HIGH ACCURACY DETECTION OF COVID-19 BASED ON NAIVE BAYES CLASSIFIER (NBC)

D Wibowo [1], M Novia [2], R N Rumaksi[3], S I Gunawan [4]

School of Business, Study Program of Actuarial Science, President University, Cikarang, Bekasi 17550, Indonesia

[1]didik.wibowo@student.president.ac.id
[2]michelle.novia@student.president.ac.id
[3]rarasati.rumaksi@student.president.ac.id
[4]steffany.gunawan@student.president.ac.id

ABSTRACT

The Coronavirus (COVID-19) outbreak has spread across the globe at such a high speed that the number of infections and deaths of people is increasing swiftly each day. Therefore, it is very important to find positive cases early for medication and control the spread of the disease. Some techniques have been tried to detect COVID-19 in several injuries, but they have limited efficiency. Here, we present application of Naïve Bayes Classifier algorithm to COVID-19 diagnosis. The algorithm only requires two probabilities in order to be executed, which are: prior probability and conditional probability. Experimental results have proven the effectiveness of NBC which could achieve a detection accuracy of more than 90%.

Keywords: *COVID-19, Naïve Bayes, Classification, NBC*

1. Introduction

COVID-19 is a pandemic disease that has been spreading globally since 2019. And it was identified as a global pandemic since 7th January 2020 by the World Health Organization (WHO) (Kang et al. 2020). COVID-19 has various symptoms that include fever, shortness of breath, cough, headache, sore throat, muscle pain, and fatigue (Huang et al. 2020). Covid-19 is mainly spreading through physical contact. The virus is spreading from the person who is carrying it into another person through mucous contact, breathe contact, or hand contact. The rapid infections of COVID-19 had resulted in a lot of damage globally, it has a negative impact on various aspect such as: daily activity, public health, and global economy. additionally, the infection of COVID-19 only takes about 4 weeks to disturb the medical system in some place once the infection started (Shaban et al. 2020).

In order to stop the global spreading of COVID-19, quarantine is also a very good option because it can help to prevent the healthy person from getting infected by the COVID-19 virus. But, to implement the quarantine, we need to be able to differentiate between the healthy person and the person who has been infected (patient). Unfortunately, it would take a long time and a lot of resource if we were to detect the COVID-19 patient by doing some medical tests. Therefore, we would need to have a different approach in order to detect the COVID-19 patient. Through this paper, we would like to suggest using one of the machine learning models, the Naïve Bayes Classifier as a mean to detect the patient of COVID-19. Based on the experiments, by using this method, we can detect the COVID-19 patient with only using a few features and it will take a shorter time to get the result with a high accuracy.

Real-time Reverse Transcription - Polymerase Chain Reaction (RT-PCR) is the best test presently used to detect COVID-19 patients (Zu et al. 2020). While the RT-PCR test is sensitive, practically fast, and reliable, it carries the risk of both false-negative and false-positive results. In consequence, the spread of COVID-19 infection is augmentative because RT-PCR assays cannot immediately characterize infected persons (Zu et al. 2020). Early detection of Covid-19 patients can use Chest radiological imaging such as Computed Tomography (CT) images and X-rays, they also have a necessary role in Covid-19 patients. But misclassification can occur in imaging features of COVID-19 and another disease (Shaban et al. 2020). With the augmentative demand for providing precise tests, the binding to CT images or RT-PCR tests as appropriate tools for COVID-19 patient discovery has decreased significantly. Because of that, quick and scrupulous detection of COVID-19 patients is highly important to avoid the infection source. Nowadays, machine learning is an additional tool for clinicians. Machine learning can help to automatically support medical diagnosis to identify and detect the novel coronavirus.

Machine learning is one of the branches of artificial intelligence, an artificial intelligence that allows systems to adapt human abilities to learn (Rabie et al. 2015; Rabie et al. 2019a). This science knowledge focuses on creating systems or algorithms that continuously learn from data and improve their accuracy over time without any particular programming. In machine learning applications, algorithms or sequences of statistical processes are trained to find specific patterns and features in large amounts of data. It aims to make a decision or prediction based on these data. The better the algorithm, the better the system's decision and prediction accuracy will be. Since machine learning now become more efficient, it's also used for covid-19 predictions, however they also have some limitations such as, low accuracy, takes lot of time and complex.

Naïve Bayes Classifier is a classification method based in Bayes theorem. The classification method used are probability and statistical, it is proposed by British scientist Thomas Bayes, which predicts future opportunities based on previous experience, is known as Bayes' theorem (Widianto 2019). In this case, it is assumed that the presence or absence of a particular event from one group is not related to the presence or absence of another event. The advantage of the use of Naive Bayes is that this method only requires a small amount of training data to determine the parameter estimation needed in the process classification. Naive Bayes often performs much better in most complex real-world situations, even though its assumption is more simplified (Dada et al. 2019; Ali and Ali 2020; Hewage et al. 2020; Lei et al. 2020). Therefore, it's good to use Naïve Bayes as Covid-19 prediction. Some evidence why using Naïve Bayes as follows: (a) It can handle both quantitative and discrete data. (b) It only requires a small amount of training data to estimate the parameters (mean and variance of the variables) required for classification. (c) Could handle missing values by ignoring instances during probability estimation calculations. (d) Could handle noise in the dataset. (e) Sturdy against irrelevant attributes. (f) Its simple and easy to implement. (g) it's enough for applications in real life for example diseases diagnoses because it's based on an already computed probabilities that make the classification could be finished fast. (h) Not required to adjust domain knowledge or parameter (Khotimah et al. 2020; Kaur and Oberoi 2020).

Recently, research on Covid-19 detection methods has been carried out extensively. An Accurate detection of Covid-19 patients based on Feature Correlated Naïve Bayes (FCNB) classification strategy has been introduced by Mansour et al. (2020) for detecting a COVID-19 Patients using Naïve Bayes Classification as its base. The FCNB consists of four phases, namely; Feature Selection Phase (FSP), Feature Clustering Phase (FCP), Master Feature Weighting Phase (MFWP), and Feature Correlated Naïve Bayes Phase (FCNBP). The FSP phase will selects the most effective features by using the Genetic Algorithm. The FCP phase will make a lot of clusters of features that is based on previously selected features from FSP phase by using the clustering technique. All of these feature clusters are named Master Features (MFs). The MFWP phase will assigns a certain weight value for each MF with a new method of weight calculation. And then, the FCNBP phase is executed to classify which is the patients using the weighted Naïve Bayes algorithm. Meanwhile, we will only use a simple Naïve Bayes Classification (NBC) as a method to predict the COVID-19 Patients. So, with this method, we would only need to calculate the prior probability, posterior probability, and the likelihood in order to get the result prediction.

2. Research Method

To introduce the Naïve Bayes Classifier (NBC) as a method to accurately detect COVID-19 Patients, we used a secondary data consisting of twenty-five patient data as a database. The database then will be separated into training data in which the Naïve Bayes Classifier model will be build and also test data which are used to test the accuracy of the prediction based on the Naïve Bayes Classifier model.

2.1. Data Selection

The data used is a dataset in the form of records consisting of 25 patient data in Ms. Excel format (Mansour et.al. 2021). Displaying data with the features of Platelet Count (PC), White blood cell counts (WBC), Monocytes Count (MC), Aspartate aminotransferase (AST), Basophils Count (BC), Lactate Dehydrogenase (LDH), and also the diagnosis result of COVID-19 in the patients based on accurate Numerical Laboratory Tests (NLTs).

Table 1. The "COVID-19" dataset with nominal values

Patient	PC	WBC	MC	AST	BC	LDH	Diagnosis
1	Low	Low	Low	High	Normal	Normal	TRUE
2	Low	Low	Normal	High	Normal	High	TRUE
3	Low	High	Normal	High	Normal	Normal	FALSE
4	Low	High	Normal	High	High	Normal	TRUE
5	Low	Normal	High	High	Normal	Normal	FALSE
6	Low	Normal	Normal	High	Normal	High	TRUE
7	Normal	Low	Low	High	Normal	Normal	TRUE
8	Normal	High	Normal	High	Normal	Normal	FALSE
9	Normal	High	Normal	High	High	High	TRUE
10	Normal	Normal	High	Normal	Normal	Normal	FALSE
11	Normal	Normal	High	Normal	Normal	High	TRUE
12	High	Low	Low	Normal	Normal	Normal	TRUE
13	High	Normal	High	Normal	Normal	Normal	FALSE
14	High	Normal	High	High	High	High	TRUE
15	High	High	Normal	High	Normal	High	TRUE
16	Low	Normal	High	Normal	High	Normal	FALSE
17	Normal	Normal	High	High	High	Normal	FALSE
18	High	Low	Low	High	Normal	High	TRUE
19	Normal	Normal	Normal	High	Normal	Normal	FALSE
20	Normal	High	Normal	High	Normal	High	TRUE
21	Normal	Low	Normal	High	Normal	High	TRUE
22	Low	High	Normal	High	High	High	TRUE
23	Low	Low	Low	High	High	High	TRUE
24	High	High	Normal	Normal	Normal	Normal	TRUE
25	High	Normal	Normal	Normal	Normal	Normal	FALSE

2.1.1. Naïve Bayes Algorithm

Naïve Bayes Classifier is a classification method that is very simple but can have a fairly high level of accuracy. This classification method is also used quite often because it only requires a small sample of data to be applied. The Naïve Bayes Classifier is based on the Naïve Bayes theorem which assumes that there is no relationship between one variable and other variables. In the theorem there is the word "Naïve" because in the real world a variable that will be used will definitely have a relationship with other variables so that the assumption is considered Naïve. Basically, the Naïve Bayes Classification method is made by calculating the probability of an object being in each class using the Naïve Bayes theorem, then the object will be classified into the class that has the highest probability. There is also the Naïve Bayes theorem algorithm as follows:

$$P(X|\mathrm{Y}) = \frac{P(Y|\mathrm{X}).P(X)}{P(\mathrm{Y})} \tag{1}$$

This algorithm is the base formula of Bayes Theorem with Y as the data with unknown class, X as the hypothesis data Y in a specific class. P(X|Y) symbol is the hypothesis probability X based on condition Y (posterior probabilities). P(X) or we can call as prior probabilities is the hypothesis probability of X. P(Y|X) symbol is probability Y under these conditions, and last P(Y) is the probability of Y.

2.2. Data Mining

In this process, we will use program R to further process the database and create training and test data, and also compute the Naïve Bayes Model and implement the model into the test data.

Table 2. A brief description of the attributes used to detect COVID-19 patients (Mansour et.al.2021)

Feature	Abbreviations	Description	Unit
White Blood Cells Counts	WBC	WBC is the total number of leukocytes or white blood cells in our body. High leukocytes means the body is fighting infection. Low leukocytes means that there is a problem with the spinal cord.	$x10^9$cells/L
Monocytes Count	MC	Monocytes are white blood cells that circulate in the blood and spleen. MC determine the number of monocytes in blood. This type of white blood cell is important for alerting the immune system to previous infections	$x10^9$cells/L
Platelet Count	PC	PC determine the number of platelet in blood. Platelets play an important role in the process of blood clotting. Platelets are also often used in screening methods (early detection) and diagnosing various diseases caused by disorders of blood clotting.	$x10^9$cells/L
Basophils Count	BC	Basophils are a type of white blood cell that plays an important role in the immune system. These basophil cells play an important role in producing an inflammatory reaction to fight infection. In addition, basophils also play a role in the emergence of allergic reactions.	$x10^9$cells/L
Lactate Dehydrogenase	LDH	LDH is an enzyme that is present in almost all cells in the body, including blood cells, muscles, brain, kidneys, pancreas, heart, and liver. LDH is responsible for converting sugars obtained from food into the energy needed by each cell.	U/L
Aspartate Aminotransferase	AST	AST is a group of transaminase enzymes found in the liver, heart, skeletal muscle, and kidneys. High AST levels indicate there is damage to the liver organs. The AST test is a blood test that checks for liver disease	U/L

2.2.1. Install and load the required packages

First, we need to install and load 4 packages related to Naïve Bayes. We use e1071, caret, readxl, and rsample packages from Rstudio. Caret and e1071 packages are use because they have the required Naïve Bayes modelling functions. Readxl is used to read our dataset from Microsoft Excel because we use xlsx type. Last is rsample packages that used for make the training and test data.

2.2.2. Import data and split data

Next step, we need to import the Covid-19 patient dataset into R using readxl package. After successfully importing the dataset, we split the dataset into training and test data. We choose 60% of the data to be the training data then extracting them into training and test data as two separate data frames.

2.2.3. Naive Bayes Model

Next, we have to do Setting seed so that the data didn't change, and also make the Naive Bayes model out of the training dataset. After that, we will get the Naïve Bayes model result.

3. Results and Discussion

After we have the Naïve Bayes model, then we can apply the model to make a prediction into the test data by using predict command. We also make the confusion matrix and statistics.

Table 3. The 15 randomly selected data from R that we are going to use as our training data

Patient	Platelet Count	White Blood Cell	Monocytes Count	Aspartate aminotransferase	Basophils Count	Lactate dehydrogenase	Diagnosis
1	Low	Low	Low	High	Normal	Normal	TRUE
2	Low	Low	Normal	High	Normal	High	TRUE
3	Low	High	Normal	High	Normal	Normal	FALSE
4	Low	High	Normal	High	High	Normal	TRUE
6	Low	Normal	Normal	High	Normal	High	TRUE
10	Normal	Normal	High	Normal	Normal	Normal	FALSE
11	Normal	Normal	High	Normal	Normal	High	TRUE
13	High	Normal	High	Normal	Normal	Normal	FALSE
15	High	High	Normal	High	Normal	High	TRUE
16	Low	Normal	High	Normal	High	Normal	FALSE
18	High	Low	Low	High	Normal	High	TRUE
19	Normal	Normal	Normal	High	Normal	Normal	FALSE
21	Low	Low	Normal	High	Normal	High	TRUE
22	High	High	Normal	High	High	High	TRUE
24	High	High	Normal	Normal	Normal	Normal	TRUE

Tabel 4. Prior probabilities from the training data

A-priori probabilities	
FALSE	0.3333333
TRUE	0.6666667

Table 5. Conditional probabilities from the training data

Conditional probabilities			
	Platelet Count		
	High	Low	Normal
FALSE	0.2	0.4	0.4
TRUE	0.3	0.5	0.2
	White Blood Cell		
	High	Low	Normal
FALSE	0.2	0.0	0.8
TRUE	0.4	0.4	0.2
	Monocytes Count		
	High	Low	Normal
FALSE	0.6	0.0	0.4
TRUE	0.1	0.2	0.7
	Aspartate Aminotransferase		
	High		Normal
FALSE	0.4		0.6
TRUE	0.8		0.2
	Basophils Count		
	High		Normal
FALSE	0.2		0.8
TRUE	0.2		0.8
	Lactate Dehydrogenase		
	High		Normal
FALSE	0.0		1.0
TRUE	0.7		0.3

Tabel 6. Confusion Matrix and Statistics

	Prediction	
	FALSE	TRUE
FALSE	3	1
TRUE	0	6

As we can see from the result, out of 10 predictions, 9 patients can be predicted correctly. This Naïve Bayes prediction can give us a 90% rate of accuracy. And based on the Naïve Bayes algorithm itself, the accuracy rate can be potentially increased by increasing the training data and/or increasing the number of features that are used to get more conditional probability so that this method can give us a higher accuracy COVID-19 detection method.

4. Conclusion and Implications

A lot of features can be used to help to predict the diagnosis of Covid-19 In a patient. This also includes Platelet Count (PC), White blood cell counts (WBC), Monocytes Count (MC), Aspartate aminotransferase (AST), Basophils Count (BC), Lactate Dehydrogenase (LDH). The above-reported results show that the Naïve Bayes models can be used as a tool in predicting Covid-19 patient diagnosis. Based upon results, we can conclude that our proposed Naïve Bayesian Classifier is simple to implement and performs well according to the overall percentage of cases that are correctly predicted by the model which indicates 90% accuracy in predicting COVID-19 patients. We would like to recommend using this Naïve Bayes Classifier as a high accuracy COVID-19 detection method.

Although the Naïve Bayes Classifier has high accuracy, it also has a defect which is: Naive Bayes models consider all the features to be completely unrelated to each other in any way. This is very unlikely in real-world applications. Thus, we suggest modifying the Naïve Bayes Classifier so that it can consider the relationship between each feature for future research.

Table 7. Detailed reports of the confusion matrix

Accuracy	**0.9**
95% CI	0.555, 0.9975
No Information Rate	0.7
P-Value [Acc > NIR]	0.1493
Kappa	0.7826
Mcnemar's Test P-Value	1.0000
Sensitivity	1.0000
Specificity	0.8571
Pos Pred Value	0.7500
Neg Pred Value	1.0000
Prevalence	0.3000
Detection Rate	0.3000
Detection Prevalence	0.4000
Balanced Accuracy	0.9286
'Positive' Class	FALSE

References

Ali ZH, Ali HA (2020) QoS provisioning framework for service-oriented internet of things (IoT). Clust Comput 23:575–591

Dada E, Bassi J, Chiroma H, Abdulhamid S et al (2019) Machine learning for email spam filtering: review, approaches and open research problems. Heliyon 5(6):1–23

Hewage P, Trovati M, Pereira E, Behera A (2020) Deep learning– based effective fne–grained weather forecasting model. Pattern Anal Appl. https://doi.org/10.1007/s10044-020-00898-1

Huang C, Wang Y, Li X, Ren L et al (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395(10233):497–506

Kang H, Xia L, Yan F, Wan Z et al (2020) Diagnosis of coronavirus disease 2019 (covid-19) with structured latent multi-view representation learning. IEEE Trans Med Imaging 39(8):2606–2614

Kaur G, Oberoi A (2020) Novel approach for brain tumor detection based on Naïve Bayes classification. In: Sharma N, Chakrabarti A, Balas V (eds) Data management, analytics and innovation. Advances in intelligent systems and computing (1042). Springer, Singapore, pp 451–462. https://doi.org/10.1007/978-981-32-9949-8_31

Khotimah B, Miswanto M, Suprajitno H (2020) Optimization of feature selection using genetic algorithm in Naïve Bayes classification for incomplete data. Int J Intell Eng Syst 13(1):334–343

Mansour, N. A., Saleh, A. I., Badawy, M., & Ali, H. A. (2021, January 15). Accurate detection of Covid-19 patients based on Feature Correlated Naïve Bayes (FCNB) classification strategy (N. A. Mansour, Ed.). Accurate detection of Covid-19 patients based on Feature Correlated Naïve Bayes (FCNB) classification strategy, 10(Ambient Intelligence and Humanized Computing), 33.

Rabie AH, Saleh AI, Abo-Al-Ez K (2015) A new strategy of load forecasting technique for smart grids. IJMTER 2(12):332–341

Rabie AH, Ali SH, Ali HA, Saleh AI (2019a) A fog based load forecasting strategy for smart grids using big electrical data. Clust Comput 22(1):241–270

Shaban W, Rabie AH, Saleh AI, Abo-Elsoud M (2020) A new COVID19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier. Knowl-Based Syst 205:1–8

Widianto Mochammad Haldi (23 December 2019) Algoritma Naive Bayes. Retrieved at 3 July 2021 from https://binus.ac.id/bandung/2019/12/algoritma-naive-bayes/

Zu Z, Jiang M, Xu P, Chen W et al (2020) Coronavirus disease 2019 (COVID-19): a perspective from China. Radiology 296(2):15–25