Simple Correlation Between Weather and COVID-19 Pandemic Using Data Mining Algorithms

Ari Fadli1,3, Azis Wisnu Widhi Nugraha1, Muhammad Syaiful Aliim1, Acep Taryana1, Yogiek Indra Kurniawan2, Widhiatmoko Herry Purnomo1

1Electrical Engineering Department, Universitas Jenderal Soedirman, Indonesia
2Informatics Department, Universitas Jenderal Soedirman, Indonesia
3arifadli@unsoed.ac.id

Abstract. Currently, Indonesia is still struggling against the Coronavirus, the number of cases of the Coronavirus is increasing every day. The virus that began to attack around December 2019 has become a pandemic and is causing serious problems because there is still no vaccine to prevent transmission. The Government of the Republic of Indonesia as the party responsible for this pandemic has made various efforts including the lockdown method in the form of a Large-Scale Social Restriction (PSBB) policy. Besides that, another effort is to present a digital information portal that can be accessed and able to present the development status of coronavirus patients. With the development of information and communication technology, this research is a find correlation between weather and Covid-19 Using Data Mining Algorithm in processing data sources from the government to produce knowledge that can be used to make decisions. The research data used came from Covid-19 surveillance data of the Ministry of Health of the Republic of Indonesia and the climatic conditions in Surabaya Indonesia in the form of the average temperature (°C), the average humidity (%), the average duration of sunlight (hour) taken from the Department of Meteorology of the Republic of Indonesia. Based on the results of the study it was found that the weather affected the number of Covid-19 patients. The highest order of correlation is the average duration of sunlight (hour), the average temperature (°C), the average humidity (%).

1. Introduction

Coronavirus (COVID-19) is a new type of virus that was first identified in Wuhan, China in December 2019. The virus which then spread rapidly throughout China and all countries in the world has been declared a pandemic by the World Health Organization (WHO) [1]. Based on a study entitled "The Novel Corona Virus Originating in Wuhan, China: Challenges for Global Health Governance.," Said that on December 31, 2019, China reported a case of pneumonia, in Wuhan, Hubei Province, China, to the World Health Organization (WHO) is caused by the 2019-nCoV virus [2].

In Indonesia, the first COVID-19 case was identified on March 2, 2020. Until March 17, 2020, there were 134 confirmed cases from eight provinces, namely Bali, Banten, DKI Jakarta, West Java, Central Java, West Kalimantan, North Sulawesi, and Yogyakarta [3]. Figure 1 shows the distribution area of COVID-19 in Indonesia.

Some research in the field of Artificial Intelligence to overcome the Covid-19 epidemic is to use a model approach to identify the administration of "drug-repurposing" drugs to infected patients [4]. Other
studies use generative models to obtain compound entities to fight coronavirus COVID-19 [5]. Other research in the field of Deep Learning to make computed tomography (CT) image processing models, the results of the study show that the model is proposed when using 499 training data and 131 test data obtained accuracy reaches 0.901 accuracies with positive predictive value 0.840 and negative predictive value 0.982 [6]. In line with this research, Deep Learning shows high performance because they can achieve an Area Under Curve (AUC) result of 0.994 with 94% sensitivity and 98% specificity (at threshold 0.5). [7].

![Geographic distribution of confirmed COVID-19 cases in Indonesia, as of 26 March 2020.](https://www.covid19.go.id/2020/03/26/2112/)

In other studies using the Naive Method algorithm, the prediction results obtained from the classification of COVID-19 infected status obtained positive and negative data with each data that is positive 55.48% of 249 data and negative 44.52% of 249 data. In other words, if classified as a number, positive classification is 139 people and negative classification is 110 people from a total of 249 data. [8]. Besides, research using the C4.5 Algorithm produced a test with a confusion matrix of 3 (three) classes from the COVID-19 supervision category, the calculation yielded an accuracy of 0.9286 (92.86%) [9].

Currently, the topic of research on the correlation between weather conditions and the spread of viruses is very interesting, some of which are researchers [10-15]. In the case of the West Nile virus in the United States and Europe [16-17] studied the relationship between climatic conditions and SARS-CoV and obtained results that climate variables can be a cause of biological interactions between SARS-CoV and humans. Other research results also state that weather correlates very significantly with changes in death rates due to pneumonia [18]. Viral transmission is caused by several factors such as temperature, humidity, and population density affected [19]. Tosepu studied the correlation between weather conditions and COVID-19 Jakarta, Indonesia, and found significant results [20].

In this study, the ID3 algorithm was used to find a correlation between the number of confirmed COVID-19 patients per day with changes in weather conditions. The number of COVID-19 patients used came from COVID-19 surveillance data of the Ministry of Health of the Republic of Indonesia and the weather change variables used were the average temperature (°C), the average humidity (%), and the average duration of sunlight (hour) in Surabaya taken from Department of Meteorology, Republic of Indonesia.
2. Material and Methods

2.1. Research Design

Data mining (knowledge discovery in databases (KDD)) is an activity of collecting historical data to find sequences, patterns, or variable relationships in large data. This data mining proceeds can be used to help make decisions in the future [21][22].

One of the advantages of the decision tree algorithm is its ability to produce easy-to-understand knowledge structures with low computational costs for classifying new cases, the ability to handle symbolic and numeric input variables, providing a clear indication of which attributes are most important for prediction or classification [23].

In this study, we used the Decision Tree ID3 algorithm to classify the data and extract the rules for the data set, as shown in Figure 2.

![ID3 Algorithm Diagram](image)

**Figure 2. ID3 Algorithm**

The decision tree resembles a tree structure. The tree is a hierarchical organization of node and link collection, where each node, except the root node, has one incoming link. Each node is a predictive feature and the link represents its respective value conditional variable. The decision rules of the decision tree are a way of representing knowledge. Based on the decision tree and based on testing can build decision rules.
The ID3 algorithm does not guarantee an optimal solution. It selects the best attributes from the given set. It then divides the dataset across each iteration. ID3 can adjust training data. As a solution for this is, rather than a larger tree, a smaller tree is preferred. Although this algorithm determines a solution, it does not always guarantee an optimal solution.

2.2. Data Collection
Figure 3 shows the distribution map of Covid-19 Patients in Surabaya.

![Figure 3. Distribution map of Covid-19 in Surabaya City](image)

The number of cases confirmed every day is collected from the official website of the city of Surabaya on page https://lawancovid-19.surabaya.go.id/visualisasi. The computerized data set at weather data was obtained from the Department of Meteorology, Ministry of the Republic of Indonesia. Data consists of a temperature average (°C), humidity average (%), duration of sunlight average (ss). The data used are from 1 June 2020 to 1 July 2020.

2.3. Study Area
This study was conducted in Surabaya. Surabaya is one of the major cities in Indonesia. The city of Surabaya as the capital of East Java Province is located on the north coast of East Java Province or precisely between 7 ° 9′ -7 ° 21 ′ South Latitude and 112 ° 36 ′ - 112 ° 54 ′ East Longitude.

2.4. ID3 Algorithm
The ID3 (Iterative DiChaudomiser 3) algorithm was first developed by J. Ross Quinlan at the University of Sydney [24]. He first presented ID3 in 1975 in a book, Machine Learning [24], volume 1, number 1. ID3 is based on the Concept Learning System (CLS) algorithm. ID3 is a supervised learning algorithm, [25] constructing a decision tree. The resulting tree is used to classify future samples. The ID3 algorithm builds a tree based on information (information acquisition) obtained from training instances and then uses the same to classify test data [25]. ID3 algorithm generally uses nominal attributes for classification without missing values. Figure 4 shows the ID3 algorithm pseudocode [26].
The selection of attributes in ID3 is used by the statistical property, which is called information gain. Gain is used to measuring how well an attribute separates training data into the target class. The attribute with the highest information will be selected. The gain value is calculated using the concept of information theory called entropy. Entropy is calculated using equation (1).

\[
\text{Entropy}(S) = -P_a \log_2 P_a - P_b \log_2 P_b \quad \ldots \ldots \ldots \ldots (1)
\]

Based on the formula above, \(P_a\) is the probability of sample S having a positive class. \(P_a\) is calculated by dividing the number of positive samples \((S+)\) by the total number of samples \((S)\) such that \(P_a = S+/S\). \(P_b\) is the probability of sample S having a negative class. \(P_b\) is calculated by dividing the number of negative samples \((S-)\) by the total sample number \((S)\) so that \(P_b = S-/S\). The leaf part of a decision tree. In the ID3 algorithm, the entropy reduction is called gain information. A comparison of sample S with attribute A can be calculated information gain using equation (2).

\[
\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{V \in \text{Null}(A)} \frac{S_V}{S} \text{Entropy}(S_V) \quad \ldots \ldots \ldots \ldots (2)
\]
3. Results and Discussion
Figure 5 shows several patients obtained from the official website of the city of Surabaya on page https://lawancovid-19.surabaya.go.id/visualisasi.

![Wheater Variable](image1)

**Figure 5. Wheater Variable**

![Number of Covid-19 Patients](image2)

**Figure 6. Number of Covid-19 Patients**

Figure 6 shows the weather variable used was obtained from the Ministry of Health of the Republic of Indonesia and climatic conditions in the city of Surabaya in the form of temperature average (°C), humidity average (%), duration of sunlight average (ss) taken from the Department of Meteorology of the Republic of Indonesia. Data will be tested using the Decision Tree method with the ID3 algorithm.
using the Rapid Miner tool. The data used are from 1 June 2020 to 1 July 2020. Table 1 shows the dataset used in this study.

Table 1. Dataset

| Tanggal   | Tavg | RH_avg | ss  | Kelas |
|-----------|------|--------|-----|-------|
| 01-06-2020 | 28 - 28.9 | 8.60   | 1.20 | No    |
| 02-06-2020 | 28 - 28.9 | 7.70   | 0.20 | Yes   |
| 03-06-2020 | 29 - 29.9 | 7.50   | 3.70 | Yes   |
| 04-06-2020 | 28 - 28.9 | 7.60   | 7.40 | No    |
| 05-06-2020 | 30 - 30.9 | 7.10   | 8.30 | Yes   |
| 06-06-2020 | 28 - 28.9 | 7.80   | 8.60 | No    |
| 07-06-2020 | 28 - 28.9 | 7.80   | 7.30 | Yes   |
| 08-06-2020 | 28 - 28.9 | 7.60   | 4.40 | Yes   |
| 09-06-2020 | 28 - 28.9 | 7.70   | 8.20 | Yes   |
| 10-06-2020 | 28 - 28.9 | 7.30   | 8.00 | No    |
| 11-06-2020 | 28 - 28.9 | 7.50   | 8.40 | No    |
| 12-06-2020 | 28 - 28.9 | 7.20   | 8.50 | Yes   |
| 13-06-2020 | 28 - 28.9 | 7.20   | 9.20 | No    |
| 14-06-2020 | 28 - 28.9 | 8.50   | 7.60 | No    |
| 15-06-2020 | 29 - 29.9 | 8.20   | 1.50 | No    |
| 16-06-2020 | 29 - 29.9 | 7.90   | 4.30 | No    |
| 17-06-2020 | 28 - 28.9 | 7.70   | 7.70 | Yes   |
| 18-06-2020 | 29 - 29.9 | 7.40   | 8.80 | Yes   |
| 19-06-2020 | 29 - 29.9 | 7.70   | 8.00 | Yes   |
| 20-06-2020 | 28 - 28.9 | 7.60   | 8.40 | Yes   |
| 21-06-2020 | 28 - 28.9 | 7.50   | 4.30 | No    |
| 22-06-2020 | 29 - 29.9 | 7.60   | 3.00 | No    |
| 23-06-2020 | 28 - 28.9 | 7.50   | 8.40 | Yes   |
| 24-06-2020 | 28 - 28.9 | 7.40   | 8.30 | Yes   |
| 25-06-2020 | 28 - 28.9 | 7.10   | 8.10 | Yes   |
| 26-06-2020 | 27 - 27.9 | 7.20   | 8.00 | No    |
| 27-06-2020 | 28 - 28.9 | 7.00   | 8.00 | No    |
| 28-06-2020 | 27 - 27.9 | 7.10   | 8.30 | Yes   |
| 29-06-2020 | 28 - 28.9 | 7.60   | 8.20 | No    |
| 30-06-2020 | 28 - 28.9 | 7.20   | 8.00 | Yes   |
| 01-07-2020 | 28 - 28.9 | 7.30   | 8.10 | Yes   |

The strong motivation for this research is because it is interesting to conduct a study on COVID-19 in Indonesia. To achieve this purpose, we are looking for a correlation between weather variables, namely the length of the sun's exposure, average temperature and average humidity for confirmed cases of COVID-19 using the ID3 algorithm. This algorithm was built to understand the variables that most influence the COVID-19 growth curve.
Based on Figure 4, the ID3 Algorithm finds that in a row the weather variables that affect the Covid-19 growth curve are the average duration of sunlight (hours), average temperature (°C), average humidity (%), The resulting decision tree is shown in Figure 7.

**Figure 7. Decision Tree**

Figure 8 is the confusion matrix used to evaluate the performance of the ID3 algorithm in producing a model to describe the correlation between weather variables and Covid-19. Based on the confusion matrix, the accuracy value is 96.67%.

**Figure 8. Confusion Matrix**

4. Conclusion
   Weather is an important factor in determining the number of confirmed Covid-19 cases in Surabaya. Weather variables such as the average duration of sunlight (hours), average temperature (°C), average humidity (%) was significantly correlated with the number of patients confirmed cured with successive gain values of 0.683, 0.450, 0.030, and with accuracy 96.77%.

5. References
   [1]. LI, Q., GUAN, X., WU, P., WANG, X., ZHOU, L., TONG, Y., REN, R., LEUNG, K.S., LAU, E.H., WONG, J.Y. AND XING, X., 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*.
   [2]. Phelan AL1,2, Katz R1, Gostin LO2. (2020), Novel Coronavirus Originating in Wuhan, China: Challenges for Global Health
   [3]. KOMPAS, 2020. UPDATE: Tambah 17 Pasien, Total Ada 134 Kasus Positif Virus Korona. [online] Tersedia pada: KOMPAS.COM <https://gaya.tempo.co/read/877228/9-jutaorang-di-indonesia-mengalami-depresi/full&view=ok>. Last Accessed on 17th March 2020
   [4]. B. R. Beck, B. Shin, Y. Choi, S. Park, and K. Kang, “Predicting commercially available antiviral drugs that may act on the novel coronavirus (2019-nCoV), Wuhan, China through a drug-target interaction deep learning model,” BioRxiv, 2020
   [5]. A. Zhavoronkov, V. Aladinskiy, A. Zhebrak, B. Zagribelnyy, V. Terentiev, D. S. Bezruk, D. Polykovskiy, R. Shayakhmetov, A. Filimonov, P. Orekhov, Y. Yan, O. Popova, Q. Vanhelen,
A. Aliper, and Y. Ivanenkov, “Potential COVID-2019 3C-like protease inhibitors designed using generative deep learning approaches.,” ChemRxiv, 2 2020.

[6]. C. Zheng, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, and X. Wang, “Deep learning-based detection for COVID-19 from chest CT using weak label,” MedRxiv, 2020

[7] O. Gozes, M. Frid-Adar, N. Sagie, H. Zhang, W. Ji, and H. Greenspan, “Coronavirus detection and analysis on chest CT with deep learning,” arXiv preprint arXiv:2004.02640, 2020.

[8]. Santoso P, Fauziah F, Nurhayati N. Application of Data Mining Classification for Covid-19 Infected Status Using Algoritma Naive Method. mantik [Internet]. 1May2020 [cited 18Jul.2020];4(1, May):267-75. Available from: http://iocscience.org/ejournal/index.php/mantik/article/view/740

[9]. Wildan Wiguna1; Dwiza Riana. DIAGNOSIS OF CORONAVIRUS DISEASE 2019 (COVID-19) SURVEILLANCE USING C4.5 ALGORITHM, Jurnal PILAR Nusa Mandiri Vol. 16, No. 1 March 2020

[10]. Wang, Y., Wang, Y., Chen, Y., & Qin Q. (2020). Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (COVID-19) implicate special control measures. Journal of Medical Virology, 92(6), 568–576. http://dx.doi.org/10.1002/jmv.25748.

[11]. Tomar, A., & Gupta, N. (2020). Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. Science of the Total Environment, 725, 138762.

[12]. Gupta, S., Naghwanshi, G. S., & Chanda, A. (2020). Effect of weather on COVID-19 spread in the US: A prediction model for India in 2020. Science of the Total Environment, 728(2020), 138860.

[13]. Ma, Y., Zhao, Y., Liu, J., He, X., Wang, B., Fu, S., et al. (2020). Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. Science of The Total Environment, 724, 138226. http://dx.doi.org/10.1016/j.scitotenv.2020.138226.

[14]. Poole, L. (2020). Seasonal influences on the spread of SARS-CoV-2 (COVID19), causality, and forecastability (March 15, 2020). SSRN Electronic Journal, http://dx.doi.org/10.2139/ssrn.3554746.

[15]. Sajadi, M. M., Habibzadeh, P., Vintzileos, A., Shokouhi, S., Miralles-Wilhelm, F., & Amoroso, A. (2020). Temperature and latitude analysis to predict potential spread and seasonality for COVID-19. SSRN Electronic Journal, http://dx.doi.org/10.2139/ssrn.3550308.

[16]. Epstein, P. R. (2001). West Nile virus and the climate. Journal of Urban Health, 78, 367–371.

[17]. Yuan, J., Yun, H., Lan, W., Wang, W., Sullivan, S. G., Jia, S., & Bittles, A. H. (2006). A climatologic investigation of the SARS-CoV outbreak in Beijing, China. American Journal of Infection Control, 34(4), 234–236. http://dx.doi.org/10.1016/j.ajic.2005.12.006.

[18]. Bull, G. (1980). The weather and deaths from pneumocemia. Lancet, 315, 1405–1408.

[19]. Dalziel, B. D., Kissler, S., Gog, J. R., Viboud, C., Bjørnstad, O. N., & Metcalf, C. J. E. (2018). Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities. Science, 362, 75–79.

[20]. Tosepu, R., Gunawan, J., Effendy, D. S., Imran Ahmad, L. O. A., Lestari, H., Bahar, H., & Asfian, P. (2020). Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. Science of the Total Environment, 725, 138436.

[21]. H. Siahaan, H. Mawengkang, S. Efendi, A. Wanto, and A. P. Windarto, “Application of Classification Method C4.5 on Selection of Exemplary Teachers,” in IOP Conference Series, 2018, pp. 1–6.

[22]. M. I. Zulfa, A. Fadli, and Y. Ramadhani, “Classification model for graduation on time study using data mining techniques with SVM algorithm,” AIP Conf. Proc., vol. 2094, 2019.

[23]. S. Sudirman, A. P. Windarto, and A. Wanto, “Data Mining Tools | RapidMiner: K-Means Method on Clustering of Rice Crops by Province as Efforts to Stabilize Food Crops In Indonesia,” IOP Conference Series: Materials Science and Engineering, vol. 420, no. 12089, pp. 1–8, 2018.
[24]. Ankur Shrivastava and Vijay Choudhary, Comparison between ID3 and C4.5 in Contrast to IDS
Surbhi Hardikar, VSRD-IJCSIT, Vol. 2 (7), 2012, 659-667.
[25]. J.R. QUINLAN, Induction of Decision Trees, 1986, Machine Learning 1:81-106
[26]. Hssina Badr, Hanane Ezzikouri Mohammed Erritali “A comparative study of decision tree ID3
and C4.5, (IJACSA) International Journal of Advanced Computer Science and Applications,
Special Issue on Advances in Vehicular Ad Hoc Networking and Applications, 2014

Acknowledgments
We would like to thank LPPM Unsoed for research funding in the scheme Riset Dosen Pemula.