Comparison Of C4.5 And Naïve Bayes Algorithm To Determine Recommendations of Patients Receiving The Covid-19 Vaccine at Cimanggis Jaya Clinic

Trevino Aristarkus Pakasi[1,a)] Lilyani Asri Utami[2,b)] Artika Surniandari[3,c)] Hilda Rachmi[4,d)] Dini Nurlela[4,e)]

[1] *Medical Education, Universitas Indonesia, Jakarta, Indonesia*
[2]*Information System, Universitas Nusa Mandiri, Jakarta, Indonesia*
[3]*Accounting Information System, Universitas Bina Sarana Informatika, Jakarta, Indonesia*
[4]*Information System, Universitas Bina Sarana Informatika, Jakarta, Indonesia*

[b)] Corresponding author: lilyani.lau@nusamandiri.ac.id
[a)]pakasitrevino@gmail.com
[c)]artika.ats@bsi.ac.id
[d)]hilda.hlr@bsi.ac.id
[e)]dini.dur@bsi.ac.id

Abstract. Providing vaccines for the general population is one of the concrete steps taken by the government to overcome the spread of the Covid-19 virus in Indonesia. As a primary care clinic, Cimanggis Jaya provides rapid antigen and PCR testing services for its patients. With the commencement of vaccines, the clinic also accepts patients who need consultation or recommendations regarding whether or not patients can be vaccinated. Therefore, it is necessary to analyze the patient whether it is recommended or should undergo further examination using comparative data mining classification methods to find out which algorithm is good for predicting vaccine administration, namely by using the C4.5 and Naïve Bayes algorithm. The results of the evaluation and validation show that the C4.5 algorithm has an accuracy value of 99.20%, while the Naïve Bayes method has an accuracy value of 98.20%. Both algorithms have AUC level with Excellent Classification diagnostics. Thus, the C4.5 algorithm is a fairly good method in predicting the recommendations of patients who are eligible for vaccines and determining who should undergo further examination. For patients who have other symptoms, further research needs to be done, if they do not have criteria that must be checked, a vaccine is recommended.

INTRODUCTION

Since the end of 2019, COVID-19 became the most terrifying disease because it was easly spread out and had infected more than 115 million people all over the world. The fatality triggered by the disease reached 2.560.647 deaths in March 2021 [1]. The Indonesian special body to tackle COVID-19 reported 1.347.026 confirmed cases. Among them, 1.160.863 were cured and 36,518 died [2]. Based on September 2020 report of the Biro Pusat Statistik (national statistical berau or BPS), the country population was 270,2 million people [3].

Therefore, Indonesia is still expecting high incidence of COVID-19 in this exponential phase of the pandemic. However, due to the limitations of the laboratories compared to the number of the population and the geographical constraints,we need to reformulate the role of primary care providers to contribute in the screening and treating of patients accordingly at the earliest possible clinical stage, properly assessing and managing the risk and referring patients appropriately whenever necessary because the number of hospitals are still very limited. The use of e-

consultation or remote consultation should be considered to screen more patients based on their reported complaints and history to minimize the physical contact but still continue assisting the patients based on their needs [4].

The country implemented the policy of large-scale social restriction that refered to the Indonesian Act no. 6 / 2018 which was further explained in the government regulation (PP no. 21/2020). Other policy for social distancing and physical distancing were applied for Indonesians since March 2020 [5]. The suggestion from our government was not obeyed because the public awarness was low.

As a preventive action, researchers is still searching for effective vaccine though it has been launched globally as well as in Indonesia. The vaccine that contains inactive virus will induce specific immunity againts specific disease [6]. In the beginning 2021, the vaccine has been available and dissminated to the targeted community with limited data about the benefit and disadvantage of the administartion due to time constraint. However there has been some consideration to classify beneficiaries based on several pre conditions that might have adverse reaction. Therefore a screening questionnaire is nesccessary to classify the population whether or not eligible to receive the vaccination.

The classificaiton techniques in statistics for mining data are the C4.5 and Naïve Bayes algorithm. We provided here three examples to prove the use of the two algorithms were excellent. A study among Hope Family Program beneficiaires found fraut among the families who receive the grant, thus the ministry had to reanalyse the data using data mining method. The study found solution to be accurate in classifying beneficiaries using the C4.5 and Naïve Bayes algorithm [7]. Other study applied the two algorithms, the Naïve Bayes Clasifier and Decision Tree (C4.5), to help predicting possibility for paying following given loan. The study calculated and found that the Decision Tree (C4.5) method had higher accuracy and efficiency (in term of duration) as compared to the Naïve Bayes Classifier method [8] to observe who would pay faster than the other group. There is one study that used Naïve Bayes Classifier to predict patients recovery, after infected by corona virus. The study result suggest to use the Naïve Bayes since it predicted 84% findings. However to be more accurate, it was suggested to combine with other method and found the most accurate and effective one [9].

The use of the C.45 algorithm and in comparison with Naive Bayes aims to produce a decision tree by dividing large patient data sets into smaller record sets through the application of established rules for vaccination, while the application of Bayes rules assumes strong independence. Besides that, Naive Bayes can also analyze the variables that most influence it in the form of opportunities [10]. Naïve Bayes Classifier is the best method for the text classification algorithm to improve performance of the classification [11].

In the study, authors compared the two algorithim classification of data mining, which were C4.5 and *Naïve Bayes* to find the highest accuracy in predicting population to receive vaccine among patients who came to a private clinic to seek COVID-19 rapid test. The private clinic which was Klinik Cimanggis Jaya further recommended them to receive vaccine for COVID-19.

METHOD

The classification process uses data mining techniques, which are a series of processes to dig up information that has not been known manually from a large database so that it is often called Knowledge Discovery Database (KDD) [1]. There are 5 (five) stages of a series of processes contained in the Knowledge Discovery Database (KDD), namely: Data Selection, Preprocessing, Data Transformation, Data Mining Process, and Evaluation [12].

a. Data Selection

   Data sets were derived from the patients medical record who came to the clinic since July until December 2020 seeking rapid antigen test. Not all data contained in the database is used in the data mining process. The data selection stage is the stage of determining or selecting the attributes that will be used in the data mining process [13]. The database structure available is Name, Gender, Mobile Number, Address, Date of Birth, Email, NIK, Province of Residence, City/Regency, District, Nearest Health Center, Rapid Antigen Results, Contact with Covid-19 Patients, Fever Symptoms, Symptoms Cough, Sore Throat Symptoms, Discomfort when Breathing, dyspnoe that limited eating and driking, consuming specific drugs in the past 14 days, Pregnant, and Comorbid. The data/attributes selected are Date of Birth, Rapid Antigen Results, Contact with Covid-19 Patients, Fever Symptoms, Cough Symptoms, Sore Throat Symptoms, Discomfort when Breathing, dyspnoe that limited eating and driking, consuming specific drugs in the past 14 days, pregnant and comorbid. While the attributes of the Province of Residence, City/Regency, District, and Nearest Health Center are not very meaningful in the data mining process. The attributes of Name, Gender, Mobile Number, Address, Email, NIK and other personal data are also not meaningful in the data mining process and are confidential.

b. Pre Processing Data

In data collection, exploration was carried out to facilitate understanding of the data. Exploration is used by looking at all existing attributes as well as anomaly data. Exploratory Data Analysis (EDA) is the initial stage of most data analysis processes and becomes a stage that will determine how well the next data analysis will be produced. The EDA component includes pre-processing, including the process of handling missing values, outliers, reduction, and data transformation. The number of patients was 526 subjects with 21 atributes. However, not all attributes may bused in the analysis. Therefore, we applied pre-processing data to extract the best available data, using the following procedure:

1) Data cleaning
   The process deleted incomplete data that contains blank cells, i.e. there was no birth date, and also other questions. There was 5,48% incomplete data to avoid missing values, resulted in 500 qualified data and may be further analysed.
2) Dimensionality reduction
   The reduction process used to get informative data. The study omited name and gender because it is irrelevant to answer the research question. Another atribute of symptoms, i.e. fever, cough, sore throat, runny nose, hard breathing, dyspnoe that limited eating and driking, and other condisition i.e. consuming specific drugs in the past 14 days, were all classified as one attribute since it had the same information.

c. Data Transformation
   The process changed birth date attribute into age, which ranged from more than 18 to less than sixty. The categorization was further applied to make the analysis easier.

   The pre-processing data ended up at 500 available sujects without any missing value, and 6 attributes were used of the next step. Table 1 shows the number of patients who were categorized as recommended and needed further consultation.

TABLE 1. Sample Dataset

Recommended	Needed further examination	Total Sample
307	193	500

d. Data Mining Process
   This study used data mining process on a dataset to compare two classification algorithms, which were C4.5 and Naïve Bayes. The data analysis was taken from a set of patients who visited a private clinic, namely Cimanggis Jaya, which result was to recommend vaccination or need to perform further examination. Using the two methods, C4.5 dan Naïve Bayes, we could decide the highest accuracy.
e. Evaluation
   The implmenetation of the two methods was followed by validation of the results using K-Fold Cross Validation method. The result was presented as confusion matrix and the Receiver Operating Characteristics (ROC) curve to measure accuracy of the two methods. Evaluation results by the confusion matrix ended up wih accuracy values, recall, and precision.

RESULTS AND DISCUSSION

The test was done using the tool RapidMiner 9.8 to implement the algorithm of C4.5 dan Naïve Bayes that produced K-Fold Cross Validation, Accuracy, Confusion Matrix dan ROC.

1. Experiment Result of The C4.5 Algorithm

FIGURE 1. Decision Tree of Classification Algorithm C4.5

The decision tree as result from Figure 1 produced a rule, The rule could be implemented to make decision on a new data. Figure 1 showed that out of all used variables, the variable 'Gejala Lain' or 'other symptom' became the root to predict recommendation for a patient to receive COVID-19 vaccine at Klinik Cimanggis Jaya because the highest gained value was found on this variable. Beside the variable of 'Gejala Lain', Algorithm C4.5 decision tree showed that 'Komorbid' (comorbidit) led to the decision for further examination. If the answer was no, then variable of 'Kontak Erat' (close contact), followed by 'Hasil Rapid Antigen' (result of antigent), 'Usia' (age), and 'Hamil' (pregnancy) were the prominent variables to decide whether a vaccine was recommended to the patient.

The implementation of C4.5 Algorithm using RapidMiner Tool obtained accuracy value of 99.20% as given in the Confusion Matrix in a following table below:

TABLE 2. Accuracy Values of the Classification Algorithm C4.5

	True Outcome		**Class precision**
Prediction	Needed further examination	Recommended	
Needed further examination	189	0	100%
Recommended	4	307	98,71%
Class recall	97,93%	100%	

*accuracy: 99.20% ± 1.03% (micro average 99,20%).

Result of the calculation was visualized in the following ROC curve for the C4.5 Algorithm that expressed the Confusion Matrix in Table 2. The horizontal line means the false positive and the vertical line was true positive.

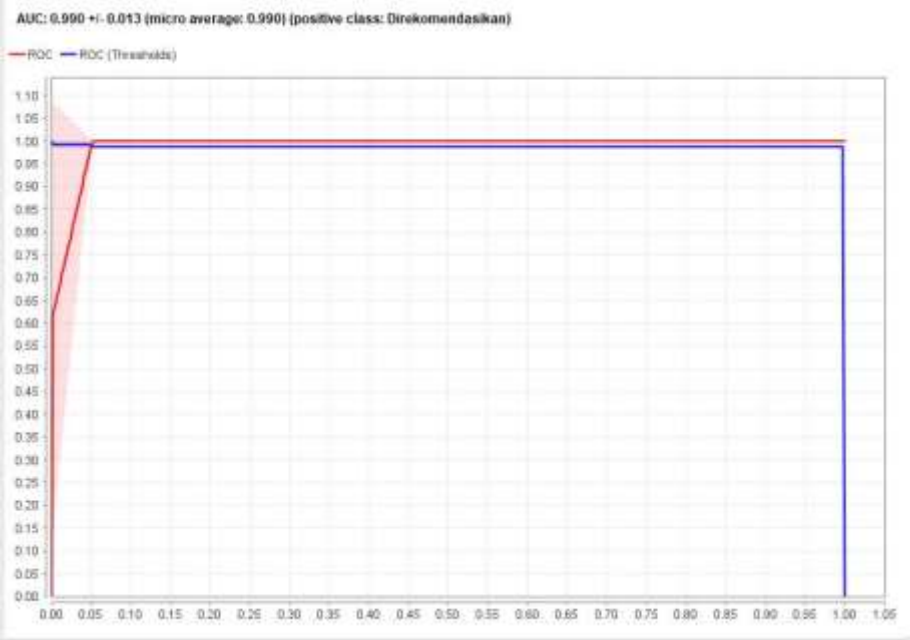

FIGURE 2. The Area Under the Curve (AUC = 0.990) of the C4.5 Algorithm

Figure 2 was the ROC of C4.5 Algorithm which had an AUC (Area Under Curve) that classified as Excellent classification since the accuracy performance of AUC reached 0.990 was between the range of 0.900-1.000. The range of performance would be categorized as Excellent Classification.

2. Experiment Result of The Naïve Bayes Algorithm

The implementation of Naïve Bayes Algorithm using RapidMiner Tool obtained accuracy value of 98.20% as evaluated by the Confusion Matrix in the following table below (Table 3):

TABLE 3. Accuracy Values of the Classification Algorithm Naïve Bayes

	True Outcome		**Class precision**
Prediction	Needed further examination	Recommended	
Needed further examination	191	7	96,46%
Recommended	2	300	99,34%
Class recall	98,96%	97,72%	

*accuracy: 98.20% ± 2.20% (micro average 98,20%).

The calculation result was visualized in the ROC curve for Naïve Bayes Algorithm in Figure 3 below, which expressed the Confusion Matrix for Table 3. The horizontal line expressed false positive and the vertical line was true positive.

FIGURE 3. The Area Under the Curve (AUC = 0.998) of the Naïve Bayes Algorithm

Figure 3 gives the ROC graph with its AUC value of 0.998, which means an Excellent Classification because it performed the accuracy of AUC 0.998 between 0.900-1.000.

3. Comparison of the Accuracies between the C4.5 and Naïve Bayes Algorithm

TABLE 4. Comparison of Accuracy Between Naïve Bayes dan C4.5 Algorithm

No	Algorithm	Accuracy
1	C4.5	99.20%
2	Naïve Bayes	98.20%

As shown in the Table 4, prediction to recommend vaccination among those who came to Klinik Cimanggis Jaya was higher when using the Algorithm C4.5 than the Algorithm Naïve Bayes (99.20%) with approximately 1% difference.

CONCLUSIONS

This study showed the test to compare two methods of data mining, which were the Algorithm C4.5 and the Naïve Bayes using a set of data from Klinik Cimanggis Jaya. Patients who were recruited came for testing antigen of COVID 19, which reached 500 data. As tested in RapidMiner Tool, we produced accuracy value on each algorithm. The result was evaluated and validated, where the Algorithm of C4.5 obtained higher accuracy of 99.20%, compared to the Algorithm Naïve Bayes (accuracy of 98.20%). Nevertheless, both algorithms of the data mining presented an AUC diagnosis of excellent classification. This means, the two methods were excellent, but the use of Algorithm C4.5 would be a better method to predict recommendation to receive vaccine and to decide who should undergo further examination.

In developing this research, authors recommended several suggestions: 1) it is important to get more data to be more precise; 2) to use the selection feature for the next study in this problem or similar studies when applies. 3) Similar study might be done to implement different mining methods. 4) the result development should adopt the prediction result as supporting data for decision process of the stakeholders. For patients who have other symptoms, further research needs to be done, if they do not have criteria that must be checked, a vaccine is recommended.

REFERENCES

1 2021, COVID-19 CORONAVIRUS PANDEMIC, *American Library Association*. [Online]. Available: https://www.worldometers.info/coronavirus. [Accessed: 01-Mar-2021].

2 2021, KOMITE PENANGANAN COVID-19 DAN PEMULIHAN EKONOMI NASIONAL, *Satuan Tugas Penanganan COVID-19*. [Online]. Available: https://covid19.go.id/. [Accessed: 01-Mar-2021].

3 2021, Badan Pusat Statistik, *BPS - Statistics Indonesia*. [Online]. Available: https://www.bps.go.id/news/2021/01/21/405/bps--270-20-juta-penduduk-indonesia-hasil-sp2020.html. [Accessed: 01-Mar-2021].

4 Pakasi T A, 2020 The Need of Trusted Primary Care: Lesson Learnt from The COVID 19 Outbreak in Indonesia *Rev. Prim. Care Pract. Educ. (Kajian Prakt. dan Pendidik. Layanan Prim.* **3**, 2 p. 3.

5 Hadiwardoyo W, 2020 Kerugian Ekonomi Nasional Akibat Pandemi Covid-19 *Baskara J. Bus. Entrep.* **2**, 2 p. 83–92.

6 Pemerintah P, 2017 *PERATURAN MENTERI KESEHATAN REPUBLIK INDONESIA NOMOR 12 TAHUN 2017 TENTANG PENYELENGGARAAN IMUNISASI* .

7 Fitriani E, 2020 PERBANDINGAN ALGORITMA C4.5 DAN NAÏVE BAYES UNTUK MENENTUKAN KELAYAKAN PENERIMA BANTUAN PROGRAM KELUARGA HARAPAN *Sist. J. Sist. Inf.* **9**, 1 p. 103–115.

8 Rosandy T, 2016 PERBANDINGAN METODE NAIVE BAYES CLASSIFIER DENGAN METODE DECISION TREE (C4 . 5) UNTUK MENGANALISA KELANCARAN PEMBIAYAAN (Study Kasus : KSPPS / BMT AL-FADHILA) **02**, 01 p. 52–62.

9 Syarifuddin F Misdram M Widodo A A Informatika P S and Pasuruan U M, 2020 KLASIFIKASI DATA SET VIRUS CORONA MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER *J. SPIRIT* **12**, 2 p. 46–52.

10 Eko P, 2012 *Data mining : konsep dan aplikasi menggunakan MATLAB* 1st ed. Yogyakarta: CV Andi Offset.

11 Aulianita R Utami L A Musyaffa N Wijaya G Mukhayaroh A and Yoraeni A, 2020 Sentiment Analysis Review of Smartphones with Artificial Intelligent Camera Technology Using Naive Bayes and n-gram Character Selection *J. Phys. Conf. Ser.* **1641**, 1 p. 1–6.

12 Wulandari R T, 2017 *Data Mining Teori dan Aplikasi Rapidminer* Yogyakarta: Gava Media.

13 Kusrini and Luthfi E T, 2009 *Algoritma Data Mining* Yogyakarta: Andi, Yogyakarta.