hilir, Pelalawan, Kampar, Rokan Hulu, Rokan Hilir, and Meranti Islands.

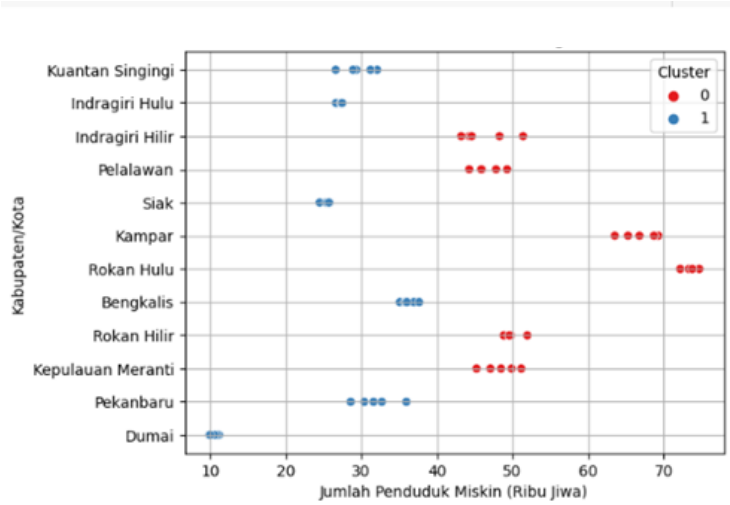

Fig. 5: Clustering 2

4.4. Evaluation of methods

The evaluation results are used to assess the method in this research, the method assessed is K-Means using the Adjusted Rand Index (ARI), Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index evaluation matrix as shown in Table 2. Evaluation results for The K-Means method shows the Adjusted Rand Index (ARI) with a value of -0.028 indicating a match that is worse than that generated randomly, the Silhouette Score with a value of 0.617 indicates that the clusters have good separation, the Davies-Bouldin Index with a value of 0.556 indicates that the clustering is good. The results were good, and the Calinski-Harabasz Index with a value of 126,206 showed that the resulting clustering was good.

Table 2: Evaluation of methods

Adjusted Rand Index	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
-0.028	0.617	0.556	126.206

5. Conclusion

This research produces poverty distribution patterns in Riau Province through mapping analysis using the K-Means method. From the results of cluster research, two clusters were found with different characteristics. Cluster 0 (C0) has a high poverty rate, while Cluster 1 (C1) has a low poverty rate. This research clearly shows differences in poverty levels at the regional level. The results highlight significant differences between clusters and provide valuable insights for policy makers. Effective grouping, which is strengthened by evaluation using the Adjusted Rand Index (ARI), Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index, shows that K-Means is successful in grouping regions with the same poverty level.

The conclusions of this research provide a strong basis for developing more targeted and effective poverty alleviation policies in Riau Province. By understanding the characteristics and challenges of each cluster, policy steps can be designed more specifically according to regional needs. Therefore, poverty level mapping analysis using K-Means not only broadens understanding of the dynamics of poverty in Riau Province, but also provides valuable guidance for sustainable and successful poverty alleviation efforts.

Suggestion

Suggestions for further research are to use additional algorithms to analyze poverty level mapping other than the K-Means algorithm. Apart from that, you can add variables to be studied, such as unemployment rate, economic level, and education distribution because some of these things can be used as supporting material related to poverty levels. By adding these variables, it will be more effective in carrying out poverty level mapping analysis.

References

[1] Nugraha, I. W. (2023). Clustering Pemetaan Tingkat Kemiskinan di Provinsi Jawa Barat Menggunakan Algoritma K-Means. Jurnal Ilmiah Wahana Pendidikan, 234-244.
[2] Nugroho Irawan Febianto, N. D. (2019). Analisis Clustering K-Means Pada Data Informasi Kemiskinan Di Jawa Barat Tahun 2018. Jurnal SISFOKOM, 130-140.
[3] Nabila Nur Fransiska R, D. S. (2018). Pengelompokkan Data Kemiskinan Provinsi Jawa Barat Menggunakan Algoritma K-Means dengan Silhouette Coefficient . Jurnal Teknologi Informasi Komunikasi (e-Journal), 29-35.
[4] Wicaksono, A. E. (2016). Implementasi Data Mining Dalam Pengelompokan Data Peserta Didik Di Sekolah Untuk Memprediksi Calon Penerima Beasiswa Dengan Menggunakan Algoritma K-Means (Studi Kasus Sman 16 Bekasi). Jurnal Teknologi Rekayasa , 206-216.
[5] Gustientiedina, M. A. (2019). Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan Pada RSUD Pekanbaru. Jurnal Nasional Teknologi Dan Sistem Informasi, 017-024.
[6] Witten, e. a. (2012). Data Mining Practical Maachine Learning Tools and Technique. San Fransisco: 2nd Edition.
[7] Tifani Amalina, D. B. (2022). Metode K-Means Clustering Dalam Pengelompokan Penjualan Produk Frozen Food. Jurnal Ilmiah Wahana Pendidikan, 574-583
[8] Mohammad Guntara, N. L. (2023). Cacah Klasterisasi dengan Algoritma K-Means Menggunakan Silhouette Coeficient. Jurnal Teknologi Informasi, 043-052.
[9] Godwin Oghuabor, U. F. (2018). Clustering Algorithm For A Healthcare Dataset Using Silhouette Score Value. International Journal of Computer Science & Information Technology, 027-037
[10] BPS Provinsi Riau URL: https://riau.bps.go.id/indicator/23/77/1/jumlah-penduduk-miskin.html