Online version available at: https://jurnal.buddhidharma.ac.id/index.php/te

JOURNAL OF TECH-E

| 2581-1916 (Online) | 2598-7585 (Printed) |

Artikel

Comparison of Data Mining Methods Using the Naïve Bayes Algorithm and K-Nearest Neighbor in Predicting Immunotherapy Success

Budi Harto[1], *Rino*[2],

[1, 2] *Universitas Buddhi Dharma, Teknik Informatika, Banten, Indonesia*

SUBMISSION TRACK

Received 20 December 2018;
Revised 12 January 2019;
Accepted 10 February 2019;
Available online 20 February 2019

KEYWORD

Immunotherapy , Data Mining K-NN, Naïve Bayes, Expert System , machine learning

KORESPONDENSI

E-mail: budiihartoo@gmail.com

A B S T R A C T

tumor or cancer is a disease that is a problem for people who are increasing every year. This disease in both the early and final stages requires attention because in this disease sufferers have a large risk of death. along with the rapid development of technology, we can use the technology to facilitate in all fields one of which is to predict success in a therapy. Data mining is one of the techniques used by the author in testing the dataset used in this study to get the best algorithm between Naïve Bayes and the K-Nearest Neighbor algorithm by using the Rapid Miner S
tudio application and applying the best algorithm into the expected application or expert system. can help users predict the success of a therapy.

INTRODUCTION

Data mining such as Naïve Bayes and K-Nearest Neighbor to dissect the data set and compare it to obtain the factors that influence and find the right algorithm from the two algorithms and then apply it to the expert system.
With the comparison of these algorithms, we will find an appropriate method to be applied or implemented in the form of an expert system to predict the success rate of a therapy, especially immunotherapy in fighting cancer cells in the cancer patient's body.

I. DATA MINING METHOD

Data mining is a term used to describe the discovery of knowledge in a database. Data mining is a process that uses statistical techniques, mathematics, artificial intelligence, and Machine Learning to extract and identify useful information and related knowledge from various large databases. [1].

II. IMMUNOTHERAPY

Immunotheraphy is a tumor or cancer therapy that is immunologically ineffective. The goal of this therapy is to obtain immunity against tumors or cancer. Immuno therapy can be done in two ways, namely specific and non-specific, where specific is done by giving Ag tumor preparations while non-specific is done by forming an immune response, especially macrophages with BCG [2].

III. Rapid Miner

Amril Mutoi Siregar and Adam Puspabhuana [3] RapidMiner is open source software.

RapidMiner is a solution for analyzing data mining, text mining and predictive analysis. RapidMiner uses a variety of descriptive and predictive techniques in providing insights to users so they can make the best decisions. RapidMiner has approximately 500 data mining operators, including operators for input, output, data preprocessing and visualization. RapidMiner is a stand-alone software for data analysis and as a data mining engine that can be integrated in its own products. RapidMiner is written using java language so that it can work on all operating systems.

IV. ALGORITHM

Algorithm as a clear procedure for solving a problem by using certain steps and limited in number to obtain the desired output from an input in a limited amount of time [4]

[5] Bayes is a simple probability-based prediction technique based on the application of the Bayes theorem (or Bayes rule) with strong (naive) independence assumptions. in other words, in naïve bayes, the model used is an "independent feature model".

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

Explanation of the formula is as follows.

P(H|E)	The conditional probability of a hypothesis H occurs if evidence is given.
P(E|H)	The probability that evidence E occurs will influence hypothesis H.
P(H)	The initial (priori) hypothesis of H hypothesis occurs regardless of any evidence.
P(E)	The initial probability (priori) of evidence E occurs regardless of the hypothesis / other evidence.

[5] Nearest Neighbor algorithm (sometimes called K-Nearest Neighbor / K-NN) is an algorithm that classifies based on the proximity of the location (distance) of a data with other data.

The simple principle adopted by the K-NN algorithm is "if an animal goes like a duck, quacking a quack a quack like a duck, and looks like a duck, that animal might be a duck".

In the K-NN algorithm, the dimensionless q data, the distance from the data to other data can be calculated. This distance value is used as a value of closeness / similarity between test data with training data. K value on K-NN means the closest K-data from the test data.

The following is the formula for K-NN prediction

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

V. CONFUSION MATRIX

[6]*Confusion matrix* Is a table to measure the performance of classification algorithms or classification models or classifiers. Here is an example *confusion matrix* :

n = 4	Prediksi class1	Prediksi class2
Aktual class1	2	1
Aktual class2	0	1

VI. JAVA

According to the Sun Microsystem definition, in the book [7] Java is the name of a collection of technologies for creating and running software on a stand-alone computer or in a networked environment.

VII. MYSQL

[8] MySQL (*My Structure Query Language*) adalah sebuah software management data bae SQL (*Database Management System*) or DBMS of many DBMS, like Oracle, MS SQL, Postagre SQL, and other. MySQL is DBMS *multithread*, *multi-user* which is free under license GNU *General Public License* (GPL).

VIII. ANALYSIS

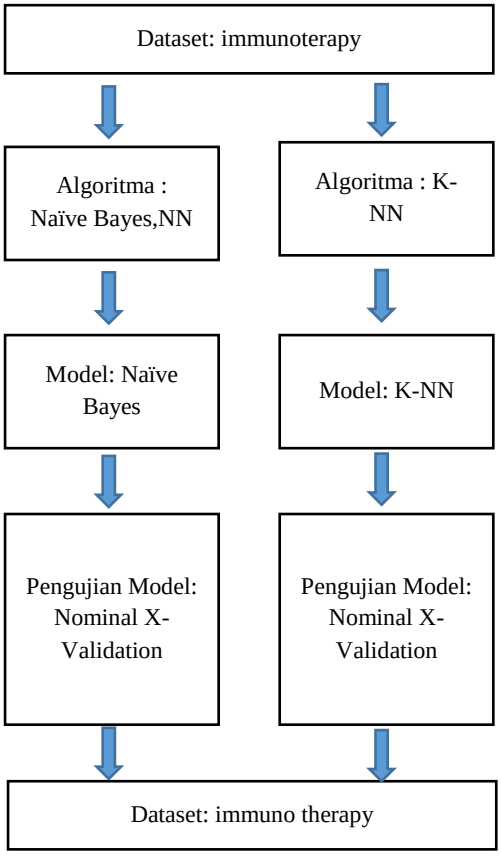
To find out between the two methods in this study, researchers need a set of data to be processed into a fast minner and entered into the system. Where from the dataset can be obtained the level of accuracy of the methods - methods or algorithms used to process the dataset. The dataset that will be scrutinized is obtained from the uci machine repository data provider site: https://archive.ics.uci.edu/ml/datasets/Immunotherapy+Dataset ‘

Where the data set has 90 records, 8 attributes and 1 of these attributes is the result attribute.

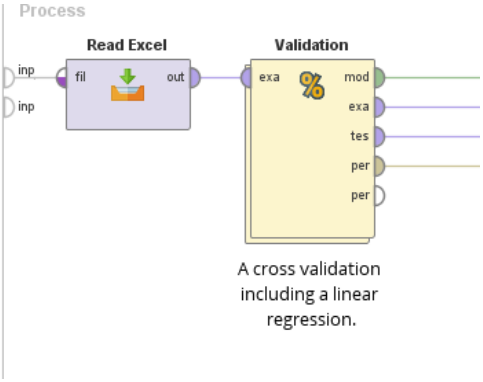
Name	*Values*
Gender	*41 Man* *49 Woman*
Age (year)	*15 – 56*
Time (month)	*0 – 12*
The number of warts	*1 – 19*
Type of wart	*1 – Common* *2 – Plantar* *3 – Both*
Surface area of the warts (mm^2)	*6 – 900*
Induration diameter of initial test (mm)	*5 – 70*
Response treatment	*1 – Yes* *0 – No*

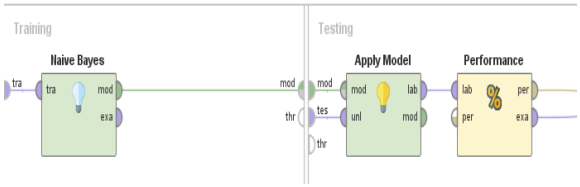
IX. HASIL

In this study, Rapid Minner is used to calculate the accuracy of an algorithm to determine the best algorithm that will be applied into the form of an expert system to predict the success of immune therapy. Following is the model that will be used to process the dataset into Rapidminer.



With the Naïve Bayes method, the data processing display looks like the image below.

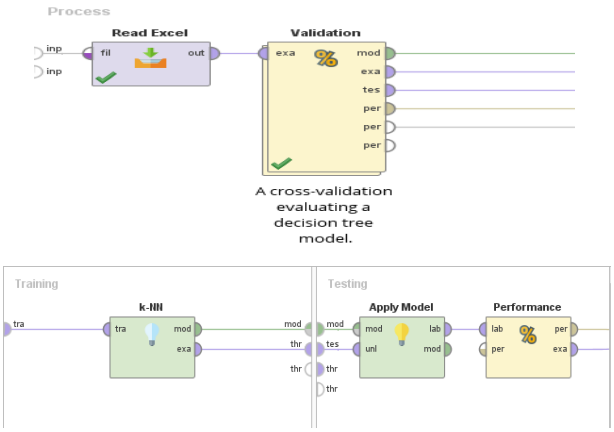




Following is the resulting Confusion Matrix table:

accuracy: 81.11%			
	True Berhasil	True Gagal	Class Precision
Pred. Berhasil	67	13	83.75%
Pred. Gagal	4	6	60.00 %
Class recall	94.37%	31.58%	

With the K-NN method, the data processing display looks like the image below.



Accuracy: 75.56%			
	True Berhasil	*True* Gagal	*Class Precision*
Pred. Berhasil	68	19	78.16%
Pred. Gagal	3	0	0.00 %
Class recall	95.77%	0.00%	

Following is the resulting Confusion Matrix table:

X. IMPLEMENTASI

Comparison of the two algorithms will get the best algorithm with the highest accuracy, from the results of the above research it is found that the naïve bayes algorithm is the algorithm with the highest accuracy and the algorithm will be applied to a simple expert system with a design as above.
In this research the implementation is carried out by making the dataset follow the rules that exist in the naïve bayes algorithm, in this case the researchers convert the dataset in the form of numbers into a label by naming each data range.
From the dataset above obtained some probability, the probability will be used to calculate the likelihood that will occur from the data to be predicted or tested. Naming the probability by using a label from the record that has been converted with a probability label. Following is an example of the probability name formed from the dataset:

P(Pria|Berhasil)
P(Pria|Gagal)
P(Wanita|Berhasil)
P(Wanita|Gagal)
P(Muda|Berhasil)
P(Muda|Gagal)
P(Tua|Berhasil)
P(Tua|Gagal)
P(Paruhbaya|Berhasil)
P(Paruhbaya|Gagal)
P(Sebentar|Berhasil)
P(Sebentar|Gagal)
P(Sedang|Berhasil)
P(Sedang|Gagal)
Dst.

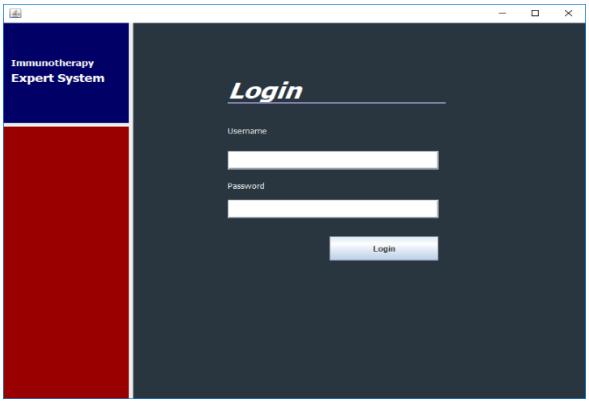
These probabilities have the value of each of the calculations in the dataset. Here is an example of how to calculate a probability.

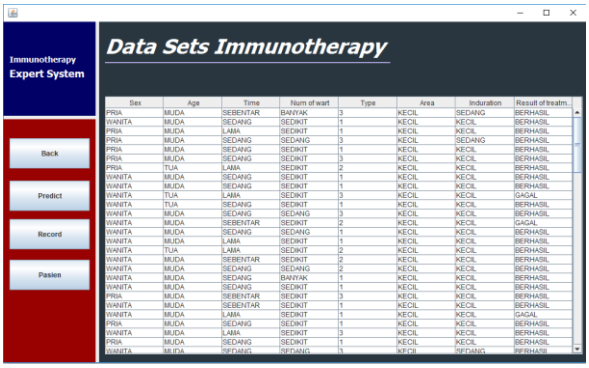
P(Label|Berhasil) = The number of labels that have the probability of success is divided by the number of chances of success

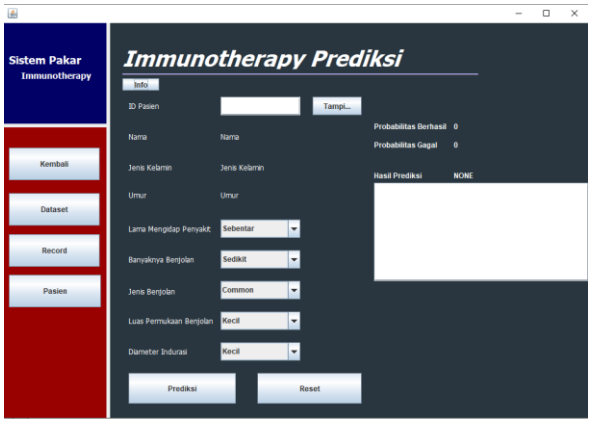
P(Label|Gagal) = The number of labels that have the possibility of failure is divided by the number of possible failures

XI. TAMPILAN PROGRAM

Sex	*Age*	*Time*	*Number of Warts*	*Type*	*Area*	*induration diameter*	*Result of Treatment*
PRIA	MUDA	SEBENTAR	BANYAK	3	KECIL	SEDANG	BERHASIL
WANITA	MUDA	SEDANG	SEDIKIT	1	SEDANG	KECIL	GAGAL
PRIA	MUDA	LAMA	SEDIKIT	1	LUAS	KECIL	BERHASIL
PRIA	PARUH BAYA	SEDANG	SEDANG	3	KECIL	SEDANG	GAGAL
PRIA	MUDA	SEDANG	SEDIKIT	1	KECIL	KECIL	BERHASIL
PRIA	MUDA	SEDANG	SEDIKIT	3	KECIL	LEBAR	BERHASIL
PRIA	TUA	LAMA	SEDIKIT	2	KECIL	KECIL	BERHASIL







XII. CONCLUSION

After making a comparison, design, manufacture and test of an expert system that has been made, the researcher concludes the following:

Naïve Bayes algorithm has a superior accuracy of 81.11% compared to the K-NN algorithm which only has an accuracy of 75.56% in managing the immunotherapy dataset.

The expert system created in this study can help in predicting the success rate or response of immune therapy to tumor or cancer.

Algorithms in the data mining algorithm are quite good in managing immunotherapy datasets because they have a high percentage of accuracy.

REFERENCES

[1] D. Nofriansyah, Konsep Data Mining vs Sistem Pendukung Keputusan, Jogjakarta: Deepublish, 2014.

[2] A. H. Sri, Immunologi dasar dan Immunologi klinis, Jogjakarta: Graha Ilmu, 2013.

[3] A. P. Amril Mutoi Siregar, DATA MINING: Pengolahan Data Menjadi Informasi dengan RapidMiner, Surakarta: CV Kekata Group, 2018.

[4] Suarga, Algoritma Pemrograman, Jogjakarta: Andi Offset, 2012.

[5] E. Prasetyo, Data Mining, Jogjakarta: C.V. Andi Offset, 2012.

[6] D. T. N. M. Reza Faisal, Belajar Data Science : Klasifikasi dengan Bahasa Pemrograman R, Banjar Baru: Scripta Cendekia, 2019.

[7] R. M. Salahudin, Pemrograman J2ME Belajar Cepat Pemrograman Perangkat Telekomunikasi Mobile, Bandung: Informatika, 2010.

[8] Anhar, PHP & MySql Secara Otodidak, Jakarta: mediakita, 2010.

Kurnia, Y., Isharianto, Y., Giap, Y. C., & Hermawan, A. (2019, March). Study of application of data mining market basket analysis for knowing sales pattern (association of items) at the O! Fish restaurant using apriori algorithm. In *Journal of Physics: Conference Series* (Vol. 1175, No. 1, p. 012047). IOP Publishing.

BIOGRAPHY

Budi Harto In 2015 Graduated from Poris Indah High School, and in 2015 continued studying at the Buddhi Dharma University and Graduated in the Informatics Engineering Study Program (S1) in the Database field, 2019. In 2019 worked at PT. Adrena Solusi Insan Muda as Software Developer.

Rino received his Bachelor of Informatics Engineering (S.Kom) from STMIK Buddhi, Indonesia in 2008 and his Master of Computer Science (M.Kom) concentration in Software Engineering from STMIK Eresha, Indonesia in 2012. He was a lecturer in the Engineering Study Program Informatics, Faculty of Science Technology & Technology, Buddhi Dharma University.