Prediction Modelling of COVID-19 Outbreak in Indonesia using a Logistic Regression Model

F W Wibowo¹, Wihayati²

¹ Informatics Department, Universitas Amikom Yogyakarta, Jl. Ring Road Utara, Condong Catur, Depok, Sleman, Yogyakarta 55283, Indonesia
² Astrophysics and Mathematics Group, Sekip Utara Bulaksumber, Yogyakarta 55281, Indonesia

Email: ferry.w@amikom.ac.id¹, wihayati@gmail.com²

Abstract. The COVID-19 outbreak has changed the world at large since it was announced by the World Health Organization (WHO). Many policies in various countries were then implemented to control its spread. Most aspects of human life and the environment are affected by this pandemic. This paper aims to determine the prediction model for the COVID-19 outbreak in Indonesia. The approach used for this modelling employs a logistic regression model. The data modeled in this paper is data on the distribution of COVID-19 sufferers and data on patients who have recovered from COVID-19 in Indonesia. The data obtained as research material were taken from March 2, 2020, to November 12, 2020. From the results of this paper, this prediction model obtained logistic regression coefficient values for data on COVID-19 sufferers in Indonesia of 8.114748 and 0.750743, while the coefficient values for data on sufferers who recovered from COVID-19 were 9.360925 and 0.788334. The results of the prediction model for sufferers and people who have recovered from COVID-19 have the same accuracy value, namely mean absolute error (MAE) of 0.02, mean square error (MSE) of 0.00, and $R^2$ of 0.99.

1. Introduction
A coronavirus is a virus that causes infection from common fever to severe acute respiratory syndrome (SARS). Several new variants of the coronavirus have emerged, such as SARS CoV-2 which causes COVID-19. The SARS CoV-2 has the potential to spread widely due to the lack of immunity from humans. Symptoms of being attacked by COVID-19 will appear from 1 to 14 days and are similar to the flu. Most of the sufferers will experience a fever above 37.8 °C and a dry cough, but it is usually accompanied by dizziness, diarrhea, and so on. About 20% of sufferers, on the seventh day, will experience shortness of breath, pneumonia, and lung inflammation [1]. About 30% of sufferers do not appear these symptoms. About 80% of sufferers have mild symptoms, while the more severe diseases are in older people, diabetics, heart disease, chronic respiratory disease, impaired immune system. Only about 1% of sufferers are at high risk. The worst risk that can occur when someone suffering from COVID-19 is acute respiratory distress syndrome, multi-organ failure, and death [2]. This is what has made the world health organization (WHO) declared that COVID-19 is a pandemic. WHO has classified the level of global risk from low risk to high risk [3]. So that the COVID-19 outbreak is a concern for all stakeholders around the world to control and anticipate its spread [4]. The government will find it difficult to overcome the coronavirus disease 2019 (COVID-19) which also has an impact on the economy at large. One possible way to deal with it is to control the death rate as low as possible [5]. Control of the COVID-19 outbreak is defined as a situation where there are no
new infections due to COVID-19 between 12 and 16 weeks after the case began. An outbreak of up to 5000 cases will be difficult to control in that time, so it is classified as an uncontrolled outbreak [6]. So that all possibilities can be done in controlling the spread of this outbreak, one of which is the predictive approach.

The phenomenon of the emergence of COVID-19 has made many researchers predict the rate of spread of this virus. The prediction results in the distribution of COVID-19 are used as one of the recommendations for controlling the spread of this virus. If it is not handled quickly, there will be a sizeable spike in the distribution in the coming period. So that it is very necessary to get attention related to worldwide problems. The accuracy of predictions with the original data is necessary to obtain a fairly intensive study. This paper addresses predictions by modelling the rate of spread of COVID-19 and the rate of recovery from COVID-19 in Indonesia. The approaching model used in this research paper is a logistic regression model. The results obtained from this approach are used to obtain the coefficient value of the logistic regression analysis model. While the test parameter used is the goodness of fit test, namely the R-square ($R^2$). Not only using $R^2$ to test the model but also to get a comparison of the accuracy of the resulting prediction models, this paper uses mean absolute error (MAE) and mean square error (MSE).

2. Related Works

The COVID-19 has had a significant impact on human life around the world. Several number of countries have implemented many regulations to control its spread. However, it could not be known yet how long this pandemic will end. Many studies have predicted trends in the rate of spread of this virus. Until now, this is still a challenge for researchers to predict the trend. Research [7] has used data before 16 June 2020 using a logistical model and predicts that the global outbreak will reach its peak in late October. Five countries are the object of the research because the largest total number of countries confirmed with COVID-19, one of which is Indonesia. Research related to the spread rate of COVID-19, rate of recovery, and mortality due to COVID-19 was conducted by [8] using the support vector regression (SVR) regression method and the Bayesian regression model. With data taken from March 6, 2020, to March 26-27, 2020, it was found that the Bayesian Ridge approach provides simulation results that are closer to the original data than using the Support Vector Machine (SVM) approach. Paper [9] used an exponential growth modelling approach to understand the distribution pattern of COVID-19 across 42 countries. Several parameters were included as variables in the predictive guided machine learning model such as infrastructure, environment, policies, and infections related to the independent variables. The machine learning models developed in this case are decision tree, logistic regression, random forest, and SVM. The accuracy values obtained were between 76.2% and 92.9%. Paper [10] has developed a COVID-19 outbreak prediction system using machine learning modelling consisting of Auto-Regressive Integrated Moving Average (ARIMA), Auto-Regressive Moving Average (ARMA), Bayesian Ridge Regressor, Holt-Winter Regressor, Linear Regression, Linear Regressor, Random Forest Regressor, Support Vector Regressor, and XGBoost Regressor. The prediction model results obtained have an average accuracy of about 88% by taking data from 10 countries with dense populations, including Indonesia. Meanwhile, the highest accuracy of 99.93% was obtained for Ethiopia using ARMA. Meanwhile, this paper uses data before 12 November 2020 and models the rate of spread of COVID-19 and the rate of recovery from COVID-19 in Indonesia. The research conducted in this paper uses the logistic regression analysis method.

3. Research Methodology

The research dataset related to the COVID-19 in Indonesia used in this paper is taken from the website: https://data.humdata.org/dataset/indonesia-covid-19-cases-recoveries-and-deaths-per-province. The dataset provided on this website is updated from data provided by Badan Nasional Penanggulangan Bencana (BNPB - Indonesia National Disaster Management Authority), which is a national COVID-19 task force, where the source is taken from the website: https://opendata.arcgis.com/datasets/db3be1cadcf44b6fa709274c12726c59_0.geojson. The dataset used in this study was taken from March 2, 2020, to November 12, 2020. The dates in this dataset are then converted into the sequence of days on which data collection on the spread of the COVID-19
virus in Indonesia began. This dataset consists of the cumulative number of positive cases of COVID-19 detected, the number of cases that have recovered from COVID-19, the number of deaths after suffering from COVID-19, the number of patients in the hospital due to being exposed to COVID-19, the number of cases of new COVID-19 sufferers per days, the number of cases recovered from COVID-19 per day, the number of cases of death due to COVID-19 per day, and the number of cases treated per day. This paper only uses two data that will be used in predictive analysis related to COVID-19, namely cases of COVID-19 infection and cases recovering from COVID-19 due to COVID-19, as shown in Figure 1.

Figure 1 shows that the cases of patients who recovered were significantly larger than those of patients who died caused by COVID-19 in Indonesia. The data obtained will be analyzed using a regression model. The regression model is an analysis that has the aim of measuring the effect of a variable on other variables. The simple linear regression model could be shown in equation 1.

\[ y = \alpha_0 + \alpha_1 \psi + \epsilon \]  

(1)

Where \( y \) is the predicted or dependent variable, \( \psi \) is the independent variable, \( \alpha_0 \) is a constant, \( \alpha_1 \) is the regression coefficient, while \( \epsilon \) is a variance error or residual.

3.1. Logistic Regression Analysis Model

A logistic regression analysis model is a statistical data analysis method to describe the relationship between the predicted variable which has two or more categories with one or more independent variables on a continuous or category scale. The logistic regression method has techniques and procedures similar to the linear regression method. If the linear regression method in estimating parameter values often uses the ordinary least squares (OLS) method, then the logistic regression method estimates the parameter values using the maximum likelihood estimation (MLE) method. The MLE method maximizes the chances of classifying observed objects into appropriate categories and then converting them into simple regression coefficients. To find the logistic regression equation, the model used is shown in equation 2.
\[ \mu(\psi) = \frac{e^{\alpha_0 + \sum_{i=1}^{n} \alpha_i \psi_i}}{1 + e^{\alpha_0 + \sum_{i=1}^{n} \alpha_i \psi_i}} \]  

(2)

Where \( \mu \) is a function. From equation 2, it can obtain \( 1 - \mu(\psi) \), as shown in equation 3.

\[ 1 - \mu(\psi) = 1 - \frac{e^{\alpha_0 + \sum_{i=1}^{n} \alpha_i \psi_i}}{1 + e^{\alpha_0 + \sum_{i=1}^{n} \alpha_i \psi_i}} \]  

(3)

So, it can be written as shown in equation 4.

\[ 1 - \mu(\psi) = 1 + e^{\alpha_0 + \sum_{i=1}^{n} \alpha_i \psi_i} - e^{\alpha_0 + \sum_{i=1}^{n} \alpha_i \psi_i} \]  

(4)

Equation 4 can be simplified, as shown in equation 5.

\[ 1 - \mu(\psi) = \frac{1}{1 + e^{\alpha_0 + \sum_{i=1}^{n} \alpha_i \psi_i}} \]  

(5)

Thus, to find \( \frac{\mu(\psi)}{1 - \mu(\psi)} \) can be shown in equation 6.

\[ \frac{\mu(\psi)}{1 - \mu(\psi)} = e^{\alpha_0 + \sum_{i=1}^{n} \alpha_i \psi_i} \]  

(6)

Then, equation 6 can be modeled as \( \zeta(\psi) \) which is shown in equation 7.

\[ \zeta(\psi) = \ln \left( \frac{\mu(\psi)}{1 - \mu(\psi)} \right) \]  

\[ \zeta(\psi) = \ln \left( e^{\alpha_0 + \sum_{i=1}^{n} \alpha_i \psi_i} \right) \]  

\[ \zeta(\psi) = \alpha_0 + \sum_{i=1}^{n} \alpha_i \psi_i \]  

(7)

It can be written as shown in equation 8.

\[ \zeta(\psi) = \ln(\text{odds}) \]  

(8)

Usually, \( \ln(\text{odds}) \) is known as logit. The types of the logistic regression model are binary logistic regression, multinomial logistic regression, and ordinal logistic regression. The binary logistic regression model is used to analyze the relationship between one response variable and several predictor variables, with the response variable in the form of dichotomous qualitative data, which is 1 to indicate the existence of a characteristic and 0 to indicate the absence of a characteristic. The binary logistic regression model is used if the response variable produces two categories of 0 and 1.

### 3.2. Testing Parameters

Testing parameters in the logistic regression model needs to be done. This test is used to determine whether the independent variable in the model is significant to the dependent variable or not. There are three models of test that can be done, namely Goodness of fit test, simultaneous test, and partial test. The goodness of fit test is carried out to find out how well the logistic regression model can explain the relationship between the dependent and independent variables. In this regression, the parameter seen in the goodness of fit test is Pseudo R\(^2\) or known also as the coefficient of determination. Pseudo R\(^2\) is an artificial R-square that is used because no one has replaced the OLS R-square in the logit model. The R-square equation can be written as shown in equation 9.
Where \( \varphi \) is a sum squared regression (SSR) which is the sum of residuals squared (see equation 10). The residual is the difference between the dependent variable \( \gamma \) and observed values \( \gamma' \). Whereas \( \delta \) is the total sum of squares (SST) which is the distance sum of the data is away from the mean all squared \( \gamma \).

\[
R^2 = 1 - \frac{\varphi}{\delta} 
\]

The simultaneous test aims to determine the effect of the independent variables simultaneously on the dependent variable. The partial test is done by testing each \( \alpha \) individually. In the logistic regression model, \( \alpha \) shows the magnitude of the difference between the value of the dependent variable when the independent variable \( \psi + 1 \) and the value of the dependent variable when the independent variable \( \psi \) is for each value of \( \psi \). For dichotomous independent variables, it is assumed that the \( \psi \) values are 0 and 1. The values of the logistic regression model for the binary independent variables \((0, 1)\) are shown in Table 1.

### Table 1. The values of the logistic regression model for the binary independent variables \((0, 1)\).

| \( \psi \) | \( \gamma = 0 \) | \( \gamma = 1 \) |
|---|---|---|
| \( \gamma = 0 \) | \( 1 - \mu(0) = \frac{1}{1 + e^{\alpha_0}} \) | \( 1 - \mu(1) = \frac{1}{1 + e^{\alpha_0 + \alpha_1}} \) |
| \( \gamma = 1 \) | \( \mu(0) = \frac{e^{\alpha_0}}{1 + e^{\alpha_0}} \) | \( \mu(1) = \frac{e^{\alpha_0 + \alpha_1}}{1 + e^{\alpha_0 + \alpha_1}} \) |

The odds value for the occurrence of the dependent variable among the independent variables that have \( \psi = 0 \) is shown in equation 11.

\[
\frac{\sigma(\gamma = 1 | \psi = 0)}{\sigma(\gamma = 0 | \psi = 0)} = \frac{\mu(0)}{1 - \mu(0)} 
\]

While the odds value for the occurrence of the dependent variable among the independent variables having \( \psi = 1 \) is shown in equation 12.

\[
\frac{\sigma(\gamma = 1 | \psi = 1)}{\sigma(\gamma = 0 | \psi = 1)} = \frac{\mu(1)}{1 - \mu(1)} 
\]

If the odds values in equations 11 and 12 are converted into log form as equation 8. For \( \psi = 0 \), it will be as shown in equation 13.

\[
\zeta(0) = \ln\left(\frac{\mu(0)}{1 - \mu(0)}\right) 
\]

Meanwhile, for \( \psi = 1 \), it will be as shown in equation 14.

\[
\zeta(1) = \ln\left(\frac{\mu(1)}{1 - \mu(1)}\right) 
\]

From the model equations obtained in Table 1, the log odds ratio \( \zeta \) is obtained. The log odds ratio is the difference in the logit value as shown in equation 15.
Thus, the value of $\xi$ is obtained, as shown in equation 16.

$$\xi = e^{\alpha_1}$$

So, this value can be written as shown in equation 17.

$$\ln(\xi) = \alpha_1$$

### 4. Result and Discussion

The common sigmoid curve method can determine the function of the COVID-19 distribution modelling from the provided dataset [11]. In this case, the model function of the affected data is shown in equation 18.

$$\gamma(y) = \frac{1440000}{1 + e^{0.019(292-y)}}$$

The approximation result of equation 18 can be simulated as shown in Figure 2.

![Figure 2. The model of the COVID-19 affected data in Indonesia.](image)

This model was then confirmed using modelling of the logistic regression method. The results of the logistic regression model coefficients from being exposed to COVID-19 and recovering from COVID-19 are shown in Table 2.

| Cases       | $\alpha_0$ | $\alpha_1$ |
|-------------|------------|------------|
| Affected    | 8.114748   | 0.750743   |
| Recovered   | 9.360925   | 0.788334   |

In this research, 80% of the entire dataset was used as the training data, the rest was used as the test data. The fit models obtained from the coefficient values of sufferers and people recovering from COVID-19 are shown in Figures 3 and 4, respectively.
In measuring the accuracy of the prediction results, there are three types of significant measurements, namely the mean square error (MSE), mean absolute percentage error (MAPE), and mean absolute error (MAE). MSE is calculated by summing the squares of all prediction errors in each period and dividing it by the number of predicted periods, in contrast to MAPE which states the percentage error of forecasting results against actual demand during a certain period. MAE is generated from the absolute value of the difference between the output model values and the actual data. In this paper, two types of measurements are used, namely MSE and MAE. Meanwhile, to see the test parameters $R^2$ will be used. The results of MSE, MAE, and $R^2$ are presented in Table 3.

Table 3. The Value of MAE, MSE, and $R^2$ for affected and recovered from COVID-19.

| Cases    | MAE | MSE | $R^2$ |
|----------|-----|-----|-------|
| Affected | 0.02| 0.00| 0.99  |
| Recovered| 0.02| 0.00| 0.99  |

Table 3 shows that the MAE value shows that it is very small for both cases, namely 0.2. This shows that the prediction model that has been done is very good. While the number of MSE is equal to zero, this indicates that a small MSE is the best prediction method and the prediction accuracy is very high. The results of $R^2$ on data where Indonesians are infected with COVID-19 and Indonesians recover show a value of 0.99. This indicates that the independent variable can explain the dependent variable by 99 percent or the variation in the dependent variable that can be explained by the model is 99 percent. In logistic regression models, the main things that must be considered are the indicators of the model's significance, the significance of the independent variables, and the direction of the coefficients of these variables.

5. Conclusion

The prediction model from this paper has been carried out and obtained logistic regression coefficient values for data on COVID-19 sufferers in Indonesia of 8.114748 and 0.750743, while the coefficient values for data on patients who recovered from COVID-19 were found to be 9.360925 and 0.788334. The prediction model accuracy has also been tested with good results. The results of the prediction model for sufferers and people who have recovered from COVID-19 have the same accuracy values, namely MAE of 2%, MSE of 0%, and $R^2$ of 99%. From this model, it can be concluded that there is an upward trend in COVID-19 sufferers and people recovering from COVID-19 in Indonesia. Future research will examine areas in Java because the sufferers from this island are quite large in Indonesia.

6. References

[1] Lauer S A, Grantz K H, Bi Q, Jones F K, Zheng Q, Meredith H R, Azman A S, Reich N G, and Lessler J 2020 The Incubation Period of Coronavirus Disease 2019 (COVID-19) From
Publicly Reported Confirmed Cases: Estimation and Application \textit{Ann Intern Med.} doi:10.7326/M20-0504

[2] Lippi G, Lavie C J, and Sanchis-Gomar F 2020 Cardiac troponin I in patients with coronavirus disease 2019 (COVID-19): Evidence from a meta-analysis \textit{Prog Cardiovasc Dis} vol 63 no 3 pp 390-391 doi:10.1016/j.pcad.2020.03.001

[3] WHO 2020 Coronavirus disease 2019 (COVID-19) \textit{Situation Report-94}

[4] Mattiuzzi C and Lippi G 2020 Which lessons shall we learn from the 2019 novel coronavirus outbreak? \textit{Ann Transl Med.} Vol 8 no 3 p 48 doi: 10.21037/atm.2020.02.06

[5] Anderson R M, Heesterbeek H, Klinkenberg D, and Hollingsworth T D 2020 How will country-based mitigation measures influence the course of the COVID-19 epidemic? \textit{The Lancet} doi: 10.1016/S0140-6736(20)30567-5

[6] Hellewell J, Abbott S, Gimma A, Bosse N I, Jarvis C I, Russell T W, … van Zandvoort K 2020 Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts \textit{The Lancet Global Health} doi:10.1016/s2214-109x(20)30074-7

[7] Wang P, Zheng X, Li J, and Zhu B 2020 Prediction of epidemic trends in COVID-19 with logistic model and machine learning techniques \textit{Chaos, Solitons & Fractals} 110058 doi:10.1016/j.chaos.2020.110058

[8] Parhusip H A 2020 Menelusuri Covid-19 di dunia dan di Indonesia dengan model regresi SVM, Bayesian dan Gaussian \textit{Jurnal Ilmiah Sains} vol 20 no 2 pp 49-57 doi:10.35799/jis.20.2.2020.28256

[9] Kasilingam D, Prabhakaran S P S, Rajendran D K, Rajagopal V, Kumar T S, and Soundararaj A 2020 Exploring the growth of COVID-19 cases using exponential modelling across 42 countries and predicting signs of early containment using machine learning \textit{Transboundary and Emerging Diseases} doi:10.1111/tbed.13764

[10] Khakharia A, Shah V, Jain S, Shah J, Tiwari A, Daphal P, Warang M, and Mehendale N 2020 Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning \textit{Ann. Data. Sci.} doi: 10.1007/s40745-020-00314-9

[11] Shen C Y 2020 Logistic growth modelling of COVID-19 proliferation in China and its international implications \textit{International Journal of Infectious Diseases} vol 96 pp 582-589 doi: 10.1016/j.ijid.2020.04.085

Acknowledgments
The authors thank Indonesia Research and Scientist Institute (IRSI) who has given chance to publicize this research paper.