OPEN ACCESS
Indo-JC
Indonesia Journal of Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

Ind. Journal on Computing
Vol. 6, Issue. 3, December 2021. pp. 35-46
doi:10.34818/indojc.2021.6.3.594

Trials and Progress Prediction of Covid-19 Vaccine Using Linear Regression and SIR Parameters

Ananda Aulia Rizky [#1], Novi Rahmawati [#2], Adil El-Faruqi [#3], Faisal Dharma Adhinata [#4*],
Nur Ghaniaviyanto Ramadhan [#5]

[#] *Department of Software Engineering, Faculty of Informatics, Institut Teknologi Telkom Purwokerto, Indonesia*

[1] 19104053@ittelkom-pwt.ac.id
[2] 19104079@ittelkom-pwt.ac.id
[3] 19104081@ittelkom-pwt.ac.id
[4*] faisal@ittelkom-pwt.ac.id
[5] ghani@ittelkom-pwt.ac.id

Abstract

This research aims to explain the effectiveness of the Covid-19 vaccine worldwide to reduce the number of Covid-19 patients. Almost all countries in the world are trying to overcome Covid-19 by imposing a lockdown system. The government is also looking for a solution to suppress the spread of Covid-19 by administering a vaccine. Vaccination is one of the efforts that are considered effective in dealing with Covid-19 in affected countries. There are at least 85 types of vaccines in the development stage. The methodology in this study uses data mining with linear regression techniques and SIR Parameters. The Linear Regression and SIR Parameters carried out from February to May 2021 got an MSE value of 5.2 for France, 0.9 for Indonesia, and 6.9 for Italy. From this study, the most widely used vaccines were Pfizer/BioNTech, Oxford/AstraZeneca, Moderna, and finally Johnson & Johnson. This research also shows that vaccines in Indonesia are effective enough to suppress the spread of COVID-19 but still need to apply strict health protocols when doing activities in public.

Keywords: Covid-19, Vaccine, Linear Regression, SIR

Abstrak

Penelitian ini bertujuan untuk menjelaskan efektivitas vaksin Covid-19 di seluruh dunia untuk mengurangi jumlah pasien Covid-19. Hampir seluruh negara di dunia berusaha mengatasi Covid-19 dengan memberlakukan sistem lockdown. Pemerintah juga mencari solusi untuk menekan penyebaran Covid-19 dengan pemberian vaksin. Pemberian vaksin merupakan salah satu upaya yang dinilai efektif mengatasi Covid-19 di negara yang terdampak. Sedikitnya ada 85 jenis vaksin dalam tahap pengembangan. Metodologi pada penelitian ini menggunakan data mining dengan teknik regresi linier dan SIR Parameters. Hasil implementasi Linear Regression dan SIR Parameters yang dilakukan dari bulan Februari hingga Mei 2021 mendapatkan nilai MSE sebesar 5.2 untuk negara France, 0.9 untuk negara Indonesia dan 6.9 untuk negara Italy. Dari penelitian ini vaksin paling banyak digunakan yaitu Pfizer/BioNTech, Oxford/AstraZeneca, Moderna, dan terakhir Johnson&Johnson. Hasil penelitian ini juga menunjukkan vaksin di negara Indonesia

Received on October, 2021. Accepted on December, 2021

sudah cukup efektif menekan penyebaran COVID-19, namun masih perlu menerapkan protocol kesehatan yang ketat ketika beraktivitas ditempat umum.

Kata Kunci: Covid-19, Vaksin, Linear Regression, SIR

I. Introduction

Coronavirus is an RNA virus with a particle size of 120-160 nm. This virus initially infects animals, one of which is bats. Before the Covid-19 outbreak, six types of coronavirus could infect humans, namely alpha coronavirus, beta coronavirus, SARS-CoV, or Severe Acute Respiratory Illness Coronavirus, and the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) [1].

Currently, the world is still shaken by the virus infection, from day to day, the number of patients infected with the Covid-19 virus is continuously increasing, plus when the long holiday arrives, cases can reach two to three times so that it is difficult to control, this requires clear and precise planning to tackle this problem. WHO gave an appeal letter to the World Community to implement a lockdown system.

In some countries like Vietnam, New Zealand, South Korea, Cuba, Germany, France, etc., cases are almost declining. However, on the other hand, in countries like the United States, India, Brazil, Russia, etc. This disease spreads very quickly. In America, daily positive cases are identified more than 100,000 cases on average [2]. Almost all over the world imposed lockdowns as a general strategy to control the spread of disease to some extent. Various countries change their lockdown strategies from time to time for their socioeconomic situation, which affects the spread of Covid-19.

One of the most effective efforts to reduce this pandemic is the provision of vaccines. Vaccines are one of the most reliable and cost-effective public health interventions ever implemented that save millions of lives every year [3]. Following the decoding of the SARS-CoV-2 genome in early 2020 and the WHO's declaration of a pandemic in March 2020 [4], scientists and pharmaceutical companies are racing against time to develop a vaccine. As of 22 December 2020, at least 85 vaccines are in preclinical development in animals, and 63 in clinical development in humans, of which 43 are in phase I, 21 in phase II, 18 in phase III, six have been approved for initial or limited use, two have been approved. For full use, and one vaccine has been abandoned [5]. Pfizer-BioNTech and Moderna mRNA vaccines have been approved for emergency use in the US.

Based on the vaccines circulating in the world, a model is needed to analyze during Covid-19 whether the vaccines circulating in the world can function adequately to reduce Covid-19 cases in the world. So we need a method to perform data processing quickly. The technology used is data mining [6]. Data mining is believed to manage large amounts of data and use different techniques and algorithms. Data mining can also produce patterns that can later make it easier to read the existing data. The technique used in data processing is Linear Regression and SIR [7]. This research is essential to do because it can produce a decision whether the vaccine circulating in the world is efficient to continue to be used or not.

Covid-19 vaccine trials and vaccine progress with predictive results of vaccines and infected people is to determine whether the currently circulating vaccines show significant results in dealing with Covid-19 or no change at all. If there is a change for the better, then this vaccine can be continued at a later date. The purpose of using linear regression is to find out the number of patients who use the Covid-19 vaccine, while the use of the SIR method is to find outpatients who are susceptible, affected, and recovered from Covid-19.

II. Research Method

Prediction of Trials and Progress of Covid-19 vaccines begins with data collection. This study focuses on data on the spread of vaccines around the world. The number of Covid-19 19 infections ranges from hundreds, even thousands, so the data must be processed first before being processed further. The preprocessed data is done by creating a model using the Linear Regression and SIR Parameters methods. The best results will be used to predict the efficiency of the circulating vaccine. The flow chart of the Vaccine Trial Prediction system is shown in Figure 1.

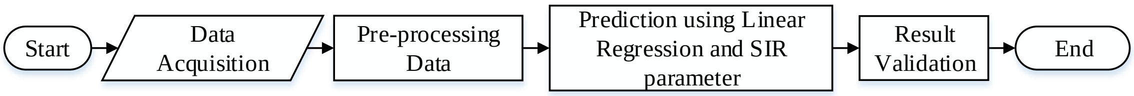


Figure 1. Vaccine trial prediction flowchart

A. Data Acquisition

This study uses the New Cases and Covid-19 Vaccination dataset [8]. from one of the portals on Research information and data to make progress against the world's biggest problems including covid 19 vaccine data around the world.

B. Pre-processing Data

This preprocessing is the first step that must be done when processing data so that the method can work properly. The pre-processing step, the data cleaning step, is carried out and used to remove unnecessary information from processing.

C. Application of Linear Regression Algorithm and SIR Parameters

1. Linear Regression

Linear regression algorithm is a statistical technique used to determine the effect of one or more variables on a variable. Variable is the value of the influencing variable called the independent variable, or explanatory variable. The variable that is affected is called the independent variable or the dependent variable [9].

Regression is a modeling technique used to predict the value of specific input data. Regression is used to determine the strength of the relationship between the dependent variable and the independent variable. The most important way to predict is to form a linear regression by looking for the relationship between one or more predictor variables (X) and response variables (Y) [10]. In its implementation, the linear regression method is divided into two, namely, simple linear regression and multiple linear regression. In this study, the method used is multiple linear regression because there is more than one dependent variable.

The goal is to adjust the Y data forecast line for X, where Y is the number of people who have been vaccinated and X is the time of initial use of the vaccine. To determine the line, this method is usually used to estimate the intercept and slope regression parameters. The multi-linear regression formula in its calculation can be explained in equation (1).

$$Y = a + bX \qquad (1)$$

where,

Y	: dependent variable
X	: independent variable
a	: constant
b	: regression coefficient

The magnitude of a constant a and b can use equations (2) and (3)

$$a = \frac{(\sum y)(\sum x^2)-(\sum x)(\sum xy)}{n(\sum x^2)-(\sum x)^2} \qquad (2)$$

$$b = \frac{n(\sum xy)-(\sum \mathrm{x})(\sum y)}{n(\sum x^2)-(\sum \mathrm{x})^2} \qquad (3)$$

2. Model SIR

The SIR (Susceptible - Infected - Recovered) model is a simple model that effectively predicts vaccines circulating in the world. In the SIR model, there are three categories: Infected, Recovered, or Vulnerable. The infected category is those who can transmit the disease. The recovered category is those who have recovered from infection [11]. In contrast, Vulnerable are those who have never been infected. The variables used are as follows:

S : Susceptible;
I : Infected;
R : Removed;
Beta : Infection rate per infected person;
Gamma : Recover rate of the infected;
k : vaccine effect factor;
The SIR model is calculated by successively changing S, I, R in equations (5), (6), (7), and (8).

$$N = S + I + R \qquad (3)$$

$$S(t+1) = S(t) - S(t)\,\beta\, I(t) - k \times N \qquad (4)$$

$$I(t+1) = I(t) + S(t)\,\beta\, I(t) - \gamma \times I(t) \qquad (5)$$

$$R(t+1) = R(t) + \gamma\, I(t) + kN \qquad (6)$$

Beta is the ability of infection in the case of an infected person to which the infected population per day recovers.

D. Result Validation

This research tested using the cross-validation method to validate specific datasets based on the formed model. Cross-validation is a method that can evaluate the predictions of a model [12]. This study evaluates the matrix using MSE. MSE (Mean Squared Error) itself is an estimated parameter to verify the accuracy of the resulting forecast [13].

$$MSE(y, \acute{y}) = \frac{1}{n} + \sum_{i=1}^{n} (y, \acute{y})^2 \qquad (7)$$

where,

y’ : predictive value

y : true value

n : number of observations

The MSE value is small so that the prediction results will be accurate.

III. Result and Discussion

A. Dataset of Vaccines

In this study, the data used came from the Kaggle Covid-19 Global Dataset. The data consists of 195 countries with 17607 data. In connection with this study, three country datasets were selected to be used in the analysis to see the effectiveness of the Covid-19 vaccine to reduce the number of Covid-19 patients. This country has a large population and may be used as a benchmark for other countries that will use the Covid-19 vaccine to see how it affects the three countries, and these three countries are France, Indonesia, and Italy.

Table 1. Dataset

No	Country	Code	Date	T_vacc	P_fu_vacc	T_vacc_pHun	P_vacc_pHun	Vaccines
1	Afghanistan	AFG	2/22/2021	NaN	NaN	NaN	NaN	AstraZeneca, Pfizer, Sinopharm
2	Afghanistan	AFG	2/23/2021	NaN	NaN	NaN	NaN	AstraZeneca, Pfizer, Sinopharm
…	…	…	…	…	…	…	…	…
17607	Zimbabwe	ZWE	5/11/2021	709772	170246	4.78	3.63	Sinopharm
17607	Zimbabwe	ZWE	5/12/2021	730365	180568	4.91	3.7	Sinopharm

B. Data Preprocessing

Some missing or damaged data from the data obtained, so data preprocessing needs to be done so that the data is easier to use. This preprocessing also aims to transform data into a more practical format to get more accurate values and reduce computation time in large-scale problems. After this preprocessing, 6343 were present, and 11264 were missing. The table of preprocessing results can be seen in Table 2.

Table 2. Data Preprocessing Results

No	Country	Code	Date	T_vacc	P_fu_vacc	T_vacc_pHun	P_vacc_pHun	Vaccines
1	Albania	ALB	2/18/2021	3049	611	0.11	0.08	AstraZeneca, Pfizer, Sinovac, Sputnik V
2	Albania	ALB	5/12/2021	632676	187921	21.98	15.45	AstraZeneca, Pfizer, Sinovac, Sputnik V
…	…	…	…	…	…	…	…	…
6342	Zimbabwe	ZWE	5/11/2021	709772	170246	4.78	3.63	Sinopharm
6343	Zimbabwe	ZWE	5/12/2021	730365	180568	4.91	3.7	Sinopharm

A. Implementation of Linear Regression and SIR Parameter

In this case, each country must have different types of vaccines ranging from AstraZeneca or Oxford, Moderna, Pfizer, Sinovac, and others, but it does not rule out that there may be similarities in using the existing vaccine types. For this study, the three countries that will be used also have different types of vaccines, which can be seen in Table 3.

Table 3. Types of Vaccines in 3 countries

Country	**Vaccine**
France	Johnson&Johnson, Moderna, Oxford/AstraZeneca, Pfizer/BioNTech
Indonesia	Oxford/AstraZeneca, Sinovac
Italy	Johnson&Johnson, Moderna, Oxford/AstraZeneca,Pfizer/BioNTech

With this difference, it can be seen how effective the species see the picture, as shown in Figure 2. The picture shows the results of a graph from France as an example, this graph starts on January 3-9 May 2021 and can then be seen that the most vaccines used in this country are Pfizer/BioNTech and the graph is steadily increasing, the Oxford/AstraZeneca type is second, followed by Moderna and finally Johnson&Johnson.

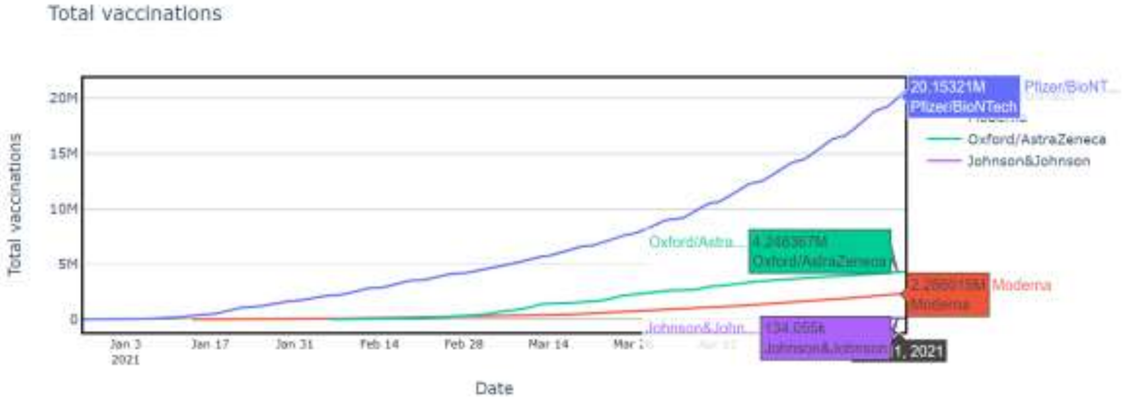


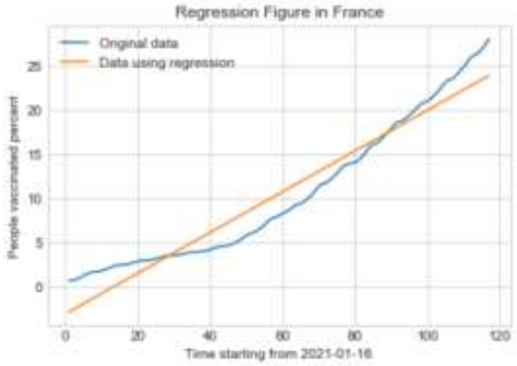
Figure 2. Graph of vaccination use in France

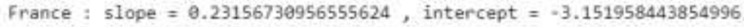
In this research, two methods will be used, namely linear regression and SIR. The parameters of these two methods are not related because this linear is intended to predict the number of vaccinated people, and SIR is to predict people infected with the vaccine.

1. Linear Regression Model

This method is intended to make predictions or estimates. In this study, the results will be analyzed and then estimate whether the use of the Covid-19 vaccine will be effective or not. Linear regression has a relationship between the dependent variable (y) and the independent variable (x). The dependent variable is the effect variable. This variable is taken from the data of people vaccinated per hundred while the dependent is the cause variable, for this variable comes from the date data. For this prediction, it is only to predict the number of people who have been vaccinated by date. The output results have no accuracy. This research only visualizes, and the date data cannot be used as a variable because what works well the type of data used also does not solve the problem correctly.

a) Vaccine use in the France





```
France : slope = 0.23156730956555624 , intercept = -3.151958443854996
```

Figure 3. Vaccine use in the France

In Figure 3, it can be seen that the regression results in France with the variable people vaccinated per hundred and date, this date starts from January 16, 2021, after being drawn by the regression graph, the graph shows stability and continues to increase even though there is a slight slope. This graph shows a slope of 0.23 and an intercept of -3.15. The prediction results that will be tested are on the 25th, and the value of a is taken from the intercept and b using the slope.

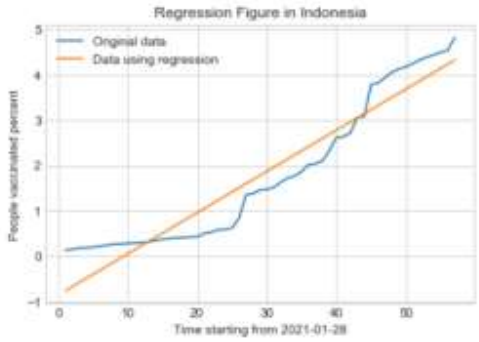
$$Y = a + bX$$

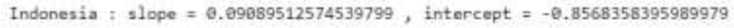
$$Y = -3{,}15 + 0{,}23(25)$$

$$= 2{,}6$$

So if people are vaccinated per hundred on the 25th, it is predicted that there will be 3.75 people vaccinated.

b) Vaccine use in the Indonesia





```
Indonesia : slope = 0.09089512574539799 , intercept = -0.8568358395989979
```

Figure 4. Vaccine use in the Indonesia

In Figure 4, it can be seen that the vaccine used in Indonesia starts from January 28, 2021. The graph shows an increase and a decrease, but the graph continues to increase steadily. The regression results in Indonesia with a value of x or the dependent variable using data on people vaccinated per hundred showed a slope of 0.09 and an intercept of -0.85. The results of the prediction that will be tested are on the 30th. It is described as follows:

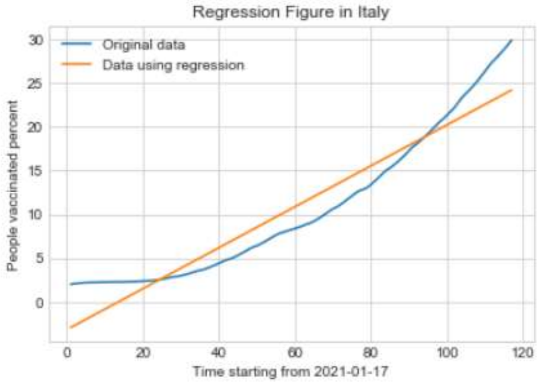
$Y = a + bX$

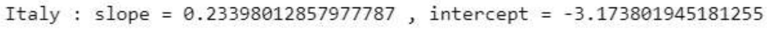
$Y = -0.85 + 0.09(30)$

$= 1.85$

So if people are vaccinated per hundred in Indonesia on the 30th, it is predicted that there will be 1.85 people vaccinated to be produced.

c) Vaccine use in the Italy





```
Italy : slope = 0.23398012857977787 , intercept = -3.173801945181255
```

Figure 5. Vaccine use in Italy

In Figure 5, it can be seen that the results of the regression in Italy with the dependent variable using data on people vaccinated per hundred and independent data date, this date starts from January 17, 2021, after the line drawn by the regression graph shows the curve curves outward and this graph shows stability and continues increase. This graph produces a slope of 0.23 and an intercept of -3.17. The prediction results that will be tested is on the 20th, described as follows:

$Y = a + bX$

$Y = -3.17 + 0.23(20)$

$= 1.43$

So if in Italy the data of people vaccinated per hundred is on the 20th, it is predicted that there will be 1.43 data generated.

2. SIR Parameters Model

Based on the available data and the parameter estimation model, a comparison of the number of recovered and susceptible infections is obtained. There are three countries, namely: France, Indonesia, and Italy using the SIR Model (Susceptible or susceptible, Infected or recovered and recovered) to predict future tendencies jumlah orang yang rentan yang terinfeksi Oleh Covid-19.

a) Predicted comparison in the France

The graphic comparison between the forecast data and the data in France is based on the actual and model data obtained from the parameters from February to May 2021. In Figure (1). There is a decrease in susceptible patients from March to May 2021 and Figure (2). There was a spike in Covid-19 infected cases, with a peak of

850,000 patients from March to May 2021, then Figure (3). Predictions of cases recovering from Covid-19 are similar from February to the end of March 2021 and an increase in May 2021 by 100,000 patients.

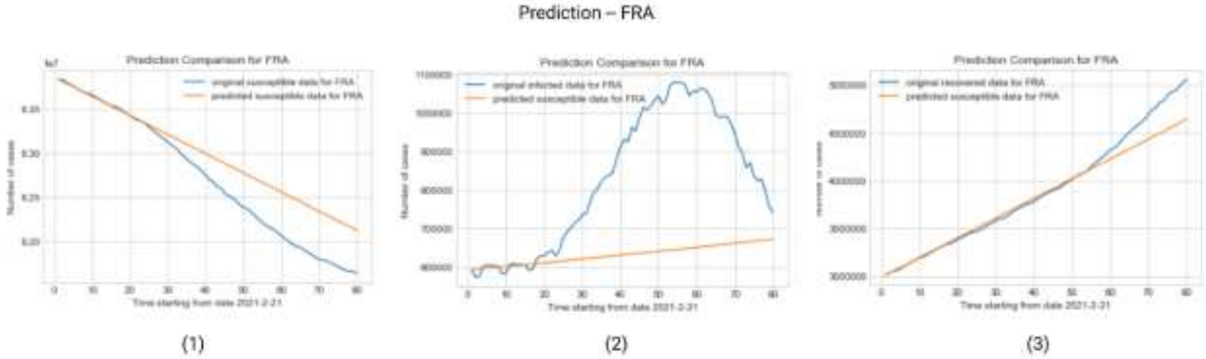


Figure 6. Graph of SIR Parameters in FRA

Based on Figure 6 shows that each parameter has a different graph, such as Figure (1) or the data of people who are susceptible to experiencing a decline in this country of France, in the following Figure, namely (2) or the data of people infected with the graph has a peak point between 40 and 50 data, the last Figure (3) or people who recovered experienced a relatively good and stable improvement.

b) Predicted comparison in IDN

For Indonesia, the comparison of the number of existing cases with the predicted data has some differences, and this data is based on actual data and models obtained from parameters starting February 2021. Figure (1) shows the number of people susceptible to Covid-19 is almost the same with predictive data, but there is an increase of 1000 from March to May 2021 and Figure (2). The number of infected people gets a varying number where there is an increase and decrease and Figure (3). This shows that the number of people recovered and decreased by 50,000 from March to May 2021.

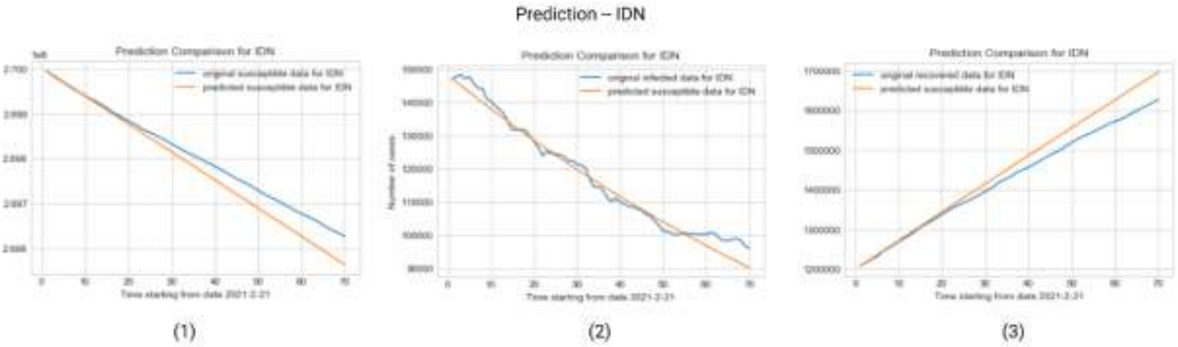


Figure 7. SIR Parameters Graph at IDN

Based on Figure 7, the graph in Indonesia shows that each parameter has different results, such as Figure (1) or the data of vulnerable people has a decreasing graph, in the following Figure (2) or the data of infected people the graph produces a decrease even though at the end it is slightly experienced the increase, the last Figure (3) or people who recovered experienced a relatively good and stable increase, meaning that Indonesia was able to overcome this Covid-19 well.

c) Prediction comparison in ITA

In Italy, the number of existing cases with predicted data is different, starting in early March 2021, as shown in Figure (1). The number of original data with predictive data on vulnerable people is almost the same, and there is an increase in May 2021 by 50,000 and Figure (2). The predicted number of infected people has decreased and peaked at 800,000 patients starting in May 2021. Finally, Figure (3). The number of recovered patients has increased by 40,000 patients from March to May 2021.

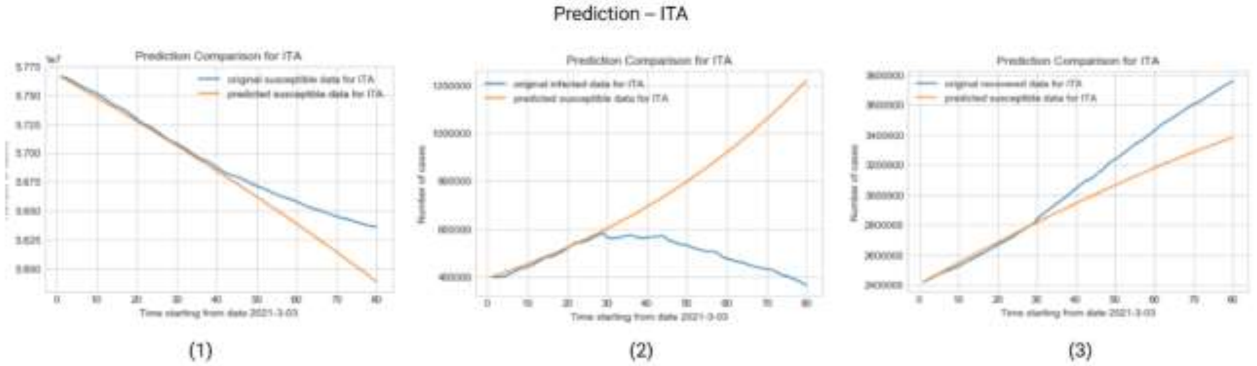


Figure 8. SIR Parameters Chart in ITA

Based on Figure 8, it shows that each parameter has different results, such as Figure (1) or the data of vulnerable people the graph produces a decrease, then the following Figure, namely (2) or the data of infected people, the graph produces an increase and then decreases, the last image is the Figure (3) or people who recovered experienced a relatively good and stable increase, meaning that Italy was able to overcome this Covid-19 even though there was an increase in infected people.

B. Result of Validation

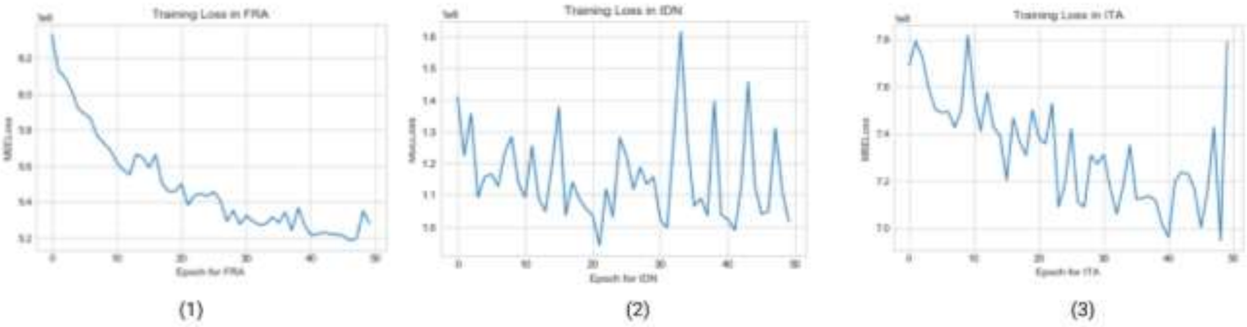


Figure 9. MSE Graph of Training Layer and Epoch 50 Results in 3 countries

In Figure 9, it can be seen that the results of the France trial, which had been carried out using seven layers and Epoch 50, produced the smallest MSE of 5.184119191524046 ≈ 5.2, for the results of the Indonesian state trial it produced the smallest MSE of 0.9397460777116299 ≈ 0.9 and for the country trial results Italy produced the smallest MSE which was 6.948037597422851 ≈ 6.9.

Table 4. The Training Results

Country	Epoch	MSELoss
France	50	5.2
Indonesia	50	0.9
Italy	50	6.9

The results in this study showed that the MSELoss was still high. The MSE formula uses the square of the difference between the predicted results and the actual results. In further research, the pre-processing normalization method can be added to make the data range smaller. Alternatively, we can also use other regression methods.

IV. CONCLUSION

This research applies Linear Regression and SIR Parameters, both of which have different goals, where the use of linear regression is to determine the number of patients using the Covid-19 vaccine, while the use of the SIR method is to determine susceptible patients, affected, and recovered from Covid-19. This research was conducted several times using 6343 data. Optimal results obtained that the MSE value achieved was 5.2 for France, 0.9 for Indonesia, and 6.9 for Italy, and then from this study, it was found that the most widely used vaccines were Pfizer/BioNTech in the first place, Oxford/AstraZeneca in second, followed by types Moderna and finally Johnson & Johnson. This research also shows that vaccines in Indonesia are effective enough to suppress the spread of COVID-19 but still need to apply strict health protocols when doing activities in public.

REFERENCES

[1] A. Susilo *et al.*, “Coronavirus Disease 2019: Tinjauan Literatur Terkini,” *J. Penyakit Dalam Indones.*, vol. 7, no. 1, p. 45, 2020, doi: 10.7454/jpdi.v7i1.415.

[2] A. Senapati, S. Maji, and A. Mondal, “Piece-wise linear regression: A new approach to predict Covid-19 spreading,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1020, no. 1, 2021, doi: 10.1088/1757-899X/1020/1/012017.

[3] I. Hajj Hussein *et al.*, “Vaccines Through Centuries: Major Cornerstones of Global Health,” *Front. Public Heal.*, vol. 3, no. January 2016, 2015, doi: 10.3389/fpubh.2015.00269.

[4] D. Cucinotta and M. Vanelli, “WHO declares Covid-19 a pandemic,” *Acta Biomed.*, vol. 91, no. 1, pp. 157–160, 2020, doi: 10.23750/abm.v91i1.9397.

[5] C. Zimmer, J. Corum, and S. L. Wee, “Coronavirus vaccine tracker. The New York Times,” 2020.

[6] S. Budi, “Data Mining:‘Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis.’” Teori dan Aplikasi. Graha Ilmu Yogyakarta. Buku, 2007.

[7] R. Aggarwal and P. Ranganathan, “Common pitfalls in statistical analysis: Linear regression analysis,” *Perspect. Clin. Res.*, vol. 8, no. 2, pp. 100–102, 2017, doi: 10.4103/2229-3485.203040.

[8] Worldmeter.com, “COVID Live Update: 186,356,010 Cases and 4,026,894 Deaths from the Coronavirus.” 2021, [Online]. Available: https://www.worldometers.info/coronavirus/.

[9] T. H. Iman Mustofa Kamal and R. Ilyas, “Prediksi Penjualan Buku Menggunakan Data Mining Di Pt. Niaga Swadaya,” *Seminar Nasional Teknologi Informasi & Multimedia*, vol. 02, no. February. pp. 49–54, 2017.

[10] H. W. Herwanto, T. Widiyaningtyas, and P. Indriana, “Penerapan Algoritme Linear Regression untuk Prediksi Hasil Panen Tanaman Padi,” *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 8, no. 4, p. 364, 2019, doi: 10.22146/jnteti.v8i4.537.

[11] I. A. M. Rosyid, R. Respatiwulan, and S. S. Handajani, “Model Penyebaran Penyakit SIR Tipe Rantai Binomial dengan Kontak Random dan Waktu Penyembuhan Bernilai Tak Hingga,” *Indones. J. Appl.*

Stat., vol. 3, no. 2, p. 132, 2021, doi: 10.13057/ijas.v3i2.44307.

[12] S. Hulu, “Analisis Kinerja Metode Cross Validation Dan K-Nearest Neighbor Dalam Klasifikasi Data,” *Fak. Ilmu Komput. dan Teknol. Inf. Univ. Sumatera Utara*, vol. 1, no. 1, pp. 1–77, 2020.

[13] A. A. Gofur and U. D. Widianti, “Sistem Peramalan Untuk Pengadaan Material Unit Injection Di Pt. Xyz,” *Komputa J. Ilm. Komput. dan Inform.*, vol. 2, no. 2, 2015, doi: 10.34010/komputa.v2i2.86.