Hybrid clustering and classification methods to find out the pattern of the spread of covid-19 in East Java province

M W F Umam\textsuperscript{1,2}, M Fatekurohman\textsuperscript{1,2} and D Anggraeni\textsuperscript{1,2}

\textsuperscript{1}Department of Mathematics, Faculty of Mathematics and Natural Science, University of Jember, Indonesia
\textsuperscript{2}Data Science Research Group Department of Mathematics, Faculty of Mathematics and Natural Science, University of Jember, Indonesia

E-mail: mfatekurohman.fmipa@unej.ac.id

Abstract. Covid-19 entered Indonesia in March 2020, which had an extraordinary impact on all aspects of life. The real impact is the economic and social sectors. Various efforts and policies from the Indonesian government were made to suppress the increase in the rate of Covid-19 cases. The spread of Covid-19 cases in Indonesia has been evenly distributed in various provinces, one of which is East Java. East Java has 38 districts/cities, where the case of each district is different, there are green, yellow, red, and black zones. To find out the pattern of spread of each district /city zone, clustering is used. The method used in this study was to use the hybrid method of clustering with K-means and classification with its algorithm Decision Tree. The purpose of this study to prepare mitigation measures to inhibit the rate of spread of Covid-19 in East Java Province. The variables used include the number of positive, the number of dead, the number of recovered, the number of suspects, and the number of probable. With this grouped data, it is expected to help make the right decision in reducing the spread and minimizing the number of positive patients. Based on the results of the study, there will be cluster results, namely 7 very high-risk areas (C0), 1 high risk district (C1), 29 moderate risk districts (C2), and 1 low-risk district (C3). It also has a validity index level of 0.356 as measured by the Davies-Bouldin Index.

1. Introduction

The year 2020 is a difficult time for life to take place in the countries of the world. All economic and social movements are hampered by the pandemic COVID 19 outbreak (Corona Virus Diseases 2019). The Covid 19 outbreak was first detected in the Chinese city of Wuhan. The virus is spreading rapidly and infecting many of the world's population. Various efforts and policies from the governments of each country are made to suppress the covid 19 case [3]. Covid-19 is a new corona virus that is producing an outbreak of the pneumonia virus. The virus can cause the death of people of all ages, particularly those suffering from chronic or elderly illnesses. In addition to health risks, it also forces the country's economy. The Covid-19 pandemic has affected unprecedented communities and economies everywhere around the world [7].

Several researchers have conducted research on Covid-19 cases and their clustering at the world level [8], Covid-19 cases of District Jember [1], and some countries [2], but more in-depth analysis is needed to find out the cases of Covid-19 spread in Indonesia [6] especially in East Java. The purpose of this study is to account for the spread of the Covid-19 Virus in East Java province based on
parameters of the number of positive cases, patients recovered, patients died, suspects, and probable. Clustering is done using the K-means method, which groups data into multiple clusters based on data similarities. Corona virus spread data that has been grouped using the K-means Clustering Algorithm is expected to help make the right decision in reducing the spread of corona virus and minimize the number of COVID-19 positive patients.

Classification is the grouping of records, observations, or observing and forming a class of objects that have similarities. This type of data mining is different from the classification is the absence of target variables in classification [5]. While K-means is a classification algorithm that is quite simple by partitioning datasets into several k clusters. The algorithm is quite easy to implement and run, relatively fast, easy to customize and widely used. The K-Means method can partition data into groups so that data of the same character is entered into the same group and data of different characteristics are grouped into other groups [4].

2. Method

This research process uses hybrid clustering and classification methods. The data used in this research include secondary data obtained from the Ministry of Health, clicked and processed from covid19.go.id. The data used the number of spread of the covid-19 pandemic in East Java on March 8th, 2021. Then the data is preprocessed according to the need by utilizing Microsoft Excel and Rapid Miner Studio software. The variable used include the number of positive cases \((c_1)\), number of recovered \((c_2)\), number of death \((c_3)\), number of suspect \((c_4)\) and number of probable \((c_5)\) Based on data obtained from east Java province.

The steps used in the research are:
1. Data collection and finding sources of literacy
2. Preprocessing data
3. Determining the number of clusters using the Davies Bouldin Index (DBI) parameter. The smaller the value of the DBI obtained \((DBI \geq 0)\), the better the cluster obtained.
4. Processing clustering and classification data using the K-Means method and the C4.5 method with the Rapid Miner Studio application
5. Perform a validity test.
6. Finish

![Figure 1. Method of Research.](image-url)
3. Results and discussion
Based on figure 1, the data clustered using the K-Means method to determine the results of cluster mapping, then the results of the cluster processed using classification methods to see the value of rules in the form of decision trees. In this case, before clustering and classifying, you must first determine how many clusters using by Davies Bouldin Index (DBI). The best cluster is indicated by the smallest DBI value. (DBI ≥ 0).

Here is a comparison of the number of clusters (c) using the Davies Bouldin Index (DBI).
Table 1. DB Value.

| Cluster (c) | Davies Bouldin Index |
|-------------|----------------------|
| 2           | 0.516                |
| 3           | 0.377                |
| 4           | 0.356                |
| 5           | 0.385                |

Based on Table 1, data are processed using 4 clusters where clusters that have a value of 4 are the best clusters. Based on this, the number of clusters used in this study is 4 labels, namely very high-risk clusters ($C_0 =$ black zone), high-risk clusters ($C_1 =$ red zone), moderate-risk clusters ($C_2 =$ yellow zone), and low-risk clusters ($C_3 =$ green zone).

After determining the number of clusters, the data is then classified to determine the mapping of district or city areas based on the level of risk that has been determined.

Figure 3. Cluster mapping results.
In figure 3, there are 7 districts with very high risk (black zone), then 1 district with high risk cluster (red zone), 29 districts with moderate risk clusters (yellow zone), and 1 district with low risk (green zone). Here are the results of cluster mapping of the spread of the Covid-19 pandemic in Indonesia.

a. Very high risk clusters ($C_0$ = black zone), includes Banyuwangi Regency, Blitar Regency, Gresik Regency, Jember Regency, Jombang Regency, Kediri Regency, and Malang City.

b. High risk clusters ($C_1$ = red zone), includes the city of Surabaya.

c. Moderate risk clusters ($C_2$ = yellow zone), includes Bangkalan Regency, Bojonegoro Regency, Bondowoso Regency, Lamongan Regency, Lumajang Regency, Madiun Regency, Magetan Regency, Malang Regency, Mojokerto City, Mojokerto Regency, Nganjuk Regency, Ngawi Regency, Pacitan Regency, Pamekasan Regency, Ponorogo Regency, Probolinggo Regency, Sampang Regency, Situbondo Regency, Sumenep Regency, Trengalek Regency, Tuban Regency, Tulungagung Regency, Batu City, Blitar City, Kediri City, Madiun City, Pasuruan City, and Probolinggo City.

d. Low risk clusters ($C_3$ = green zone), includes Sidoarjo Regency.

The determination of clusters ($cluster_0$, $cluster_1$, $cluster_2$ and $cluster_3$) based on the final centroid results shown by the following image:

Figure 4. Centroid table.

If the results of cluster mapping have been obtained, then the results of the mapping are processed by the C4.5 method for decision tree.

Figure 5. Part of decision tree.

In the decision of tree results, it can be taken an information that the black zone occurs in the number of positive cases smaller than 8348 and greater than 3638 (3638 $> C_0 < 8348$). For more details, watch the picture below.
The last step is to perform a validity test, or in the RapidMiner application program called a performance test using a distance performance cluster operator. Based on the assessment of the Davies-Bouldin Index. Where in this study has an optimal DBI value of 0.356.

**Figure 6.** Description of decision tree.

The value obtained from the decision tree for the cluster is very high (C₀ = black zone) based on C4.5 is if the number of positive cases is smaller than 8348 and greater than 3638 (3638 > C₀ < 8348).

**Figure 7.** Validity test.

### 4. Conclusion

Based on the results of the study it can be concluded that the combination of clustering and classification methods obtained the results of mapping the region in the form of clusters obtained by 7 black zone districts (cluster_0), 1 red zone district (cluster_1), 29 yellow zone districts (cluster_2), and 1 green zone district (cluster_3). The value obtained from the decision tree for the cluster is very high (C₀ = black zone) based on C4.5 is if the number of positive cases is smaller than 8348 and greater than 3638 (3638 > C₀ < 8348).

### Acknowledgment

The authors thank the partners who have helped complete this research so that this study can go well and the University of Jember Research and Community Service Institute (LP2M) which has provided funding for the 2021 fiscal year to the Data Science Research Group, Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Jember. The author also expresses deep gratitude to Allah swt. for still being given health during the work of this research during the covid-19 pandemic.

### References

1. Audina B and Fatekurrohman M 2020 Analisis Survival pada Data Pasien Covid 19 di Kabupaten Jember *Berkala Saintek* 8(4) pp 118-121

2. Azarafza M and Akgun H 2021 Clustering method for spread pattern analysis of coronavirus (COVID-19) infection in Iran *Journal of Applied Science, Engineering, Technology, and Education* 3(1) 1-6

3. Diksa I G B N 2021 Peramalan Gelombang Covid 19 Menggunakan 4 Hybrid Non-Linear Regression Logistic – Double Exponential Smoothing di Indonesia dan Prancis *Jambura Journal of Mathematics* 3(1) 37-51

4. Prasetyo 2012 *Data Mining – Konsep dan Aplikasi Menggunakan Matlab* (Yogyakarta: Andi)

5. Prayoga Y 2020 *Pengertian dan Macam Macam Data Mining* https://tiarway.com/data-mining/macam-macam-data-mining (Accessed April 8th, 2021 at 2:18 PM)
[6] Rembulan G D, Wijaya T, Palullungan D, Alfina K and Qurthuby M 2020 Kebijakan Pemerintah Mengenai Corona virus Disease (COVID-19) di Setiap Provinsi di Indonesia Berdasarkan Analisis Klaster Journal of Industrial and Management System 13(2) 74-86

[7] Yonar H, Yonar A, and Tekindal M A 2020 Modeling and Forecasting for the number of cases of the COVID19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods Eurasian Journal of Medicine and Oncology 4(2) 160-165

[8] Zarikas V, Poulopoulos S G, Gareiou Z, and Zervas E 2020 Clustering Analysis of Countries Using the COVID-19 Cases Dataset Data Brief. 31 2352–3409