MACHINE LEARNING FOR PREDICTING SPREAD OF COVID-19 IN INDONESIA

Nur Hayati[1*]; Eri Mardiani[2]; Fauziah[3]; Toto Haryanto[4], Viktor Vekky Ronald Repi[5]

Department of Information and Communication Technology [1,2,3]
Nasional University
https://www.unas.ac.id
nurhayati@civitas.unas.ac.id[1*], eri.mardiani@civitas.unas.ac.id[2], fauziah@civitas.unas.ac.id[3]

Department of Computer Science[4]
IPB University
https://www.ipb.ac.id
totoharyanto@apps.ipb.ac.id[4]

Department of Engineering Physics, Faculty of Engineering[5]
Nasional University
https://www.unas.ac.id
vekky_repi@civitas.unas.ac.id[5]

(*) Corresponding Author

Abstract*—In previous research, we carried out an analysis using the FBProphet model to predict the COVID-19 outbreak in Indonesia. The application of the FBProphet model to time series data is considered quite good because it produces a MAPE of 22.60% with a linear distribution. Additionally, based on the pattern in the previous dataset and the total number of active cases currently stands at 2,606, in this research we tried to use the Linear Regression (LR) model as a comparison with the FBProphet model by using additional data from the same data source, KAWALCOVID19 website. Data collection started from March 2, 2020 to December 19, 2021. The aim of this research is the same as previous research, namely predicting the spread of COVID-19. The analysis process is carried out by preprocessing the data by validating missing data and validating the format of the data variables. Then carry out descriptive analysis and data visualization so that it can be seen that in this 657 data there is a fluctuates data that non-periodically from July to August 2021. Next, model analysis is carried out using FBProphet and LR and validating the results of each model. The research results are in the form of evaluation metrics where the LR model gets better RMSE, MAE and MAPE values compared to FBProphet, namely 292.91; 178, 81 and 12.79%.*

Keywords*: covid-19, evaluation metrics, fbprophet, linear regression.*

Intisari—*Pada penelitian sebelumnya, kami melakukan analisis menggunakan model FBProphet untuk memprediksi wabah COVID-19 di Indonesia. Penerapan model FBProphet pada data time series dinilai cukup baik karena menghasilkan MAPE sebesar 22,60% dengan distribusi berbentuk linear. Selanjutnya, berdasarkan pola pada dataset sebelumnya dan total kasus aktif saat ini sudah mencapai 2.606 kasus maka pada penelitian kali ini kami mencoba untuk menggunakan model Linear Regresi (LR) sebagai pembanding dengan model FBProphet dengan menggunakan data tambahan dari sumber data yang sama yaitu dari website KAWALCOVID19. Pengambilan data dimulai dari tanggal 2 Maret 2020 hingga 19 Desember 2021. Tujuan dari penelitian ini sama dengan penelitian sebelumnya yaitu memprediksi penyebaran COVID-19. Proses analisis dilakukan dengan preprosesing data dengan memvalidasi data missing dan memvalidasi format dari variabel data. Kemudian melakukan analisis deskriptif dan visualisasi data sehingga terlihat bahwa pada 657 data ini terdapat data yang berfluktuasi non periodik pada bulan Juli hingga Agustus 2021. Selanjutnya dilakukan analisis model dengan menggunakan FBProphet dan LR serta memvalidasi hasil dari masing-masing model tersebut. Hasil penelitian berupa metrik evaluasi dimana model Regresi Linier mendapatkan nilai RMSE, MAE dan MAPE yang lebih baik dibandingkan dengan FBProphet yaitu 292.91; 178, 81 dan 12,79%.*

Kata Kunci*: covid-19, metrik evaluasi, fbprophet, linear regresi.*

VOL. 9. NO. 2 FEBRUARY 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v9i2.5174

INTRODUCTION

Even though the COVID-19 pandemic has turned into an endemic after almost three years, there are still additional new cases in Indonesia. On December, 30 2023 in Indonesia there were 318 active cases with 1 death. Meanwhile, the total active cases as of the end of December 2023 reached 2,606 [1]. Although this increase isn't as worrying as before, where deaths and seriously ill patients being hospitalized only reached less than 5 cases per week. However, it is possible that there will be an increase in 2024. This could be due to a lack of vaccine supply and the vaccine policy that must be paid for starting January, 1 2024 [2].

Based on Kementrian Kesehatan vaccine website, it is known that the vaccination target in Indonesia is 234,666,020. However, in reality as of February, 2 2024, vaccination has only reached 86.88% for dose 1, 74.56% for dose 2, 39.08% for dose 3 and 2.02% for dose 4 [3]. So further analysis is needed related to predicting the spread of COVID-19.

Prediction of COVID-19 cases is used as a method to estimate the increase in the number of cases in the future. Several prediction methods that have been used include technical analysis, medical analysis, and the use of intuition, but a better method is using machine learning in the case of predicting COVID-19 cases [4].

In previous research [5], we carried out an analysis using the FBProphet model to predict the COVID-19 outbreak in Indonesia. The results of the analysis in the form of MAPE for confirmed, dead and recovered patients were 22.60%, 21.67% and 22.53%. The dataset used in the research was taken from March 2, 2020 to February 2021. The distribution of data in the previous dataset was increasing from day to day and there was no fluctuating data. The application of the FBProphet model to the time series data is considered quite good because the MAPE results are <50%. So in further research, we also use the FBProphet model with the same data source from the KAWALCOVID19 website by adding a dataset up to December 19, 2021. In this 657 data there is non-periodic fluctuating data in July to August 2021 [6]. The use of this model is based on the advantages of FBProphet itself that can handle data that tends to be non-linear well and allowing the model to accommodate more complex fluctuations.

Beside FBProphet, there are several previous studies on COVID-19 predictions were carried out using machine learning [7], [8], [9], [10], [11], [12] and deep learning algorithms [13], [14], [15]. In Brazil, the Boltzmann function method was used to estimate the COVID-19 case [16], the results of the study showed a substantial increase in infection cases. In Jakarta, Indonesia, the impact of COVID-19 death was analyzed using the ARIMA method [17]. In Egypt, Pandemi COVID-19 predictions use Nonlinear Auto Regressive Artificial Neural Networks (NARANN) and Autoregressive Integrated Moving Average (ARIMA) [18]. The NARRAN model shows a better performance than ARIMA based on different statistical criteria. Furthermore, in research using four machine learning models, namely Linear Regression (LR), Lasso, Support Vector Machine (SVM), and Exponential Smoothing (ES) which predict the number of new infected cases, the number of deaths, and the number of recovered cases of COVID-19 patients throughout the world for the next 10 days [19] shows that the LR and LASSO models also show good performance in estimating the death rate and confirmed positive cases, to some extent. In Nigeria, a LR model was used to predict the prevalence of COVID-19 [20] based on travel and contact history, and the predictions showed to be very close.

Based on the literature review, it can be concluded that LR shows good performance in predicting COVID-19 cases. Besides that, in previous research [5], data on the spread of COVID-19 was linearly distributed. So apart from using the FBProphet model, this research will also use the LR model. The advantage of the LR model itself is that it can be quite resistant to outliers if the amount of data is large enough and isn't too influenced by extreme data. The results of the metric evaluation in this prediction model are RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error) values by calculating the difference between the predicted value and the actual value, where if the MAPE value has a smaller percentage then the accuracy will be better.

MATERIALS AND METHODS

Apart from that, in previous research, the trend analysis of the spread of COVID-19 was carried out using the FBProphet algorithm and resulted in a fairly good MAPE. The use of the FBProphet algorithm itself is considered capable of handling data that tends to be non-linear well and allows the model to accommodate more complex fluctuations such as in the data for this research.

FBProphet is an algorithm used to make predictions from time series data based on an additive model where non-linear trends are matched in time series annually, weekly and daily, with holiday effects. This algorithm works very well with time series data that have strong seasonal effects and datasets that have a lot of data. Prophet is robust to missing data and trend movements, and can also handle outliers well. Basically, Prophet is an additive regression model with four main components, namely:

1. Linear or logistic growth curve. Prophet can automatically detect trend changes by selecting change points from the data.
2. Annual trend components modeled using the Fourier Series.
3. Weekly trend component using dummy variables.
4. List of important holidays, provided by the user.

So FBProphet is very suitable for predicting time series data such as the spread of COVID-19. Meanwhile, the following is the basic formula from Prophet [21]:

$$k(t) = tr(t) + se(t) + ho(t) + id(t) \qquad (1)$$

where:
$tr(t)$ = trend
$se(t)$ = seasonal
$ho(t)$ = holiday
$id(t)$ = individual

Furthermore, the Linear Regression (LR) algorithm as a supervised machine learning algorithm used to predict continuous target variables based on one or more predictor variables is used in this research. The reason is because the nature of linear regression can tolerate the presence of outlier or extreme data, so it is suitable for non-periodic fluctuating data which occurs in the COVID-19 dataset in the current research. The following is a detailed explanation of the Linear Regression algorithm [22]:

$$Y = \alpha + \beta X + \varepsilon \qquad (2)$$

where:
X = independent variable
α = intercept
β = slope
ε = error

The α and β values can be found using the Least Square method which functions to minimize the error value between actual data and predicted data. Given a sample of data values n with points $(X_1, Y_1), (X_2, Y_2), \dots (X_n, Y_n)$, the regression coefficient can be found using the following formula [22]:

$$\beta = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \qquad (3)$$

$$\alpha = \bar{y} - \beta\bar{x} \qquad (4)$$

where:
$\bar{x}$ = average of $x_1, x_2, \dots, x_s$
$\bar{y}$ = average of $y_1, y_2, \dots, y_s$

The research method can be seen in Figure 1.

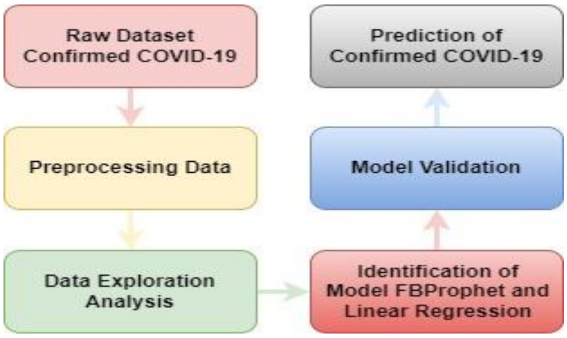

Figure 1. Research Methods

A. Raw Dataset Confirmed COVID-19

The dataset used in this research was taken from the KAWALCOVID19 website. The data used is data on confirmed COVID-19 patients starting from March 2 2020 to December 19 2021. The variables used in this study consist of the “date” variable and the number of “confirmed” COVID-19 patients on that date. The file format obtained is a CSV file.

```
data = pd.read_excel('/content/covid-19_confirmed.xlsx')
data
```

Figure 2. Raw Dataset Confirmed COVID-19

B. Preprocessing Data

Perform data preprocessing to clean data from noise and remove irrelevant information. The process begins by identifying missing data and validating the entry format for each variable (“date“ for date variable and “numeric” for confirmed variable).

```
data.info()
data.isnull().sum()
```

Figure 3. Identifying Missing Data

```
data_missval_cols = data[['Date', 'Confirmed']]
data_missval_cols.boxplot()
plt.title('Boxplot - To detect outliers')
plt.xlabel('Variable')
plt.ylabel('Values')
plt.show()
```

Figure 4. Identifying Outlier

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 9. NO. 2 FEBRUARY 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v9i2.5174

```
data['Date'] = pd.to_datetime(data['Date'], format='%Y/%m/%d')
x = data['Date']
date = pd.DataFrame(x, columns=['Date'])
df = data.loc[:,'Confirmed']
merge = pd.concat([date, df], axis=1)
confirmed_df = pd.DataFrame(merge.loc[:,'Date':'Confirmed'])
confirmed_df.head()
confirmed_df['Time'] = np.arange(len(confirmed_df.index))
confirmed_df.head()
```

Figure 5. Validating Data

C. Data Exploration Analysis

Exploratory Data Analysis (EDA) was carried out to understand the characteristics and structure of the data through visualization of the dataset distribution and descriptive analysis. Identification of small fluctuations or noise in the data can help in selecting appropriate filtering methods or noise handling techniques to clean the data before building the model.

```
confirmed_df.describe()
```

Figure 6. Descriptive Analysis

```
fig = plt.figure(dpi = 100)
ax = plt.axes()
ax.set(xlabel = 'Date',ylabel = 'Count of Cases',title = 'Confirmed Cases')
merge.plot(x='Date',y='Confirmed',label='Confirmed',legend=True,figsize=(10,6),ax=ax,lw=2)
```

Figure 7. Visualization of Distribution of Dataset

D. Identification of Model FBProphet and Linear Regression

To identifying model of FBProphet and LR, we need to create a dataframe containing the future dates we want to predict. The resulting data framework can be used to compare the performance of the two models in predicting future value based on historical data.

```
prophet = Prophet(daily_seasonality=True)
prophet.fit(confirmed_df)
future_dates = prophet.make_future_dataframe(periods=365)
predictions = prophet.predict(future_dates)
from prophet.plot import plot_plotly
plot_plotly(prophet, predictions)
```

Figure 8. FBProphet Model

```
confirmed_df['Lag_1'] = confirmed_df['Confirmed'].shift(1)
confirmed_df.head()
X = confirmed_df.loc[:, ['Lag_1']]
X.dropna(inplace=True)
y = confirmed_df.loc[:, 'Confirmed']
y, X = y.align(X, join='inner')

model = LinearRegression()
model.fit(X, y)

y_pred = pd.Series(model.predict(X), index=X.index)
```

Figure 9. Linear Regression Model

E. Model Validation

Validate the prediction model using evaluation metrics such as RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error). RMSE is used to measure how close the prediction model is to the actual value. Lower RMSE values indicate better performance.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(y_t - y\_pred_t)}{n}} \quad (5)$$

Meanwhile, MAE is used to influence each deviation by the same amount.

$$MAE = \sum_{t=1}^{n} \frac{|y_t - y\_pred_t|}{n} \quad (6)$$

Next, MAPE is used to measure the average percentage of prediction error relative to the actual value. MAPE is measured in percentages and lower values indicate better performance.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{|y_t - y\_pred_t|}{y_t}\right| x100 \quad (7)$$

where:
n = lots of data
y_t = actual data
y_pred_t= prediction data

```
rmse = np.sqrt(mean_squared_error(y, y_pred))
mae = mean_absolute_error(y, y_pred)
mape = calculate_mape(y, y_pred)
```

Figure 10. Model Validation

F. Prediction of Confirmed COVID-19

After obtaining the best model between FBProphet and LR as seen from the results of the metric evaluation analysis, predictions are then made for the next data.

Although LR isn't specifically designed to model time series data, in some situations, the use of linear regression models on time series data can be justified or provide adequate results. One of the reasons for using this model is that the dataset has a simple structure and there are no elements of temporal dependency so that linear regression might be able to produce quite good prediction values. However, it doesn't rule out the possibility that this model is not suitable for the type of dataset (time series).

RESULTS AND DISCUSSION

A. Dataset

The dataset used was 657 which came from the KAWALCOVID19 website starting from March 2,

VOL. 9. NO. 2 FEBRUARY 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v9i2.5174

2020 to December 19, 2021. The dataset is made into two columns consisting of the 'Date' and 'Confirmed' attributes. Then the 'Date' data is converted into datetime64[ns] with the yyyy-mm-dd format and the 'Confirmed' data into a numeric format as a predictive variable. The dataset used in this research can be seen in Table 1.

Table 1. Dataset confirmed COVID-19

	Date	Confirmed
0	2020-03-02	2
1	2020-03-03	0
2	2020-03-04	0
3	2020-03-05	0
4	2020-03-06	2
...	...	...
652	2021-12-15	205
653	2021-12-16	213
654	2021-12-17	291
655	2021-12-18	232
656	2021-12-19	164

The dataset in Table 1 has been cleaned and will be used to create a model and analyze it. This data is a patients who has confirmed COVID-19. The descriptive analysis of the dataset can be seen in Table 2.

Table 2. Descriptive analysis of confirmed COVID-19

	Confirmed
count	657
mean	6216
std	9,631.82
min	0
25%	922
50%	3989
75%	6725
max	59,532

Based on Table 2, the average number of confirmed patients is 6,516. However, the standard deviation value is quite large, namely 9,631.82 with a min value of 0 and a max value of 59,532. The resulting mean, standard deviation and min-max ranges identify data distribution that isn't normally distributed (fluctuations occur). Next, we can see the shape of the data distribution through graphic visualization in Figure 11.

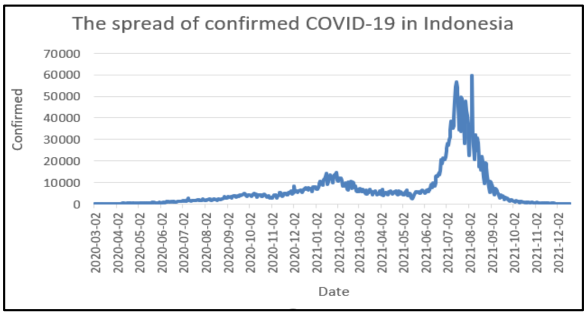

Figure 11. The Spread of Confirmed COVID-19

Based on Figure 11, there was a fluctuating (extreme) distribution in July to August 2021.

B. Prediction Model of FBProphet

The dataset is divided into two parts, namely 80% training data and 20% test data from 657 days. Then we trained the model with 525 data to predict the next 156 data. Next, plot the results of the predictions and then evaluate them. Figure 12 below is the prediction result using the FBProphet model.

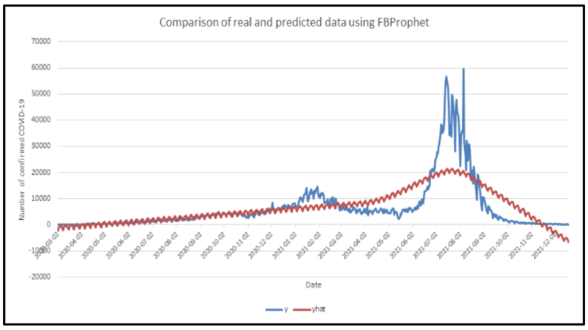

Figure 12. Prediction Model of FBProphet for Confirmed COVID-19

From Figure 12, the FBProphet model is able to fit the train data but fails to predict the test data. This happens because the model overfits and produces negative predictions. For details of the distribution of train data against the model, it can see in Figure 13 and test data in Figure 14.

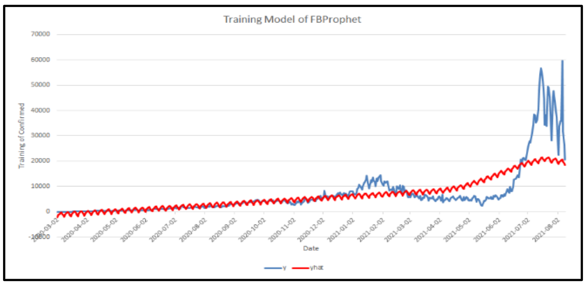

Figure 13. Training Model of FBProphet for Confirmed COVID-19

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)
VOL. 9. NO. 2 FEBRUARY 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v9i2.5174

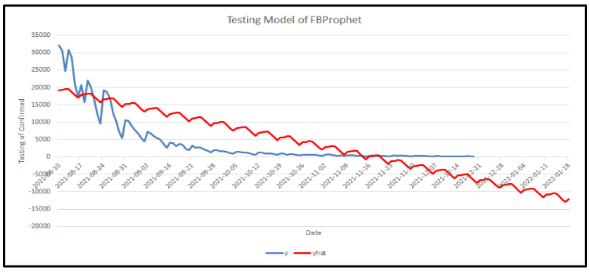

Figure 14. Testing Model of FBProphet for Confirmed COVID-19

In Figure 13, the model training is almost close to the spread of the actual data, while in Figure 14, the model predictions gradually decrease until they reach negative values (not close to the actual data). In fact, the predicted value is increasingly negative for the next month from December 20, 2020 to January 18, 2023. Meanwhile, a comparison of the real dataset and model predictions can be seen in Table 3.

Table 3. Actual vs prediction model FBProphet for Confirmed COVID-19

	Date	y (Real)	yhat (Prediction)
0	2020-03-02	2	-2224
1	2020-03-03	0	-1207
2	2020-03-04	0	-888
3	2020-03-05	0	-450
4	2020-03-06	2	-242
....	....	....	....
652	2021-12-14	190	-5329
653	2021-12-15	205	-5229
654	2021-12-16	213	-5010
656	2021-12-17	291	-5021
657	2021-12-18	232	-5848

Overfitting and negative predictive value can occur for several reasons:

1. High Model Complexity: If FBProphet is configured with many seasonal components and changepoints that exceed the complexity in the real data, the model can overfit the training data, thereby failing to accurately predict future data.
2. Unrepresentative Training Data: If the training data doesn't cover enough variation or enough periods to capture repeating patterns, the model can rely too much on noise or specific patterns from the training set (there is non-periodic fluctuating data in July to August 2021)
3. Lack of Regularization: Regularization helps prevent overfitting by adding a penalty to model complexity. The lack of regularization in FBProphet can cause the model to overfit with detailed noise from the training data.
4. Unstationary: Make sure time series data is stationary, or consider using methods such as differentiation to eliminate unwanted trends or seasonality.

Strategies to overcome this can be:

1. Reduce model flexibility by adjusting hyperparameters, such as reducing the number of changepoints or adjusting other hyperparameters to limit model complexity.
2. Use cross-validation to evaluate the model's performance on the test dataset and ensure that the model isn't overly specific to a particular training dataset.

Hyperparameter adjustments and stationary processes weren't carried out in this research because the researchers wanted the data trends to match the original and there was no data engineering that was too far from the original data. Meanwhile, the cross-validation process had been carried out previously but still obtained inappropriate values.

C. Model Prediction of Linear Regression

We use previous values as features (lag variables) to train a linear regression model. After training the model, then we make predictions for the test data. Finally, the predicted results and the actual values are visualized to compare them. The visualition of prediction results of the Linear Regression (LR) model can be seen in Figure 15.

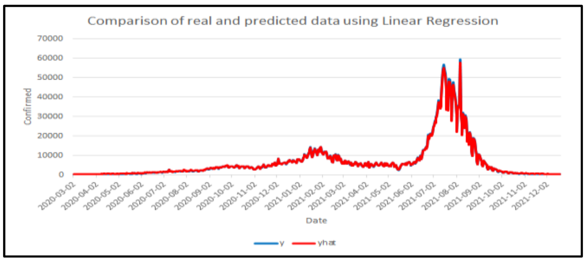

Figure 15. Prediction Model of Linear Regression for Confirmed COVID-19

In Figure 15, the prediction results of the training model using the LR model are almost close to the actual data. Likewise with the results of testing the data. Meanwhile, a comparison of the real dataset and model predictions can be seen in Table 4

Table 4. Actual vs prediction model Linear Regression for Confirmed COVID-19

	Date	y (Real)	yhat (Prediction)
0	2020-03-02	2	201
1	2020-03-03	0	199
2	2020-03-04	0	199
3	2020-03-05	0	199
4	2020-03-06	2	201
....	....	....	....
652	2021-12-14	190	398
653	2021-12-15	205	406
654	2021-12-16	213	481
656	2021-12-17	291	424
657	2021-12-18	232	425

Linear Regression Models can often have good performance in predicting COVID-19 cases compared to FBProphet, especially in some specific cases. There are several reasons why this could happen:

1. Model Interpretation, LR provides a direct and intuitive interpretation of the relationship between an independent variable (e.g., time) and a dependent variable (number of COVID-19 cases). This can help in better understanding what factors influence the spread of the disease.
2. Model Simplicity, LR Models have great simplicity compared to models like FBProphet which have many components such as seasonal trends, holidays, and changepoints. In the context of COVID-19 predictions, this simplicity can be useful because not all COVID-19 data has clear seasonal patterns or significant point changes. In other words, Linear Regression can more easily adapt to data that may be more linear in nature.
3. Flexibility in Data Processing, LR can easily handle data transformations such as log-transform or outlier handling, which can help improve model performance and overcome problems such as negative or extreme values in predictions.
4. FBProphet Limitations, Although FBProphet is a powerful tool for time series modeling, it also has limitations in terms of datasets that require special transformations or have patterns that do not match the assumptions of the FBProphet model. If the data doesn't have a clear seasonal component or significant point changes, FBProphet may not be able to produce accurate predictions.

D. Metric Evaluation

Metric evaluation is the final stage used to determine the difference between the predicted value between the model and the observed truth value. Based on the evaluation that has been carried out, Table 5 is a comparison of the results of RMSE, MAE and MAPE from the FBProphet and Linear Regression models.

Table 5. Evaluation Metric of RMSE, MAE and MAPE using FBProphet and Linear Regression

	RMSE	MAE	MAPE (%)
FBProphet	6263.26	3620.04	66.79
Linear Regression	292.91	178.81	12.79

In Table 5, the Linear Regression algorithm gets better RMSE, MAE and MAPE values compared to FBProphet, namely 292.91; 178, 81 and 12.79%. The smaller the result of the evaluation metric, the better the prediction model.

The analysis results using the LR model are better than FBProphet possibly because the dataset used doesn't require additional model complexity. FBProphet, although powerful for some time series forecasting cases, may be too complex for data that doesn't have highly seasonal patterns or doesn't require special handling of holidays and special events. Linear regression models are likely to be more efficient if the data has a relatively simple structure.

CONCLUSION

This research was conducted to see the differences in predictions using the FBProphet and Linear Regression models for time series data which are fluctuating but not periodic. Even though FBProphet is known as a model that is suitable for fluctuating data, in this research the FBProphet model received poor evaluation results when compared to the Linear Regression model. This happens because the model is overfit and produces negative predictions. The data used in this research is also follow-up data from previous research. The dataset obtained from the KAWALCOVID19 website consisted of 657 data starting from March 2 2020 to December 19 2021. Preprocessing was carried out by identifying missing data and formatting predictor variables. Next, the data is divided into two parts, namely 80% train data which is used to train the model and predict 20% test data. The research results showed that the Linear Regression algorithm obtained better RMSE, MAE and MAPE values compared to FBProphet, namely 292.91; 178, 81 and 12.79%. The analysis results using the LR model are better than FBProphet, possibly because the dataset used doesn't require additional model complexity. Limited data obtained can also cause this to happen. As can be seen in the graph of the spread of COVID-19, where initially the data was linear, but at certain times (in July-August 2021 data there was an unusual spike in data and only occurred in one time period). So the linear assumption occurs on this dataset and causes simple calculations using LR to be better than using FBProphet which may be too complicated for data that doesn't have seasonal patterns or require special handling of holidays and special events. In further research, it can be carried out by adding datasets so that we can see the distribution of further data distribution (whether there is still overfitting or not). If overfitting doesn't occur in the subsequent data, the hyperparamer adjustment process or cross validation can be

VOL. 9. NO. 2 FEBRUARY 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v9i2.5174

carried out to produce better predictions. In future research, you can use the Neural FBProphet model or other time series machine learning models.

REFERENCE

[1] Romys Binekasri, “Kasus Covid RI Naik di Nataru 2024, Tak Cuma Varian Baru,” CNBC Indonesia. Available: https://www.cnbcindonesia.com/tech/20231231091838-37-501574/kasus-covid-ri-naik-di-nataru-2024-tak-cuma-varian-baru. [Accessed: Jan. 05, 2024].

[2] DPR RI, “Kasus Varian Covid-19 Meningkat, Kebijakan Vaksin Berbayar per 1 Januari 2024 Waktunya Tidak Pas.” Available: https://www.dpr.go.id/berita/detail/id/48185/t/Kasus+Varian+Covid-19+Meningkat%2C+Kebijakan+Vaksin+Berbayar+per+1+Januari+2024+Waktunya+Tidak+Pas. [Accessed: Feb. 06, 2024].

[3] Kementrian Kesehatan, “Vaksinasi COVID-19 Nasional.” Available: https://vaksin.kemkes.go.id/#/vaccines. [Accessed: Feb. 02, 2024].

[4] L. Zhang, J. Zhu, X. Wang, J. Yang, X. F. Liu, and X. K. Xu, “Characterizing COVID-19 Transmission: Incubation Period, Reproduction Rate, and Multiple-Generation Spreading,” *Front Phys*, vol. 8, Jan. 2021, doi: 10.3389/fphy.2020.589963.

[5] N. Hayati, Fauziah1, D. R. Poetra, and D. Wandi, “Trend of the spread of COVID-19 in Indonesia using the machine learning prophet algorithm,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 3, pp. 1780–1788, Dec. 2021, doi: 10.11591/ijeecs.v24.i3.pp1780-1788.

[6] “Dataset COVID-19,” kawalcovid19. Available: http://www.kawalcovid19.id/. [Accessed: Jan. 09, 2024].

[7] A. Solanki and T. Singh, “COVID-19 epidemic analysis and prediction using machine learning algorithms,” *Emerging Technologies for Battling Covid-19*, 2021, doi: 10.1007/978-3-030-60039-6_3.

[8] M. Mele and C. Magazzino, “Pollution, economic growth, and COVID-19 deaths in India: a machine learning evidence,” *Environmental Science and Pollution Research*. Springer, 2021. doi: 10.1007/s11356-020-10689-0.

[9] P. Wang, X. Zheng, J. Li, and B. Zhu, “Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics,” *Chaos Solitons Fractals*, vol. 139, Oct. 2020, doi: 10.1016/j.chaos.2020.110058.

[10] R. Mattera, “A weighted approach for spatio-temporal clustering of COVID-19 spread in Italy,” *Spat Spatiotemporal Epidemiol*, 2022, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877584522000247

[11] M. Zivkovic, N. Bacanin, K. Venkatachalam, A. Nayyar, A. Djordjevic, I. Strumberger, and F. Al-Turjman, COVID-19 cases prediction by using hybrid machine learning and beetle antennae search approach, Sustainable Cities and Society, Volume 66, 2021, 102669, ISSN 2210-6707, https://doi.org/10.1016/j.scs.2020.10266.

[12] S. Tuli, S. Tuli, R. Tuli, and S. S. Gill, “Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing,” *Internet of Things (Netherlands)*, vol. 11, Sep. 2020, doi: 10.1016/j.iot.2020.100222.

[13] S. Sah, B. Surendiran, R. Dhanalakshmi, S. N. Mohanty, F. Alenezi and K. Polat, “Forecasting COVID-19 Pandemic Using Prophet, ARIMA, and Hybrid Stacked LSTM-GRU Models in India,” *Computational and Mathematical Methods in Medicine,* vol 2022, May 2022, doi: 10.1155/2022/155602.

[14] V. Aakash, S. Sridevi, G. Ananthi, and S. Rajaram, “Forecasting of Novel Corona Virus Disease (Covid-19) Using LSTM and XG Boosting Algorithms,” *Data Analytics in Bioinformatics: A Machine Learning Perspective*, 2021, doi: 10.1002/9781119785620.ch12.

[15] A. Tiwari, R. Gupta, and R. Chandra, “Delhi air quality prediction using LSTM deep learning models with a focus on COVID-19 lockdown,” *arXiv preprint arXiv:2102.10551*, 2021, [Online]. Available: https://arxiv.org/abs/2102.10551

[16] G. C. de Melo, R. A. de Araújo Neto, and K. C. G. M. de Araújo, “Forecasting the rate of cumulative cases of COVID-19 infection in Northeast Brazil: A Boltzmann function-based modeling study,” *Cad Saude Publica*, vol. 36, no. 6, 2020, doi: 10.1590/0102-311X00105720.

[17] F. Fadly and E. Sari, “An Approach to Measure the Death Impact of Covid-19 in Jakarta using Autoregressive Integrated Moving Average (ARIMA),” *Unnes Journal of Public Health*, vol. 9, no. 2, pp. 108–116, Jul. 2020, doi: 10.15294/ujph.v9i2.38460.

[18] A. I. Saba and A. H. Elsheikh, “Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks,” *Process Safety and*

Environmental Protection, vol. 141, pp. 1–8, Sep. 2020, doi: 10.1016/j.psep.2020.05.029.
[19] F. Rustam *et al.*, "COVID-19 Future Forecasting Using Supervised Machine Learning Models," *IEEE Access*, vol. 8, pp. 101489–101499, 2020, doi: 10.1109/ACCESS.2020.2997311.
[20] R. O. Ogundokun, A. F. Lukman, G. B. M. Kibria, J. B. Awotunde, and B. B. Aladeitan, "Predictive modelling of COVID-19 confirmed cases in Nigeria," *Infect Dis Model*, vol. 5, pp. 543–548, Jan. 2020, doi: 10.1016/j.idm.2020.08.003.
[21] "Prophet Forecasting at Scale." Available: https://facebook.github.io/prophet/. [Accessed: Jan. 09, 2024].
[22] Jeremy Arkes, *Regression Analysis a Practical Introduction second edition*. 2023. Available: https://imuacid-my.sharepoint.com/personal/naidi_imu_ac_id/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fnaidi_imu_ac_id%2FDocuments%2FDATA_NAIDI%2CS%2ESOS%2E%2CMM%2FDATA_EBOOK%2FEBOOK_001%2F06%20Regression%2EAnalysis%2EA%2EPrac_Intro%2E2nd%2EEd%2Epdf&parent=%2Fpersonal%2Fnaidi_imu_ac_id%2FDocuments%2FDATA_NAIDI%2CS%2ESOS%2E%2CMM%2FDATA_EBOOK%2FEBOOK_001&ga=1. [Accessed: Jan. 09, 2024].