A Comparison of Naive Bayes Methods, Logistic Regression and KNN for Predicting Healing of Covid-19 Patients in Indonesia

2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT) | 978-1-6654-0514-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/EIConCIT50028.2021.9431845

1st Manzilur Rahman Romadhon
Master of Informatics Engineering
State Islamic University of Maulana Malik Ibrahim
Malang, Indonesia
19841010@student.uin-malang.ac.id

2nd Fachrul Kurniawan
Master of Informatics Engineering
State Islamic University of Maulana Malik Ibrahim
Malang, Indonesia
fachrulk@ti.uin-malang.ac.id

***Abstract*—Since it was declared a global pandemic by the World Health Organization (WHO), the number of cases of Covid-19 patients who died has continued to increase. One of the countries with the highest death rate in the world is Indonesia. On Saturday, April 4, 2020, Indonesia reached the highest death rate for Covid-19 patients, around 9.11%. This number must be suppressed so that there are no more victims. For this reason, it is necessary to know actually the factors that can reduce the risk of death and predict the chance of curing Covid-19 patients. In data mining, there are several methods that can be used to predict a patient's recovery rate by considering several variables. The variables used in this study were age and gender. Naive Bayes Method, logistic regression, and K-Nearest Neighbor (KNN) are the methods to be chosen in this study to analyze their most accurate performance. The result shows that KNN has the highest accuracy, which is 0.750 compared to logistic regression which has a value of 0.703 as well as Naive Bayes which has the same value. Meanwhile, the level of precision of the three models shows that KNN also has the highest value, namely 0.750 than logistic regression and Naive Bayes which have the same value, namely 0.700. The recall value of the three also shows that the kNN remains the highest with 0.750 compared to the two comparison models which have the same value, namely 0.708.**

Keywords*—*KNN, Naive Bayes, logistic regression, covid-19.

I. Preliminary

The covid-19 virus was first declared a global pandemic by WHO on March 12, 2020 by WHO General Director Dr Tedros Adhanom Ghebreyesus recalled the alarming extent and severity of the virus[1]. After that the virus patients continued to increase in every country. One of the countries affected is Indonesia. Indonesia is one of the countries in the world that was ranked the highest in the percentage of deaths of Covid-19 patients, namely on Saturday, April 4, 2020 with a percentage of 9.11%[2]. However, this number continues to fluctuate.

The high percentage of covid-19 patient deaths in Indonesia is certainly worrying. For this reason, the government and society must work together to reduce or at least reduce the percentage of deaths. This should also be of concern to academic activists to conduct research related to the Covid-19 virus so that factors can be analyzed that can reduce the high mortality rate of Covid-19 patients in Indonesia. President Jokowi has also expressed high support for research related to the handling of covid-19 [3]. Currently on the internet there is a lot of data about Covid-19 that can be accessed both in tables and graphics. Several government and private institutions have published data such as data from the task force to accelerate the handling of covid-19 [3], from Institute of Electrical and Electronics Engineers (IEEE) Xplore [4], and from kawalcovid19 [3].

From these data, a pattern or knowledge can be found through data mining techniques. Data Mining is a process that applies mathematical, statistical, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge based on large databases[5]. Artificial intelligence is also one of the elements in the Internet of Things (IoT) technology. Technology Internet of Things (IoT) is a phenomenon in which an entity has the ability to transmit data over a network without the need for human or human intervention to a device where data produced by the instruments (sensors) connected to the chosen object is generated[6]. One of the analysis methods found in data mining is classification. Classification is a technique of grouping or categorizing data, where the data has a class or label. The function of the classification method can be used to predict trends in the future data.

There are three classification methods among several methods that are often used in data mining. Namely Naive Bayes, K-Nearest Neighbour and Logistic Regression. Several researchers have conducted research related to these three methods. Among them was done by Nova Tri Romadloni [7] et al who compared Naive Bayes, KNN and Decision tree to analyse KRL transport sentiment, Daniela Xhemali et al [8] who examined the performance of Naive Bayes for the classification of training web pages, and Van Der Heide et al [9] who compared logistic regression, Naive Bayes and random forest methods to predict animal survival. From these studies, no research has been found that analyses the performance of the Naive Bayes method, KNN and logistic regression to predict patients with high prevalence of Covid-19. For this reason, this study will conduct an analysis of the three performance of these methods to predict the cure rate for Covid-19 patients in Indonesia. The results of this study are expected to contribute in the health sector to find the factors that affect the patient's recovery rate and actions that can be taken in the future.

II. Base theory

A. Naive Bayes

Naive Bayes is a classification algorithm for calculating probability by calculating the frequency and combination of values in a data [10]. The Naive Bayes algorithm was put forward by the British scientist Thomas Bayes. Naive Bayes will predict future opportunities based on previous experience so that it is known as the Bayes Theorem. The main characteristic of this Naive Bayes Classifier is the result of a very strong (Naive) prediction of independence from each condition or event. A conditional probability is an occurrence probability calculation, A, when another event, B, has occurred, which is noted as P(A|B), which incorporates both A and B probabilities. This theory is used to measure the likelihood of a data set joining a particular category on the basis of the inferential data available [11].

The advantage of using Naive Bayes is that this algorithm only requires a small amount of training data (training data) to select the estimated parameters needed in the classification process. Because it is assumed to be an independent variable, only the variance based on a variable in a class is needed to determine the classification, not the entire covariance matrix. The steps in the Naive Bayes process are counting the number of classes or labels, counting the number of possibilities per class, multiplying all class variables, and comparing the products per class [12]. Bayes' theorem finds the probability of events occurring based on other events that have occurred. Bayes' theorem is expressed mathematically as the following equation:

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)} \tag{1}$$

Where:

P (A | B)= is the probability that A will occur when event B occurs.

P (A) = probability of occurrence of event A.

P (B | A)= the probability that event B occurs when event A occurs

P (B) = probability of occurrence of event B [13].

B. Logistic regression

Logistic regression is a type of regression that connects one or more independent variables (independent variables) with the dependent variable of the type of category; can be 0 and 1, true or false, big or small. The type of independent variable in the form of this category is what distinguishes logistic regression from multiple regression or other linear regression. The logistic regression equation is stated as follows:

$$Ln\left(\frac{p}{1-p}\right) = B_0 + B_1X \tag{2}$$

Where:

B0 = constant

B1 = the coefficient of each variable

The p value or probability (Y = 1) can be found by the equation:

$$p = \frac{e^{(B0+B1X)}}{(1+e^{(B0+B1X)})} \tag{3}$$

This equation can be used to calculate the probability of a respondent having a variable value that has been defined in the equation, the final result of the p value will of course be in the range of 0-1 [14].

C. K-Nearest Neighbour

K-Nearest Neighbour (kNN) is an algorithm in machine learning that includes supervised learning. Supervised learning aims to find new patterns in data by connecting existing data patterns with new data[15]. KNN is used to classify an object, based on the k training data that is closest to the object. The condition for the value of k is that it cannot be greater than the amount of training data, and the value of k must be odd and more than one[16]. The purpose of the kNN algorithm is to classify new objects based on attributes and training samples. Where the results of the new test samples are classified based on most of the categories on the KNN. The steps in the KNN algorithm are as follows:

- determine the K parameter (number of closest neighbours)
- calculate the square of the object's Euclidean distance to the given training data
- sort the results of point b in ascending order
- collect category Y (nearest neighbour classification based on k value)
- by using the most majority nearest neighbour category, the object category can be predicted.

One of the advantages of the KNN algorithm is that it is very nonlinear because it is a nonparametric machine learning model. The advantage of nonparametric models is that the resulting class decision lines can be very flexible and nonlinear. In addition, KNN is easy to understand and implement. However, the weakness of this method is that it needs to determine the K parameter first, and also the high computation value because it has to calculate the distance between x and all other instances in the dataset. So that it will slow down the computer process.[17]

D. Model evaluation

The algorithm model used in this study will be evaluated to determine the best algorithm that can be used to predict the recovery of Covid-19 patients in Indonesia. The best model is chosen by considering the accuracy, precision and recall values of each algorithm. Accuracy is the level of closeness between the predicted value and the actual value. Precision is the level of accuracy found between the requested information and the response given by the system. Meanwhile, recall is the level of success of the system in retrieving information. The calculation of the value of accuracy, precision and recall is stated in the following formula:[18]

$$Akurasi = \frac{TP+TN}{TP+FN+FP+TN} \tag{4}$$

$$Presisi = \frac{TP}{TP+ FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

Where:

TP : True positive

TN: True negative

FP : False positive

FN : False negative

E. *Confusion matrix*

The configuration matrix is a table consisting of the number of rows of test data that are predicted to be true and false by the classification model used. A confusion matrix table is needed to select the best performance from a classification model [19]. This method is often used with multiple classifiers or more than two classes. The confusion matrix is quite suitable for use in research that measures how accurate the classification results are based on the model that has been made. The following is a configuration matrix Table I [16].

III. Results and Discussion

A. *Dataset*

Dataset is a collection of data. In tabular data problems, a data set corresponds to one or more database tables, where each table column represents a specific variable, and each row corresponds to a specific record of the data set in question[20]. The dataset used in this study is the dataset of covid-19 patients sourced from the kawalcovid19.com site and the kawalcorona.com site. The dataset to be used is pre-processing first by eliminating unnecessary variables. Table II shows the identity of the dataset of Covid-19 patients before pre-processing. Then in Table III the dataset is displayed after pre-processing and ready for analysis.

Fig 1 shows a comparative simulation of the Naive Bayes algorithm, logistic regression and KNN to determine the best algorithm used to predict the recovery of Covid-19 patients who have tested positive. The tool used to simulate the algorithm model is Orange version 3.25.0. Orange is a tool in data mining techniques such as RapidMiner which is commonly used for modelling data mining algorithms. The simulation flow starts with patient training data sent to the test and score widget, then in the test and score widget, the learning process is carried out using the Naive Bayes algorithm, logistic regression and KNN as shown in the picture. The logistic regression model used lasso regularization with a strength of C=1. While the metric used in KNN is Euclidean with neighbour 5.

TABLE I. CONFIGURATION MATRIX

		Predicted	
		Positive	**Negative**
Actual	**Positive**	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

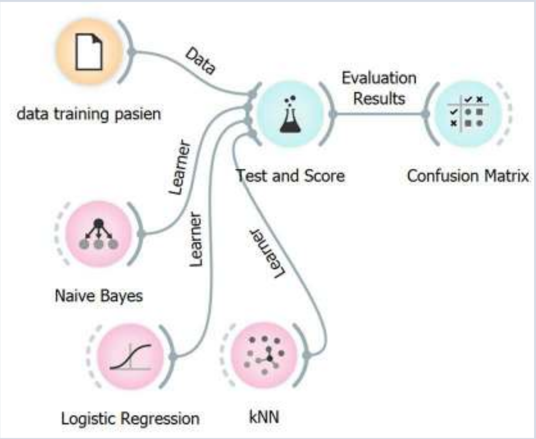


Fig. 1. simulation of the learning process for the Naive Bayes classification model, logistic regression and KNN

TABLE II. IDENTITIES OF THE DATASET OF COVID-19 PATIENTS IN INDONESIA

No.	Attribute	Type
1	Gender	Category
2	Age	Numeric
3	WN	String
4	Province	String
5	Status	Category
6	Source of contact	Numeric
7	Age group	Numeric
8	MD (died)	Numeric
9	S (Cured)	Numeric
10	DP (Under care	Numeric

TABLE III. IDENTIFIES THE DATASET OF COVID-19 PATIENTS AFTER PRE-PROCESSING

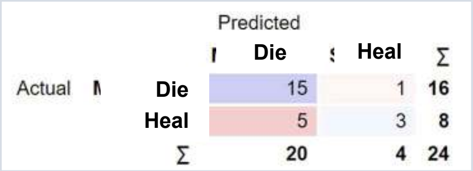
No.	Attribute	Type
1	Gender	Category
2	Age	Numeric
3	Province	Category
4	Status	Category

TABLE IV. RESULTS OF THE EVALUATION OF THE KNN MODEL, LOGISTIC REGRESSION AND NAIVE BAYES

Model	AUC	CA	F1	Precision	Recall
KNN	0.801	0750	0.722	0750	0750
Logistic regression	0.785	0.708	0.703	0.700	0.708
Naive Bayes	0.805	0.708	0.703	0.700	0.708

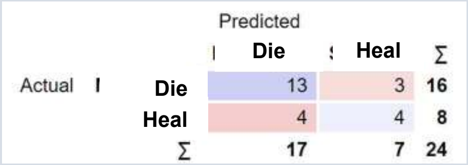
Table IV shows the results of the tests and scores carried out for the testing process for the Naive Bayes model, logistic regression and KNN. The test was carried out with the sample type 3-fold validation. Tests may have different results by changing the value of the k-fold validation. The results of the comparison of the three models show the value of Classification Accuracy (CA), Area under ROC Curve (AUC), precision and recall. The result shows that KNN has the highest accuracy, which is 0.750 compared to logistic regression which has a value of 0.703 as well as Naive Bayes which has the same value. Meanwhile, the level of precision of the three models shows that KNN also has the highest value, namely 0.750 than logistic regression and Naive Bayes which have the same value, namely 0.700.

In Fig 2, 3 and 4 show the confusion matrix table of the KNN, logistic regression and Naive Bayes models. The KNN model has the highest level of accuracy, precision and recall as evidenced in Table IV because KNN has classified and predicted the dataset of Covid-19 patients with the closest level of accuracy. It is proven by the results of the confusion matrix in Fig 2 which shows that the true positive (TP) and true negative (TN) values, namely 15 and 3, are the highest compared to the TP and TN values in logistic regression and Naive Bayes. So most predictions made by KNN match between reality and predictions. Whereas for false negative (FN) and false positive (FP), the kNN model produces the lowest value, namely 6. While the FN and FP for the other two models have the same value, namely 7. The higher the TP and TN values, the more accurate they are. However, the higher the FP and FN values, the more accurate they are, but the less accurate they are.



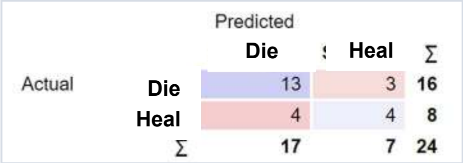
		Predicted		
		Die	Heal	Σ
Actual	Die	15	1	16
	Heal	5	3	8
	Σ	20	4	24

Fig. 2. Confusion matrix KNN



		Predicted		
		Die	Heal	Σ
Actual	Die	13	3	16
	Heal	4	4	8
	Σ	17	7	24

Fig. 3. Confusion matrix of Naive Bayes



		Predicted		
		Die	Heal	Σ
Actual	Die	13	3	16
	Heal	4	4	8
	Σ	17	7	24

Fig. 4. onfusion matrix logistic regression

CONCLUSION

Based on the comparison of the three classification algorithm models, namely Naive Bayes, logistic regression and KNN, it can be concluded that these three models can be used to predict the cure rate for Covid-19 patients in Indonesia. However, the best model of the three is k-Nearest Neighbour (kNN) because it has the highest accuracy, 0.750, compared to logistic regression and Naive Bayes which have the same score of 0.703. The variables that affect the cure rate for Covid-19 patients include variables of age, sex, and patient province. This research is actually a research involving very simple variables. Of course, it can still be developed further by adding patient data variables which are used as training data. Such as comorbidities, diet, travel history and other variables. This of course can further improve the accuracy of the modelling results. The limited data on covid-19 patients in Indonesia is one of the simple problems of the supporting variables in this study.

REFERENCE

[1] A. A. M. D. Silva, "On the Possibility of Interrupting the Coronavirus (COVID-19) Epidemic Based on The Best Available Scientific Evidence." SciELO Public Health, 2020.

[2] M. D. Rahiem, "The Emergency Remote Learning Experience of University Students in Indonesia Amidst The COVID-19 Crisis," *Int. J. Learn. Teach. Educ. Res.*, vol. 19, no. 6, pp. 1–26, 2020.

[3] R. Djalante *et al.*, "Review and Analysis of Current Responses To COVID-19 in Indonesia: Period of January to March 2020," *kawal covid-19*. p. Progress in Disaster Science 6, 2020.

[4] N. Ramadijanti and A. Basuki, "Comparison of Covid-19 Cases in Indonesia and Other Countries for Prediction Models in Indonesia Using Optimization in SEIR Epidemic Models," in *2020 International Conference on ICT for Smart Society (ICISS)*, 2020, pp. 1–6, doi: 10.1109/ICISS50791.2020.9307543.

[5] E. Turban, J. E. Aronson, and T. P. Liang, *Decision Support Systems and Intelligent System,(Sistem Pendukung Keputusan dan Sistem Cerdas)*, Ed. 7. Jld. Yogyakarta: Andi Offset, 2005.

[6] F. Kurniawan, H. Nurhayati, Y. M. Arif, S. Harini, S. M. S. Nugroho, and M. Hariadi, "Smart Monitoring Agriculture Based on Internet of Things," in *2018 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, 2018, pp. 363–366, doi: 10.1109/EIConCIT.2018.8878510.

[7] N. T. Romadloni, I. Santoso, and S. Budilaksono, "Comparison of Naive Bayes, KNN And Decision Tree Methods To Sentiment Analysis Of Commuter Line KRL Transport," *J. IKRA-ITH Inform.*, vol. 3, no. 2, pp. 1–9, 2019.

[8] D. Xhemali, C. J-Hinde, and R. G-Stone, "Naive Bayes Vs. Decision Trees Vs. Neural Networks in The Classification of Training Web Pages," *Int. J. Comput. Sci.*, vol. 4, no. 1, pp. 16–23, 2009.

[9] E. M. M. Van der Heide, R. F. Veerkamp, M. L. Van Pelt, C. Kamphuis, I. Athanasiadis, and B. J. Ducro, "Comparing Regression, Naive Bayes, And Random Forest Methods in The Prediction Of Individual Survival To Second Lactation In Holstein Cattle," *J. Dairy Sci.*, vol. 102, no. 10, pp. 9409–9421, 2019, doi: 10.3168/jds.2019-16295.

[10] [G. Dimitoglou, J. A. Adams, and C. M. Jim, "Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability," *arXiv preprint*. p. 1206.1121, 2012.

[11] S. Suhartono, "Identification of Virtual Plants Using Bayesian Networks Based on Parametric L-System," *Int. J. Adv. Intell. Informatics*, vol. 4, no. 1, pp. 40–52, Mar. 2018.

[12] W. D. Septiani, "Komparasi Metode Klasifikasi Data Mining Algoritma C4. 5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis," *J. Pilar Nusa Mandiri*, vol. 13, no. 1, pp. 76–84, 2017.

[13] M. M. Saritas and A. Yasar, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *Int. J. Intell. Syst. Appl. Eng.*, vol. 7, no. 2, pp. 88–91, 2019.

[14] A. F. Hayes and J. Matthes, "Computational Procedures for Probing Interactions in OLS and logistic regression: SPSS and SAS Implementations," *Agung Budi Santoso*, vol. 41, no. 3, pp. 924–936, 2009.

[15] I. Gazalba and N. G. I. Reza, "Comparative Analysis Of K-Nearest Neighbour and Modified K-Nearest Neighbour Algorithm for Data Classification," in *In 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2017, pp. 294–298, doi: 10.1109/-ICITISEE.2017.8285514.

[16] M. Rivki and A. M. Bachtiar, "Implementation of K-Nearest Neighbor Algorithm in Classification of Follower Twitter Using Indonesian Language," *J. Sist. Inf. (Journal Inf. Syst.*, vol. 13, no. 1, pp. 31–37, 2017, doi: 10.1017/CBO9781107415324.004.

[17] A. G. Novianti and D. Prasetyo, "Penerapan Algoritma K-Nearest Neighbor (K-NN) untuk Prediksi Waktu Kelulusan Mahasiswa," in *In Seminar Nasional APTIKOM (SEMNASTIKOM)*, 2017, pp. 108–113.

[18] K. Tampubolon, H. Saragih, B. Reza, K. Epicentrum, and A. Asosiasi, "Implementasi Data Mining Algoritma Apriori Pada Sistem Persediaan Alat-Alat Kesehatan," *Maj. Ilm. Inf. dan Teknol. Ilm.*, vol. 1, no. 1, pp. 93–106, 2013.

[19] C. O. Freitas, J. M. De Carvalho, J. Oliveira, S. B. Aires, and R. Sabourin, "Confussion Matrix Disagreement for Multiple Classifiers," *In Iberoamerican Congress on Pattern Recognition*. pp. 387–396, 2007.

[20] H. Gjoreski *et al.*, "The University of Sussex-Huawei Locomotion and Transportation Dataset for Multimodal Analytics With Mobile Devices," in *IEEE Access*, 2018, vol. 6, pp. 42592-42604., doi: 10.-1109/ACCESS.2018.2858933.