SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10150
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparison of C4.5 Algorithm and Naïve Bayes for Last Information on ICU Patients

Sumpena
STMIK Nusa Mandiri
Jakarta, Indonesia
sumpenasuhandi@gmail.com

Yuma Akbar
STMIK Nusa Mandiri
Jakarta, Indonesia
Yuma.pjj@gmail.com

Nirat
STMIK Nusa Mandiri
Jakarta, Indonesia
mrnirat@gmail.com

Mario Henky
STMIK Nusa Mandiri
Jakarta, Indonesia
mhenky@gmail.com

Abstract— Intensive Care Unit (ICU) is part of an independent hospital that is structurally under Anesthesia and Reanimation Installation. Critical patients need care and supervision by the medical team at the Intensive Care Unit (ICU) in ICU patient care using tools such as ventilators, monitors and Central Venous Pressure (CVP). With the complexity of the problem in the Intensive Care Room which consists of various types of components and medical device equipment to support care activities for critical patients. To find out information on ICU patients, various studies need to be conducted in accordance with the needs of decision makers. The information to be presented is information to determine the prediction of patients treated in ICU by utilizing the C4.5 algorithm and Naïve Bayes algorithm and processed with the Rapid Miner application. From the results of this study by analyzing patient data using a ventilator, Central Venous Pressure (CVP) and also a diagnosis of sepsis then the data is processed and classified. ICU patient accuracy results obtained, namely C4.5 algorithm has an accuracy of 81.25% and AUC 0.623, while Naïve Bayes has an accuracy of 80.66% and AUC 0.795. From these results, the C4.5 algorithm has a better edge of 0.59% than the Naive Bayes algorithm.

Keywords—Intensive, Care, C4.5, Naïve Bayes, Algorithm.

I. Introduction

Intensive Care Unit (ICU) is part of an independent hospital that is structurally under Anesthesia and Reanimation Installation, which consists of staff who have special skills and are equipped with special equipment intended for observation, treatment and therapy of patients suffering from the disease, injury or complication – Complications that are life threatening or potentially life threatening. The ICU provides capabilities and facilities, infrastructure and special equipment to support vital functions using the skills of medical staff, nurses and other staff who are experienced in managing these conditions (Kepmenkes, 2010).

To find out information about ICU patient care requires patient data which includes the use of medical devices such as Ventilators, Monitors, CVP, medical supplies and also the diagnosis of special patients with sepsis. This data will be used to determine the final prediction of patients treated in the ICU.

Diagnosis of Sepsis according to the latest consensus is a state of life-threatening organ dysfunction caused by disruption of the body's response to infection. The use of SIRS criteria to identify sepsis is considered not helpful anymore (Irvan, Febyan, 2018).

For most patients, mechanical ventilation is a short-term therapy used to support oxygen exchange until the cause of respiratory failure is resolved (Sahetya, Allgood, Gay, & Lechtzin, 2016).

Central Venous Pressure (CVP) is an invasive hemodynamic monitoring method. Central Venous Pressure (CVP) is often used in intensive care rooms especially in patients who experience impaired fluid balance, heart failure, evaluation of therapeutic response and media for giving therapy or fluids hypertonic(Lesmana, 2018).

With the complexity of the problems in the Intensive Care Room which consists of various kinds of components and equipment of medical devices to support the activities of care for critical patients, to find out information in the intensive care implementation, it is necessary to conduct various studies in accordance with the needs of decision makers.

To meet the need for information about ICU care patients, this study tried to conduct research on ICU patient care, which in this study only discussed the problem of indications of care discharged from patients treated in the ICU (C.W. Cheng, N. Chanani, J. Venugopalan, K. Maher, 2013)**.**

The information that will be presented is information to determine the prediction of patients treated in ICU by utilizing the C4.5 algorithm method and Naïve Bayes algorithm so that it can be known the comparison of the two algorithms in determining decision making (Fridayanthie, 2015).

In using data available in the ICU in real-time, Artificial Intelligence Experts can accurately predict the onset of sepsis in ICU patients 4-12 hours before clinical recognition. A prospective study is needed to determine the clinical utility of the proposed sepsis prediction model. (Nemati et al., 2017).

According to research on diabetes disease (Fatmawati, 2016) with the title Comparison of C4.5 and Naïve Bayes classification of data mining algorithms for the prediction of diabetes with 768 data has resulted in comparison results, namely the Naïve Bayes algorithm model with higher accuracy of 1.83% than the C4.5 algorithm.

II. Literature Review

A. Data Mining

Data mining is the *process* of discovering interesting patterns and knowledge from *large* amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically (Han, Kamber, & Pei, 2012)**.**

Based on their duties, data mining is grouped into 6 namely is description, estimation, prediction, classification, clustering, and association. Classification (taxonomy) is the process of placing certain objects (concepts) in a set of categories, based on each object (concept) property. The classification process is based on four fundamental components, namely class, predictor, training set, and dataset testing (Septiani, Studi, & Informatika, 2017).

The successful application of data mining in highly visible fields like e-business, marketing and retail have led to the popularity of its use in knowledge discovery in databases (KDD) in other industries and sectors. Among sectors that are just discovering data mining are the fields of medicine and public health. Applying data mining in the medical field is a very challenging process due to the idiosyncrasies of the medical profession.

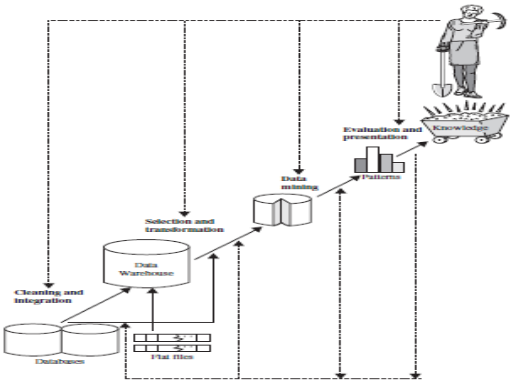


Figure 1. Steps in KDD Process
Source : Han & Kember

Data mining techniques help extract valuable information from health service data. Patient status prediction is a very sensitive issue to avoid the worst complications that can cause patient damage, especially due to chronic illness. Therefore, health professionals need a predictive model that is simple, general, interpretable, and can be trusted to make decisions faster for patients (Alotaibi & Sasi, 2015).

B. Grouping of Data Mining Techniques.

Data Mining is divided into groups based on tasks that can be carried out (Fridayanthie, 2015)**.**

1. Classification

Classification is a technique can classify a new data by manipulating existing data that has been classified and by using the results to provide a number of rules. One easy and popular example is Decision tree, which is one of the most popular classification methods because it is easy to interpret. Decision tree is a predictor model using tree structures or hierarchical structures.

2. Association

Used to recognize the behavior of a specific event or process where an association relationship arises at each event. One example is Market Basket Analysis, which is one of the association methods that analyze the possibility of customers to buy several items simultaneously.

3. Clustering

Used to analyze different data groupings, similar to classification, but grouping has not been defined before the data mining tool is run. Usually use neural network methods or statistics. Clustering divides items into groups based on the data mining tools found.

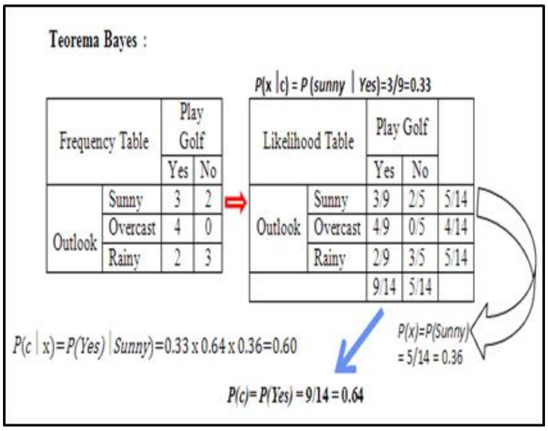
C. Classification Algorithm

Among the most popular classification models : Decision / Classification Trees, Bayesians Classifiers / Naïve Bayes Classifiers, Neural Networks, Statistical Analysis, Genetic Algorithms,
Rough Sets, K-Nearest Neighbor Classifier, Memory Based Reasoning, Support Vector Machines (Vercellis, 2011).

D. Algorithm Decision Tree (C4.5)

The Decision Tree resembles a flowchart structure, each internal node is declared as a test attribute, each branch represents the output of the test, and each node (terminal node) determines the class label. The top node of a tree is node. The decision tree algorithm in the decision tree formation can be a decision tree C4.5. (Han et al., 2012).

Gain (S, A) is the acquisition of information from attribute A relative to the data output of S. The acquisition of information is obtained from the output data or the dependent variable S which is grouped by attribute A, denoted by the gain (S, A).


Teorema Bayes :
P(x | c) = P (sunny | Yes)=3/9=0.33
P(c | x)=P(Yes) | Sunny)=0.33 x 0.64 x 0.36=0.60
P(x)=P(Sunny)
= 5/14 = 0.36
P(c)= P(Yes) = 9/14 = 0.64

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

Likelihood Table		Play Golf		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

Figure 2. Description of Bayes Theorem
Source : https://informatikalogi.com/algoritma-naive-bayes

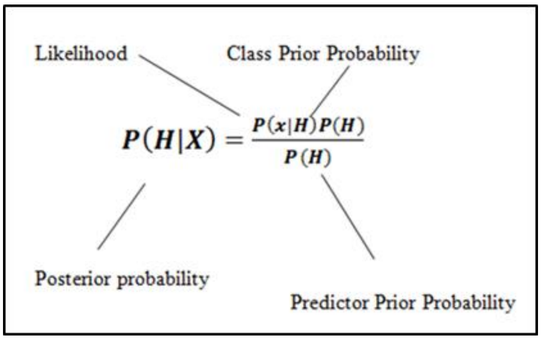

Likelihood
Class Prior Probability
$P(H|X) = \frac{P(x|H)P(H)}{P(H)}$
Posterior probability
Predictor Prior Probability

Figure 3. Theory of Bayes Theorem
Source : https://informatikalogi.com/algoritma-Naïve Bayes

Information:

X : Data with unknown classes
H : The data hypothesis is a class Specific
P (H|X) : Hypothesis probability based on Conditions (Posteriori probability)
P (H) : Hypothesis probability (prior probability)
P (X|H) : Probability based on conditions at Hypothesis
P (X) : Probability

E. Rapid Miner

Rapid Miner is a system which supports the design and documentation of an overall data mining process. It's not only an almost comprehensive set of operators, but also structures that express the control of the process (Crc & Hofmann, 2014).

Rapid Miner is an open source Rapid Miner is a solution for analyzing data mining, text mining and predictive analysis. Rapid Miner uses a variety of descriptive and predictive techniques in providing insights to users so they can make the best decisions. Rapid Miner has approximately 500 data mining operators, including operators for input, output, data preprocessing and visualization. Rapid Miner is a stand-alone software for data analysis (Wicaksana, I Wayan Simri, Baskoro, & Ambarwati, 2013).

III. Research Method

Indication of patients leaving the ICU is part of the output of the ICU care process that is needed for patient care that is continuously carried out until the patient can recover. In this research method, the authors are interested in analyzing and studying data in the intensive care section of a hospital so that the results of this study can be useful for information for medical or paramedical personnel who need information in caring for ICU patients and can also be used as material a basis for decision making for users involved in the ICU ward in a hospital. In this study the authors examined limited ICU patients with parameters on medical devices and severe diagnoses called "Sepsis" so that it can be known :

1. The extent to which the method used can find out the best level of accuracy for prediction of indications of patients who leave / move ICU.
2. The method used is to use C4.5 Algorithm, Naïve Bayes.
3. Testing of the performance of both methods. Performed using the confusion matrix and ROC curves. As for applications, used are Rapid Miner tools. The stages carried out in this study.

A. Data Preparation

In the stage of data preparation , it is a stage to prepare data to be applied in modeling which previously came from the initial raw data to the classification stage. In this stage is process known as pre-processing is a stage that contains many activities as follows (Kurniawan, Surakarta, & Bayes, 2018).

B. Data collection

a. The data taken came from the ICU Central Army Hospital Jakarta patient register report data source. The data collected is taken from the patient's register and according to the needs of the study, the data used for analysis is sorted first and the data is taken according to the research needs.

b. In this study comparative analysis will be performed using two data mining Classification methods. The proposed method for processing student data is the use of the C4.5 Algorithm and Naïve Bayes.

c. Data processing is done by using the Rapid Miner application so that the problem solving models expected by the author will be generated and finally it can be concluded well and accurately.

d. The framework in this study is based on the theories already established there before like in the picture below :

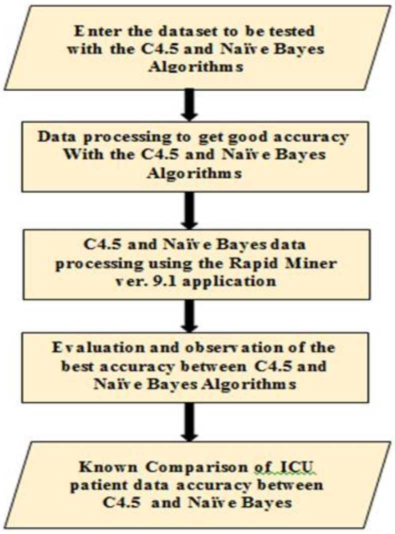

Enter the dataset to be tested with the C4.5 and Naïve Bayes Algorithms
Data processing to get good accuracy With the C4.5 and Naïve Bayes Algorithms
C4.5 and Naïve Bayes data processing using the Rapid Miner ver. 9.1 application
Evaluation and observation of the best accuracy between C4.5 and Naïve Bayes Algorithms
Known Comparison of ICU patient data accuracy between C4.5 and Naïve Bayes

Figure 4. Research Method
Source : Research result (2019)

IV. Result and Discussion

A. Data processing

Data prepared for this study consisted of 343 patients taken from the patient registers in December 2018 and January 2019. The data must still be scrutinized and completed if there is insufficient data. As for the incomplete data, it will be completed by searching the data source from the manual register.

Table 1. ICU Patients Dataset

AGE	GENDER	DIAGNOSIS OF SEPSIS	VENTILATOR	CVP	LENGTH OF ICU STAY	INFORMATION
90.5	MALE	TRUE	TRUE	TRUE	<3 days	DIED
53.6	FEMALE	TRUE	TRUE	TRUE	>3 days	DIED
70.8	MALE	FALSE	FALSE	FALSE	<3 days	MOVE
65.9	MALE	FALSE	FALSE	FALSE	>3 days	MOVE
56.8	MALE	FALSE	FALSE	TRUE	>3 days	DIED
73.4	FEMALE	FALSE	FALSE	FALSE	<3 days	MOVE
78.3	MALE	FALSE	FALSE	FALSE	<3 days	MOVE
64.8	FEMALE	FALSE	FALSE	FALSE	<3 days	MOVE
....	....	....	....	....	....	....
....	....	....	....	....	....	....
Number of patients: 343 patients						

Source : research result (2019)

B. Analysis with C4.5 Algorithm

This ICU care patient data has been classified using Microsoft Excel which consists of 343 patient data for December 2018 until Januari 2019 with the attributes :

a. Age
b. Gender
c. Diagnosis of Sepsis
d. Ventilator (breathing apparatus)
e. CVP. (Liquid and Drug Input Aid)
f. Length Of ICU Stay
g. Information

After analyzing the data by entering the ICU patient excel data set into Rapid Miner using the arrangement shown in Figure 4, it produces a decision tree like the image below :

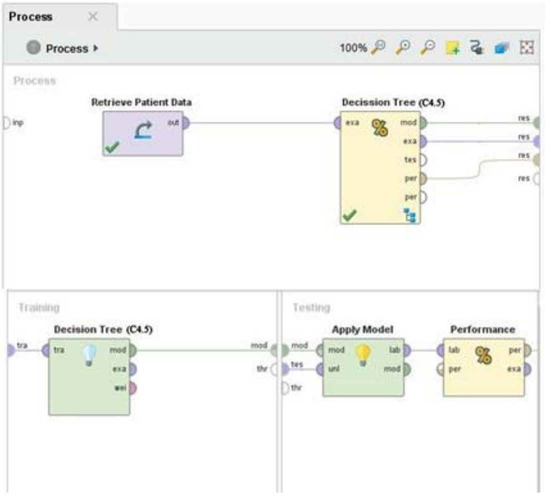


Figure 5. Model C4.5 in Rapid Miner
Source : Research result (2019)

After executing it will produce Accuracy level as in table 2

Table 2 Accuracy Levels of C4.5 Algorithms

Accuracy : 81.25 %

	True Died	True Move	Class Precision
Pred Died	**23**	**13**	**63.89%**
Pred Move	**51**	**255**	**83.33%**
Class Recall	**31.08%**	**95.15%**	

And also could be known about AUC from C4.5 like figure 6 :

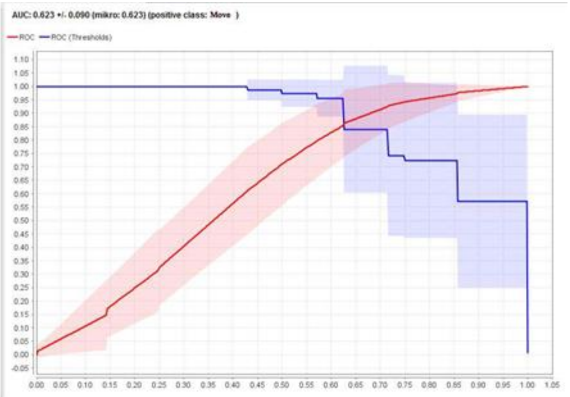


Figure 6 AUC Value C4.5 Algorithm.
Source : Research result (2019)

Based on testing these data it can be seen that, the level of accuracy using the C4.5 algorithm is 81.25%, AUC is 0.623 (Positive class: Move).

C. Analysis with Naïve Bayes Algorithm

To find out the accuracy level of ICU patient data better, the writer tries to make a comparison of the accuracy level using the Naïve Bayes algorithm.

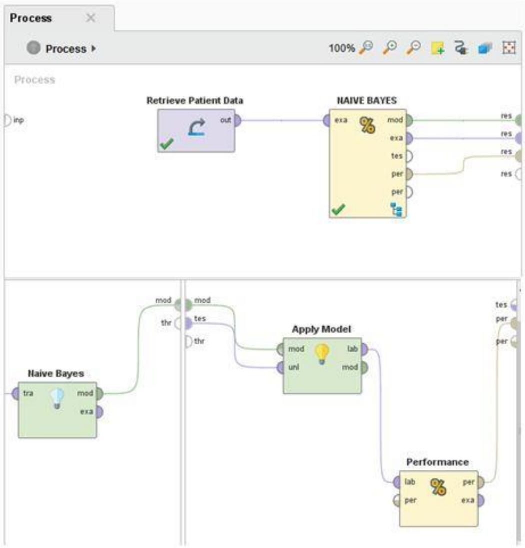


Figure 7. Model Naïve Bayes in Rapid Miner
Source : Research result (2019)

Table 3. Accuracy Levels of Naïve Bayes Algorithms

Accuracy : 80.66 %

	True Died	True Move	Class Precision
Pred Died	**32**	**24**	**57.14 %**
Pred Move	**42**	**244**	**85.31 %**
Class Recall	**43.24 %**	**91.04**	

Could be known about AUC from Naïve Bayes like figure 8 :

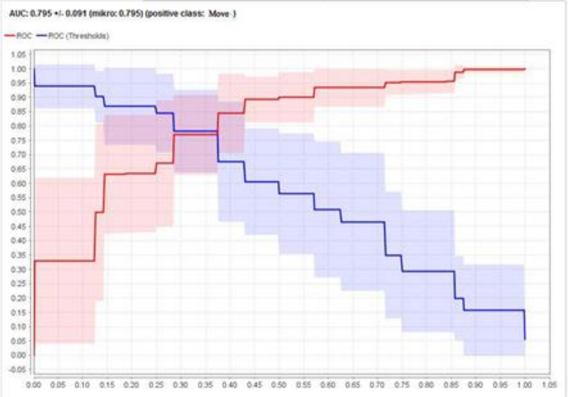


Figure 8 AUC Value Naïve Bayes Algorithm.
Source : Research result (2019)

After testing the ICU patient data using the Naive Bayes algorithm, 80.66% accuracy is generated, AUC is 0.795 (positive class move).

Based on data analysis using the C4.5 algorithm and the Naïve Bayes algorithm the difference in accuracy is as shown in table 4

Table 4. Comparison of performance C4.5 and Naïve Bayes algorithm.

COMPARISON PERFORMANCE	C4.5	NAIVE BAYES
ACCURACY	81.25%	80.66 %
AUC	0.623	0.795

Source : Research result (2019)

V. Conclusion and Suggestion

A. Conclusion

The conclusion of the research and analysis with the use of the C4.5 algorithm and Nave Bayes Algorithm can be seen the comparison of the two algorithms, namely C4.5 algorithm has an accuracy of 81.25% and the AUC 0.623 while Naïve Bayes has an accuracy of 80.66% and the AUC 0.795, which means that the results for this case study C4.5 algorithm are 0.59 % higher than the Naïve Bayes algorithm.

B. by suggestion

So that this research is enhanced by suggestions:

a. This research is expected to be used by medical authorities as consideration for ICU patient prediction.

b. This research can be developed with other optimization methods such as Ant Colony Optimization (ACO) and others.

VI. References

Alotaibi, N. N., & Sasi, S. (2015). Predictive Model for Transferring Stroke In-Patients to Intensive Care Unit, 848–853.

C.W. Cheng, N. Chanani, J. Venugopalan, K. Maher, and M. D. W. (2013). IcuARM-An ICU Clinical Decision Support System Using Association Rule Mining. *IEEE Journal of Translational Engineering in Health and Medicine*, *1*(1), 4400110. https://doi.org/10.1109/JTEHM.2013.2290113

Crc, H., & Hofmann, M. (2014). *Data Mining and Knowledge Discovery Series Edited by*.

Fatmawati. (2016). Perbandingan Algoritma Klasifikasi Data Mining Model C4 . 5 Dan Naive Bayes Untuk Prediksi Penyakit Diabetes. *Jurnal Techno Nusa Mandiri*.

Fridayanthie, E. W. (2015). ANALISA DATA MINING UNTUK PREDIKSI PENYAKIT HEPATITIS DENGAN MENGGUNAKAN METODE NAIVE BAYES DAN SUPPORT VECTOR MACHINE. *Journal of Applied Microbiology*.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concept and Techniques. San Fransisco: Mofgan Kaufan Publisher. Data Mining.*

https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Irvan, Febyan, S. (2018). Sepsis dan Tata Laksana Berdasar Guideline Terbaru, *X*, 62–73.

Kepmenkes-no-1778-tahun-2010-tentang-pedoman-pelayanan-icu-di-rumah-sakit.pdf. (2010).

Kurniawan, Y. I., Surakarta, U. M., & Bayes, N. (2018). COMPARISON OF NAIVE BAYES AND C . 45 ALGORITHM IN DATA MINING, *5*(4), 455–464. https://doi.org/10.25126/jtiik

Lesmana, H. (2018). AKURASI PENGUKURAN TEKANAN VENA SENTRAL (Central Venous Pressure) Pendahuluan Central Venous Pressure (CVP) atau tekanan vena sentral merupakan salah mode Positive End Ekspiratory Pressure). Hal-hal yang dapat mempengaruhi Cardiovaskuler Care Unit , , *1*(1), 1–13.

Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., & Buchman, T. G. (2017). Supplementary Digital Content for: An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Journal of Biological Chemistry*, *39*(5), 561–563. https://doi.org/1-4244-1484-9/08/$25.00

Sahetya, S., Allgood, S., Gay, P. C., & Lechtzin, N. (2016). Long - Ter m Mechanical Ventilation. *Clinics in Chest Medicine*. https://doi.org/10.1016/j.ccm.2016.07.014

Septiani, W. D., Studi, P., & Informatika, M. (2017). KOMPARASI METODE KLASIFIKASI DATA MINING ALGORITMA C4.5DAN NAIVE BAYES UNTUK PREDIKSI PENYAKIT HEPATITIS, *13*(1), 76–84.

Vercellis, C. (2011). *Business Intelligence: Data Mining and Optimization for Decision Making*. *Methods*.

Wicaksana, I Wayan Simri, D. A. C., Baskoro, D. A., & Ambarwati, L. (2013). Belajar Data Mining dengan Rapid Miner.