Forecasting COVID-19 Vaccination Trends in Indonesia using Machine Learning

Ahmad Fauzan Aqil[1], Hsi-Chieh Lee[2], Sofi Ismarilla Wardani[3]
[1,2,3]National Quemoy University, Taiwan

ABSTRACT
The ongoing COVID-19 pandemic requires much research to deal with this problem. Medical treatment has resulted in vaccine findings that work as an immune system to block the COVID-19 reaction process. However, many of these developments are still undergoing improvement and periodic testing to found better results for humans. Therefore, forecasting trends of the COVID-19 vaccine in Indonesia is carried out to regularly predict vaccines' effectiveness by adjusting conditions. This forecasting uses the time-series forecasting method by prioritizing a machine learning process in predicting probably future forecasts. Based on the highest vaccine used, we propose ARIMA and Facebook Prophet as machine learning models to predict vaccine trends in each country. The Prophet model results achieved an RMSE score of 0.176, which these results contained vaccines distributed in Indonesia. Besides that, the ARIMA model achieved an RMSE score of 0.453 using the same dataset. The results obtained from this method can be considered a policy for the government to deal with the effective use of vaccines according to future needs. As a further development, this research can be reviewed by paying attention to external aspects such as social and economic factors affecting the COVID-19 vaccination. The results obtained are more comprehensive and representative than this research based on conditions that provide policies for handling COVID-19.

CONTACT
afauzanaqil@gmail.com

KEYWORDS
Vaccine COVID-19, Future prediction, Time-series forecasting, Evaluation metrics, Vaccine distribution

INTRODUCTION

Pandemic COVID-19 that took place during the past year is affecting all aspects of human life. This influence changes social culture and behaviour for humans in every activity. It was recorded that until May 24 2021, John Hopkins University released the total number of COVID-19 cases globally, reaching 166 million people [1]. These results indicate that COVID-19 cases continue to emerge and continue to increase in several countries. Researchers are racing to find a suitable vaccine for tackling the spread of the pandemic level with a current phenomenon. Vaccination is believed to be the right solution, like other diseases that already have their immune vaccines.

Researchers who develop vaccines have their respective preferences according to research procedures made either through the test material, the type used, or based on the genetic tests made. This procedure follows the review conducted by Kaur et al. regarding vaccine-making strategies carried out by other researchers [2]. The results of this study led to a large number of vaccine variants obtained so that the distribution process of the vaccine has varied time. Therefore, the use of vaccines in each country has its differences following applicable country policies and regulations.

As happened in the United Kingdom (UK), the government chose to distribute its vaccine variants, namely AstraZeneca and Pfizer-BioNTech [3]. The government has such confidence that the vaccine's findings are compatible with the physiology of its citizens. The successful use of vaccines in the UK by utilizing genetics from monkeys can prevent being adopted by several countries to supply the vaccine made in the UK [4]. Countries that do not produce vaccine variants independently carry out several imports from countries that produce vaccine variants. This case occurred in several countries with high cases of COVID-19 with minimal research development or several countries that had collaborated to bring in vaccine variants quickly to cope with the increasing spread of cases. Most of these countries come from Asia and Africa, which import vaccine variants from other countries so that one country can have a wide variety of distribution of vaccine variants. Therefore, this case is based on the agreement of two countries working together as importers and recipients of vaccine distribution.

Based on these policies, each country has different preferences according to the needs and interests of their country. Vaccine development is currently encouraging various countries to become vaccine importers from vaccine developing countries so that the majority of vaccines distributed depend on vaccine providers for importing countries. The existence of this condition causes the distribution of vaccines around the world to be

10.52162/3.2021118

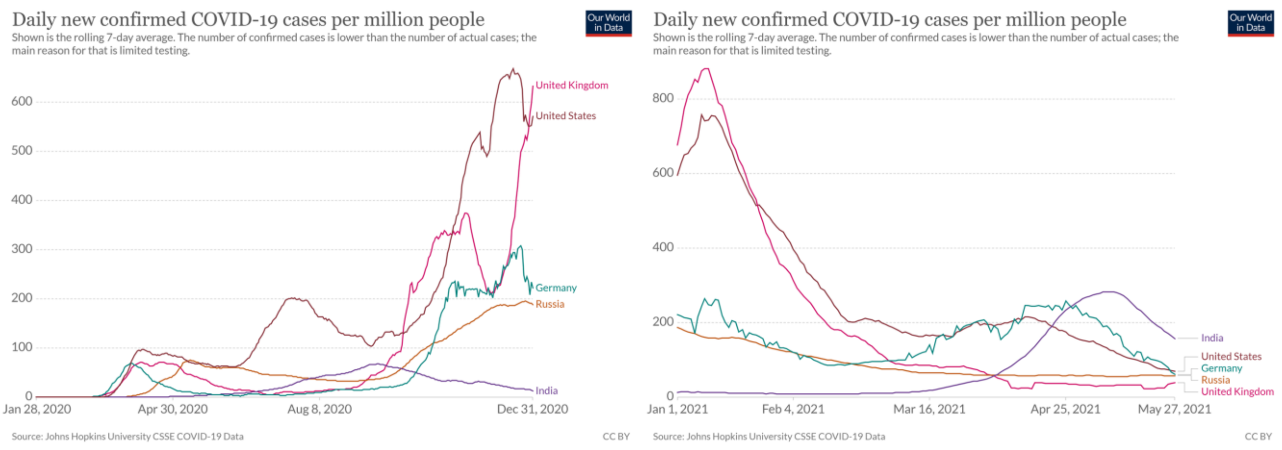


Figure 1. Total COVID-19 cases per Daily. (Left) COVID-19 cases before distribution vaccines and (Right) COVID-19 cases before distribution vaccines.

very diverse and have their respective existence values throughout the world. Therefore, the distribution of each vaccine can be analyzed based on the number of vaccines distributed in Indonesia [5]. This analysis is intended as a future vaccine development in providing the proper vaccine dosage based on trends in vaccine use and the growing number of positive cases of COVID-19.

The intensity of vaccine distribution has decreased COVID-19 cases in several countries, especially developing vaccines. It was noted that since the vaccine was available in early 2021, John Hopkins University released a COVID-19 decline that occurred up to 30%, as shown in Figure 1 [1]. The positive impact generated in encouraging COVID-19 cases has given people the confidence to use vaccines early to increase the demand for vaccines in each country. Forecasting is needed to find out how much demand will come following current demand trends to know vaccination needs in Indonesia as one of the countries that many COVID-19 cases now.

Using a machine learning approach, this forecasting processes statistical data resulting from time-series data as needed. The data series used depends on how significant the influence is on the time-series trend used. Knowing this, forecasting research using machine learning is often the proper reference in knowing future progress. This research requires several trials to produce the best results. Therefore, we took two methods that can be used for this research, namely Prophet and ARIMA. These three methods examine time-series data on vaccine distribution from December 2020 to May 2021 to see trends that occur during the initial distribution of vaccines over the last five months.

It can be concluded that the main contribution to this research is,

1. Knowing how the distribution trends of COVID-19 vaccines in Indonesia are based on the data obtained
2. Examine the comparison of the original and predicted values generated in the COVID-19 vaccine trend forecasting study
3. Describing the future forecasting results from the data obtained into a statistical model.
4. Evaluate the accuracy and precision of the tested data with the prediction of the resulting data, considering the study's final results.

METHODS

Forecasting statistical data is very dependent on the data input process provided so that several stages are needed that affect the processing of data before the testing process. Machine learning processes help data processing performance as a suitable data setting in training and testing data. Therefore, the application of a suitable method in applying the inputted data is essential to produce the best value.

Materials

In this study, we take data from open access data taken from Kaggle that has been compiled by Preda [6]. The data has a detailed time series starting from Dec. 02 2020, to May 19 2021, with varying values for each column. The data requires preprocessing to conduct forecasting so that the data becomes proportional according to the characteristics of the correct data to be processed. The data preprocessing carried out includes filtering, sorting, checking data normalization, and checking data differencing. The details of vaccine data before preprocessing the data are shown in Table 1.

This method has characteristics that can determine how optimally the machine learning process runs. In addition, other supporting data in analyzing the vaccine more deeply we adjust to produce an optimal data presentation. These data include GDP from all countries to adjust the vaccine data per continental taken from Pramod [7]. Then continental data are taken from Olteanu as a grouping of vaccine data per continent [8]. We researching by selecting total vaccinations on Indonesia per daily case. Indonesia has been used AstraZeneca and Sinovac vaccines as vaccine type that is distributed to all its citizens. We only consider the trend conditions that have occurred during the vaccine distribution period in Indonesia as a reference for forecasting vaccine needs in the future.

Table 1. Details of Vaccines Data that representative until 19th May 2021

No	Columns	Non-Null Count	Dtype
1	Country	15666 non-null	Object
2	Iso_code	15666 non-null	Object
3	Date	15666 non-null	Datetime64[ns]
4	Total_vaccinations	15666 non-null	Float64
5	People_vaccinated	15666 non-null	Float64
6	People_fully_vaccinated	15666 non-null	Float64
7	Daily_vaccinations_raw	15666 non-null	Float64
8	Daily_vaccinations	15666 non-null	Float64
9	Total_vaccinations_per_hundred	15666 non-null	Float64
10	People_vaccinated_per_hundred	15666 non-null	Float64
11	People_fully_vaccinated_per_hundred	15666 non-null	Float64
12	Daily_vaccinations_per_million	15666 non-null	Float64
13	Vaccines	15666 non-null	Object
14	Source_name	15666 non-null	Object
15	Source_website	15666 non-null	Object

We use cross-validation on total vaccination data developed through actual data and predictive data as a comparative data analysis that can be used for future forecasting. This method serves to prove the relationship between actual data and predict data generated by machine learning. To find out that these results are related, we take an evaluation method in the form of metrics evaluation results from the learning model. We took the root mean square error (RMSE), mean absolute percentage error (MAPE) and mean absolute error (MAE) as evaluations of the error rate in the model. Meanwhile, we evaluate the correlation of the data using the R-square (R^2) value [9][10][11].

Facebook Prophet

Facebook Prophet has data processing features that prioritize measurements on DateTime series and variable values. Datetime series is used as a measurement tied to the variable's value that determines the seasonality trend. In addition, this series can be used to determine daily, weekly, and yearly seasonality as a review of the results of predictions made in specific forecasts. Meanwhile, the loaded variable values are natural numbers and decimal numbers that meet the statistical data collected. Prophet allows learning with intuitive parameters that can be set without knowing the underlying model [8].

Forecasting with the prophet model makes the data process flexible by assuming trends that have occurred before. The use of Prophet helps to learn data performance by generating data quickly and interpreting data parameters efficiently. Our approach uses the Prophet according to trend data based on daily, weekly, and yearly seasonality to present data based on the data obtained.

In this method, we use data on total vaccinations distributed per day as data processing. This data is sorted by distribution date, which is added up per day according to the number of vaccines distributed. As a forecasting model, a prophet's ability to analyze data is based on the scale of the inputted data. In general, the Prophet analyzes data based on the seasonality used [9]. Seasonality is a forecasting result obtained from the prevailing trend from input data to produce certain graphs that can influence trends that occur in the future. It can be concluded that the mathematical equation of seasonality used in the Prophet is as follows:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

Here $g(t)$ is trend function value which non-periodic changes in time series value, $s(t)$ representative periodic changes data either weekly, daily, or yearly and $h(t)$ represents the holiday condition of the data series with ϵ_t is function customarily distributed data, the data presented are approximate data that considers the upper and lower limit values that are influenced by the trend of the input data to enable the resulting data output to be accurate.

10.52162/3.2021118

Auto Regression Integrated Moving Average (ARIMA)

The most frequently used model for forecasting time series in statistical data research is ARIMA. The ARIMA model framework makes it easy for users to take advantage of the desired deviation values and distributed series values to produce regressions that match the input trend data [10]. However, ARIMA cannot translate non-stationary data so that the resulting data predictions do not match the data trends that occur. Therefore, it is necessary to do stationarity by differencing the data used if the data is detected to be non-stationary. In addition, the differencing data can be performed several times until the data reaches stationary. Differentiating data is a step to distinguish time series data in a series framework so that the resulting data has different characters. The results of the data differencing carried out can be seen in Figure 2.

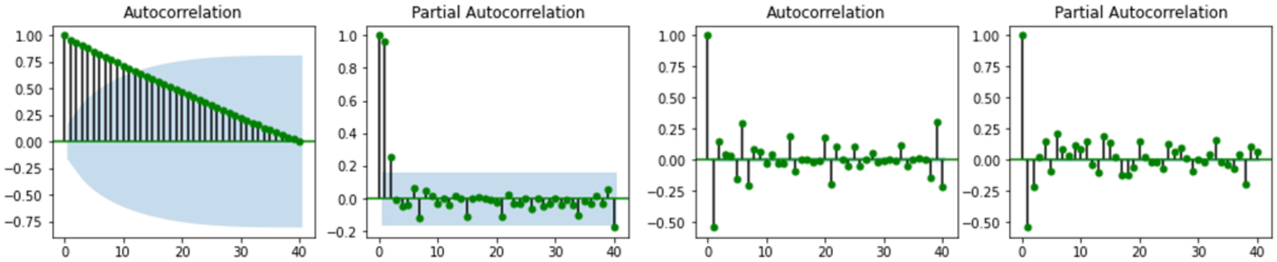


Figure 2. Autocorrelation and Partial Autocorrelation based vaccine distribution in Indonesia on Total Vaccinations per Hundred. (Left) Before Differencing Data and (Right) After Differencing Data.

The result of differencing is in the form of autocorrelation of data in which the relationship between data series is described by a significant similarity between complete correlation and partial correlation. This equation becomes a reference for data processed in ARIMA. We take data on total vaccinations per hundred as the primary data in the learning process. Furthermore, this data will be adjusted to the regression value of the ARIMA model characteristics by choosing the order of p, d, and q where p is the order (number of time lags) of the autoregressive model, d is the degree of differencing, and q is the order of the model moving average.

RESULTS AND DISCUSSIONS

Forecasting the distribution carried out in this study contains data that can be adequately explored. The input data obtained have sufficient details as data that can be explored further. To achieve this, we need some preprocessing data according to data objectives exploration according to the research objectives. We explore the primary data connected to the supporting data where this data has a commensurate contribution to produce in-depth data analysis.

Exploration Data Analysis

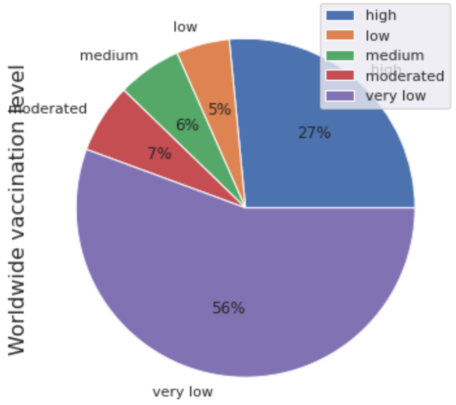
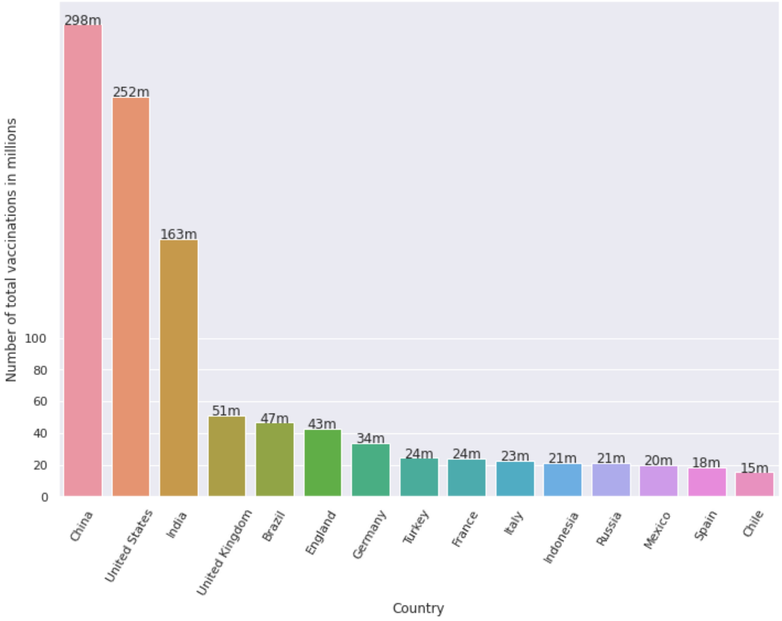


Figure 3. Worldwide total accumulative vaccinations level.

Filtering and sorting were carried out to present datasets on this research. Based on the data collected up to May 06 2021, we found that vaccine data distributed by country of use is shown in Figure 3. Based on the number of total vaccinations, China is the country with the highest vaccine users, with a total number of vaccinations of 298 million people. It can be concluded that as the country that was first infected with COVID-

19, China developed a vaccine to overcome the pandemic. We classify each country based on the rating of the use of vaccines that have been distributed according to Figure 4, where the current majority of vaccine use is still deficient overall with a percentage of 56%.

Figure 4. The number of total vaccinations per country per May 06 2021.



At this stage, we make forecasts on each model that we take on the processed data. We review the comparison between the actual data that occurs with the resulting forecast so that the prediction data assesses the actual data as reference data for forecasting. In our model experiment, we took samples from several data series on different vaccine data due to the data adjustment to the learning model architecture that was run. However, every run model has the same learning performance for forecasting, so that these models can be compared based on the evaluation results of forecasting metrics.

Forecasting on Facebook Prophet

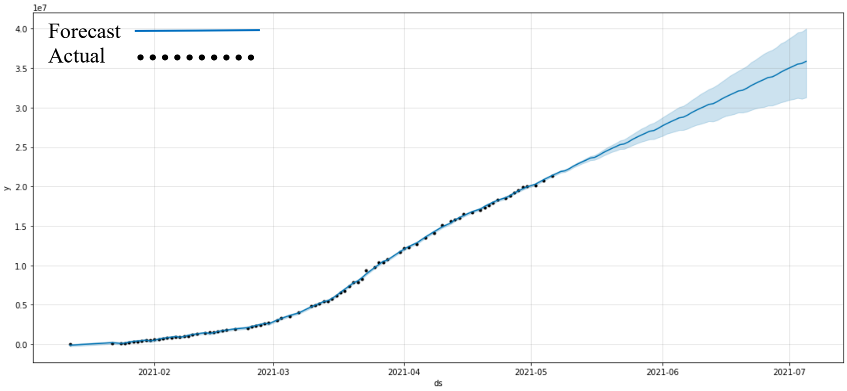


Figure 5. Comparison Graph Picture between Actual and Forecast based Total Vaccinations per Hundred on Prophet model

We use the total vaccinations per hundred data series as a forecasting test for the Prophet model. We filtered the data by the distributed vaccine in Indonesia. In this model setting, we first sort the data series based on the date data for which they correspond to the data series of total vaccinations per hundred. The process of reading the Prophet in learning uses the ds variable as the date data series and y as the data series being tested [11]. The two data series are processed in the Prophet model to produce several predictive output variables; namely, $yhat$ contains the forecast value from the time series, $yhat\_lower$ contains the lower part of the confidence interval for forecasting, and $yhat\_upper$ contains the upper part of the confidence for forecasting. Forecasting results from the Prophet can be seen in Figure 6, where the blue line is the data forecast value while the black dots are the actual data value. We conducted a forecasting experiment until June 1, 2021, as a comparison with the actual data.

10.52162/3.2021118

In the forecasting results, Facebook Prophet has a value that can describe suitability between the actual data and the forecast data. The effect of seasonality on the Prophet function which uses daily seasonality allows the Prophet to run training depending on the daily data series. On the other hand, if the data series used and the functions used are not the same, the results of the Prophet will reach a high error value. Therefore, Prophet needs to recognize what data series is used before forecasting to produce optimal values. Figure 5 shows a significant deviation between the actual data and the forecasted data, so this result needs to be evaluated based on the accuracy of the data and the value of the data error. By applying the predicted lower and upper limit values, the Prophet has a confidence value of 95% on uncertainty interval with a critical value loss of 0.057 based on the predicted data results. Thus, the predictive value of the Prophet model proves that the results are valid. Therefore, forecasting from the Prophet can be continued as a prediction of future values for vaccination trends in Indonesia. However, the similarity between forecasting results with actual data explains that the Prophet's forecasting results are close to the truth of the data. Therefore, the Prophet's forecasting results allow further evaluation as a machine learning model in vaccine distribution forecasting.

Forecasting on ARIMA

In the previous section, using the ARIMA model requires differencing if the data is not stationary, so we use preprocessed data series in this discussion. We used total vaccinations data as the processed data series based on vaccine distribution in Indonesia. We divide the train data and testing data with a scale of 80:20 to compare the forecasting and actual data. The comparison results are shown in Figure 7, where we can observe the difference in the blue line as the forecast value and the orange line as the actual value. The difference is influenced by the order value of ARIMA, which has a particular value where we take the order value of p, d, q, which is 0,1,1, by choosing the basic exponential smoothing model [12]. These results are based on input data obtained until May 6, 2021. In terms of the order from ARIMA, the results obtained on the comparison of actual data and predict data indicate that there is data that is not following the actual trend that occurs. This result is evidenced by the forecast value which has a high-value deviation in several data series. Even though it has a fairly high upper and lower limit, the uncertainty of the interval shown from the ARIMA prediction is quite large.

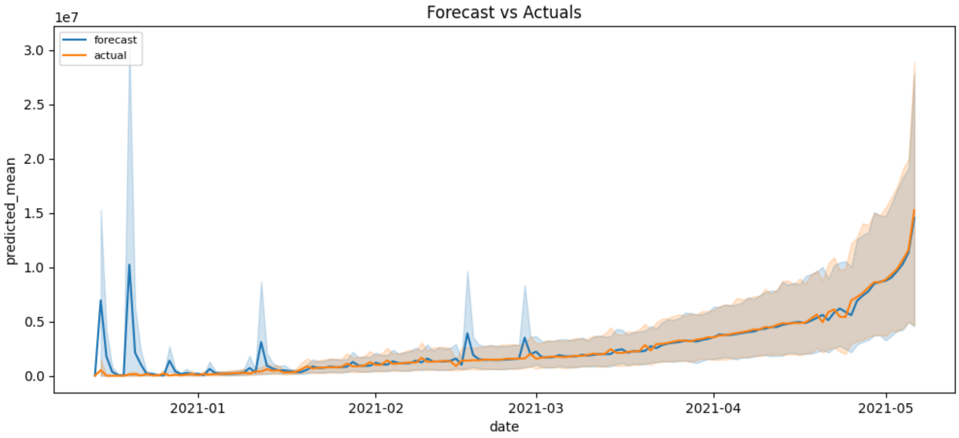


Figure 7. Comparison Graph between Forecast and Actual based Total Vaccinations on ARIMA model

Based on the forecasting results from ARIMA, we know how appropriate the order selection is to the processed data series. This result is based on a significance level of 97% on uncertainty interval with a critical loss value of 0.039 so that the level of confidence of this model can be validated. We observed how ARIMA studied the vaccination DateTime series by giving different treatments for each order according to the actual data. Thus, we find the best results approach the actual data by utilizing the regression value on the mean integration of the DateTime series. Therefore, ARIMA predictions have a good match with the actual data. Basic exponential smoothing affects the forecasting value of the actual data so that the forecasting data shows a graph similar to the actual data graph. This result is due to the p-value of 0, which means that no auto-regression occurs to affect the forecasting value where the number of time lags has a small value. Therefore, changes in the degree of time series data in forecasting can adjust the actual value of the data.

We experimented with predicting the future forecast by taking a few days ahead. This experiment determines how large the quantity of vaccine distribution will be in the future by estimating the total vaccinations that occur in all countries based on the vaccine distributed in Indonesia. Therefore, future results predictions become one of the references for estimating vaccinations in the future.

10.52162/3.2021118

Future Prediction on Facebook Prophet

In the Prophet model, we take the following 60 periods based on daily forecasting by utilizing daily seasonality in the forecasting process. Thus, the results of future predictions from the Prophet start from May 7, 2021, to July 5, 2021. The results of some predictions from the Prophet can be seen from the data in Table 2, which shows the values of $yhat$, $yhat\_lower$, and $yhat\_upper$. The increasing $yhat$ value followed by the increasing $yhat\_lower$ and $yhat\_upper$ values can be interpreted that the resulting prediction is possible within the scope of the distance between $yhat\_upper$ and $yhat\_lower$. However, the predicted trend will continue to rise based on the resulting $yhat$ value.

Table 2. Future Forecasting Predictions on Prophet based Total Vaccinations

ds	yhat	yhat_lower	yhat_upper
2021-07-01	1.969836e+07	1.955696e+07	1.985028e+07
2021-07-02	1.993938e+07	1.979447e+07	2.008535e+07
2021-07-03	2.028423e+07	2.013788e+07	2.042362e+07
2021-07-04	2.087085e+07	2.073125e+07	2.101529e+07
2021-07-05	2.139996e+07	2.154747e+07	7.535284e+08

The Prophet has the advantage of producing future forecasting that is similar to the actual value by paying attention to the forecast value from the comparison of the actual data value. The simplicity of the Prophet using the daily seasonality algorithm as a learning process produces data that can estimate future dataseries values at intervals that are not much different from the actual data. However, the Prophet's attempt to do forecasting still has an error value with the interval between the lower and upper limits on the forecast value.

Comparison of actual data and future forecasting is influenced by forecasting data from the input data process. The discrepancy between forecasting data and actual in Figure 6 is caused by an error value from the learning process, causing the future forecasting output to have a value that is not commensurate with the actual data. However, the results of learning data from the actual data shown by the future prediction are close to the accuracy of the data to the actual data value. Therefore, we discuss the error rate obtained from the Prophet to evaluate how much effectiveness and accuracy is obtained based on the evaluation metrics vaSlue achieved.

Future Predictions on ARIMA

The discussion of ARIMA future forecasting is very dependent on stationary data and the selected order value. We conducted a trial from May 5 to July 05, 2021, based on total vaccinations from vaccine distribution in Indonesia. We choose the same order as the forecasting data for p, d, q, i.e. 0, 1, 1. Based on the data initiation, we get that future forecasting results are not following the previous actual data trend, where this future forecasting has the same value in every processed data series. These results are evidenced in Table 3, which shows that the value of total vaccinations on the last day of inputted data has the same value as each future forecasting data. Therefore, the learning process from ARIMA has a significant error in carrying out future forecasting. However, we consider the results of the forecasting data as fairly accurate data output for ARIMA.

Forecasting learning conducted by ARIMA has shortcomings in identifying the value of the comparison between the actual value and the predicted value. Some data points have a large deviation from the actual value so that the value of the next series has a value that is influenced by the value of the deviation of the previous series. Therefore, it can be seen how the values generated in Table 3 have a stagnant value without any changes in the data. However, as a forecasting model, ARIMA still has the advantage of predicting the accuracy of the actual value with the forecasting value of the processed data.

Table 3. Future Forecasting Predictions on ARIMA based Total Vaccinations

ds	pred
2021-07-01	600350.388343
2021-07-02	600350.388343
2021-07-03	600350.388343
2021-07-04	600350.388343
2021-07-05	600350.388343

Evaluation

Forecasting the data generated in our experimental model is evaluated based on the evaluation metrics collected [13]. The results of this evaluation are shown in Table 4 based on several types of evaluations that we chose by comparing the Prophet and ARIMA models.

Table 4. Comparison Evaluation Metrics between Facebook Prophet and ARIMA

Model	RMSE	MAPE	MAE	R^2
Facebook Prophet	0. 1757902	0. 00978718	0. 1503154	0.998761345
ARIMA	0. 4533622	0.071731997	0. 445328	0.929188967

The smaller the error value obtained, the prediction results obtained are more in line with the actual data. In Table 4, we find that Prophet has a better metrics evaluation value than ARIMA. RMSE of Prophet value obtained 0.1757 has a difference of 0.2775 compared to ARIMA with a value of 0.4533 so that the RMSE of the Prophet is lower than ARIMA. This result is proven to be in line with the value of other evaluation metrics where each MAPE and MAE from the Prophet have lower values than ARIMA,with score 0.00978 and 0.1503.

Based on these results, we confirm the results of the evaluation metrics by using the R^2 value as a measure of how well the model synchronizes between predictions and actual data in reading trends. R^2 of Prophet is 0.9988, which is higher than ARIMA with a value of 0.9292, which has a difference of 0.0045. Therefore, the precision value of Prophet is more accurate than ARIMA.

CONCLUSIONS

Since the distribution of vaccines was carried out, the trend that has occurred was the increasing supply of vaccines in various countries to encourage lower rates of COVID-19 cases. We use Prophet and ARIMA as forecasting models where the results shown from both models have predictive values that are pretty accurate with actual data. Furthermore, future forecasting results obtained from the two models have significant differences where ARIMA cannot predict future forecasting well because it has a stagnant value. However, Prophet was able to produce a better future forecasting value because the prediction increased total vaccinations. This result is evidenced by the lower metrics evaluation value of the Prophet than ARIMA so that the Prophet is more suitable as a forecasting model for COVID-19 vaccination.

REFERENCES

[1] C. for S. S. E.(CSEE), "COVID-19 Map," 2021. https://coronavirus.jhu.edu/map.html (accessed May 24, 2021).

[2] S. P.Kaur andV.Gupta, "COVID-19 Vaccine: A comprehensive status report," *Virus Res.*, vol. 288, 2020, [Online]. Available: https://doi.org/10.1016/j.virusres.2020.198114.

[3] A. de Figueiredo, "Forecasting sub-national trends in COVID-19 vaccine uptake in the UK," *medRxiv*, pp. 1–17, 2020.

[4] S.Loomba, A.deFigueiredo, S. J.Piatek, K.deGraaf, andH. J.Larson, "Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA," *Nat. Hum. Behav.*, vol. 5, no. 3, pp. 337–348, 2021, doi: 10.1038/s41562-021-01056-1.

[5] K. K. R.Indonesia, "Vaksinasi COVID-19 Nasional," 2021. https://vaksin.kemkes.go.id/#/vaccines (accessed Jul. 22, 2021).

[6] G.Preda, "COVID-19 World Vaccination Progress," 2021. https://www.kaggle.com/gpreda/covid-world-vaccination-progress (accessed May 19, 2021).

[7] S.Pramod, "Gap_minder_GDPgrowth," 2019. https://www.kaggle.com/sriharipramod/gap-minder-gdpgrowth (accessed May 03, 2021).

[8] A.Olteanu, "Country Mapping - ISO, Continent, Region," 2020. https://www.kaggle.com/andradaolteanu/country-mapping-iso-continent-region (accessed May 10, 2021).

[9] T.Chai andR. R.Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014, doi: 10.5194/gmd-7-1247-2014.

[10] Y.Dong andH.Jiang, "Global Solar Radiation Forecasting Using Square Root Regularization-Based Ensemble," *Math. Probl. Eng.*, vol. 2019, 2019, doi: 10.1155/2019/9620945.

[11] J.Zhu, B.Shen, A.Abbasi, M.Hoshmand-Kochi, H.Li, andT. Q.Duong, "Deep transfer learning artificial intelligence accurately stages COVID-19 lung disease severity on portable chest radiographs," *PLoS One*, vol. 15, no. 7 July, pp. 1–11, 2020, doi: 10.1371/journal.pone.0236621.