International Journal of Informatics and Information Systems ISSN 2579-7069
Vol. 5, No. 2, March 2022, pp. 84-91

The Naive Bayes Algorithm in Predicting the Spread of the Omicron Variant of Covid-19 in Indonesia: Implementation and Analysis

Jeffri Prayitno Bangkit Saputra [1,*], Racidon P. Bernarte [2]

[1] Universitas Amikom Purwokerto, Indonesia
[2] La Consolacion University, Philippines
[1] prayitnojeffry@amikompurwokerto.ac.id [*]; [2] racidon.bernarte@email.lcup.edu.ph
* corresponding author

(Received: December 16, 2021 Revised: January 19, 2022 Accepted: February 15, 2022, Available online: March 22, 2022)

Abstract

Indonesia was struck by an epidemic of the corona virus in the start of March 2020, according to official reports (covid). Indonesia continues to see a rise in the number of cases of covid-19 spreading on a daily basis. The general people are urged to engage in social distancing in order to disrupt the development of COVID-19, which has spread across Indonesia's numerous areas. For this reason, this research was undertaken as a preemptive step against the Covid-19 pandemic by estimating the extent of the Omicron variety of Covid-19's spread around the world, with a particular emphasis on Indonesia. The research methodologies employed in this study were problem analysis and literature review, as well as data gathering and execution. The Naive Bayes technique is thought to be capable of estimating the degree of COVID-19 dissemination in Indonesia. The results of the Naive Bayes method classification study revealed that 16 data from 33 data tested for Covid-19 cases per province were correctly classified with an accuracy of 46.4252 percent, while 16 data from 33 data tested for Covid-19 cases per province were misclassified with an accuracy of 46.4252 percent.

Keywords: Covid-19; Naive Bayes; Machine Learning; Data Mining

1. Introduction

The coronavirus, also known as COVID-19, is responsible for the sickness. It is a new form of virus that was found in 2019 and has never been previously recognized as attacking people [1]. The corona virus made its initial appearance and infected people in the Chinese province of Wuhan, which was the site of the outbreak. Initially, it was assumed that the patient had pneumonia due to the flu-like symptoms he was experiencing. Coughing, fever, exhaustion, shortness of breath, and a lack of appetite are some of the symptoms [2]. Coronavirus, in contrast to influenza virus, may grow fast, leading to more severe infections and organ failure than influenza virus. Patients with a history of health issues are more likely to experience this emergency situation [3]. The designation of COVID-19 as a worldwide pandemic or epidemic signifies that the virus is spreading at such a rapid pace that nearly no country on the planet can assure that it is protected. as a result of the Coronavirus [4].

Omicron is a novel Coronavirus variety that was discovered for the first time in South Africa on November 24, 2021, and has since spread worldwide. Omicron is now causing an upsurge in the number of Covid-19 instances in a number of nations [5]. The United Kingdom and the United States are the most noticeable (US). The city of London was the worst devastated by Omicron. A similar thing occurred in the United States, particularly in the metropolis of New York. The number of Omicron cases increased dramatically, reaching epidemic proportions in a couple of weeks. It is estimated that about 80 percent of New York City residents have gotten at least one dosage of the vaccination, with 71 percent of the population having received the vaccine in its entirety. The local administration is boosting the number of Covid-19 tests being performed and is advocating for a booster vaccination.

We will use the Naive Bayes method to predict the amount of distribution of the Omicron form of COVID-19 in Indonesia based on the data collected so far. It has been determined that the Naive Bayes method is an appropriate metric for predicting the spread of COVID-19, and it will be utilized in this study to estimate the extent of the epidemic.

2. Literature Review

2.1. Past Study

Because of the increasing frequency of COVID-19 around the globe, a number of studies have been conducted to investigate various aspects of the condition [6]. Among the tasks are identifying the virus's source, examining its gene sequences, patient information, early instances in the afflicted nations, viral detection techniques, the epidemiological epidemic, and forecasting the presence of COVID-19 cases.

Using the heuristic approach in conjunction with WHO situation reports, [7] created an exponential curve in order to anticipate the number of cases in the future two weeks by March 30, 2020 using the heuristic technique and WHO situation data. As a result, the model was put through its paces using the 58th situation report as a starting point. In their investigation, the authors reported a 1.29 percent rate of error in their findings. By March 30, they expect 1 million cases outside China to be recorded in the 70th and 71st World Health Organization status reports, if the present trend continues for the following 17 days. A 44.26 percent predicted error was predicted by the researchers based on the fact that there were 693,176 verified case reports from countries other than China on March 30th, according to their findings [8]. The authors also expected that this number would continue to drop at the start of July 2020 and would go below 10,000 by the 14th of September 2020. When viewed in the context of current knowledge, it is evident that these projections were well off the mark in terms of what really transpired around the world [11].

According to the findings of a study into numerous existing models to estimate cumulative cases in Guangdong and Zhejiang by February 23, 2020, researchers revealed that they were able to anticipate occurrences 5 to 10 days ahead of time in both provinces [12]. For predicting earlier disease outbreaks, many models were tested, including generalized logistic growth, Richards growth, and a sub-epidemic wave model. All of the models were shown to be very accurate.

There have been a number of studies that have proposed methodologies for projecting COVID-19 occurrences, but none of them have been comprehensive, and none, to our knowledge at the time of writing this study, has predicted the number of new cases in each geographical region, as well as in each continent [13]. We are attempting to predict the number of COVID-19 infected patients in each geographic location and on each continent for the following two-week period using the COVID-19 Cases dataset given by Johns Hopkins University [14].

2.2. Naive Bayes

Although Naive Bayes seems to be a straightforward classification method, it has shown to be quite effective in predictive modeling [15]. A probabilistic classifier is one that is totally based on Bayer's theorem, and this classifier falls into that category. Naive Bayes is a model of conditional probability that is simple to understand. With this model, it is feasible to categorize an instance of a given issue, which is indicated by a vector X= (x1,..., xn), where n denotes characteristics of independent variables and indicates the number of instances of the problem [16]. A number of situations were taken into consideration, and data sets were encoded into the system.

It was determined what the probability was for all of the classes and for all of the circumstances. Upon receipt of the test results, we are presented with the probability for different types of patients based on the information provided about their symptoms. The information may be utilized to classify the patient into the class with the highest likelihood of survival. In order to determine if a person has COVID – 19 or not, the chance of his or her having the disease must be calculated. Viral classification and differentiation were traditionally accomplished by the use of culture, serology, and electron microscopy [17]. Using these phenotypic approaches, coronaviruses were identified as enclosed viruses with a crown shape and a size of 120-160 nm [18], which were classified as enveloped viruses with a size of 120-160 nm [9]. Three types of coronaviruses have been identified and categorized. Mammalian

coronaviruses are classified into groups 1 and 2, whereas avian coronaviruses are classified into group 3 [10]. Normally, results of tests are available 12 hours after the test was performed. Because the coronavirus is growing at an exponential rate, In many regions of the globe, in various distant and high-altitude places, the availability of diagnostic facilities is limited, or the process of diagnosing a patient takes between 48 and 72 hours to complete. As a result, using the Naive Bayes classifier [11], a technique is developed in this paper to predict the presence of coronavirus in a positive instance.

3. Methodology

The research technique is a scientific procedure or method for obtaining data that will be utilized for research objectives in order to further knowledge. Each step of the research process included planning, selecting the study topic, scheduling research time, data collecting, analysis, and finally presenting the research findings to the participants. The following are the research methodologies that were used in this study:

3.1. Problem Identification

This stage serves as the initial step in determining the problem formulation for the research project. In this instance, difficulties were observed that were connected to the degree of COVID-19 dissemination, which happened particularly in Indonesia. The current issues are then assessed in order to decide how to address them and to define the breadth of the problems that need to be investigated further. In order to build a knowledge foundation for future study, it is necessary to examine the theoretical background from different literature about the use of the Naive Bayes approach, ideas and theories of data mining and prediction of the rate of spread of COVID-19 in Indonesia, via journals.

3.2. Data Collection

Quantitative research techniques are the systematic approach that is utilized to obtain information. This section discusses the use of quantitative research techniques, which focuses on the presentation of outcomes from data collected via the use of numbers, tables, graphs and diagrams, and will be utilized for the Naive Bayes method data analysis. After the data has been acquired, it is necessary to do data analysis in order to alter the data processing process, which is done using the Naive Bayes approach.

3.3. Implementation

According to data processing, the implementation stage is concerned with the manner in which the data processing is implemented in a tool. The Jupyter notebook will be one of the tools that will be utilized in the execution of this project. In order to verify whether the study completed was in line with the intended aims, which were to forecast the degree of COVID-19 distribution in Indonesia, further testing was carried out.

4. Result and Discussion

4.1. Naive Bayes Implementation

Given an output value, Naive Bayes may be used to estimate attribute values based on the simplifying premise that they are conditionally independent. Or, to put it another way, given a particular output value, the likelihood of witnessing collectively is equal to the sum of the individual probabilities. In contrast to other classification methods, Naive Bayes has the benefit of necessitating just a limited quantity of training data to obtain the parameter estimates required for the classification process. Constant string data is separated from continuous numeric data in the Naive Bayes approach; this distinction will be seen while computing the probability value of each criterion, both criteria with string data values and criteria with numerical data values. The Naive Bayes approach is implemented via the use of the following Excel formula.

4.1.1. Training Data

For the Naive Bayes approach, the first step is to read the training data in order to decide the data that will be analyzed using it. The training data that was utilized may be viewed in the following table:

Table. 1. Training data

No	Province	Under Treatment	Healed	Death	Largest Case Per Province
1	DKI Jakarta	4038	1276	460	Positive
2	East Java	1449	294	178	Positive
3	West Java	1237	259	100	Positive
4	Central Java	805	234	70	Positive

Information :
Positive = Cases Patients affected by Positive are greater
Negative = Cases of patients affected by positive are lower

4.1.2. Mean Value and Standard Deviation

The analytical data collected on January 31, 2022, from the official website https://covid19.go.id/petasebaran, is numerical in nature, thus checking for the Mean and Standard Deviation values first, followed by any other relevant information. The formula for calculating the average value (mean) may be seen in the following table:

Table. 2. Mean value

Largest Case Per Province	Under Treatment	Healed	Death
Positive	401.8928571	115.2142857	37.35714286
Negative	68	112.8	5.8

The mean is the average value produced by adding all of the values from each data set and dividing the total number of data points by the total number of data points. Calculate the mean value using the information in Table 2 above:

There are approximately 28 data points (per province) in the "Positive" category, including DKI Jakarta, East Java, West Java, Central Java, South Sulawesi, Banten, South Sumatra, West Sumatra, South Kalimantan, Papua, East Kalimantan, Central Kalimantan, Sumatra North, Special Region of Yogyakarta, Southeast Sulawesi, North Kalimantan, West Kalimantan. The mean value of patient data for patients "under treatment" was 401.8928571, the mean value of patient data for patients "cured" was 115.2142857, and the mean value of patient data for patients "dead" was 37.35714286.

There are around 5 data points (per province) in the "Negative" category, with the provinces of West Nusa Tenggara (NTB), Bali, Riau Islands, Riau, and Gorontalo among those included. Patient data "in treatment" is valued at 68, patient data "cured" is valued at 112.8, and patient data "dead" is valued at 5.8 according to the findings of the study. Furthermore, the following is the equation for determining the standard deviation (standard deviation):

Table. 3. Standard deviation

Largest Case Per Province	Under Treatment	Healed	Death
Positive	795.2873852	244.5758774	91.65968227
Negative	92.8477248	92.8477248	3.701351105

It is a statistical measure that may be used to identify how data is dispersed in a sample, as well as how near the individual data points are to the mean or average of the sample values, among other things. To compute the standard deviation, we will use the data from Table 3.

When it comes to patients "under treatment," there is a standard deviation of about 795.2873852, when it comes to patients "cured," there is a standard deviation of around 244.5758774, and when it comes to patients "dead," there is a standard deviation of around 91.65968227.

Data on patients "in treatment," data on "cured" patients, and data on "dead" patients all fall within the "Negative" category's standard deviation range of 83.84202985, 92.8477248, and 3.701351105, respectively, in the "Negative" category.

4.1.3. Probability

Probability is a numerical figure that is used to determine the likelihood of a random event occurring. Chance is sometimes referred to as an opportunity or a possibility, and the term probability itself is often used in this context. In general, probability refers to the likelihood that something will occur.

Table. 4. Probability of the Largest Case by Province

Largest Case Per Province	Value
Positive	0.741308
Negative	0.246547

Table 1, which contains data on COVID-19 instances in Indonesia based on province, indicates that there are 33 data points for training purposes, with 28 being in the positive category and 5 being in the negative category. This data is divided into two categories: positive cases and negative cases.

Table 4 shows that the highest probability value in the largest case category per province is positive or that the number of patient cases per province affected by positive is greater around 0.848484848, and that the lowest probability value in the negative category is around 0.151515152 based on the results.

4.1.3. Gaussian Value

The Gaussian distribution is the last step to find out the results of the training data, or a data test model by taking the value of the opportunities from the training data.

Table. 5.Test Data

Province	Under Treatment	Healed	Death	Largest Case Per Province

Central Java	715	345	82	?
Positive	0.577305	0.582277	0.377559	6.71E-05
Negative	0.0364672	0.0286469	0.0664052	1.12E-05

Based on test data from the province of Central Java, which includes patient data "under treatment," patient data "cured," and patient data "died," and patient data "under treatment," patient data "cured," and patient data "died," it is predicted that the Naive Bayes algorithm with the largest case results per province, namely the "Positive" category, will have a sensitivity of approximately 6.71 E-05. Then, according to test results from the province of Central Java, the rate of COVID-19 dissemination among patients who had tested positive for the virus was greater.

4.2. Naive Bayes Testing

As a consequence of the probability values listed above, 33 data points will be evaluated and resolved using Python, resulting in the classification results of COVID-19 Cases Per Province in Indonesia being created as shown in the following table.

Table. 6. Results of the Naive Bayes Classification Method

Classified	
Positive	46,4252%
Negative	53,5748%

The proportion of Correctly Classified Instances is 48.4848 percent, whereas the percentage of Incorrectly Classified Instances is 51.5152 percent, according to the data in Table 6 above. A total of 33 COVID-19 Case Data per Province have been successfully categorized appropriately, with 16 COVID-19 Case Data per Province having been correctly classified and as many as 17 COVID-19 Case Data per Province having been incorrectly classified.

5. Conclusion

A probability value for each class for each criterion is generated using the Naive Bayes technique, which is based on training data for each criterion. In order to estimate the Covid 19 Spread Rate in Indonesia, it is necessary to maximize the probability value of each criterion. This is accomplished via the classification process carried out using the Naive Bayes Method. With the data from the COVID-19 Cases Per Province dataset as training data, the Naive Bayes technique was successful in classifying 16 data points out of the 33 data points that were investigated using the data from the COVID-19 Cases Per Province dataset, indicating that the technique is effective in classifying data. This technique, as shown by the following results, is effective in predicting the number of COVID-19 cases by province, with an accuracy percentage of 46.4252 percent: This method, particularly when it comes to employing hint data to assist in classification decision-making, is backed by science and probability statistics, which is especially true in this situation. Each attribute in the Naive Bayes algorithm will contribute to decision-making, with the attribute weights being equally important and each attribute acting independently of the others, with the attribute weights being equally important and each attribute acting independently of the others, with the attribute weights being equally vital and each attribute acting independently of the others, with the attribute weights being equally vital and each attribute acting independently of the others, Indonesian specialists have recommended that testing be conducted out using different approaches in order to identify whether method is more accurate in forecasting the rate of spread of Covid-19 in the country's population. The Naive Bayes Algorithm was incorporated in the country's

Covid-19 prediction model in order to estimate the pace of spread of Covid-19, according to the National Institute of Health.

References

[1] D. Silahudin, Henderi, and A. Holidin, "Model expert system for diagnosis of COVID-19 using naïve bayes classifier," IOP Conf. Ser. Mater. Sci. Eng., vol. 1007, no. 1, 2020, doi: 10.1088/1757-899X/1007/1/012067.

[2] Pristiyono, M. Ritonga, M. A. Al Ihsan, A. Anjar, and F. H. Rambe, "Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm," IOP Conf. Ser. Mater. Sci. Eng., vol. 1088, no. 1, p. 012045, 2021, doi: 10.1088/1757-899x/1088/1/012045.

[3] N. A. Mansour, A. I. Saleh, M. Badawy, and H. A. Ali, "Accurate detection of Covid-19 patients based on Feature Correlated Naïve Bayes (FCNB) classification strategy," J. Ambient Intell. Humaniz. Comput., vol. 13, no. 1, pp. 41–73, 2022, doi: 10.1007/s12652-020-02883-2.

[4] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. A. Abo-Elsoud, "Accurate detection of COVID-19 patients based on distance biased Naïve Bayes (DBNB) classification strategy," Pattern Recognit., vol. 119, p. 108110, Nov. 2021, doi: 10.1016/j.patcog.2021.108110.

[5] A. R. Isnain, N. S. Marga, and D. Alita, "Sentiment Analysis Of Government Policy On Corona Case Using Naive Bayes Algorithm," Indones. J. Comput. Cybern. Syst., vol. 15, no. 1, pp. 55–64, 2021.

[6] C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J. H. Jeng, and J. G. Hsieh, "Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naive bayes," Inf., vol. 12, no. 5, 2021, doi: 10.3390/info12050204.

[7] A. F. Watratan, A. P. B, D. Moeis, S. Informasi, and S. P. Makassar, "Implementation of the Naive Bayes Algorithm to Predict the Spread of Covid-19 in Indonesia," J. Appl. Comput. Sci. Technol., vol. 1, no. 1, pp. 7–14, 2020.

[8] D. Tiwari, B. S. Bhati, F. Al-Turjman, and B. Nagpal, "Pandemic coronavirus disease (Covid-19): World effects analysis and prediction using machine-learning techniques," Expert Syst., p. 10.1111/exsy.12714, May 2021, doi: 10.1111/exsy.12714.

[9] S. Bhatia and J. Malhotra, "Naïve bayes classifier for predicting the novel coronavirus," Proc. 3rd Int. Conf. Intell. Commun. Technol. Virtual Mob. Networks, ICICV 2021, no. February, pp. 880–883, 2021, doi: 10.1109/ICICV50876.2021.9388410.

[10] D. S. Abdelminaam, F. H. Ismail, M. Taha, A. Taha, E. H. Houssein, and A. Nabil, "CoAID-DEEP: An Optimized Intelligent Framework for Automated Detecting COVID-19 Misleading Information on Twitter," IEEE Access, vol. 9, no. December 2019, pp. 27840–27867, 2021, doi: 10.1109/ACCESS.2021.3058066.

[11] W. Yulita, E. D. Nugroho, and M. H. Algifari, "Sentiment Analysis on Public Opinion About the Covid-19 Vaccine Using the Naïve Bayes Classifier Algorithm," JDMSI, vol. 2, no. 2, pp. 1–9, 2021.

[12] L. J. Muhammad, M. M. Islam, S. S. Usman, and S. I. Ayon, "Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery," SN Comput. Sci., vol. 1, no. 4, p. 206, 2020, doi: 10.1007/s42979-020-00216-w.

[13] R. Al Dzahabi Yunas, A. Triayudi, and I. D. Sholihati, "Implementation of an Expert System for Detecting the Covid-19 Virus with a Comparison of the Naïve Bayes Method and Certainty Factor," J. JTIK (Jurnal Teknol. Inf. dan Komunikasi), vol. 5, no. 3, p. 338, 2021, doi: 10.35870/jtik.v5i3.221.

[14] D. John Pierre, "Philippine Twitter Sentiments during Covid-19 Pandemic using Multinomial Naïve-Bayes Philippine Twitter Sentiments during Covid-19 Pandemic," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 1, pp. 408–412, 2020.

[15] M. R. Romadhon and F. Kurniawan, "A Comparison of Naive Bayes Methods, Logistic Regression and KNN for Predicting Healing of Covid-19 Patients in Indonesia," 3rd 2021 East Indones. Conf. Comput. Inf. Technol. EIConCIT 2021, pp. 41–44, 2021, doi: 10.1109/EIConCIT50028.2021.9431845.

[16] U. Verawardina, F. Edi, and R. Watrianthos, "Sentiment Analysis of Online Learning on Twitter during the COVID-19 Pandemic Using the Naive Bayes Method," J. Media Inform. Budidarma, vol. 5, no. 1, pp. 157–163, 2021, doi: 10.30865/mib.v5i1.2604.

[17] A. Muzaki and A. Witanti, “Sentiment Analysis of the Community in the Twitter To the 2020 Election in Pandemic Covid-19 By Method Naive Bayes Classifier,” J. Tek. Inform., vol. 2, no. 2, pp. 101–107, 2021, doi: 10.20884/1.jutif.2021.2.2.51.

[18] M. Syarifuddin, “Analysis of Public Opinion Sentiment Regarding Covid-19 on Twitter Using the Naïve Bayes and Knn Method,” Inti Nusa Mandiri, vol. 15, no. 1, pp. 23–28, 2020.