International Journal Of Artificial Intelegence Research
Vol 4, No 2, December 2020, pp. 151 - 161

ISSN: 2579-7298

Predicting the Spread of the Corona Virus (COVID-19) in Indonesia: Approach Visual Data Analysis and Prophet Forecasting

Amir Mahmud Husein[a,1*], Jefri Poltak Hutabarat [a,2], Jeckson Edition Sitorus[a,3], Tonazisokhi Giawa[a,4], Mawaddah Harahap[a,5]

[a] *Univeritas Prima Indonesia, Fakultas Teknologi dan Ilmu Komputer, Program Studi Teknik Informatika*
[1] *amirmahmud@unprimdn.ac.id**
* *corresponding author*

ARTICLE INFO

Article history:
Received 23 Sep 2020
Revised 21 Oct 2020
Accepted 16 Dec 2020

Keywords:
Exploratory Data Analysis
Prophet
Coronavirus
COVID-19
Forecasting

ABSTRACT

The development trend of the coronavirus pandemic (COVID-19) in various countries has become a global threat, including in Southeast Asia, such as Indonesia, the Philippines, Brunei, Malaysia, and Singapore. In this paper, we propose an Exploratory Data Analysis (EDA) model approach and a time series forecasting model using the Prophet method to predict the number of confirmed cases and cases of death in Indonesia in the next thirty days. We apply the EDA model to visualize and provide an understanding of this pandemic outbreak in various countries, especially in Indonesia. We present the trends in the spread of epidemics from the countries of China from which the virus originates, then mark the top ten countries and their development and also present the trends in Asian countries. We present an analytical framework comparing the predicted results with the actual data evaluated using the MAPE and MAE models, where the prophet algorithm produces good performance based on the evaluation results, the relative error rate of our estimate (MAPE) is around 6.52%, and the model average false 52.7% (MAE) for confirmed cases, while case mortality was 1.3% for the MAPE and MAE models around 236.6%. The results of the analysis can be used as a reference for the Indonesian government in making decisions to prevent its spread in order to avoid an increase in the number of deaths

I. Introduction

The coronavirus pandemic (COVID-19) is one of the viruses transmitted through respiratory infections that can lead to death, this virus was first detected in China, Wuhan City, Hubei Province in December 2019, and on January 30, 2020, the COVID-19 outbreak. declared a Health Emergency by WHO [1]. As of April 2020, as many as 120 countries have reported approximately 2 million cases with 195,755 people dying and more than 781,109 people recovered, in Indonesia this case first appeared in March 2020, recorded more than 1,000 confirmed cases, with a mortality rate. up to 8.8% spread across 34 Provinces and have enacted travel restrictions, school closings to break the chain of the spread of this pandemic. The National Disaster Management Agency, Task Force for the Acceleration of Handling COVID-19 of the Republic of Indonesia reported that dated April 12 2020 at 16.00 WIB, the spread of this virus in Indonesia was 4,241 confirmed, under treatment 3,509, 359 people recovered and died 373, this will have a severe economic impact, thus requiring stricter policies and plans for predicting confirmed cases in the coming days to limit the growth factors associated with the increase in the number of cases.

The application of a mathematical model to estimate the growth of the spread of COVID-19 has attracted the interest of many researchers by proposing various approaches, such as [2] predicting the severity of transmission rates in various regions of Italy[3], studying the factors[4] proposed a

DOI: 10.29099/ijair.v5i1.192

Machine Learning approach to predict outbreak activity in China and [5] proposed a SIR model to study epidemic development in India, the Susceptible-Infectious-Quarantined Model -Recovered (SIQR) was proposed [6] for data analysis from the Brazilian Ministry of Health and Zhou et al [7] proposed a Logistics model and an SEIR model. Gupta & Pal [8] proposed an ARIMA model based on exploratory data analysis for prediction of outbreak trends in India, whereas [9] proposed a timely and short-term forecast [10] by proposing a correlation model for the growth of the legal power of COVID-19 on four continents and the inefficiency of quarantine strategies. Recently [1] proposed a Neural network (NN) approach to the Long short-term memory (LSTM) model to predict the parameters, risks, and effects of an epidemic, whereas [12] applying a modified ANFIS model.

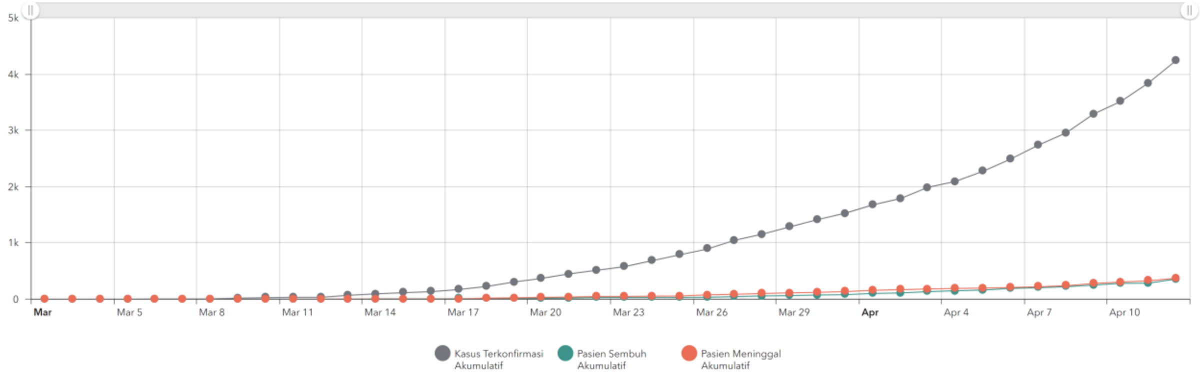


Fig. 1. Coronavirus Case Statistics in Indonesia

Source: http://covid19.bnpb.go.id

Several research results have concluded that countries that do not implement quarantine the number of cases will grow exponentially [13], but predicting a pandemic requires an analysis of contributing factors to make relatively accurate estimates such as the available dataset. The current number of confirmed cases in Indonesia still needs to be analyzed because of the limited testing data available and, likely, more cases have not been diagnosed than have already been diagnosed. In this paper, we aim to analyze the trend situation shortly of the outbreak in Indonesia by applying the Exploration Data Analysis (EDA) model approach, then building a time series forecasting model using the prophet method. The EDA model is a data analysis model using various techniques to maximize insight into the data set as a step in making hypotheses before the data analysis step to understand the data used [14]-[17]. The EDA model is applied to estimate various parameters in the model, then simulations are carried out to see what will happen under the various scenarios. The results of the analysis will be applied to a time series prediction model based on grouping individuals in a population of confirmed cases, deaths, and cures.

II. Method and Data

A. Prophet Method

The proposed forecasting method in this paper is the Prophet model [18] which is the liquid of one of the model's overnight series when based on the procedure where the non-linear trend is adaptive with the seasonal fit, Kraton, and daily, weekly effects plus. Propheat models have better performance with the current series that has a strong seasonal effect and a few seasons of IDI data, strong for data loss and trend shifting, and usually handles outliers well. Prophet provides a practical approach to forecasting on a scale " which intends to automate the general characteristics of the current series by providing simple and customizable methods. This approach begins by modeled on a series when using the parameters specified by Muhamad, generates an estimate, and then evaluates it. In general, the Prophet model is formulated as follows:

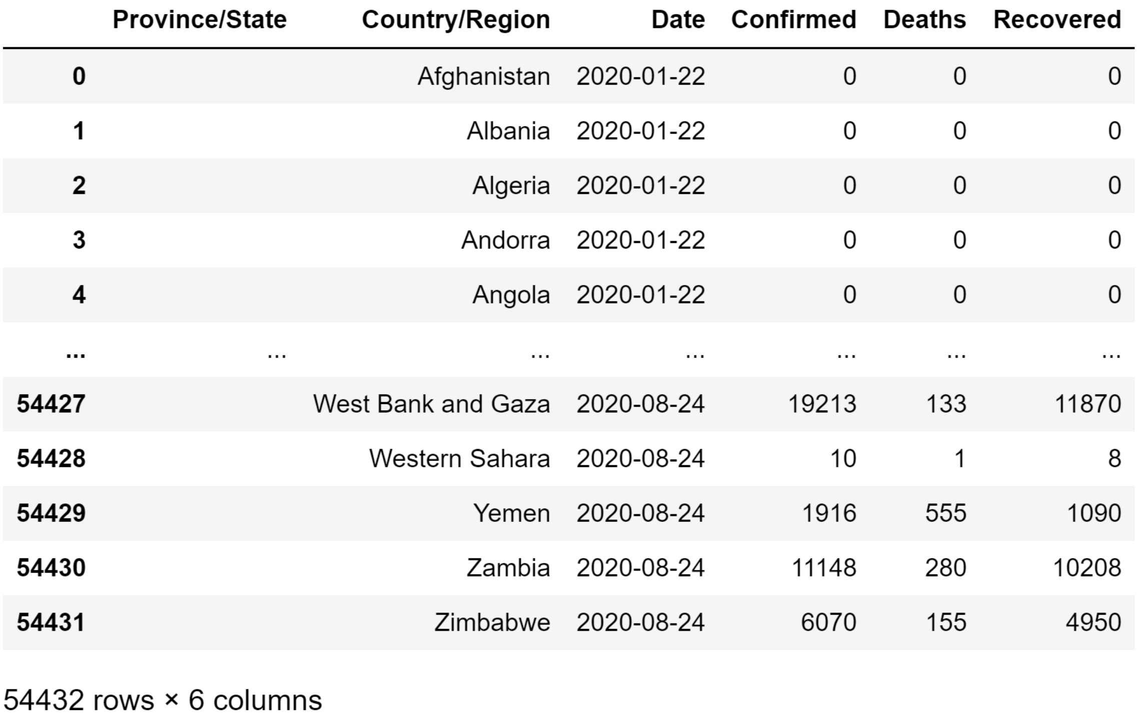
$$Y (t) = g (t) + S (t) + H (t) + \varepsilon t \quad (1)$$

Where: g(t) trend model, which describes the long-term increase or decrease in the data. s(t) models seasonality with the Fourier series, which describes how data is affected by seasonal factors such as time of year, h(t) models the effect of holidays or major events affecting business time series and ϵ_t represents the term irreducible error.

B. Data

We use open-source data (https://github.com/CSSEGISandData/COVID-19) to estimate the various parameters in the model and then simulations are carried out to see what will happen under the various scenarios. We collect data on cases reported every day up to April 29, 2020, then estimate the cross-cutting trend of outbreaks in various worlds starting from China, then illustrating a time series of confirmed COVID-19 cases from the top ten countries for confirmed cases, deaths, and recovered, also, we evaluate the trend of outbreak developments in Southeast Asian countries which focus on Indonesia, the Philippines, Brunei, Malaysia, and Singapore. The results of the analysis will be applied to the Prophet model to predict confirmed COVID-19 cases and death cases in Indonesia.

Table 1. Dataset COVID-19



	Province/State	Country/Region	Date	Confirmed	Deaths	Recovered
0		Afghanistan	2020-01-22	0	0	0
1		Albania	2020-01-22	0	0	0
2		Algeria	2020-01-22	0	0	0
3		Andorra	2020-01-22	0	0	0
4		Angola	2020-01-22	0	0	0
...	...	...	...	...	...	...
54427		West Bank and Gaza	2020-08-24	19213	133	11870
54428		Western Sahara	2020-08-24	10	1	8
54429		Yemen	2020-08-24	1916	555	1090
54430		Zambia	2020-08-24	11148	280	10208
54431		Zimbabwe	2020-08-24	6070	155	4950

54432 rows × 6 columns

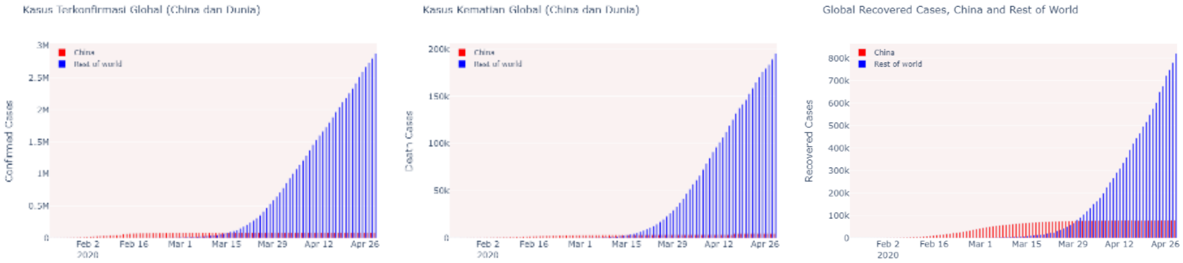
III. RESULT AND DISCUSSION

A. Exploratory Data Analysis

In this paper, an Exploratory Data Analysis (EDA) approach is proposed and visualizes the 2019-nCoV open dataset provided by Johns Hopkins University to provide insights into this virus outbreak on various continents, especially in Southeast Asia. We used the dataset for January-April 2020, data analysis was divided into several phases, namely analysis of trends in China and the world, analysis of global data, analysis of trends in the number of cases in Indonesia, then applying the Prophet model to predict outbreak trends in Indonesia for confirmed cases, recovered and cases of death within the next 30 days. The first discovery of this coronavirus pandemic was found in China, Hubei Province, so it is necessary to analyze how the development of this virus and its spread in countries around the world is very useful to understand the global trend of increasing the number of cases over time. There are always patterns in any data, but what is of concern is how strongly the data follows a pattern that spreads exponentially.

1) Outbreak Trends in China and the World

China's Hubei province has recorded the highest number of cases (67k). The highest number of cases in China is in Hubei province, where the virus is believed to have originated. In Hubei, the city of Wuhan noted the number of cases confirmed the highest, followed by Zhejiang, Guangdong, and Hunan. Province has recorded a surge of 30% in the case of new which was confirmed by 14.840 on 13th February 2020.



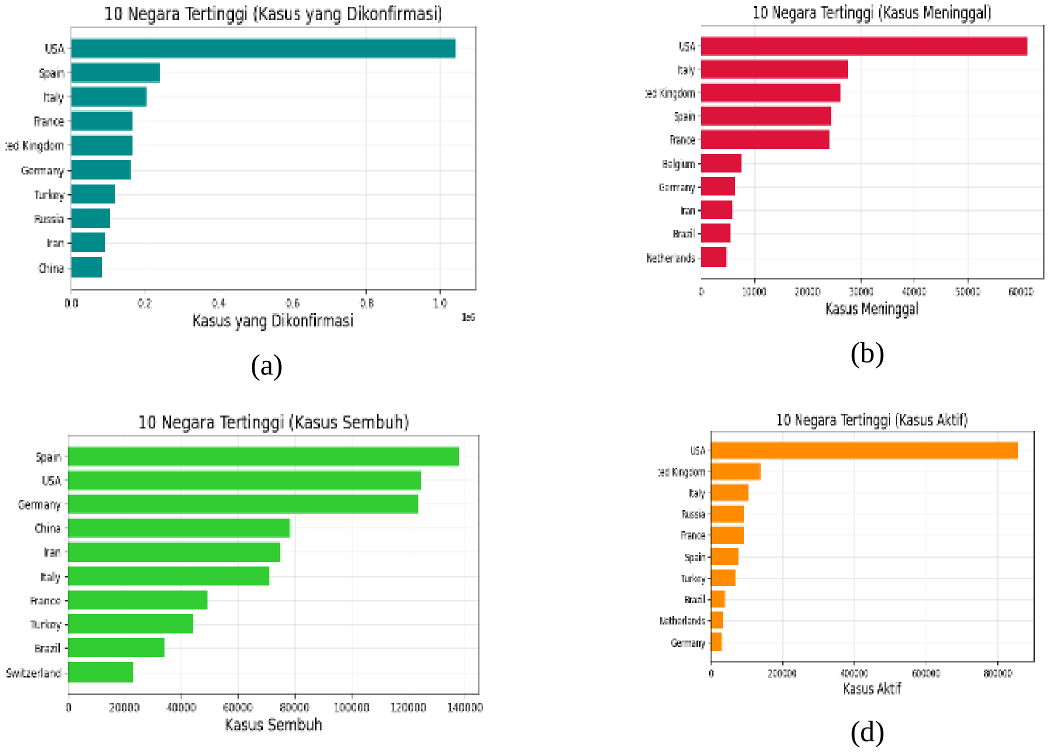
(a) (b) (c)

Fig. 2. Outbreak Treb in China and the World, (a) Confirmed Cases, (b) Deaths and (c) Cured Cases.

The cases began to be reported in other countries outside the country of China from the date of 8 February. Since it is, the cases increased drastically in other countries. But at the end of February, the cases that have been stable in the State China, but the situation has worsened in the whole world seen in picture 1 (a) cases confirmed, while the number of death continue to increase in the whole world seen in picture 1 (b) Cases Died.

2) Trends in the Top 10 Countries

The next stage, analyzing the trend of outbreak development in the ten countries with the highest prevalence of confirmed cases, deaths and recoveries is presented in Figure 2 which is a visual display of the 10 countries with the highest number of cases outside China, where the USA is the highest country for confirmed cases, deaths, and cases. active, while for recovered cases Germany is in the top rank. And the USA is in position 4 of 10 countries



(a) (b)

(c) (d)

Fig. 3. Trends in the Top 10 Countries in (a) Confirmed Cases, (b) Deaths, (c) Recovering and (d) Active Cases

Tren outbreaks in the whole world have increased significantly in January 2020 until April 2020 with the number of confirmed cases as many as 3,193,886 and 2,276,381 deaths are presented in fig 4.

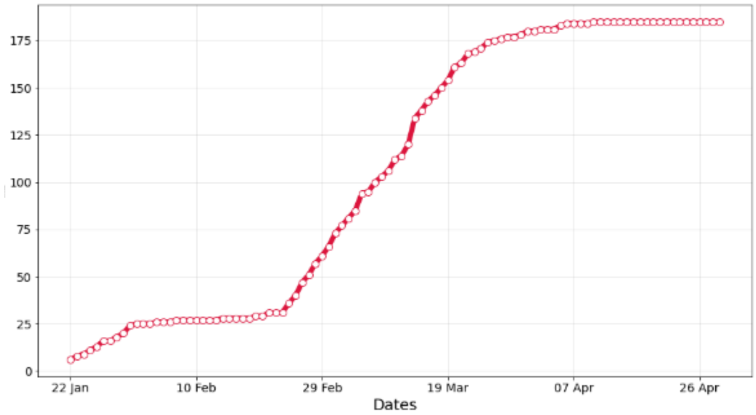
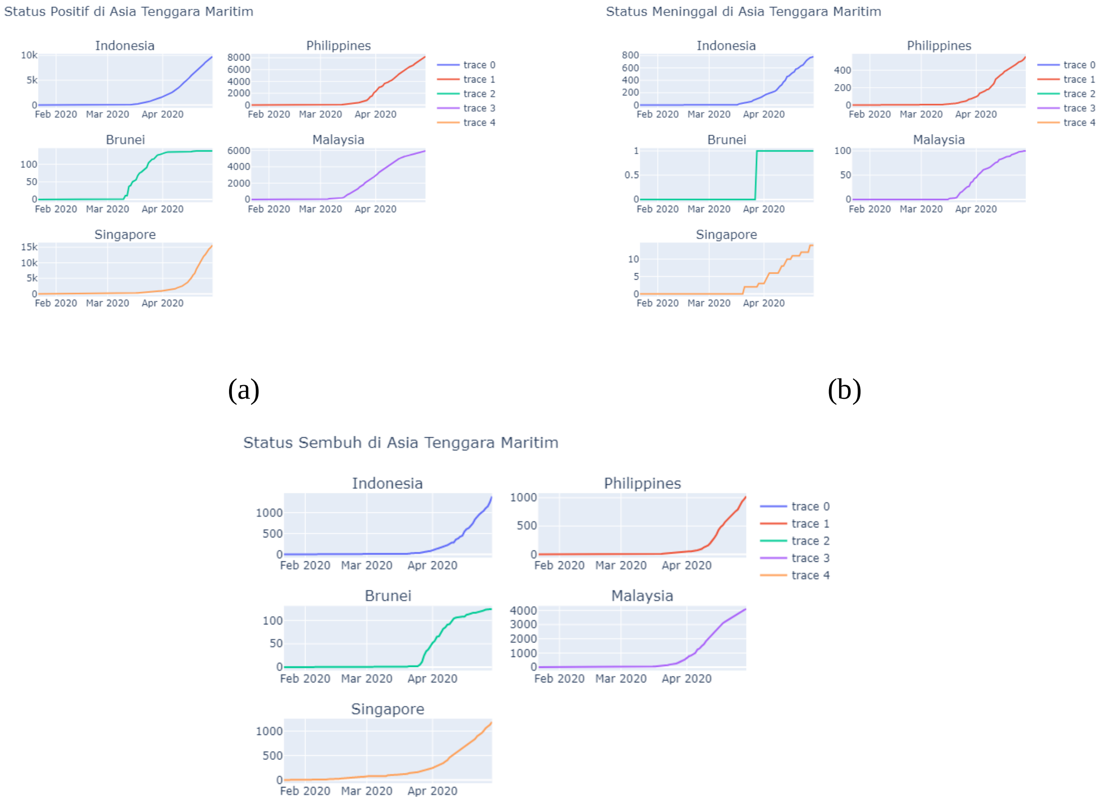


Fig. 4. Global Outbreak Trends

3) Outbreak Trends in Southeast Asia

Further analysis is how the development of the outbreak in the state of Southeast Asia, in this case, we focus on 5 (five) countries that are Indonesia, Philippines, Brunei, Malaysia, and Singapore with confirmed cases, died, and recovered in Fig 5



(a)

(b)

(c)

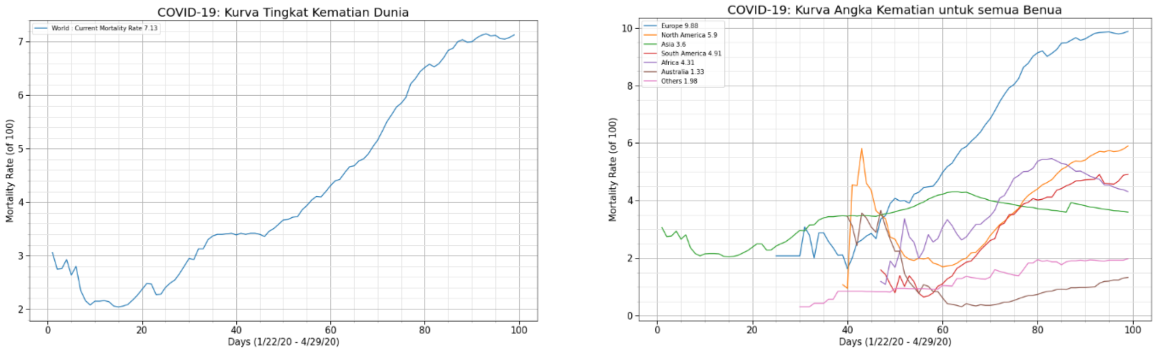
Fig. 5. Outbreak Trends in Southeast Asia in (a) Confirmed cases, (b) Died and (c) Recovered

In the confirmed cases of the outbreak in Figure 5 (a), Indonesia until April 29, 2020, there were 9,771 cases, the Philippines as many as 8,212, Brunei 138, Malaysia 5,945 and Singapore with 15,641 at most and Indonesia was in the second position of the 5 countries. The state of Indonesia died in the first place as many as 784, while the cases recovered by the State of Indonesia were in the first place as many as 1,391 then the state of Singapore, which is presented in table 1

Table 2. Outbreak Trends in Southeast Asia

Country	Case		
	Confirmed	Died	Healed
Indonesia	9,771	784	1,391
Philippines	8,212	558	1,023
Brunei	138	1	124
Malaysia	5,945	100	4,087
Singapore	15,641	14	1,188

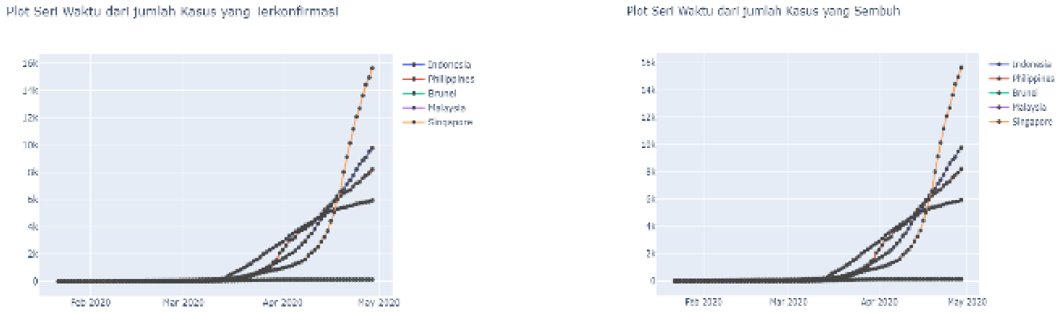
We used predictive analysis, to estimate how many confirmed cases and deaths could be expected in the near future. For each epidemic, the most important evaluation is the Death Rate. It is a measure of the number of deaths in a given population over a specified interval. Figure 6(a) shows how the death rates varied from January 22 2020 to April 2020 worldwide and Figure 6(b) shows the variation in mortality rates across different continents over time.



(a) (b)

Fig. 6.(a) Death rate curve case around the world, (b) death curve of the whole continent

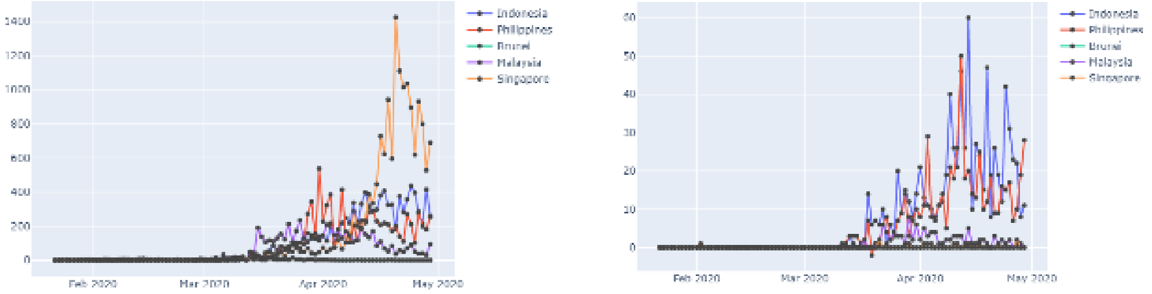
Based on an image of 6 worldwide death curves of 7.13 where continental Europe has the highest level of 9.88 and the lowest in the Australian continent by 1.33. While analysis of the time series trend outbreaks in Southeast Asian countries specifically Indonesia, Philippines, Brunei, Malaysia, and Singapore can be seen in the picture 7 trend cases confirmed while the picture 8 cases cured



(a) (b)

Fig. 7.Time Series Trends in the outbreak of Southeast Asian countries (a) confirmed cases, (b) Cured cases

In figure 8 plague trends we present in a visual form for everyday development where there is an increase in cases die each day.



(a) (b)

Fig. 8.Time Series Trend outbreak COVID-19 in southeast Asia every day (a) confirmed cases, (b) Cases of death

The plot above shows the number of confirmed cases reported per day, so we are currently looking at the cumulative number. The number of cases has increased exponentially in Southeast Asian countries since the first cases in China. The trend of increasing death cases in Indonesia occurred on April 14, 2020, as many as 60 people, while in the month of the most confirmed cases in April 2020 this shows that in April there was an increase in cases in Indonesia.

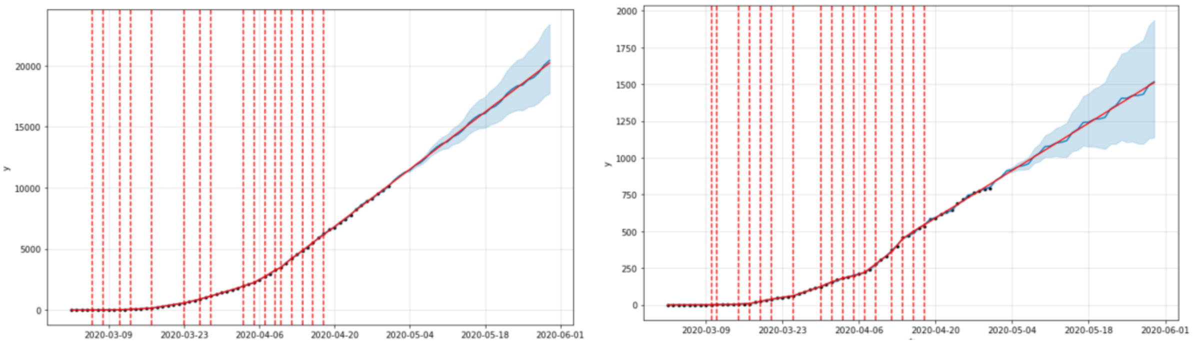
B. Analisys Tren and Forecasting

In this study, we built a model to estimate the number of confirmed cases and deaths in Indonesia within the next 30 days using the Prophet algorithm based on available data up to 29 April 2020, we applied a time series forecasting model to a dataset detailing the number and location of cases. confirmed pandemics, including people who recovered, and those who died. Algorithms prophets are used to predict future values based on previously observed values. We also consider parameter setting to optimize the predicted results which are evaluated by Mean Absolute Percentage (MAP) while the Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MEA) models are used for the evaluation of prediction errors

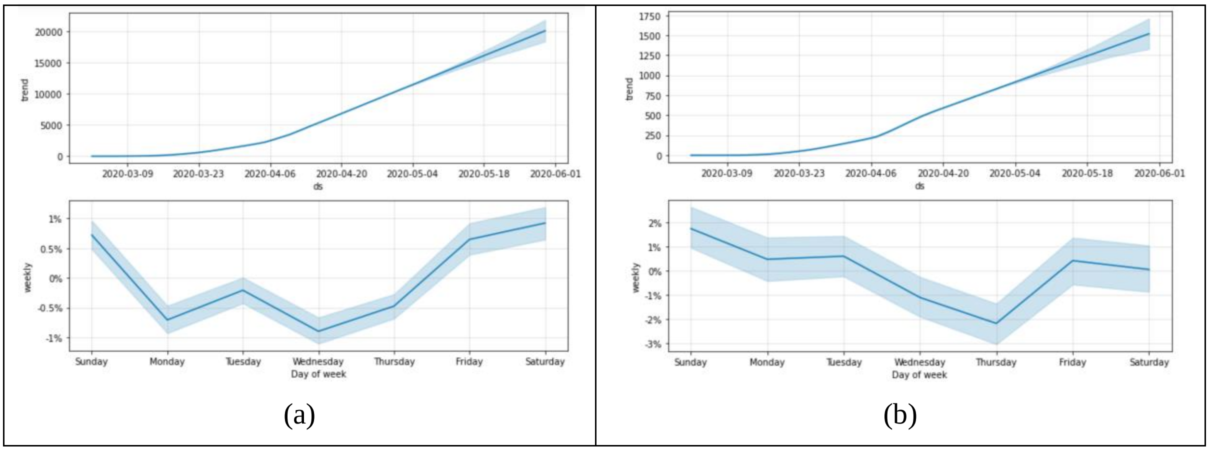
Table 3. dataset Ina_COVID

Date	Confirmed	Deaths	Recovered
2020-03-02 00:00:00	2	0	0
2020-03-03 00:00:00	2	0	0
2020-03-04 00:00:00	2	0	0
2020-03-05 00:00:00	2	0	0
2020-03-06 00:00:00	4	0	0
..	..	..	..
..	..	..	..
2020-04-27 00:00:00	9096	765	1151
2020-04-28 00:00:00	9511	773	1254
2020-04-29 00:00:00	9771	784	1391
2020-04-30 00:00:00	10118	792	1522

In table 3, it is part of the dataset used for prediction, the predictive analysis framework is used by dividing the historical data for the last month for prediction needs then comparing the prediction results with the actual data. By default the prophet algorithm determines the ds and y variables to build a prediction model, so the first step is to change the Date variable to ds and Confirmed to y for the prediction of confirmed cases, while the death cases death becomes y and will produce an estimated value ('yhat') lower limit ('yhat_lower') and upper limit ('yhat_upper') estimate. The results of the prediction of confirmed cases and deaths are presented in Figure 10 while Figure 11 is a visualization of the trend and weekly analysis for both cases.



(a) (b)



(a) (b)

The trend in the development of confirmed cases for the State of Indonesia which has increased until May 2020 is around 20,424 estimated results, while the actual data is 25,773 cases, for death cases, the estimated results are around 1,517 while the actual data is 1,573, so it can be concluded that it needs to be considered especially in implementing policies to break the chain of distribution. The results of the comparison between predictions and actual data are shown in table 3.

Table 4. Prophet algorithm prediction results

Date	Confirmed		Deaths	
	Predict	actual	Predict	actual
01/05/2020	10597.23	10551	846.77	800
02/05/2020	10961.09	10843	870.19	831
03/05/2020	11263.05	11192	910.30	845
04/05/2020	11432.69	11587	918.36	864
05/05/2020	11822.88	12071	938.50	872
06/05/2020	12064.09	12438	946.41	895
07/05/2020	12442.09	12776	959.53	930
08/05/2020	12957.37	13112	1008.46	943
09/05/2020	13326.99	13645	1031.95	959
10/05/2020	13621.4	14032	1075.13	973
11/05/2020	13757.04	14265	1080.46	991
12/05/2020	14158.72	14749	1100.08	1007
13/05/2020	14382.16	15438	1105.45	1028
14/05/2020	14768.92	16006	1116.99	1043
15/05/2020	15317.5	16496	1170.16	1076

16/05/2020	15692.89	17025	1193.71	1089
17/05/2020	15979.76	17514	1239.97	1148
18/05/2020	16081.39	18010	1242.56	1191
19/05/2020	16494.56	18496	1261.66	1221
20/05/2020	16700.22	19189	1264.48	1242
21/05/2020	17095.75	20162	1274.45	1278
22/05/2020	17677.64	20796	1331.85	1326
23/05/2020	18058.79	21745	1355.47	1351
24/05/2020	18338.12	22271	1404.80	1372
25/05/2020	18405.75	22750	1404.66	1391
26/05/2020	18830.4	23165	1423.24	1418
27/05/2020	19018.29	23851	1423.51	1473
28/05/2020	19422.58	24538	1431.91	1496
29/05/2020	20037.78	25216	1493.55	1520
30/05/2020	20424.7	25773	1517.22	1573

Based on the comparison of the predicted results with the actual data presented in table 3, the prophet algorithm has a good performance with the evaluation result of our estimated relative error rate (MAPE) of about 6.52%, and the average of our model is wrong 52.7 (MAE) for confirmed cases, whereas case mortality was 1.3% for the MAPE and MAE models around 236.6%. Our experimental results using the prophet algorithm to do the job of predicting confirmed cases and deaths in Indonesia within the next 30 days resulted in a good error rate by adjusting for the acceleration of the increase in the number of confirmed cases and deaths.

IV. Conclusion

In this paper, we present a data visualization of the trend of the coronavirus outbreak in confirmed cases, deaths, and cures based on data reported from January to April 29, 2020. The Exploratory Data Analysis (EDA) model approach is applied to provide an understanding of the trend of outbreaks that started in China, the ten highest countries, trends in Southeast Asian countries, and time series experience using the Prophet algorithm for the next 30 days in Indonesia. Based on the comparison of predicted and actual data on confirmed cases and deaths in Indonesia, the prophet algorithm does a good job where the result of our estimation error rate (MAPE) evaluation is around 6.52%, and our model average is wrong 52.7 (MAE) for confirmed cases, while the case mortality is around 1.3% for the MAPE and MAE models around 236.6%, besides that the results of the experience result in the conclusion that there will be a significant increase in the trend of the spread of this pandemic in Indonesia for confirmed cases and cases of death. increase in the next month, this can be used as input in the policy-making process to limit the spread of the virus, besides that, we still need to consider conditions of government policies such as limiting access to communication in future research.

References

[1] Anastassopoulou, C., Russo, L., Tsakris, A., & Siettos, C. (2020). Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLoS One, 15(3), e0230405. https://doi.org/10.1371/journal.pone.0230405

[2] Turiel, J., & Aste, T. (2020). Wisdom of the crowds in forecasting COVID-19 spreading severity. 9–10. Retrieved from http://arxiv.org/abs/2004.04125

[3] Li, L., Yang, Z., Dang, Z., Meng, C., Huang, J., Meng, H., … Peng, H. (2020). Propagation analysis and prediction of the COVID-19. MedRxiv, 12, 2020.03.14.20036202. https://doi.org/10.1101/2020.03.14.20036202

[4] Liu, D., Clemente, L., Poirier, C., Ding, X., Chinazzi, M., Davis, J. T., … Santillana, M. (2020). A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models. (d). Retrieved from http://arxiv.org/abs/2004.04019

[5] Singh, R., & Adhikari, R. (2020). Age-structured impact of social distancing on the COVID-19 epidemic in India. (March). Retrieved from http://arxiv.org/abs/2003.12055

[6] Crokidakis, N. (2020). Data analysis and modeling of the evolution of COVID-19 in Brazil. Retrieved from http://arxiv.org/abs/2003.12150

[7] Zhou, X., Hong, N., Ma, Y., He, J., Jiang, H., & Liu, C. (2020). Forecasting the Worldwide Spread of COVID-19 based on Logistic Model and SEIR Model.

[8] Gupta, R., & Pal, S. K. (2020). Trend Analysis and Forecasting of COVID-19 outbreak in India. MedRxiv, 2020.03.26.20044511. https://doi.org/10.1101/2020.03.26.20044511

[9] Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J. M., … Chowell, G. (2020). Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infectious Disease Modelling, 5, 256–263. https://doi.org/10.1016/j.idm.2020.02.002

[10] Manchein, C., Brugnago, E. L., da Silva, R. M., Mendes, C. F. O., & Beims, M. W. (2020). Strong correlations between power-law growth of COVID-19 in four continents and the inefficiency of soft quarantine strategies. 1–10. Retrieved from http://arxiv.org/abs/2004.00044

[11] Pal, R., Sekh, A. A., Kar, S., & Prasad, D. K. (2020). Neural network based country wise risk prediction of COVID-19. d, 1–9. Retrieved from http://arxiv.org/abs/2004.00959

[12] Al-qaness, M. A. A., Ewees, A. A., Fan, H., & Abd El Aziz, M. (2020). Optimization Method for Forecasting Confirmed Cases of COVID-19 in China. Journal of Clinical Medicine, 9(3), 674. https://doi.org/10.3390/jcm9030674

[13] Elmousalami, H. H., & Hassanien, A. E. (2020). Day Level Forecasting for Coronavirus Disease (COVID-19) Spread: Analysis, Modeling and Recommendations. Retrieved from http://arxiv.org/abs/2003.07778

[14] Bezerra, A., Silva, I., Guedes, L. A., Silva, D., Leitão, G., & Saito, K. (2019). Extracting value from industrial alarms and events: A data-driven approach based on exploratory data analysis. Sensors (Switzerland), 19(12). https://doi.org/10.3390/s19122772

[15] Khan, A. M., Siddiqi, M. H., & Lee, S. W. (2013). Exploratory data analysis of acceleration signals to select light-weight and accurate features for real-time activity recognition on smartphones. Sensors (Switzerland), 13(10), 13099–13122. https://doi.org/10.3390/s131013099

[16] Ma, D., Fan, H., Li, W., & Ding, X. (2019). The state of mapillary: An exploratory analysis. ISPRS International Journal of Geo-Information, 9(1). https://doi.org/10.3390/ijgi9010010

[17] Ma, X., Hummer, D., Golden, J. J., Fox, P. A., Hazen, R. M., Morrison, S. M., … Meyer, M. B. (2017). Using visual exploratory data analysis to facilitate collaboration and hypothesis generation in cross-disciplinary research. ISPRS International Journal of Geo-Information, 6(11). https://doi.org/10.3390/ijgi6110368

[18] Taylor, S. J., & Letham, B. (2017). Forcast at Scale Prophet. PeerJ Preprints, 1–25. https://doi.org/10.7287/peerj.preprints.3190v2