Application of Data Mining to Group the Spread of Covid-19 in Indonesia Using the K-Means Algorithm

Moh. Agus Efendi [a1], Zaehol Fatah[a2],
[a] Information System, Faculty of Science and Technology, Ibrahimy University, Indonesia
[b] Information System, Faculty of Science and Technology, Ibrahimy University, Indonesia
e-mail: [1]aguse6147@gmail.com, [2]zaeholfatah@gmail.com,

Abstrak

Pandemi COVID-19 yang dimulai pada akhir 2019 telah berdampak signifikan pada sektor kesehatan, ekonomi, dan sosial di seluruh dunia, termasuk di Indonesia. Kondisi geografis dan demografis yang beragam di Indonesia menghadirkan tantangan unik dalam memahami penyebaran kasus COVID-19. Variasi dalam distribusi kasus dipengaruhi oleh faktor-faktor seperti kepadatan penduduk, mobilitas antar wilayah, dan akses layanan kesehatan. Pendekatan berbasis data menjadi krusial untuk mengidentifikasi pola penyebaran dan mengembangkan strategi mitigasi yang efektif.Algoritma clustering K-Means digunakan sebagai metode data mining untuk mengelompokkan wilayah di Indonesia berdasarkan kesamaan karakteristik kasus COVID-19. Data sekunder, termasuk jumlah kasus, tingkat kesembuhan, dan angka kematian, diproses menggunakan aplikasi RapidMiner. Analisis menghasilkan tiga kluster: wilayah dengan prevalensi tinggi, sedang, dan rendah. Hasil tersebut konsisten dengan perhitungan manual, yang menunjukkan efektivitas algoritma serta keandalan RapidMiner dalam menangani data skala besar dan kompleks.Hasil clustering memberikan wawasan berharga tentang pola distribusi COVID-19, sehingga memungkinkan perumusan kebijakan mitigasi yang lebih spesifik dan terarah. Selain itu, penerapan algoritma K-Means menunjukkan potensi untuk membangun sistem peringatan dini dalam mengantisipasi lonjakan kasus di masa depan. Pendekatan ini menegaskan pentingnya pengambilan keputusan berbasis data dalam pengelolaan pandemi dan strategi alokasi sumber daya.

Kata kunci: *Algoritma, COVID-19, Data Mining, Indonesia, K-Means, Pengelompokan Wilayah.*

Abstract

The COVID-19 pandemic that began in late 2019 has had a significant impact on health, economic, and social sectors around the world, including in Indonesia. The diverse geographic and demographic conditions in Indonesia present unique challenges in understanding the spread of COVID-19 cases. Variations in case distribution are influenced by factors such as population density, mobility between regions, and access to health services. A data-driven approach is crucial to identify patterns of spread and develop effective mitigation strategies.The K-Means clustering algorithm was used as a data mining method to group regions in Indonesia based on similar characteristics of COVID-19 cases. Secondary data, including the number of cases, recovery rates, and mortality rates, were processed using the RapidMiner application. The analysis produced three clusters: areas with high, medium, and low prevalence. The results are consistent with manual calculations, demonstrating the effectiveness of the algorithm and the reliability of RapidMiner in handling large-scale and complex data.The clustering results provide valuable insights into the distribution patterns of COVID-19, allowing for the formulation of more specific and targeted mitigation policies. In addition, the application of the K-Means algorithm shows the potential to build an early warning system in anticipating future spikes in cases. This approach emphasizes the importance of data-driven decision-making in pandemic management and resource allocation strategies.

Keywords: *Algorithms, COVID-19, Data Mining, Indonesia, K-Means, Regional Clustering.*

1. Introduction

The COVID-19 pandemic that began in late 2019 has had a significant impact on various aspects of human life, both in terms of health, economy, and social. The rapid spread of the SARS-CoV-2 virus in Indonesia presents various challenges, especially due to the geographical and demographic diversity that exists in this archipelagic country [1]. The variation in the spread of COVID-19 cases that occurs in various regions is influenced by various factors such as population density, mobility between regions, and the availability of health services in each region [2].

In an effort to reduce the impact of the COVID-19 pandemic, a data-based approach plays a very important role in understanding the spread pattern and supporting the adoption of effective mitigation policies [3]. One relevant method is data mining, with the K-Means algorithm that can be used to group regions based on the similarity of the characteristics of the spread of COVID-19 cases [4]. Through the application of this algorithm, the spread pattern can be mapped based on the factors that play a role, which can ultimately be used as a basis for formulating more specific and effective policies.

In addition to grouping the spread based on the same characteristics, the K-Means algorithm also has the potential to support the creation of an early warning system for future spikes in cases. With its ability to process large-scale and complex data, data mining using the K-Means algorithm can help the government and stakeholders in designing more appropriate and efficient mitigation measures [4], [5]. Therefore, understanding the pattern of the spread of COVID-19 through data analysis with the K-Means algorithm approach is important to support more appropriate and data-based policy making in dealing with this pandemic. Based on this background, it is important to answer research questions related to the pattern of the spread of COVID-19 cases in Indonesia and the application of the K-Means algorithm to identify the characteristics of its spread

2. Research Methods

2.1. Data Collectiont

his research uses secondary data obtained from official reports related to COVID-19 in Indonesia. The data used includes the number of cases, number of recovered patients, and number of deaths due to COVID-19 in various provinces in Indonesia. This information is taken from trusted sources, such as the Ministry of Health or the COVID-19 Handling Task Force, with a focus on data from a particular year. These data were used as a basis for the epidemiological analysis for this study.

2.2. Data Maining

A simple definition of data mining is the extraction of important or interesting information or patterns from data in large databases. In scientific journals, data mining is also known as Knowledge Discovery in Database (KDD) [6].

2.3. K-Means

K-Means is an algorithm commonly used for document clustering. The main principle of K-Means is to construct k prototypes or centroids from a set of n-dimensional data (Aryan, 2010). Before applying the k-means algorithm, the data will be preprocessed first. The K-Means algorithm is included in partitioning clustering which separates data into k separate regions. The K-Means algorithm is very popular because of its ease and ability to cluster large data and outliners very quickly [7].

2.3. Clustering

Clustering analysis is a technique used to identify similar objects or individuals by paying attention to several criteria (Kuncono, 2003; 242). Clustering analysis is an analysis to group similar elements as research objects into groups (clusters) that are located and mutually exclusive [8].

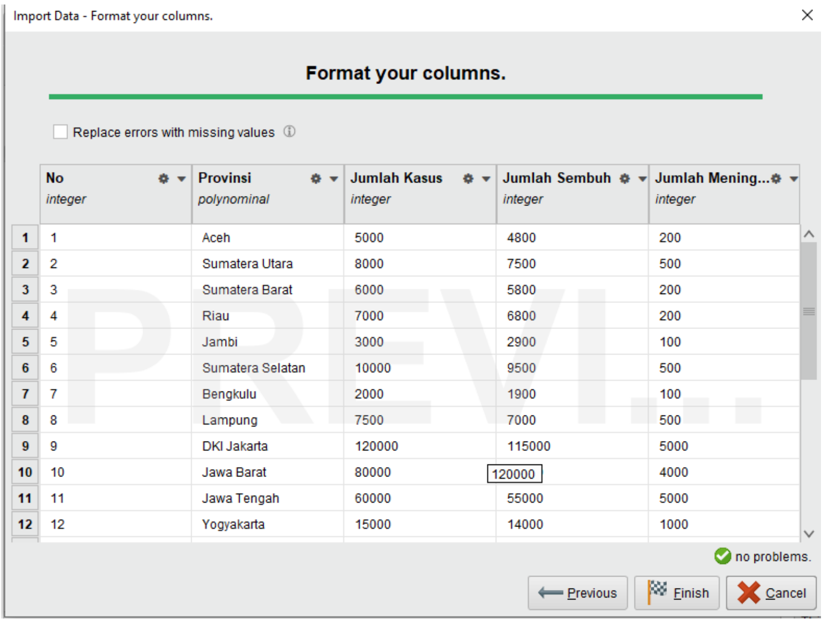
2.4. Covid 19

According to WHO (2020), the COVID-19 virus is transmitted during close contact through breathing (such as form) and vomiting. Therefore, to limit transmission of the virus, WHO (2020) continues to recommend frequent hand hygiene, using respiratory protection, regularly cleaning and disinfecting surfaces, maintaining physical distance and avoiding people with respiratory or respiratory symptoms [9].

3. Results And Analysis

3.1. Data Selection

In the initial stage, the available dataset includes a number of attributes, such as number, name of province, number of cases, number of recoveries, and number of deaths, with a total of 25 entries. The first step in the data processing process is to import data related to COVID-19 which is stored in Excel (xlsx) format into the RapidMiner application. To carry out this import, users can select the "import data" option or use the configuration wizard on the Read Excel operator, which functions to read data that has been uploaded. The following is an image that shows data selection.



Import Data - Format your columns.

Format your columns.

□ Replace errors with missing values

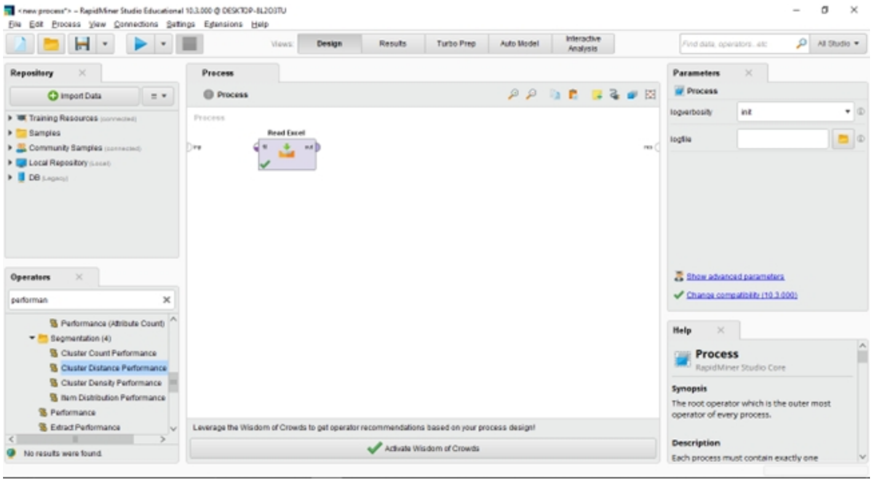
	No integer	Provinsi polynominal	Jumlah Kasus integer	Jumlah Sembuh integer	Jumlah Mening... integer
1	1	Aceh	5000	4800	200
2	2	Sumatera Utara	8000	7500	500
3	3	Sumatera Barat	6000	5800	200
4	4	Riau	7000	6800	200
5	5	Jambi	3000	2900	100
6	6	Sumatera Selatan	10000	9500	500
7	7	Bengkulu	2000	1900	100
8	8	Lampung	7500	7000	500
9	9	DKI Jakarta	120000	115000	5000
10	10	Jawa Barat	80000	120000	4000
11	11	Jawa Tengah	60000	55000	5000
12	12	Yogyakarta	15000	14000	1000

✓ no problems.

Figure 1. Display of Covid 19 Data Input

3.2. System Processing

At this stage, we will explain the steps for using the K-Means algorithm in RapidMiner using imported data. The first step begins by importing an Excel file containing the data to be processed, as shown in the image below.



<new process*> – RapidMiner Studio Educational 10.3.000 @ DESKTOP-8L2O3TU
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators, etc
All Studio
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
DB (Legacy)
Operators
performan
Performance (Attribute Count)
Segmentation (4)
Cluster Count Performance
Cluster Distance Performance
Cluster Density Performance
Item Distribution Performance
Performance
Extract Performance
Process
Read Excel
Parameters
Process
Show advanced parameters
Change compatibility (10.3.000)
Help
Process
RapidMiner Studio Core
Activate Wisdom of Crowds

logverbosity: init

logfile:

No results were found.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Synopsis

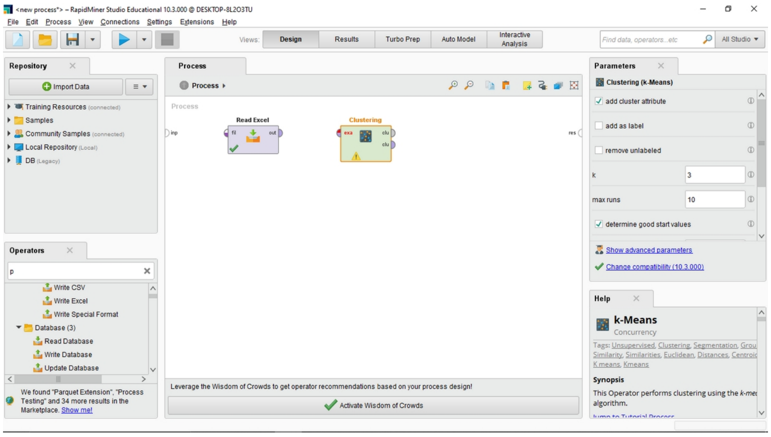
The root operator which is the outer most operator of every process.

Description

Each process must contain exactly one

Figure 2. Processing Stage 1

Next, click the Clustering and Segmentation option and select the K-Means method to perform grouping. Data must first be connected to the Clustering operator so that the process can run.



Clustering (k-Means)

☑ add cluster attribute

□ add as label

□ remove unlabeled

k 3

max runs 10

☑ determine good start values

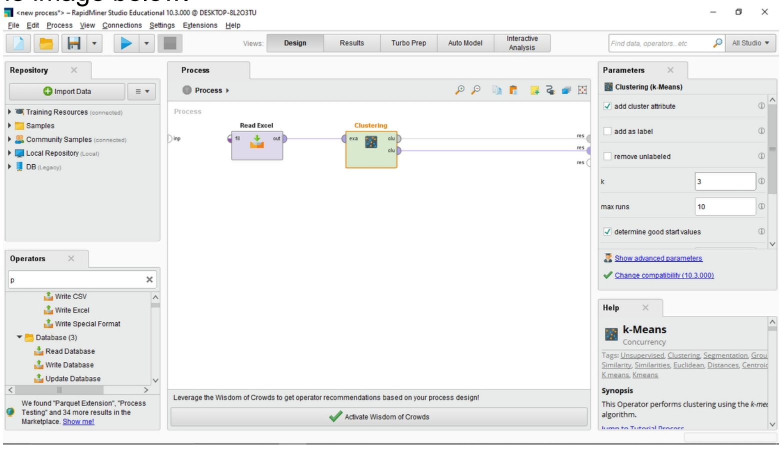
k-Means

Synopsis

This Operator performs clustering using the k-means algorithm.

Figure 3. Processing Stage 2

After that, connect Excel's Read operator with the Clustering operator, as shown in the image below.



Clustering (k-Means)

☑ add cluster attribute

□ add as label

□ remove unlabeled

k 3

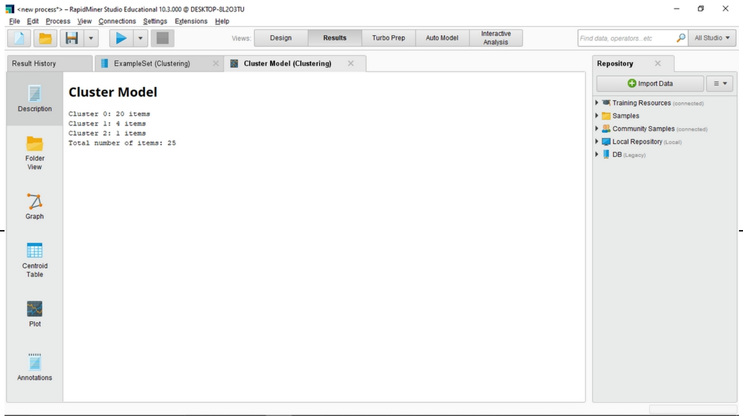
max runs 10

☑ determine good start values

Figure 4. Processing Stage 3

3.3. RapidMiner System Output

To obtain grouping results, the next step is to click the blue arrow icon located in the top center or on the Toolbar. This step will display the final results and is the final stage in using the RapidMiner tool, as seen in the image below.



Cluster Model

```
Cluster 0: 20 items
Cluster 1: 4 items
Cluster 2: 1 items
Total number of items: 25
```

Figure 5. RapidMiner Model Cluster Values

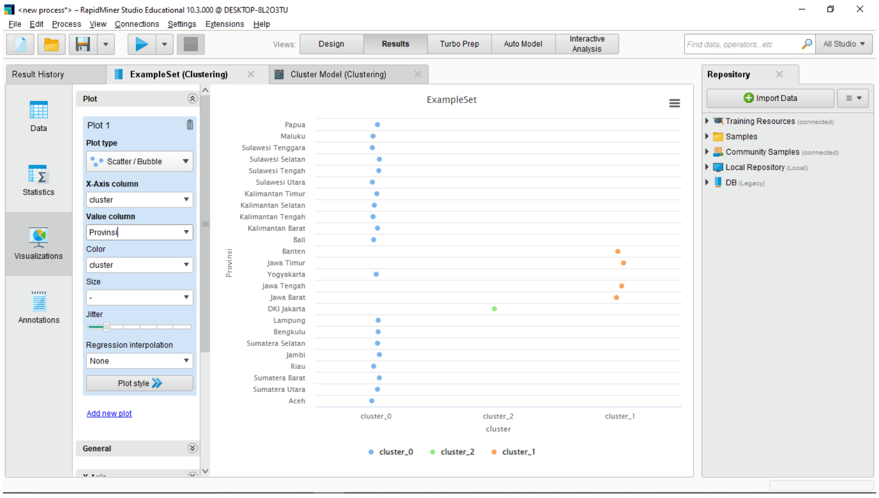
Information:

a) The number of Cluster 0 (high) is 4 items

b) The number of Cluster 2 (medium) is 1 item

c) The number of Cluster 1 (low) is 20 items

Thus, the grouping results produced by RapidMiner can be seen in the following image.



<new process*> – RapidMiner Studio Educational 10.3.000 @ DESKTOP-8L2O3TU

Plot 1

Plot type: Scatter / Bubble

X-Axis column: cluster

Value column: Provinsi

Color: cluster

Size: -

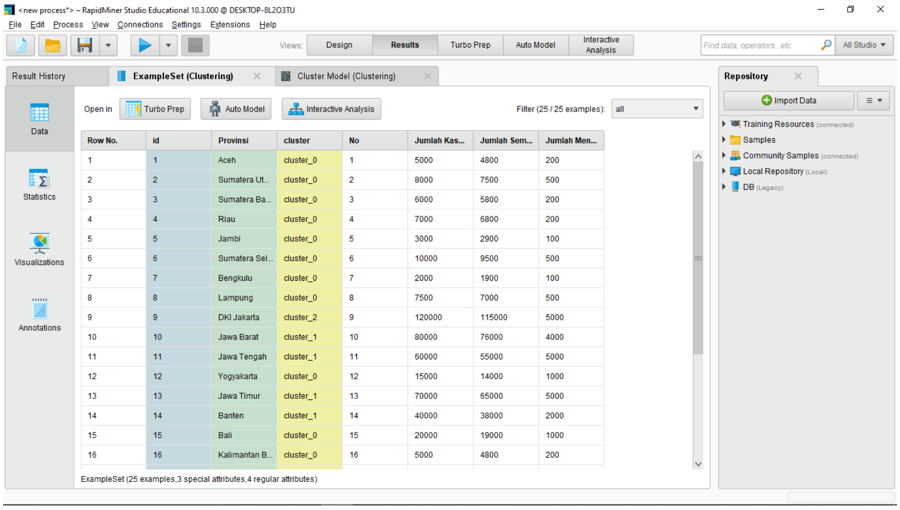
Regression interpolation: None

Figure 6. RapidMiner Clustering Results

Based on the picture. It can be seen that the high group has 1 item, the medium group has 2 items, while the low group has 0 items.

3.4. Discussion

Based on the explanation of the stages of use and the results that have been displayed, it can be explained that there is a connection between the results of manual calculations using the K-Means algorithm and the results obtained using tools such as Rapid Miner. The results of manual calculations using the K-Means algorithm, when carried out using applications such as Microsoft Office Excel 2007, show similarities to the results produced by RapidMiner. This shows that manual calculation results can be integrated with the RapidMiner application for further analysis.



<new process*> – RapidMiner Studio Educational 10.3.000 @ DESKTOP-8L2O3TU
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis
Find data, operators...etc
All Studio
Result History
ExampleSet (Clustering)
Cluster Model (Clustering)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model Interactive Analysis
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
DB (Legacy)

Filter (25 / 25 examples): all

Row No.	id	Provinsi	cluster	No	Jumlah Kas...	Jumlah Sem...	Jumlah Men...
1	1	Aceh	cluster_0	1	5000	4800	200
2	2	Sumatera Ut...	cluster_0	2	8000	7500	500
3	3	Sumatera Ba...	cluster_0	3	6000	5800	200
4	4	Riau	cluster_0	4	7000	6800	200
5	5	Jambi	cluster_0	5	3000	2900	100
6	6	Sumatera Sel...	cluster_0	6	10000	9500	500
7	7	Bengkulu	cluster_0	7	2000	1900	100
8	8	Lampung	cluster_0	8	7500	7000	500
9	9	DKI Jakarta	cluster_2	9	120000	115000	5000
10	10	Jawa Barat	cluster_1	10	80000	76000	4000
11	11	Jawa Tengah	cluster_1	11	60000	55000	5000
12	12	Yogyakarta	cluster_0	12	15000	14000	1000
13	13	Jawa Timur	cluster_1	13	70000	65000	5000
14	14	Banten	cluster_1	14	40000	38000	2000
15	15	Bali	cluster_0	15	20000	19000	1000
16	16	Kalimantan B...	cluster_0	16	5000	4800	200

ExampleSet (25 examples,3 special attributes,4 regular attributes)

Figure 7. RapidMiner Calculation Data Display

4. Conclusion

The analysis results show that the K-Means algorithm is an effective tool for clustering regions based on the level of COVID-19 prevalence. Grouping data into three main categories of prevalence: high, medium, and low. This grouping provides deeper insight into the differences in distribution characteristics between regions and supports the development of more specific and targeted policies.

The use of RapidMiner as an analysis tool has been proven to facilitate the processing of large and complex data. The clustering results produced by RapidMiner are consistent with manual calculations using the K-Means algorithm, which shows the reliability of this tool for more efficient analysis. These results also open up opportunities for wider application, such as developing more effective containment strategies in allocating medical resources and prioritizing high-risk areas.

In addition, this data mining-based approach has great potential to support early warning systems for future spikes in COVID-19 infections. This approach, which can analyze data on a large and complex scale, makes a significant contribution to understanding case distribution patterns and helping to develop more efficient and evidence-based policies.

References

[1] R. Sitohang dan A. S. Dewi, *Data Mining: Teori dan Aplikasi*, Bandung: Informatika, 2016.

[2] S. Suyanto, *Machine Learning dan Data Mining: Mengolah Data Menjadi Informasi Berharga*, Yogyakarta: Andi, 2018.

[3] S. R. Girsang dan J. Simarmata, *Algoritma dan Pemrograman Data Mining*, Jakarta: Elex Media Komputindo, 2017.

[4] A. Kusrini dan E. Luthfi, *Algoritma Data Mining*, Yogyakarta: Andi, 2009.

[5] H. Wibisono, *Pengolahan Data dengan Data Mining Menggunakan Algoritma K-Means*, Jakarta: Bumi Aksara, 2021.

[6] S. K. M. K. D. A. N. A. P. S. K. M. K. Amril Mutoi Siregar, Data Mining: Data Processing into Information with Rapidminer. CV Kekata Group. [Online].

[7] K-Means Modeling Algorithms and Big Data Analysis (Mustahiq Data Mapping). Pascal Books, 2022. [Online]. Available: Https://Books.Google.Co.Id/Books?Id=_Bjmeaaaqbaj

[8] Anxiety During the Covid-19 Pandemic: Jariah Publishing. Jariah Publishing Intermedia, 2020. [Online]. Available: Https://Books.Google.Co.Id/Books?Id=4rukeaaaqbaj

[9] M. K. Rahayu Mayang Sari, Clustering Method for Data Analysis of Dangerous Diseases. Serasi Media Teknologi, 2024. [Online]. Available: Https://Books.Google.Co.Id/Books?Id=Pwooeqaaqbaj