JOSEIT JOURNAL OF SYSTEMS ENGINEERING AND INFORMATION TECHNOLOGY
http://jurnal.iaii.or.id/index.php/JOSEIT

Comparative Analysis of Naïve Bayes and Decision Tree Algorithms in Data Mining Classification to Predict Weckerle Machine Productivity

Fried Sinlae[1*], Anugrah Sandy Yudhasti[2], Arief Wibowo[3]
[123]Universitas Budi Luhur, Jakarta, Indonesia

[*]*sinlaefried@gmail.com*

Abstract

The level of data accuracy in everyday life is necessary because it is reflected in the ever-advancing development of information technology. Analysis of data processing in information that can provide knowledge with the help of data mining systems. Algorithms commonly used for prediction are Naive Bayes and Decision Trees. The purpose of this study is to compare the Nave-Bayes algorithm and the decision tree algorithm in terms of the accuracy of predicting the productivity of the Weckerle machine at PT XYZ. The method used is a literature study from various related sources and understanding of the data in the source related to the subject of the classification method of the Naive Bayes algorithm and the decision tree into the data mining system. The results of this study are a classification using the Nave-Bayes algorithm with a higher level of confidence than the decision tree algorithm.

Keywords: Naïve Bayes; Decision Tree; Algorithm; Data Mining

1. Introduction

The level of data accuracy is needed in everyday life as it keeps driving the development of information technology. According to [1], when determining any decision under certain conditions, the information must be taken into account, therefore the availability of information has become a medium to analyze and summarize knowledge from data, which is useful for decision-making. However, knowledge from data about information is not enough to make decisions. Data analysis is also required to generate a consideration from the information provided. The analysis of data management in information that can provide knowledge is carried out using a data mining system [2].

One problem can be predicted from the trend in terms of rules or estimates into the future using data mining. There are various real-life applications of steps and techniques in data mining, one of which is the classification technique. According to [3], this classification is a basic form of data analysis, while according to [4] the classification is a technique for group membership according to existing data. Previously, there have been many studies on predicting closure using classification methods, one of which is the Nave Bayes algorithm and the decision tree algorithm. According to [3], the Nave-Bayes algorithm itself is currently popular because the accuracy rate of these two algorithms is high. Therefore, this study aims to compare the accuracy of the Nave-Bayes algorithm and the decision tree algorithm in terms of the accuracy of predicting the productivity of the Weckerle machine at PT XYZ.

The research the author will be conducting at this point is a comparison of Nave Bayes and decision tree methods in data mining classification to predict the productivity of the Weckerle machine at PT XYZ.

2. Method

In this study, there are guidelines to help the research achieve maximum results and avoid falling short of the research goals. The research steps are as follows:

a. Problem Identification. This stage carries out an explanation of the research problem by explaining the important points of the problem and then the author gets a foundation for describing the research problem.
b. Literature Review. The research was conducted in literature from various related references and understanding the data in the references related to the topic of the Naïve Bayes and Decision Tree algorithm classification methods into a data mining system. There was also an assessment of theories related to research topics that have been carried out in existing research as well as the development of various current theories and references.
c. Analysis. At this stage, the process of studying, describing and solving a problem and research objectives is carried out.
d. Paper Implementation. At this stage, the application of the analysis results into a paper or journal is carried out.

Accepted by editor: 25-01-2021 | Final revision: 15-09-2022 | Online publication: 25-09-2022

3. Result and Discussion

In this research, we analyze by comparing two methods namely decision tree algorithm and Nave Bayes algorithm method in a data mining system. Here we compare the two methods using data from two sources addressing the issue of predicting the productivity of the Weckerle machine at PT XYZ. The first source of Weckerle machine productivity data below uses the Decision Tree algorithm computation. The process of processing this data using the decision tree method works to build a decision tree.

There are several processes, namely:

a. Calculating the results of data summation, this data summation is based on the number of result attributes by fulfilling the predetermined conditions.
b. Determine the attribute and use it for Node. Node is one of the attributes with the highest gain value from other attributes.
c. Create branching for each member of the Node.
d. Check if any member of the Node has a zero value, but if the result has a zero value, then determine which one is appropriate to become a leaf of the decision tree. Continue until the entire entropy value of the members of the Node has a value of zero so that the process stops.
e. If there is more than zero entropy value coming from one member of the Node, then repeat the previous process from the beginning until all Nodes have zero value.

Table 1. Data Sample

Finish Goods	Down Time	Reject	Output	Productive
High	Available	Many	Many	Yes
Medium	Available	Many	Many	Yes
Medium	No Available	Many	Many	Yes
Low	Available	Many	Many	Not Productive
Low	Available	Many	Little	Not Productive
High	Available	Many	Many	Yes
High	No Available	Little	Little	Not Productive
High	No Available	Many	Many	Yes
Medium	Available	Little	Many	Yes
Low	Available	Many	Many	Not Productive
Low	Available	Many	Many	Not Productive
Low	No Available	Many	Little	Not Productive
Medium	No Available	Little	Little	Not Productive
Medium	No Available	Little	Many	Productive
High	No Available	Many	Many	Productive
High	Available	Many	Little	Not Productive
Low	No Available	Many	Little	Not Productive
High	Available	Many	Many	Productive
Medium	Available	Many	Many	Productive
High	Available	Many	Many	Productive
High	Available	Many	Many	Productive
Low	No Available	Many	Little	Not Productive
Medium	Available	Many	Many	Productive
High	Available	Many	Little	Not Productive
Medium	No Available	Little	Little	Not Productive
Low	No Available	Little	Many	Not Productive
High	Available	Many	Many	Productive
High	Available	Little	Many	Productive
High	Available	Many	Many	Productive
Medium	Available	Little	Many	Productive
Medium	Available	Little	Many	Productive
Medium	No Available	Many	Many	Productive
High	Available	Many	Many	Productive
Low	Available	Little	Little	Not Productive
High	Available	Many	Many	Productive

After getting sample data, go through the process of calculating data amount, entropy, and gain. The results are in the following table:

Table 2. Calculation Of the Amount of Data, Entropy, and Gain

Node			Sum	Productive	No Productive	Entropy	Gain
1	Total		35	21	14	0,97095	
	Finish Goods						0,446569331
		High	15	12	3	0,72193	
		Medium	11	9	2	0,68404	
		Low	9	0	9	0,00000	
	Down Time						0,052411577
		Available	23	16	7	0,88654	
		No Available	12	5	7	0,97987	
	Reject Product						0,011891174
		Many	25	16	9	0,94268	
		Little	10	5	5	1,00000	
	Output Products						0,517872341
		Many	25	21	4	0,63431	
		Little	10	0	10	0,00000	

After running through several calculation processes using the decision tree method from the data, one obtains the results of the highest gain value, namely the production output. Then the production output is put here in the root of the decision tree. And the finished goods production becomes a factor that determines the productivity of the weckerle machine. Seen on the following picture:

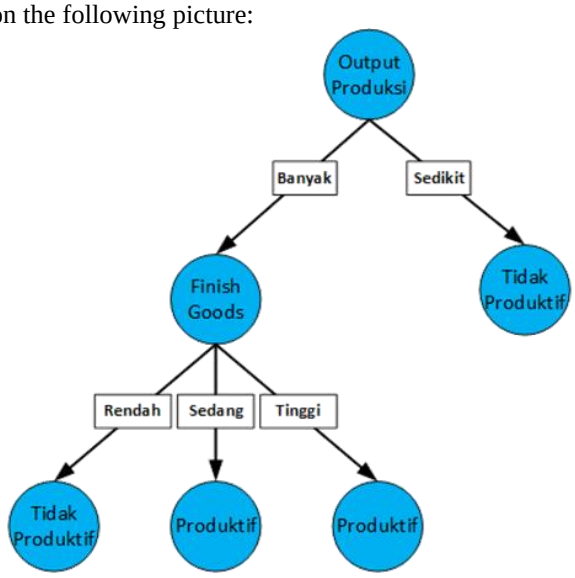


Figure 1. Decision Tree

Figure 1 explains that there are two types of members, namely many members (productive) and few members (unproductive). There are 3 members in Finish Goods, namely low (unproductive), medium (productive), high (productive). And why this decision tree only reaches Finish Goods, that's because the value between productive and unproductive members has a value of 0, so the decision can be obtained directly. This also shows that rejects and downtimes do not affect the productivity of the Weckerle PT XYZ machine.

The next test concerns sample data using tools in the Excel application, namely Rapidminer tools, starting with the connection process between the sample database and the operator and subsequent validation as shown in Figure 2 below:

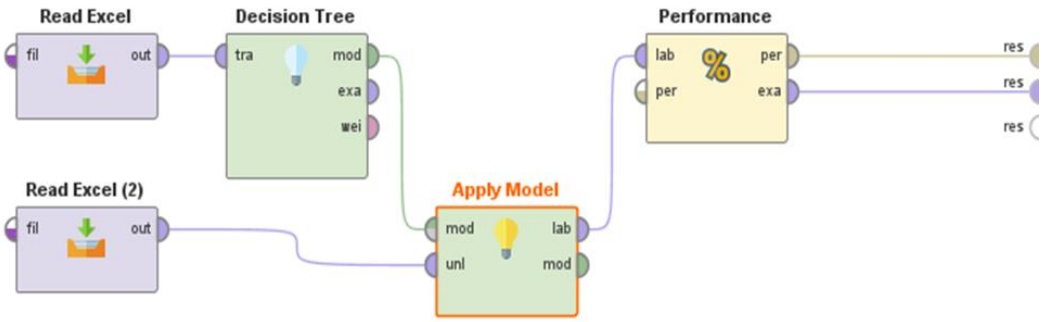


Figure 2. Decision Tree Connection in RapidMiner Tools

From the RapidMiner connection process above, the results obtained are the same as from the manual calculation process in Figure 2 to obtain the decision tree results as below. This chapter explains the process carried out in this study. There are several steps used in this research.

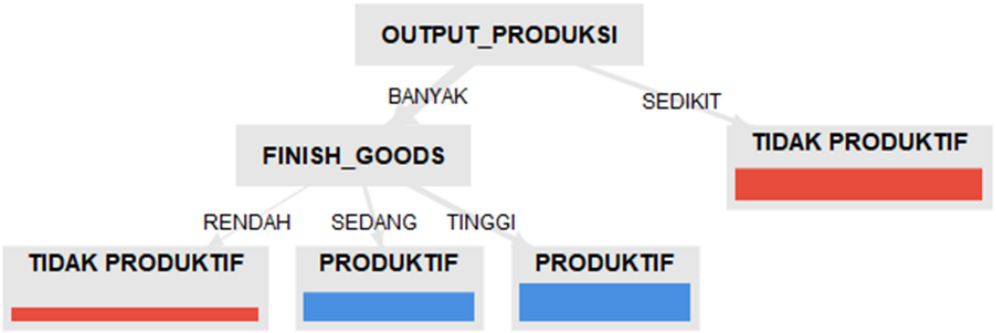
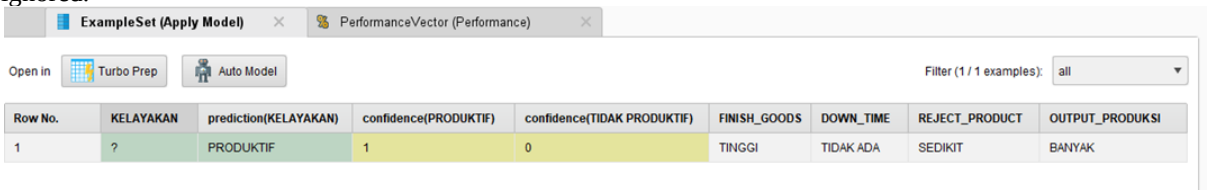


Figure 3. Decision Tree in RapidMiner Tools

The following is a screenshot of the measurement results of the test data against the Apply Decision Tree algorithm model in predicting the productivity of the Weckerle machine. It can be seen that the trust is productive 1 and unproductive 0 as it follows the decision tree in Figure 4. Downtime and scrap products are ignored.



Row No.	KELAYAKAN	prediction(KELAYAKAN)	confidence(PRODUKTIF)	confidence(TIDAK PRODUKTIF)	FINISH_GOODS	DOWN_TIME	REJECT_PRODUCT	OUTPUT_PRODUKSI
1	?	PRODUKTIF	1	0	TINGGI	TIDAK ADA	SEDIKIT	BANYAK

Figure 4. Prediction of Decision Tree Test Data on RapidMiner Tools

In this second case study, the Nave Bayes method is applied, with the data processing still based on the same topic. With this method, there is a set application testing process with RapidMiner. With this RapidMiner application, it is grouped by the selected attributes, which are the same attributes and data sets as the decision tree method.

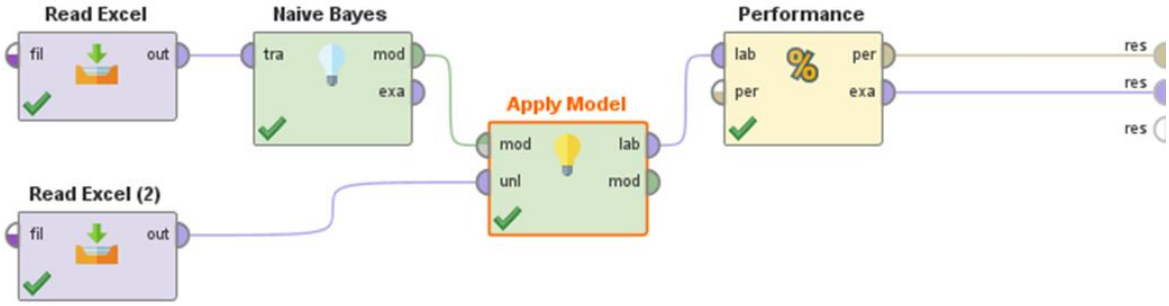
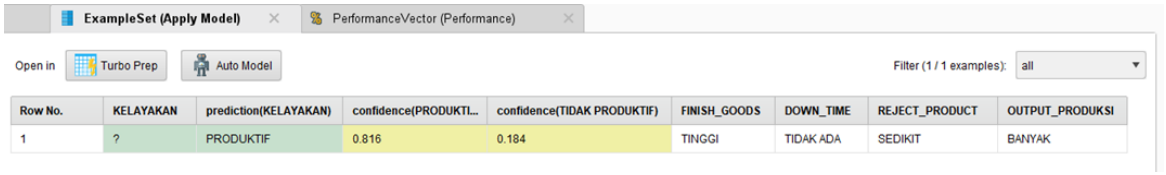


Figure 5. Naïve Bayes Connection in RapidMiner Tools

The following is a screenshot of the measurement results of the test data against the Apply Nave Bayes algorithm model in predicting the productivity of the Weckerle machine. It can be seen that the confidence is productive 0.816 and unproductive 0.184 because it follows the probability formula in Nave Bayes that it can be implemented in numbers unlike the decision tree algorithm which just follows the decision tree and ignores attributes that do not go into the calculation.



Row No.	KELAYAKAN	prediction(KELAYAKAN)	confidence(PRODUKTI...	confidence(TIDAK PRODUKTIF)	FINISH_GOODS	DOWN_TIME	REJECT_PRODUCT	OUTPUT_PRODUKSI
1	?	PRODUKTIF	0.816	0.184	TINGGI	TIDAK ADA	SEDIKIT	BANYAK

Figure 6. Prediction of Naïve Bayes Test Data in RapidMiner Tools

4. Conclusion

From the results of the explanation in the above description it can be concluded that the use of the Nave-Bayes algorithm method evaluates the accuracy of the data to show the confidence of the processed test data applied to display decimal numbers. Meanwhile, the decision tree algorithm gets the results of measuring the confidence of test data in predicting the productivity of the Weckerle machine, which only shows the number 1 for productive and 0 for unproductive. So, this shows that the Nave Bayes algorithm for predicting the deal has a higher confidence level of accuracy than the decision tree algorithm.

References

[1] N. M. Huda, APLIKASI DATA MINING UNTUK MENAMPILKAN INFORMASI TINGKAT KELULUSAN MAHASISWA (Skripsi), Semarang: Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Diponegoro, 2010.
[2] A. Fikri, Penerapan Data Mining Untuk Mengetahui Tingkat Kekuatan Beton Yang Dihasilkan Dengan Metode Estimasi Menggunakan Linear Regression (Skripsi), Semarang: Fakultas Ilmu Komputer Universitas Dian Nuswantoro, 2013.
[3] Y. I. Kurniawan, "PERBANDINGAN ALGORITMA NAIVE BAYES DAN C.45 DALAM KLASIFIKASI DATA MINING," Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), pp. 455-463, 2018.
[4] M. S. S. G. Arpit Bansal, "Improved K-mean Clustering Algorithm for Prediction Analysis using Classification Technique in Data Mining," International Journal of Computer Applications, pp. 35-40, 2017.
[5] P. B. Davies, Database Systems Third Edition, New York: Plgrave Macmillan, 2004.
[6] R. J. Roiger, Data Mining: A Tutorial-Based Primer, CRC Press, 2017.
[7] A. Saleh, "KLASIFIKASI METODE NAIVE BAYES DALAM DATA MINING UNTUK MENENTUKAN KONSENTRASI SISWA (STUDI KASUS DI MAS PAB 2 MEDAN)," Konferensi Nasional Pengembangan Teknologi Informasi dan Komunikasi, pp. 200-208, 2015.
[8] Bustami, "PENERAPAN ALGORITMA NAIVE BAYES UNTUK MENGKLASIFIKASI DATA NASABAH ASURANSI," JURNAL INFORMATIKA, pp. 884-898, 2018.
[9] A. K. N. I. A. Ayung Candra Padmasari, "Penerapan Model Decision Tree untuk Rancangan Game Multiplayer Berbasis Jaringan (Uka-Uka Tresure Hunter)," Jurnal Edsence Vol. 1 No. 1, pp. 19-24, 2019.
[10] B. Unhelkar, Software Engineering with UML, Boca Raton: Auerbach Publications/CRC Press, 2018.
[11] Z. Syahputra, "Penerapan Permodelan UML Sistem Informasi Perpustakaan pada Universitas Islam Indragiri Berbasis Client Server," Jurnal SISTEMASI, pp. 57-64, 2015.
[12] J. Martin, Rapid Application Development, New York: Macmillan Publishing, 1991.
[13] Nik, M. N., Nor, A. A., & Hazlifah, M. R., "Implementing Rapid Application Development," in Implementing Rapid Application Development (RAD) Methodology in Developing Practical Training Application System, Kuala Lumpur, Malaysia, IEEE, 2010, pp. 1664-1667.
[14] T. Lucidchart, "4 Phases of Rapid Application Development Methodology," 23 May 2018. [Online]. Available: https://www.lucidchart.com/blog/rapid-application-development-methodology. [Accessed 7 11 2020].
[15] D. D. A. R. Ricardo M. Bastos, Extending UML Activity Diagram for Workflow Modeling in Production Systems, Hawaii: IEEE, 2002.
[16] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," Jurnal Edik Informatika, pp. 213-219, 2016.
[17] E. Elisa, "Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti," Jurnal Online Informatika, pp. 36-41, 2017.