# ANALYSIS OF COVID-19 GROWTH TRENDS THROUGH DATA MINING APPROACH AS DECISION SUPPORT

**Mohamad Ilyas Abas[1)*], Irawan Ibrahim[2)], Syahrial[3)], Rizal Lamusu[4)], Umar Sako Baderan[5)], Riklan Kango[6)]**
[1,2,3,4,5)]Universitas Muhammadiyah Gorontalo, [6)]Politeknik Balikpapan, Indonesia
[1)]ilyasabas@umgo.ac.id, [2)]irawan_ibrahim@umgo.ac.id, [3)]syahrial@umgo.ac.id, [4)]riza_lamusu@umgo.ac,id, [5)]usbaderan@umgo.ac.id, [6)]riklan.kango@poltekba.ac.id

**Submitted** : Oct 21, 2022 | **Accepted** : Nov 5, 2022 | **Published** : Jan 1, 2023

**Abstract:** This study aims to analyze the growth trend of covid-19 using prediction algorithms in data mining for covid-19 data throughout Indonesia. This can be used as a decision support to analyze several government policies towards regulatory intervention so far. The method used is the best prediction method in time series data, including Neural Network, SVM, Linear Regression, K-Neirest Neighborn and optimizes it with optimization algorithms. This research is focused on the application of these applications. It is hoped that this research will produce an analysis of the growth trend of Covid cases every day, in addition to its contribution so that it can assist the government in determining the best policy direction and also as an education to the public. in addition, the research will contribute to science in the field of predictive analysis by finding the best RMSE formulation. The results of this study show that Neural Network-Particle Swarm Optimization has the smallest Roort Mean Square Error which is 265,326, and the two Neural Network Genetic Algorithm are 266.801, Neural Network Forward Selection is 275,372 and Neural Network without optimization has the largest RMSE which is 297.204. These results can be used as a reference for the use of similar algorithms in time series data, both Covid-19 data and other data.

**Keywords:** Covid-19; Data mining algorithm; forecasting; Neural Network; Indonesia;

## INTRODUCTION

The spread of the corona in Indonesia is quite surprising, starting with only 2 cases, now there are 2,491 as of October 30, 2021. Of course this is not good news because in the reflection of Wuhan, they have implemented a strict lockdown so that currently they have succeeded in anticipating the corona virus, which is named covid 19. Corona or covid This 19 was first spread in one of the provinces in China, namely Hubei Province, precisely in the city of Wuhan. Initially, it was suddenly reported that this virus occurred through the habits of the people of Wuhan who consumed one of the animals, namely bats. Viral contact that occurs through bats to humans. But the name of the virus is indeed evolved to be between humans and humans and can even survive in the air. At that time, the recommendation to wear gloves and masks was mandatory for all people in Wuhan. According to data monitored through the Indonesian Ministry of Health, 33 provinces have been infected with this dangerous virus. The most cases of contact or infection are dominant in Jakarta. Because Jakarta is the capital city of Indonesia, it is natural for large numbers of contacts to occur (Pitoyo, Edy Prihantoro, 2021).

Departing from this, the researcher made a study to analyze the growth trend of COVID-19 using several prediction algorithms in data mining. With this research, in addition to providing scientific contributions in terms of algorithms, it also provides a new picture of the combination of algorithms analyzed. The data obtained at this time amounted to 36 data records. The data was obtained through the long kawalcovid19.id which does provide statistical data voluntarily for every day starting March 2, 2020 until the time the research will begin. Until now, Covid-19 has not been controlled and we have entered the "New Normal" era in the midst of the Covid-19 growth conditions which are not declining but are gradually increasing. Just imagine that in just 1 month there have been 2,491 cases, if you multiply it by 2 months, there are approximately 5,000 cases.

*name of corresponding author

Some researchers have predicted that if this pandemic is handled in an orderly and serious manner, it will end around the end of May or early June 2020. Some researchers in Indonesia also say that it will end at the end of April or early May. Of course this must be accompanied by being obedient to the government by staying at home(Winata et al., 2021). If not, of course, all people in Indonesia do not want this pandemic to continue for a long time and even get worse. Therefore, this research on the prediction of the spread of positive cases in Indonesia will provide a little overview of the analysis of the Covid-19 growth trend that will be able to provide decision support to the government and the public. These predictions can also provide information to the government about the trend of growth charts and contacts that result in the number of positive patients increasing. The methods used are NNPSO, NNGA, NNFS and NN in order to contribute to knowledge in the use of the best prediction algorithm in forecasting, these methods are very good at making predictions which often get error values or often called the smallest RMSE in making predictions.

Research on the prediction of the number of passengers at Djalaludin airport in Gorontalo. The pattern of each passenger will be the result of an analysis of management improvements at the airport and see the arrival pattern of each passenger. This research uses neural network algorithm and genetic algorithm. The results of this study show that the smallest RMSE value is 0.092 from the normalized data(Abas et al., 2017).

Research on previous predictions has been mostly done on other datasets such as(Abas & Lasarudin, 2019; Sheikh et al., 2016)about prediction and how the prediction algorithm works and the results show the smallest RMSE of the experimental results in the data used.

Research on the covid case uses a neural network backpropation and fuzzy Tsukamoto algorithm. The purpose of this research is to find a model in predicting the addition of Covid-19 cases in Indonesia according to time series data every day. The results of this study indicate that the resulting model has a correlation coefficient value of 0.84278 and the smallest MSE is 1.632337 in normalized data.(Arianto, 2020).

Other research on Covid-19, namely:(Hikmawan et al., 2020)Sentiment Analysis on Corona Virus Pandemic Using Machine Learning Algorithm. This study collects public sentiment through twitter data related to tweets about the Corona Virus. The algorithm used is Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) to build a tweet classification model on sentiment whether it has positive, negative or neutral popularity. The result is that the SVM algorithm has a high accuracy value of 76.21% with a precision value of 78.04% and a recall value of 71.42%.

The same research on Covid19 on how to determine the level of student understanding of social distancing using the C4.5 algorithm. The data that were collected were 287 students aged 18-25 years, the classification process using the C4.5 algorithm or known as the decision tree resulted in 93.73% accuracy with class precision who understood social distancing 96.97%, students who understood but kept active 100% and those in doubt 75.71% (Sudipa et al., 2020).

Research on Implementation of Data Mining Algorithm for Predicting Popularity of Playstore Games in the Pandemic Period of Covid-19. In his research, it describes activities in filling time at home by doing various activities, especially playing games, and the category of games that are always played. This study was conducted to predict the level of popularity of games on the Playstore application in order to find out how many games are often played. The method used is nave Bayes and C4.5. The results obtained, namely the C4.5 algorithm, show 73 games are classified as popular and 12 games are not popular with an accuracy value of 85.83%. Meanwhile, Naive Bayes shows 23 popular games and 62 unpopular games with 80% accuracy. The results of the ROC Curve evaluation of the nave Bayes algorithm have an AUC value of 0,(Sulistyowati et al., 2020).

Research on the Implementation of the Naive Bayes Algorithm to Predict the Spread of Covid-19 in Indonesia. This research was conducted as an anticipatory measure against the Covid-19 pandemic by predicting the level of its spread, especially in Indonesia. The method used in this research is problem analysis and literature study. The algorithm used is nave Bayes to predict the rate of spread of Covid-19. The result is that this algorithm can classify 48.4848% in the sense that it shows that 16 data have been successfully classified out of 33 tested data(Watratan et al., 2020).

Public Analysis Sentiment Against Joko Widodo on the Covid-19 outbreak using the Machine Learning Method. Sentiment analysis is indeed done by utilizing one of the sciences in data mining, namely text mining. The analysis process carried out in this study is with the keywords "Jokowi" and "Covid" by utilizing one of the social media namely twitter. The three classification methods used are nave Bayes, support vector machine (SVM) and k-NN. From the comparison of the three methods, it shows that SVM is the best algorithm in the classification with an accuracy of 84.58%.(Hikmawan et al., 2020).

Based on some of these studies, researchers also make predictions on time series data as has been done by several previous researchers. The algorithms used are time series prediction algorithms such as neural networks, KNN, LR and so on and use optimization algorithms for both preprocessing such as Forward Selection and other parameter optimization algorithms, namely Genetic Algorithm and Particle Swarm Optimization so that research will contribute to knowledge of the comparison of several data mining algorithms. especially for time series

*name of corresponding author

prediction. The urgency for government policy, which will have implications for the successful prediction of the positive number of Covid through the analysis carried out, will at least make it easier for the government to choose strategic steps in making decisions after. Decision making based on data is certainly better than without data. Therefore, the importance of this research is to provide an overview in addition to algorithm analysis as well as the important role of time series data analysis, especially Covid in the context of decision making.

## LITERATURE REVIEW

### Data Mining

(R, Turban, 2005) Big data that is owned and stored for years by an educational institution, is a potential source of knowledge and information for the progress of the institution if it is able to explore and explore it. Data mining is a concept used to find knowledge hidden in databases. Data mining is a semi-automatic process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify potential and useful knowledge information stored in large databases. Data mining is part of the KDD (Knowledge Discovery in Databases) process which consists of several stages such as data selection, pre-processing, transformation, data mining, and evaluation of results.(Maimon, 2000). In another sense, data mining is also defined as the process of obtaining useful information from a large database warehouse. Data mining can also be interpreted as the process of extracting new information taken from large chunks of data that helps in decision making(Tan, 2006). Data mining is a series of processes to explore added value from a data set

### Time series data

Time Series data is a type of data that consists of an object but covers several time periods such as daily, weekly, monthly, yearly and others.(Aggarwal, 2015). Time Series data can be used to predict future events. Because learning patterns that existed in the past will be repeated in the future. There are two types of analysis used for forecasting, namely qualitative analysis and quantitative analysis. Qualitative analysis is a forecasting technique that is carried out based on the opinion of a party so that the data cannot be used as an express representation of a value. While quantitative analysis is forecasting based on data in the past and can be made in the form of numbers which are usually time series data(Wei & Hamilton, 1994)

### *K-Fold Cross Validation*

(North, n.d.) *K-fold cross validation*is a technique to assess/validate the accuracy of a model that is built based on a particular dataset. Modeling usually aims to predict and classify new data that may never appear in the dataset. The data used in the model development process is called training data, while the data used to validate the model is called test data. One of the popular cross-validation methods is K-Fold Cross Validation. In this technique the dataset is divided into a number of K-partitions randomly. Then a number of K-experiments were carried out, where each experiment used the K-th partition data as testing data and used the remaining partitions as training data.

## METHOD

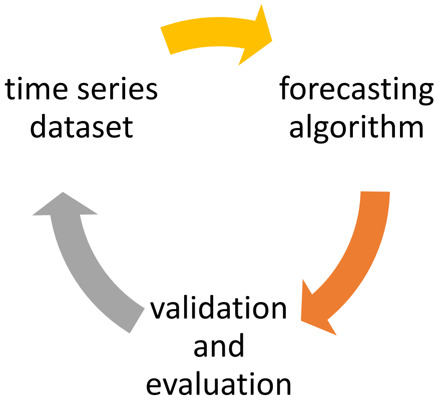

Figure 1. Research method (Abas & Lasarudin, 2019)

### Data collection

Collecting covid-19 data through the kawalcovid19.id website and processing data in a time series. The covid-19 dataset is collected from the beginning of covid-19 until the preparation of the final (latest) report October 30, 2021. The data is then pre-processed data as shown in the following table:

*name of corresponding author

Table 1. Pre-processing data

| xt-2 | xt-1 | xt |
|---|---|---|
| **4029** | 3565 | 2897 |
| **3520** | 4029 | 3565 |
| **3222** | 3520 | 4029 |
| **3732** | 3222 | 3520 |
| **4070** | 3732 | 3222 |
| **4369** | 4070 | 3732 |
| **4432** | 4369 | 4070 |
| **4267** | 4432 | 4369 |
| **3602** | 4267 | 4432 |
| **3373** | 3602 | 4267 |

**Prediction and optimization algorithm**

Conducting analysis using prediction algorithms and their optimization so as to contribute to knowledge in the framework of the best algorithm in predicting the growth trend of COVID-19.

The algorithms tested were NN, NNGA, NNPSO and NNFS which were then validated and evaluated the performance of these algorithms.

**Validation & Evaluation**

Validation and evaluation of the model uses several alternatives such as cross validation, to find the best accuracy or RMSE value. Evaluating the model by looking at the accuracy value or RMSE value found from each combination of prediction algorithms.

## RESULT

The expected results in this study are to contribute knowledge on how to process time series datasets, use prediction algorithms and optimize several prediction algorithms. The results of this study will show the performance of each algorithm and the use of optimization algorithms. This will be seen through the evaluation model, namely accuracy and RMSE.

Table 1. Case data

| data | daily case data |
|---|---|
| 2-Mar | 2 |
| 3-Mar | 0 |
| 4-Mar | 0 |
| 5-Mar | 0 |
| 6-Mar | 2 |
| 7-Mar | 0 |
| 8-Mar | 2 |
| 9-Mar | 13 |
| 10-Mar | 8 |
| 12-Mar | 0 |
| 13-Mar | 35 |
| 14-Mar | 27 |
| 15-Mar | 21 |
| 16-Mar | 17 |
| 17-Mar | 38 |
| 18-Mar | 55 |
| 19-Mar | 82 |
| 20-Mar | 60 |
| 21-Mar | 81 |

| | |
|---|---|
| 22-Mar | 64 |
| 23-Mar | 65 |
| 24-Mar | 106 |
| 25-Mar | 105 |
| …. | …. |
| 30 oct | 2897 |

Table 2. preprocessing data

| xt-2 | xt-1 | xt | actual | prediction |
|---|---|---|---|---|
| 4029 | 3565 | 2897 | 2897 | 3878 |
| 3520 | 4029 | 3565 | 3565 | 3835 |
| 3222 | 3520 | 4029 | 4029 | 3071 |
| 3732 | 3222 | 3520 | 3520 | 3193 |
| 4070 | 3732 | 3222 | 3222 | 2169 |
| 4369 | 4070 | 3732 | 3732 | 1699 |
| 4432 | 4369 | 4070 | 4070 | 2045 |
| 4267 | 4432 | 4369 | 4369 | 1877 |
| 3602 | 4267 | 4432 | 4432 | 1545 |
| 3373 | 3602 | 4267 | 4267 | 1112 |
| 4105 | 3373 | 3602 | 3602 | 1228 |
| 4301 | 4105 | 3373 | 3373 | 776 |
| 4301 | 4301 | 4105 | 4105 | 533 |
| 4411 | 4301 | 4301 | 4301 | 589 |
| 4127 | 4411 | 4301 | 4301 | 507 |
| 3906 | 4127 | 4411 | 4411 | 430 |
| 3267 | 3906 | 4127 | 4127 | 313 |
| 4497 | 3267 | 3906 | 3906 | 296 |
| 4294 | 4497 | 3267 | 3267 | 152 |
| 4094 | 4294 | 4497 | 4497 | 123 |
| 4850 | 4094 | 4294 | 4294 | 125 |
| 4538 | 4850 | 4094 | 4094 | 92 |
| 4056 | 4538 | 4850 | 4850 | 56 |
| …. | …. | ….. | | |
| 0 | 0 | 2 | | |

Table 3. Neural network analysis data

| actual | prediction | Error difference |
|---|---|---|
| 2897 | 3878 | -981 |
| 3565 | 3835 | -270 |
| 4029 | 3071 | 958 |
| 3520 | 3193 | 327 |
| 3222 | 2169 | 1053 |
| 3732 | 1699 | 2033 |
| 4070 | 2045 | 2025 |
| 4369 | 1877 | 2492 |
| 4432 | 1545 | 2887 |

*name of corresponding author

| | | |
|---|---|---|
| 4267 | 1112 | 3155 |
| 3602 | 1228 | 2374 |
| 3373 | 776 | 2597 |
| 4105 | 533 | 3572 |
| 4301 | 589 | 3712 |
| 4301 | 507 | 3794 |
| 4411 | 430 | 3981 |
| 4127 | 313 | 3814 |
| 3906 | 296 | 3610 |
| 3267 | 152 | 3115 |
| 4497 | 123 | 4374 |
| … | … | … |
| 0 | 207 | -207 |

## DISCUSSIONS

From the results of the algorithm analysis can be described as follows:

***Neural Network***

root_mean_squared_error: 297,204 +/- 53,832

***Neural Network + Genetic Algorithm***

Parameters set:
performance:
PerformanceVector [
-----root_mean_squared_error: 266,801 +/- 38,255 (micro: 269,529 +/- 0.000)
-----squared_error: 72646,149 +/- 20941,536 (micro: 72646.149 +/- 146592.651)
Neural Net.training_cycles= 94
Neural Net.learning_rate= 0.03944128770643769
Neural Net.momentum= 0.7076297693548458

***Neural Network + Particle Swarm Optimization***

root_mean_squared_error: 265,326 +/- 52,958 (micro: 270,559 +/- 0.000)

***Neural Network + Forward Selection***

root_mean_squared_error: 275,372 +/- 41,887 (micro: 278,539 +/- 0.000)

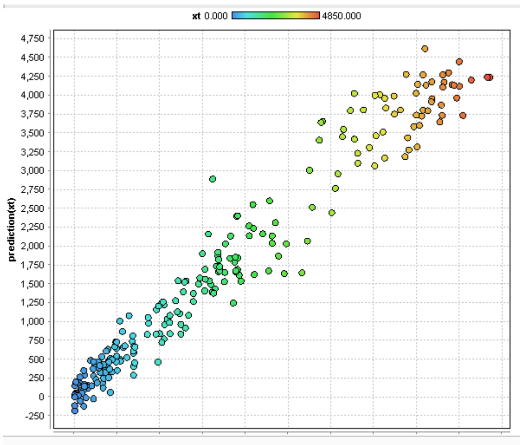

Figure 2. Scatter data processing results

*name of corresponding author

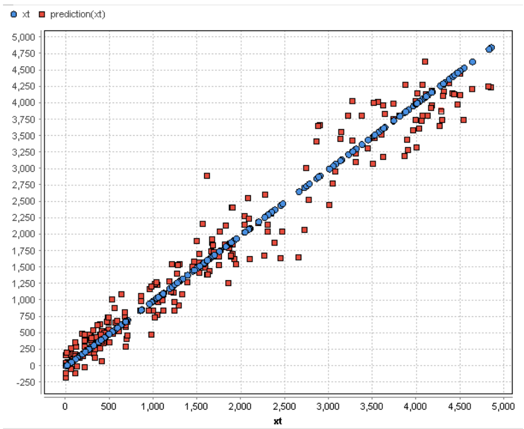

Figure 3. Scatter multiple xt and prediction (xt)

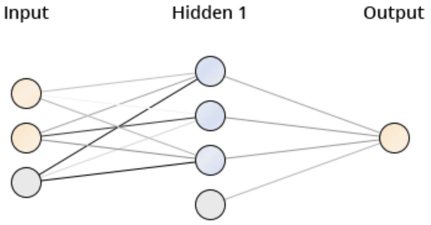

Figure 4. Neural Network Architecture

## CONCLUSION

The results of this study indicate that NNPSO has the smallest RMSE which is 265,326, and the two NNGAs are 266,801, NNFS is 275,372 and NN without optimization has the largest RMSE which is 297,204. From these results will then be analyzed so as to find predictions with high accuracy values. The results of this study have implications for the government in the next policy direction. The well-predicted Covid-19 certainly makes it easier for the government in terms of implementing regulations such as PSBB, working hours regulations and tightening recommendations for the use of masks and so on. The government can also measure how long the positive spike and the impact of the pandemic on the community.

## ACKNOWLEDGMENT

- Pimpinan pusat Muhammadiyah
- Lazismu Pusat (Risetmu)

## REFERENCES

Abas, M. I., & Lasarudin, A. (2019). *Prediction of arrival domestic and foreign tourists based on regions using neural network algorithm based on genetic algorithm Prediction of arrival domestic and foreign tourists based on regions using neural network algorithm based on genetic algorithm*. https://doi.org/10.1088/1742-6596/1175/1/012045

Abas, M. I., Syukur, A., & Soeleman, M. A. (2017). Prediksi Rentet Waktu Jumlah Penumpang Bandara Menggunakan Algoritma Neural Network Berbasis Genetic Algorithm. *Jurnal Teknologi Informasi*, *13*, 101–114.

Aggarwal, C. C. (2015). Data Mining: The Textbook. In *Springer International Publishing*. https://doi.org/10.1007/978-3-319-14142-8

Arianto, F. S. D. (2020). Prediksi Kasus COVID-19 di Indonesia Menggunakan Metode Backpropagation dan Fuzzy Tsukamoto. *Jurnal Teknologi Informasi*, *4*(1), 120–127. https://doi.org/10.13140/RG.2.2.34286.02885

Hikmawan, S., Pardamean, A., & Khasanah, S. N. (2020). Sentimen Analisis Publik Terhadap Joko Widodo terhadap wabah Covid-19 menggunakan Metode Machine Learning. *Jurnal Kajian Ilmiah*, *20*(2), 167–176. https://doi.org/10.31599/jki.v20i2.117

Maimon, O. and L. M. (2000). *Knowledge Discovery and Data Mining*. Kluwer Acamdemic.

North, M. (n.d.). *Data Mining for the Masses*.

Pitoyo, Edy Prihantoro, N. R. O. (2021). *Makna Zona Merah Covid 19 Di Dki Jakarta ( Studi Semiotika Charles Sander Peirce Berita Kompas . Com ) Meaning of the Red Zone Covid 19 in Dki Jakarta*. *15*(1), 85. https://scholar.google.com/scholar?hl=id&as_sdt=0%2C5&q=MAKNA+ZONA+MERAH+COVID+19+DI+DKI+JAKARTA+%28STUDI+SEMIOTIKA+CHARLES+SANDER+PEIRCE+BERITA+KOMPAS.COM%29+&btnG=

R, Turban, R. R. and P. R. (2005). *Introduction to Informatio Technology*.

Sheikh, F., Karthick, S., Malathi, D., & ... (2016). Analysis of data mining techniques for weather prediction. In *Indian Journal of …*. academia.edu. http://www.academia.edu/download/60805050/prediksi_hujan_c4520191005-55415-1e1j7bn.pdf

Sudipa, I. G. I., I Nyoman Alit Arsana, & Made Leo Radhitya. (2020). Penentuan Tingkat Pemahaman Mahasiswa Terhadap Social Distancing Menggunakan Algoritma C4.5. *SINTECH (Science and Information Technology) Journal*, *3*(1), 1–7. https://doi.org/10.31598/sintechjournal.v3i1.562

Sulistyowati, D. N., Yunita, N., Fauziah, S., & Pratiwi, R. L. (2020). *Implementation of Data Mining Algorithm for Predicting Popularity of Playstore Games in the Pandemic Period of Covid-19*. *6*(1), 95–100. https://doi.org/10.33480/jitk.v6i1.1425

Tan, p et all. (2006). *Introduction to Data Mining*. Pearsion Education.

Watratan, A. F., Puspita, A., & Moeis, D. (2020). Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia. *Journal of Applied Computer Science and Technology ( Jacost )*, *1*(1), 7–14.

Wei, W. W. S., & Hamilton, J. D. (1994). Time series analysis. In *Prentice Hall New Jersey 1994: Vol. SFB 373* (Issue Chapter 5, pp. 837–900). https://doi.org/10.1016/j.ijforecast.2004.02.001

Winata, K. A., Zaqiah, Q. Y., Supiana, & Helmawati. (2021). Kebijakan Pendidikan di Masa Pandemi. *Https://Jurnal.Um-Palembang.Ac.Id/Jaeducation/Article/View/3338*, *4*, 1–6. http://dx.doi.org/10.1016/j.encep.2012.03.001

*name of corresponding author