ISSN No:-2456-2165

Performance Analysis of Data Mining Classification Method Using Naïve Bayes Algorithm to Predict Student Graduation Timeliness

[1]Nurul Abdillah, [2]Syaiful Zuhri Harahap, [3]Ade Parlaungan Nasution
1 Health Information Management, STIKES Syedza Saintika 2 Informasi System, Faculty of Science & Technology, Labuhanbatu University 3 Management, Faculty of Economics & Business, Labuhanbatu University

Abstract:- Graduation rate is one of the parameters of the effectiveness of educational institutions. The decrease in student graduation rate affects college accreditation. University database stores student administration and academic data, if explored appropriately using data mining techniques, it can be known patterns or knowledge to make decisions. The naive bayes algorithm aims to measure the level of accuracy to be applied in the case of student graduation timeliness. The Naive Bayes method is a classifier with probability and statistical methods to predict future opportunities based on past experience. This research uses student data of Informatics Engineering Education program of Padang State University class of 2011. The variables used in this study were: NIM, name, gender, entry status, GPA, area of origin and employment status. Based on the test results by measuring the performance of the method, it is known that naive bayes has a good accuracy value of 93.48%. From the accuracy value can be concluded that the algorithm naive bayes have a good performance in predicting the timeliness of student graduation.

I. INTRODUCTION

Timely graduation is an important thing that needs to be treated wisely by a college. Graduation rate is one of the parameters of the effectiveness of educational institutions. The decrease in student graduation rate will affect the accreditation of universities. Therefore, it is necessary to monitor and evaluate the tendency to graduate students on time or not. The database of universities stores administrative and academic data of students, such data if explored appropriately using Data Mining techniques then can know the pattern or science to make decisions.

The use of Data Mining classification method to predict the timeliness of student graduation by using Naïve Bayes algorithm can provide information on the accuracy of student graduation timeliness. Data Mining is the process of analyzing data and summarizing the results into useful information. Technically, Data Mining is a process to find correlations between many fields in large datasets[1]. Data Mining has several methods, one of which is the classification method which is a learning technique to classify data items into predetermined class labels. Classification method has several algorithms one of them is Naive Bayes.

The Naïve Bayes method is a simple probabilistic helper that calculates a set of probabilities by calculating the frequency and combination of values in a given data set [3]. Thus the use of Data Mining method will provide the best accuracy results in data classifying. Previous research has examined the performance comparison of several Data Mining classification methods by comparing Decision Tree and Naive Bayes algorithms. The study aims to predict which students drop out. From the results of accuracy testing using both, the highest accuracy is obtained in decision tree algorithms.

Research on the use of Decision Tree algorithms such as J48, Naïve Bayes, Random Tree, and Decision Stump to identify students who are weak and likely to fail high exams. From the tests obtained that J48 algorithm is an algorithm that has the highest accuracy compared to the four algortima used [3].

II. RESEARCH METHODS

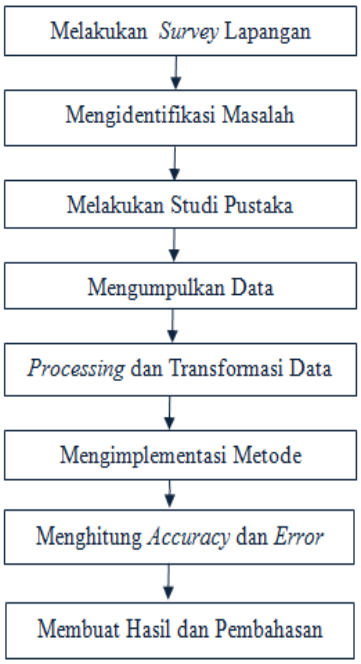

Figure 3.1 Framework

Based on the framework in figure 3.1, each step can be described as follows:

1. Conduct a Field Survey Before starting the research first conducted a survey in the field to get a qualitative picture of the accuracy of student graduation at Padang State University.
2. Identifying Problems The stage of problem identification is the stage at which the research object formulates the problem..
3. Conducting a Literature Study To achieve the objectives to be determined, it is necessary to study a literature used.
4. Collecting Data As for data collection is done in several ways, namely:

a. Direct observation method
b. Interview method
c. Library study method
d. Browsing method

5. Processing and Data Transformation At the Processing and Data Transformation stage, raw data will be converted and combined into the same format to be processed into Data Mining.
6. Implementing the Method After the analysis process, then the next stage of testing is carried out. In testing required computer hardware and software. At this stage will be done implmentation method that has been proposed before, will be tested using RapidMiner Software
7. Calculating Accuracy and Error At this stage will be calculated accuracy and error values from the algorithm Naive Bayes to evaluate the accuracy and error value of the measurement against the actual value or the value is considered correct.
8. Making Results and Discussion Results and discussion aims to provide an overview and results obtained from this research.

III. LITERATURE

3.1 Classification

Classification is a process for finding a model or function that describes or distinguishes concepts or data classes with the aim of estimating the class of an object whose label is unknown. It can also be said to be a learning (classification) that maps an element (item) of data into one of several classes already defined [4].

$$R_{\propto}(T(\propto)) = \min_{T<T_{Max}} R_{\propto}(T) \quad \ldots\ldots\ldots\ldots\ldots\ldots(2.1)$$

Classification is a technique by looking at the behavior and attributes of a defined group. This technique can provide classification to new data by manipulating existing data that has been classified and by using the results to provide a number of rules. These rules are used in new data for classified [4].

3.2 Naïve Bayes

Naive Bayes Clasifier method is one of the algorithms contained in classification techniques. Naive Bayes is a classifier of probability and statistical methods issued by the British scientist Thomas Bayes, a simple probabilistic helper who calculates a set of probabilities by calculating the frequency and combination of values in a given data set. This algorithm uses Bayes Theorem and assumes all attributes to be independent given the class variable value [5].

Bayes' theorem is combined with Naive where it is assumed the conditions between attributes are mutually free. The classification of Naive Bayes assumed that there is or is not a particular tra feature of a class has nothing to do with the characteristics of the other class. Naïve bayes is a simplification of the bayes method. Bayes' theorem is simplified to:

$$P(H|X) = P(X|H)P(X) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.4)$$

Where:

X : Data with unknown class
H : X data hypothesis is a specific class.
P(H| X) : Probability hypothesis H based on condition X (posteriori probability)
P(H) : Probability hypothesis H (prior probability)
P(X|H) : Probability X based on condition on H hypothesis
P(X) : Probability of X

Naïve Bayesian Clasifier can be described as a cluster method based on probability theory and Bayesian Theorem assuming that each variable or decision-making parameter is free (independence) being the existence of each variable has nothing to do with the existence of other attributes[6].

The flow of the Naive Bayes method is as follows:

1. Calculates the chance value of a new case from each hypothesis with an existing class (label) "P(XK| Ci)" .
2. Calculates the accumulated opportunity value of each klas "P(X|Ci)"
3. Calculates the value P(X|Ci) x P(Ci)
4. Specifies the class of the new case.

IV. ANALYSIS AND DESIGN

4.1 Data Mining Analysis Is a series of processes that include the collection, use of data, historically to find regularity, patterns or relationships in large data sets.

4.2 Data Collection In this study the data used is student data of Padang State University Informatics and Computer Engineering Education Study Program in the class of 2011 and 2012. The data used amounted to 46 records.

4.2.1 Variable Selection From student data, which is taken as a variable the decision is to pass on time and late. While taken as the determining variable in the formation of decisions are gender, entry status, GPA, area of origin and job starus.

4.2.2 Pre-Process After selecting a variable, the data format will be transformed based on the selected variables.

4.3 Classification Method The classification results obtained can provide information, about the accuracy level and errors in the timeliness of graduation of students of Padang State University. The use of Naïve

ISSN No:-2456-2165

Bayes Algorithm is done with several stages to get the desired information.

4.3.1 Classification Process Using Naive Bayes

Table 4.2 Results of Probability Calculation Right and Late

No	NIM	Probabilitas		Prediksi
		Tepat	Terlambat	
1	1102628	0,016	0	Tepat
2	1102631	0,003	0	Tepat
3	1102632	0,004	0	Tepat
4	1102638	0,001	0,003	Terlambat
5	1102644	0	0,014	Terlambat
6	1102650	0,001	0,003	Terlambat
7	1102651	0,004	0	Tepat
8	1102656	0,004	0	Tepat
9	1102663	0	0,014	Terlambat
10	1102664	0	0,009	Terlambat
11	1102668	0	0,014	Terlambat
12	1102672	0,008	0	Tepat
13	1102675	0,008	0	Tepat
14	1102676	0	0,004	Terlambat
15	1102678	0	0,004	Terlambat
16	1102687	0	0,01	Terlambat
17	1102688	0,01	0	Tepat
18	1102691	0	0,01	Terlambat
19	1102692	0,01	0	Tepat
20	1102696	0,014	0,018	Terlambat
21	1102697	0,017	0,007	Tepat
22	1102698	0,012	0	Tepat
23	1102703	0,002	0,004	Terlambat
24	1102705	0	0,003	Terlambat
25	1102707	0,002	0,004	Terlambat
26	1106999	0	0,017	Terlambat
27	1107001	0	0,005	Terlambat
28	1107016	0,007	0	Tepat
29	1107017	0,002	0,004	Terlambat
30	1107025	0,008	0,012	Terlambat
31	1107033	0,003	0,006	Terlambat
32	1202175	0,012	0,001	Tepat
33	1202183	0,012	0,001	Tepat
34	1202191	0,012	0	Tepat
35	1202196	0,002	0	Tepat
36	1202197	0,002	0	Tepat
37	1203244	0,015	0,002	Tepat
38	1203237	0,007	0,003	Tepat
39	1203238	0,003	0,001	Tepat
40	1203239	0,003	0,001	Tepat
41	1206507	0,016	0	Tepat
42	1206519	0,006	0	Tepat
43	1206520	0	0,005	Terlambat
44	1206522	0	0,042	Terlambat
45	1206538	0	0,008	Terlambat
46	1206545	0,017	0,007	Tepat

4.3 Accuracy and Error Rate of Naive Bayes Algorithm. In naive bayes accuracy calculation, obtained accuracy rate Naive Bayes algorithm has an accuracy of 93.48%.

Table 4.3 Comparison Table of Accuracy and Error Algorithms C4.5 and Naive Bayes

Algoritma	Akurasi	*Error*
Naive Bayes	93,48%	6,52 %

In naive bayes error calculation, obtained error value naive bayes algorithm has by 6.52 %.

V. IMPLEMENTATION AND RESULTS

In Implementation and Results will be explained Implementation or testing to find out the results of manual calculations with results using software supporting algorithm Naïve Bayes. This aims to see whether the data analyzed and processed is correct or not. The software used is Rapidminer Studio 7.5.3. Rapidminer Studio is an open source Data Mining application. In the case of predicting the timeliness of graduation of these students, the data to be used on Rapidminer amounted to 92 records.

5.1 Naive Bayes Algorithm Accuracy and Error Rates

a. Naive Bayes

Naive Bayes Accuracy Rate In naive bayes accuracy calculation obtained accuracy of 93.48% because it produces 86 correctly classified data.

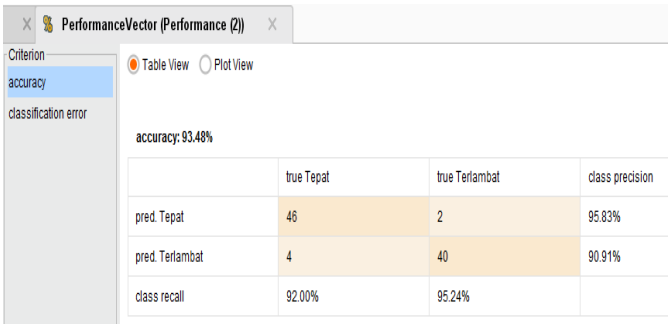

Figure 5.1 Accuracy of Naive Bayes

Naive Bayes Error Rate In Naive Bayes Error calculation obtained accuracy of 6.52% because it produces 6 incorrectly classified data.

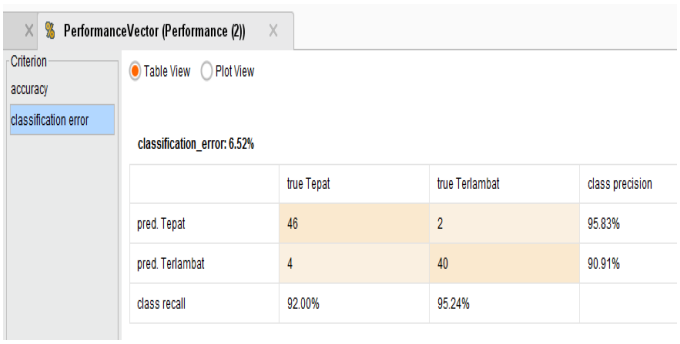

Figure 5.2 Naive Bayes Error

VI. CONCLUSION

1. Measurement of kinejra accuracy of Naive Bayes classification method resulted in an accuracy value of 93.48%.
2. The error rate measurement in Naive Bayes algorithm results in an error rate of 6.52 %.
3. From the tests that have been done, Naive Bayes Algorithm has a good performance because C4.5 has a high accuracy value, the higher the accuracy value, the more accurate the data classifying the closer to correct. Naive Bayes algorithm also has a lower error value, the lower the error value, the classifying the closer it is to true.

REFERENCES

[1]. S. Z. Harahap and M. H. Dar, "APLIKASI DAN PERANCANGAN SISTEM INFORMASI PEMESANAN PADA UPI CONVENTION CENTER DENGAN MENGGUNAKAN BAHASA PEMROGRAMAN PHP DAN MYSQL," *INFORMATIKA*, vol. 6, no. 3, pp. 24–27, 2018.

[2]. M. H. Dar and S. Z. Harahap, "IMPLEMENTASI SNORT INTRUSION DETECTION SYSTEM (IDS) PADA SISTEM JARINGAN KOMPUTER," *INFORMATIKA*, vol. 6, no. 3, 2018.

[3]. M. Siddik and S. Z. Harahap, "APLIKASI PENDUKUNG KEPUTUSAN PUPUK NON SUBSISDI DENGAN METODE STRING MATCHING (STUDI KASUS CV. FAMILY GROUPS LABUHANBATU SELATAN)," *U-NET J. Tek. Inform.*, vol. 3, no. 3, pp. 12–17, 2019.

[4]. A. Nastuti and S. Z. Harahap, "Amelia Nastuti, Syaiful Zuhri Harahap," *Tek. DATA Min. UNTUK PENENTUAN PAKET HEMAT SEMBAKO DAN KEBUTUHAN Hari. DENGAN MENGGUNAKAN Algoritm. FP-GROWTH (STUDI KASUS DI ULFAMART LUBUK ALUNG)*, vol. 7, no. 3, pp. 111–119, 2019.

[5]. S. Samsir, D. Indra, G. Hts, and S. Z. Harahap, "SPK Untuk Pemilihan Kepala Sekolah Menggunakan Metode Saw dan Profile Matching," *U-NET J. Tek. Inform.*, vol. 4, no. 1, pp. 7–12, 2020.

[6]. S. Samsir and S. Z. Harahap, "Application Design Resume Medical By Using Microsoft Visual Basic . Net 2010 At The Health Center Appointments," *Int. J. Sci. Technol. Manag.*, vol. 1, no. 1, pp. 14–20, 2020.

[7]. R. Novita and S. Z. Harahap, "PENGEMBANGAN MEDIA PEMBELAJARAN INTERAKTIF PADA MATA PELAJARAN SISTEM KOMPUTER DI SMK," *INFORMATIKA*, vol. 8, no. 1, 2020.

[8]. M. Nasution, S. Pohan, and S. Z. Harahap, "Implementasi Obrim (Option-Based Risk Management) Sebagai Framework Investasi Teknologi Informasi Perguruan Tinggi (Studi Kasus : Amik Labuhan Batu)," *INFORMATIKA*, vol. 8, no. 1, pp. 26–35, 2020.

[9]. P. Iwan, S. Z. Harahap, and A. A. Ritonga, "RANCANG BANGUN TEMPAT SAMPAH OTOMATIS PADA UNIVERSITAS LABUHANBATU," *INFORMATIKA*, vol. 8, no. 2, pp. 1–5, 2020.

[10]. S. Z. Harahap and S. Samsir, "Application Design The Data Collection Features Of The Hotel Shades Of Rantauprapat Using VBNET," *Int. J. Sci. Technol. Manag.*, vol. 1, no. 1, pp. 1–6, 2020.