

Machine Learning Algorithms for Predicting the Spread of Covid-19 in Indonesia

Syafri Arlis¹, Sarjon Defit¹

Universitas Putra Indonesia YPTK, Padang, Indonesia

Abstract - Coronavirus 2019 or Covid-19 is a major problem for health, and it is a global pandemic that has to be controlled. Covid-19 spread so fast to 196 countries, including Indonesia. The government has to study the pattern and predict its spread in order to make policies that will be implemented to tackle the spread of some of the existing data. Therefore this research was conducted as a precautionary measure against the Covid-19 pandemic by predicting the rate of spread of Covid-19. The application of the machine learning method by combining the k-means clustering algorithm in determining the cluster, k-nearest neighbor for prediction and Iterative Dichotomiser (ID3) for mapping patterns is expected to be able to predict the level of spread of Covid-19 in Indonesia with an accuracy rate of 90%.

Keywords – machine learning, k-means, k-nearest neighbor, Iterative Dichotomiser

1. Introduction

Coronavirus 2019 or Covid-19 is a major problem for health and is a global pandemic that must be controlled. This disease is characterized by symptoms of fatigue, fever, dry cough, sore throat and shortness of breath because it attacks the respiratory system acutely [1], [2], [3].

DOI: 10.18421/TEM102-61

<https://doi.org/10.18421/TEM102-61>

Corresponding author: Syafri Arlis,
Universitas Putra Indonesia YPTK Padang, Indonesia.


Email: syafri_arlis@upiyptk.ac.id

Received: 02 December 2020.

Revised: 09 April 2021.

Accepted: 19 April 2021.

Published: 27 May 2021.

 © 2021 Syafri Arlis & Sarjon Defit; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

The article is published with Open Access at www.temjournal.com

This disease is caused by the corona virus which is a new type of virus which was initially discovered in 2019 in the province of Wuhan, China and has never been identified before as attacking humans (World Health Organization, 2019). This corona virus can develop very quickly, resulting in a more severe infection of organ failure in patients who have a previous medical history [4].

Covid-19 spreads so fast to 196 countries, including Indonesia. Covid-19 was first confirmed on March 2, 2020 and to date as of August 20, 2020, it has reached 147 thousand confirmed positive (cumulatively), of which 6,418 died cases and recovered 101,000 cases.

Based on these conditions and data, every country, including Indonesia, has to be alerted to the threat of the Covid-19 virus pandemic, because there has not been a drug or vaccine for Covid-19. The government has to study the pattern and predict its spread to make policies that will be implemented to tackle the spread of some of the existing data and the tendency of each country or region to have different distribution patterns [5].

Machine learning is to extract and identify useful information and related knowledge from large databases [6], [7]. The application of Machine Learning (ML) and Artificial Intelligence (AI) encourages researchers to provide new perspectives to fight the Covid-19 outbreak [8], [9]. Several studies related to the prediction of Covid-19 spread patterns with machine learning methods, including [10] which utilize machine learning such as the Neural Network algorithm and Naive Bayes in predictions to model the spread of Covid-19. Furthermore, research [11] where the method used was a combination of Support Vector Machine (SVM) and Bayesian Ridge regression where both methods were successful in predicting the total cases of Covid-19 in several countries in the world and the Bayesian Ridge results were better. This research was continued by predicting the Covid-19 pandemic using linear regression and svm methods [12]. Research conducted by Kavadi [13] proposes linear regression and nonlinear machine learning (PDR-NML) methods for predicting a global pandemic from Covid-19.

A study that integrated the Covid-19 epidemiological data into a logistic model and the FbProphet model [14] carried out machine learning-based predictions to obtain an epidemic curve. In the research, R. Sujath et al. [15] proposed a linear regression method, multilayer perceptron and vector auto regression methods in predicting the spread of Covid-19. Determination of the Covid-19 epidemic cluster using the k-means algorithm produces clustering so as to provide a solution in slowing the spread [16] and optimization of decision tree, KNN and SVM methods for prediction in differentiating DNA samples between MERS, SARS, and Covid-19 [17].

In this study, the researchers implemented machine learning with a combination of the k-means, k-nearest Neighbor (KNN) and ID3 algorithms in predicting the spread of Covid-19 in Indonesia, with the aim of providing insight to local governments in making and making policies that will be enforced for can minimize and prevent the spread of more.

2. Materials and Method

The prediction of the pattern of the spread of Covid-19 in Indonesia uses six stages. The stages are problem analysis, data collection, data processing, K-means clustering, prediction with k-Nearest Neighbor, distribution mapping with ID3 and analysis stage. The research method used in this research can be seen in Figure 1.

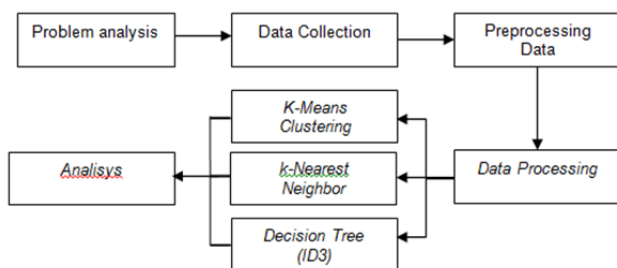


Figure 1. Stage of process in the research

Problem Analysis

Problem analysis is carried out as a reference in making the system, namely predicting the pattern of the spread of Covid-19 in Indonesia to determine the cluster for each province. The method used is a combination of machine learning techniques, namely the k-means algorithm to obtain a mapping or cluster, after obtaining a cluster from the previous data distribution pattern which is used as training data, after that an estimate or forecast is carried out by adding testing data with the k-nearest neighbor algorithm (KNN) and the ID3 algorithm to see the distribution mapping decision tree, where the results of this analysis will be made to make a prediction system for the distribution of Covid-19 for each province belonging to which zone or cluster.

Data Collection

The data used were obtained from the national Covid-19 task force which was accessed on 23 August 2020. From the data obtained, zone or cluster mapping must be carried out. The data used in this study can be seen in Table 1.

Table 1. Data Spread of Covid-19 in Indonesia

No	Province Name	Positive	Recovered	Died
1	Aceh	1211	191	32
2	Bali	4513	3953	52
3	Bangka Belitung	228	201	2
4	Banten	2544	1749	103
5	Bengkulu	288	164	24
6	Central Java	12476	7989	846
7	East Java	30315	23632	2172
8	West Sumatera	1633	1043	49
9	Central Kalimantan	2410	1805	104
10	East Kalimantan	3101	1910	121
11	North Kalimantan	349	311	2
12	Riau Ilands	745	465	31
13	Lampung	362	283	14
14	DI Yogyakarta	1193	826	33
15	DKI Jakarta	33470	23567	1097
16	Gorontalo	1959	1515	50
17	West Kalimantan	566	448	4
18	South Kalimantan	7777	5383	333
19	Jambi	274	131	5
20	West Java	9283	5668	259
21	Maluku	1669	997	31
22	North Maluku	1774	1494	61
23	NTB	2582	1879	143
24	NTT	171	149	2
25	Papua	3567	2277	42
26	West Papua	649	518	8
27	Riau	1237	767	19
28	West Sulawesi	352	237	7
29	South Sulawesi	11470	8461	349
30	Central Sulawesi	238	206	8
31	Southeast Sulawesi	1323	900	22
32	North Sulawesi	3552	2421	149
33	South Sumatera	4125	2820	227
34	North Sumatera	6129	3140	279

Data Preprocessing

In this study, data analysis was based on cases of the spread of Covid-19, confirmed positive cases, data recovered and data on deaths in Indonesia.

K-Means Clustering Implementation

Data analysis using the K-Means algorithm to obtain clustering information is as follows:

1. Determine the value of k as the number of clusters to be formed;
2. Generate a random initial k centroid (cluster center point);

3. Calculate the distance of each data to each centroid using the correlation formula between two objects, namely Euclidean Distance;
4. Group each data based on the closest distance between the data and the centroid;
5. Determine the position of the new centroid (C_k) by calculating the average value of data on the same centroid

$$C_k = \left(\frac{1}{n_k}\right) \sum d_t$$

where n_k is the number of documents in cluster k and d_t is the documents in cluster k ;

6. Return to step 3 if the position of the new centroid with the old centroid is not same.

The manual completion stages of the data in Table 1 are:

1. Determine the initial center of the cluster "Centroid".

For the determination of the initial center is taken from the value:

- a. Cluster Center 1 : {1633; 1043; 49}
- b. Cluster Center 2 : {1669; 997; 31}
- c. Cluster Center 3 : {3552; 2421; 149}

Based on these things, the number of clusters used in this study are 3 labels, namely low cluster ($C1 =$ green zone), alert cluster ($C2 =$ yellow zone), high cluster ($C3 =$ red zone)

2. Calculation of cluster center distance
Euclidian Distance is used to measure the distance between the data and the cluster center, then the distance matrix will be obtained as follows:

$$D_e = \sqrt{(M_{ix} - C_{ix})^2 + (M_{iy} - C_{iy})^2}$$

Determine the position of the new centroid (C_k) by calculating the average value of the data on each of the same centroids.

$$C_x = \left(\frac{1}{n_k}\right) \sum d_t$$

K-Nearest Neighbor (KNN) and ID3

KNN is one of the supervised machine learning methods which is capable of solving various problems flexibly [18]. The results of mapping or clusters using the k-means method can be used as needed so as to gain knowledge such as estimating or predicting by combining machine learning

techniques such as the k-nearest neighbor algorithm and making a decision tree to obtain information with the ID3 algorithm. On the Table 3 is data testing for testing from the previous cluster mapping data with the K-Means algorithm. Analysis of the data testing stages uses the k-Nearest Neighbor (KNN) algorithm to obtain estimates or predictions.

Table 2. Data Testing

Province Name	Positive	Recovered	Died	Cluster
West Sumatera	6833	4943	549	?
Lampung	2362	1283	254	?
Jambi	1874	1631	189	?

The next step is to normalized the data with formula:

$$\text{Normalization} = \frac{\text{data}_x - \text{data}_{\min}}{\text{data}_{\max} - \text{data}_{\min}}$$

Table 3. Results of the Data Training Normalization Process

Positive	Recovered	Died
0.03	0.00	0.01
0.13	0.16	0.02
0.00	0.00	0.00
0.07	0.07	0.05
0.00	0.00	0.01
0.37	0.33	0.39
0.91	1.00	1.00
0.04	0.04	0.02
∴	∴	∴
0.18	0.13	0.13

Table 4. Results of the Data Testing Normalization Process

Positive	Recovered	Died
0.10	0.08	0.07
0.04	0.05	0.05
0.03	0.04	0.02

After normalizing the data, then calculating the euclidean distance, using the formula:

$$d(x,y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

3. Result and Discussion

Table 5 shows the results of the clustering of the k-means clustering algorithm using the Covid-19 spread data for all provinces in Indonesia with 4 iterations.

Table 5. Results of the clustering of the k-means clustering algorithm using the Covid-19 spread data.

No	Province Name	Positive	Recovered	Dead	Cluster	No
1	Aceh	1211	191	32	1	1
2	Bali	4513	3953	52	1	2
3	Bangka Belitung	228	201	2	1	3
4	Banten	2544	1749	103	1	4
5	Bengkulu	288	164	24	1	5
6	Central Java	12476	7989	846	2	6
7	East Java	30315	23632	2172	3	7
8	West Sumatera	1633	1043	49	1	8
9	Central Kalimantan	2410	1805	104	1	9
10	East Kalimantan	3101	1910	121	1	10
11	North Kalimantan	349	311	2	1	11
12	Riau Ilands	745	465	31	1	12
13	Lampung	362	283	14	1	13
14	DI Yogyakarta	1193	826	33	1	14
15	DKI Jakarta	33470	23567	1097	3	15
16	Gorontalo	1959	1515	50	1	16
17	West Kalimantan	566	448	4	1	17
18	South Kalimantan	7777	5383	333	2	18
19	Jambi	274	131	5	1	19
20	West Java	9283	5668	259	2	20
21	Maluku	1669	997	31	1	21
22	North Maluku	1774	1494	61	1	22
23	NTB	2582	1879	143	1	23
24	NTT	171	149	2	1	24
25	Papua	3567	2277	42	1	25
26	West Papua	649	518	8	1	26
27	Riau	1237	767	19	1	27
28	West Sulawesi	352	237	7	1	28
29	South Sulawesi	11470	8461	349	2	29
30	Central Sulawesi	238	206	8	1	30
31	Southeast Sulawesi	1323	900	22	1	31
32	North Sulawesi	3552	2421	149	1	32
33	South Sumatera	4125	2820	227	1	33
34	North Sumatera	6129	3140	279	2	34

Based on Table 5, it can be seen that the mapping where the medium cluster (c0) consists of 5 provinces, the low cluster (c1) consists of 27 provinces and the high cluster (c2) consists of 2 provinces. Following are the complete results of cluster mapping on the number of cases of the Covid-19 pandemic in Indonesia:

- a. Medium clusters (c0 = yellow zone) are Central Java, South Kalimantan, West Java, South Sulawesi and North Sulawesi;
- b. The low cluster (c1 = green zone) is Aceh, Bali, Bangka Belitung, North Kalimantan, Banten, Bengkulu, West Sumatra, West Kalimantan, Central Kalimantan, East Kalimantan, Riau Islands, Lampung, East Nusa Tenggara, DI Yogyakarta, Gorontalo, Sulawesi West, Jambi, Maluku, Riau, North Maluku, West Nusa Tenggara, Papua, West Papua, Central Sulawesi, Southeast Sulawesi, North Sulawesi and South Sumatra;
- c. The High Cluster (c2 = red zone) is DKI Jakarta and East Java.

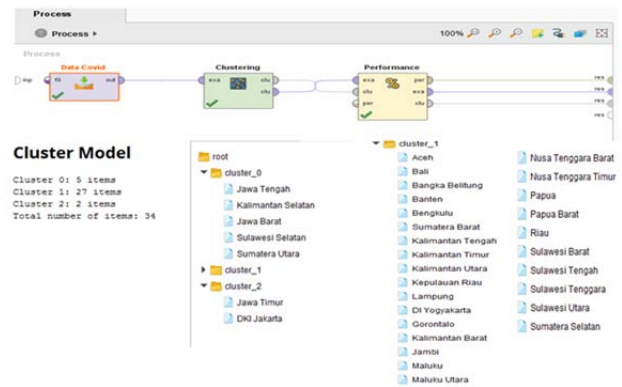


Figure 2. Testing the K-Means Method with Rapidminer

In cluster mapping (c0, c1 and c2) based on the final centroid values are shown in Figure 3.

Attribute	cluster_0	cluster_1	cluster_2
Positif	9427	1578.333	31892.500
Sembuh	6128.200	1098.519	23599.500
Meninggal	413.200	49.815	1634.500

Figure 3. Cluster mapping based on final centroid values

accuracy: 90.00%

Row ...	prediction(Label)	confidence(Cluster-1)	confidence(Cluster-2)	confide...	Nama Propi...	Positif	Sembuh	Meninggal
1	Cluster-2	0.370	0.630	0	Sumatera Ba...	6833	4943	549
2	Cluster-1	1	0	0	Lampung	2362	1283	254
3	Cluster-1	1	0	0	Jambi	1874	1631	189

Figure 4. Estimation Results with the KNN

Based on the results of the estimation of the testing data, it can be concluded that positive data ≥ 6129 are classified in cluster-2 and < 2544 are classified as cluster-1. The following is the test results of estimation or prediction analysis with Rapid Miner 9.5.001, entering training data from the results of previous k-means data analysis on new data or testing and combining the mapping results with the ID3 method in Figure 5.

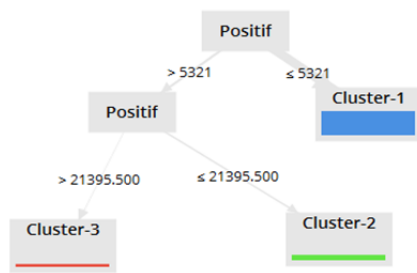


Figure 5. Mapping with the ID3 Method

4. Conclusion

The results of the research from a combination of machine learning methods using the k-means algorithm in determining clusters, then prediction and mapping of distribution patterns with k-nearest neighbor (KNN) and Iterative Dichotomiser (ID3) can be applied to the case of the Covid-19 pandemic spread in Indonesia with 90% accuracy rate. Various attempts were made to improve accuracy, especially in data mining. This research can be developed to obtain a better accuracy value by using other methods or algorithms, both in determining clusters, predicting or finding mapping patterns.

References

- [1]. Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., ... & Peng, Z. (2020). Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *Jama*, 323(11), 1061-1069.
- [2]. Liu, K. C., Xu, P., Lv, W. F., Qiu, X. H., Yao, J. L., Gu, J. F., & Wei, W. (2020). CT manifestations of coronavirus disease-2019: a retrospective analysis of 73 cases by disease severity. *European journal of radiology*, 126, 108941.
- [3]. Ahamad, M. M., Aktar, S., Rashed-Al-Mahfuz, M., Uddin, S., Liò, P., Xu, H., ... & Moni, M. A. (2020). A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. *Expert systems with applications*, 160, 113661.
- [4]. Watratan, A. F., & Moeis, D. (2020). Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia. *Journal of Applied Computer Science and Technology*, 1(1), 7-14.
- [5]. Santosh, K. C. (2020). COVID-19 prediction models and unexploited data. *Journal of medical systems*, 44(9), 1-4.
- [6]. Saifudin, A. (2018). Metode Data Mining untuk Seleksi Calon Mahasiswa pada Penerimaan Mahasiswa Baru di Universitas Pamulang. *Jurnal Teknologi*, 10(1), 25-36.
- [7]. Beunza, J. J., Puertas, E., García-Ovejero, E., Villalba, G., Condes, E., Koleva, G., ... & Landecho, M. F. (2019). Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). *Journal of biomedical informatics*, 97, 103257.
- [8]. Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons & Fractals*, 110059.
- [9]. Pinter, G., Felde, I., Mosavi, A., Ghamisi, P., & Gloaguen, R. (2020). COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach. *Mathematics*, 8(6), 890.
- [10]. Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., ... & Atkinson, P. M. (2020). Covid-19 outbreak prediction with machine learning. *Algorithms*, 13(10), 249.
- [11]. Parhusip, H. A. (2020). Study on COVID-19 in the World and Indonesia Using Regression Model of SVM, Bayesian Ridge and Gaussian. *Jurnal Ilmiah Sains*, 20(2), 49-57.
- [12]. Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B. W., Aslam, W., & Choi, G. S. (2020). COVID-19 future forecasting using supervised machine learning models. *IEEE access*, 8, 101489-101499.
- [13]. Kavadi, D. P., Patan, R., Ramachandran, M., & Gandomi, A. H. (2020). Partial derivative nonlinear global pandemic machine learning prediction of covid 19. *Chaos, Solitons & Fractals*, 139, 110056.
- [14]. Wang, P., Zheng, X., Li, J., & Zhu, B. (2020). Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos, Solitons & Fractals*, 139, 110058.
- [15]. Sujath, R., Chatterjee, J. M., & Hassanien, A. E. (2020). A machine learning forecasting model for COVID-19 pandemic in India. *Stochastic Environmental Research and Risk Assessment*, 34, 959-972.
- [16]. Maugeri, A., Barchitta, M., & Agodi, A. (2020). A Clustering Approach to Classify Italian Regions and Provinces Based on Prevalence and Trend of SARS-CoV-2 Cases. *International Journal of Environmental Research and Public Health*, 17(15), 5286.
- [17]. B. Al Kindhi, "Optimization of machine learning algorithms for predicting infected COVID-19 in isolated DNA," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 4, pp. 423-433, 2020.
- [18]. Chen, S. B., Xu, Y. L., Ding, C. H., & Luo, B. (2018). A Nonnegative Locally Linear KNN model for image recognition. *Pattern Recognition*, 83, 78-90.