RESEARCH ARTICLE

the International Journal of Education, Theology, and Humanities

ISSN: 2828-5492

Data Mining Analysis Using Naive Bayes Algorithm and Knn to Predict Graduation of D3 Students Department of Information Management

Jonas Franky Rudianto Panggabean, Kamson Sirait, Frainskoy Rio Naibaho

ABSTRACT

In this study, the specific objective of this research is to obtain the results of decisions in predicting students of the informatics management study program whether they can graduate on time for 3 years or more in the specified time and with a minimum GPA of 3.00. Student graduation is one of the Internal Quality Assurance Standards on campus or college. To achieve a decent quality of graduation, a study is needed to be able to predict the graduation rate with predetermined standards, so as to reduce and anticipate problems in the academic field that occur. In this study, data mining methods are used to predict the passing rate and standard GPA with a classification function. The algorithm used in this study is Naïve Bayes and K-NN, the stages used in the application of this research are KDD starting with selecting, preprocessing, transformation, data mining and evaluation/interpretation stages. The final result of this study using the K-NN and Naïve Bayes algorithms is that K-NN produces an accuracy rate of 94.81% and Nave Bayes produces an accuracy rate of 90.49%, so it can be concluded that the K-NN algorithm is better used to predict graduation of D3 students majoring in informatics management.

Keywords: graduation, K-NN, KDD, Naïve Bayes, Data Mining

Published Online: December 2022

ISSN: 2828-5492

Jonas Franky Rudianto Panggabean*

AMIK Medicom Medan
Email:jonasfrankypanggabean@gmail.com

Kamson Sirait
AMIK Medicom Medan
Email: kamsonsirait@gmail.com

Frainskoy Rio Naibaho
IAKN/Fakultas Ilmu Pendidikan Kristen
Email: frainskoy.rio.naibaho@gmail.com

Corresponding Author

I. INTRODUCTION

Student graduation is the most important thing in tertiary institutions, where graduation is one of the standards of success for colleges and alumni, to achieve graduation at a university, they must follow the Internal Quality Standards (SPMI), one of the standards set by the Management study program. AMIK Medicom informatics is to produce graduates on time for 3 years or 6 semesters with a minimum GPA of 3.00. There is a difference between the number of students entering and leaving, so a decision-making research is needed in predicting graduation times and standard grades to achieve the specified GPA. Graduating on time is very useful not only for students but also for the campus, where the party will increase the accreditation and quality of the campus, as well as for students it will be easier and faster to find work or continue their education to a higher level [1].

In this study using data mining, where the variable determined is the graduation time for 3 years with a minimum standard GPA of 3.00. The data mining method uses

mathematics, artificial intelligence, machine learning that is useful in identifying data on a large scale [2]. Then search for supporting values that influence each other in graduation [3]. So the suitable algorithms are KNN and Naives Bayes. The KNN algorithm works based on the shortest or closest distance [4]. While the algorithm In the application of previous research conducted by [5], the prediction accuracy using the KNN algorithm method was 70%. Meanwhile, Naive Bayes is the most efficient and effective in data mining [6].

II. THEORETICAL BASIS

A. Data Mining

The data mining method uses mathematics, artificial intelligence, machine learning that is useful in identifying data on a large scale [7].

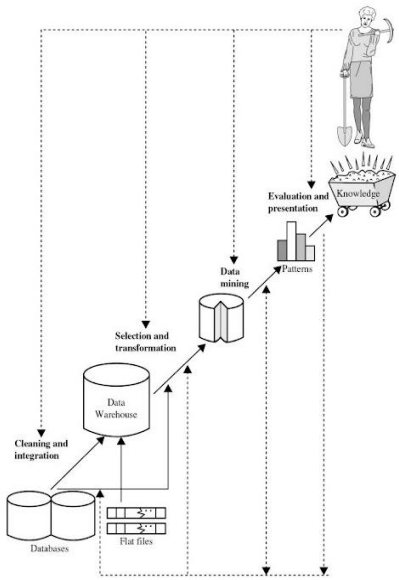
Data mining has functions in several areas, namely:

1. Description
2. Estimate
3. Prediction
4. Classification

B. Knowledge Discovery in Database (KDD)

When you submit your final version, after your paper has been accepted, prepare it in two-column format, including figures and tables.

Fig. 1. Stages of the KDD Process.



C. K-Nearest Neighbor (KNN)

The K-Nearest Neighbor (KNN) method is a method used in classifying objects with the closest rarity so that a new object classification can be generated. The formula for equation (1) is the distance formula used in KNN [8]:

$$dEuclidian\,(x,y) = \sqrt{\sum}\,(xi - yi)\,2\,i \qquad (1)$$

D. Naïve Bayes

This classification algorithm is for predicting return values, where the effect of these values determines the value of other attributes independently. Naive Bayes formula [9]. shown in the following equation (2):

$$P(Y \mid X) = P(Y) \prod P(Xi \mid Y) \qquad (2)$$

Where :

P(X | Y) : vector X in class Y as probability

P(Y) : class Y in vector X as the start of independent probability class

E. Confusion Matrix

The confusion matrix consists of data sets, namely positive class and negative class

III. RESEARCH METODHS

This research method uses knowledge discovery in database (KDD). The following is an explanation of each stage of the research from figure 2.

A. Selection

This stage selects graduates who are on time for 3 years and have a minimum GPA of 3.00 which will be the target variable and gender, GPA semesters 1 to 4 will be the predictor variable

B. Prepocessing

The data is taken based on the student database and data is cleaned from duplicate data

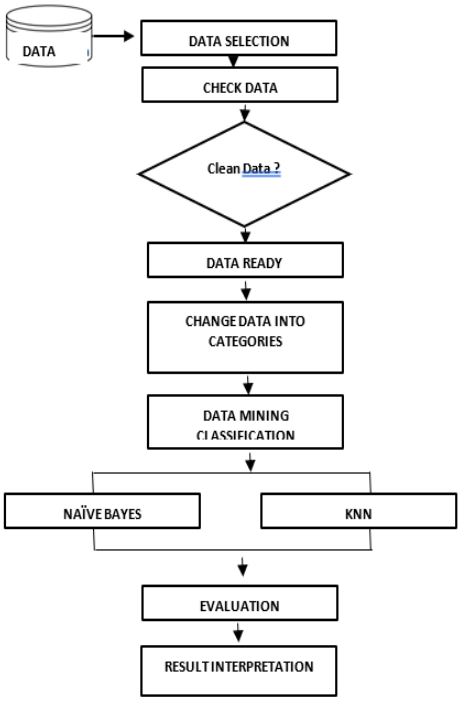


Fig. 2. Stages of research using the KDD method.

C. Transformation

At this stage there are no more data errors found, so at this stage what is done is data transformation where the types of data will be grouped by category. In the table there are 2 categories, namely target and predictor variables.

TABLE I : TARGET AND PREDICTOR VARIABLES

Category Target Variables	Category Target Variables
Graduation and GPA Appropriate (Graduated 3 years with GPA >= 3.00)	Graduation and GPA Appropriate (Graduated 3 years with GPA >= 3.00)
It is not in accordance with	It is not in accordance with
Category Predictor Variables	Category Predictor Variables
Male Gender	Male Gender

D. Data Mining

At this stage, the data technique is selected according to the classification function on the KNN and Naïve Bayes algorithms.

E. Evaluation

This stage aims to evaluate the prediction results that have been generated by the two algorithms. Confusion Matrix method is used to evaluate, while the performance values used are accuracy and error.

IV. RESULT AND ANALYSIS

A. Selection

Sources of data taken from student databases such as data on grades and graduation of the department of informatics management at AMIK Medicom from 2010 to 2015.

B. Prepocessing

Compiling and sorting the data again so that there is no duplicate data, so that it can

produce the highest value taken.

TABLE II : EXAMPLES OF STUDENT INITIAL DATA THAT HAVE NOT BEEN TRANSFORMED

No	NIM	GENDER	I-1	I-2	I-3	I-4	GPA	STATUS	APPROPRIATE	>= 3,00	DESCRIPTION
1	153100001	M	3,2	3,4	3,1	3,5	3,3	Graduated	3 yrs	✓	In accordance
2	153100002	LM	3,0	3,0	3,3	3,2	3,125	Graduated	3 yrs	✓	In accordance
3	153100003	W	2,6	2,5	2,7	2,8	2,65	Graduated	3 yrs	✓	In accordance
4	153100004	W	3,7	3,9	3,9	3,9	3,85	Graduated	3 yrs	✓	In accordance
5	153100005	W	3,5	3,2	3,2	3,1	3,25	Graduated	3 yrs	✓	In accordance
6	153100006	W	2,4	2,0	0	0	1,1	Not pass	3 yrs	×	Not in accordance
7	153100007	M	3,5	0	0	3,1	1,65	Not pass	3 yrs	×	Not in accordance
8	153100008	M	3,1	3,1	3,2	3,1	3,125	Graduated	3 yrs	✓	In accordance
9	153100009	W	3,5	3,6	3,4	3,7	3,55	Graduated	3 yrs	✓	In accordance
10	153100010	W	3,9	3,8	3,9	3,9	3,875	Graduated	3 yrs	✓	In accordance

C. Transformation

The transformation is done by making the initial data into a table and then it will be processed into several types of data starting from the overall GPA from semester 1 to semester 6. The processed dataset is in Table 3.

TABLE III : TRANSFORMED DATA

No	NIM	GENRE	I-1	I-2	I-3	I-4	DESCRIPTION
1	153100001	L	✓	P	P	P	In accordance
2	153100002	L	P	P	P	P	In accordance
3	153100003	P	×	O	O	O	In accordance
4	153100004	P	✓	P	P	P	In accordance
5	153100005	P	P	P	P	P	In accordance
6	153100006	P	×	O	O	O	Not in accordance
7	153100007	L	P	O	O	P	Not in accordance
8	153100008	L	✓	P	P	P	In accordance
9	153100009	P	P	P	P	P	In accordance
10	153100010	P	P	P	P	P	In accordance

D. Data Mining

The application used to manage data is Rapidminer with a lot of 500 data. The data transformation process uses the excel application, then cross validation is carried out with the k-Fold Cross Validation process.

K-Fold Cross validation is a method used for performance evaluation where the data is divided into two, namely training data and test data. With the value of k taken 20 fold so that 500 data becomes 25 subsets of data to be tested.

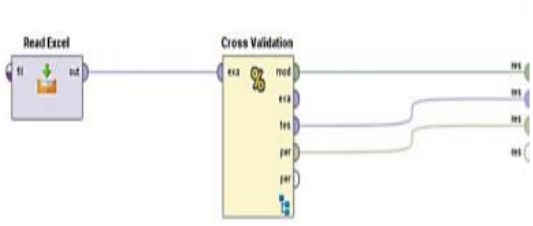


Fig. 3. Data retrieval process for Naive Bayes algorithm and kNN

E. Classification Calculation Results with RapidMiner

Algoritma Naïve Bayes. The classification technique applied to the Naïve

Bayes algorithm training data is shown in Figure 4.

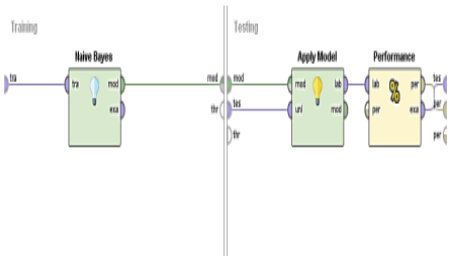


Fig. 4. Naive bayes algorithm classification.

The output results in this algorithm are the "suitable" and "incompatible" classifications made in the prediction table 5 confusion matrix

TABLE IV : CONFUSION MATRIX NAVE BAYES ALGORITHM

	PREDICTIONS ON TIME	PREDICTIONS NOT ON TIME
TRUE ON TIME	408	42
TRUE NOT ON TIME	22	28

Information:

1) Number of data true on time with prediction on time = 408
2) Number of data true on time with predictions not on time = 42
3) The number of true data not on time with predictions not on time = 22
4) The number of true data not on time with predictions not on time = 48

The evaluation used in this method is the value of accuracy and error.

Accuracy results are:

$$Accuracy = \frac{408 + 28}{408 + 42 + 22 + 28} = 87{,}2\%$$

$$Error = \frac{22 + 42}{408 + 42 + 22 + 28} = 12{,}8\%$$

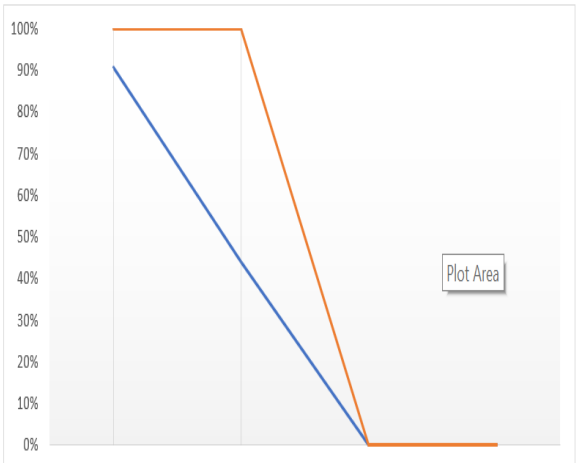


Fig. 5. Graph of the accuracy of the nave Bayes algorithm.

Algoritma KNN

The classification technique on the training data used in the KNN algorithm with K taken as many as 8.

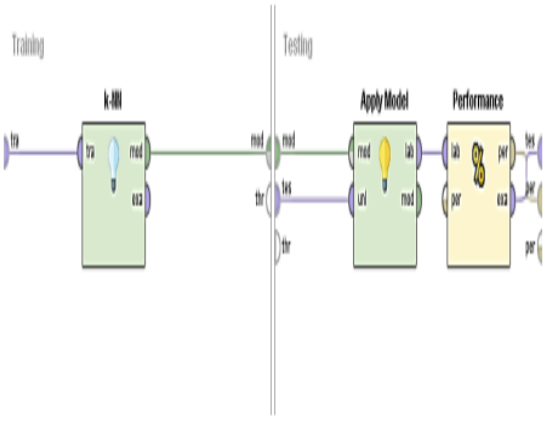


Fig. 6. KNN algorithm classification.

The ouput result in this algorithm are the " suitable" and "incompatible" classifications made in the prediction table 5 confusion matrix.

TABLE V : CONFUSION MATRIX NAVE BAYES ALGORITHM

	PREDICTIONS ON TIME	PREDICTIONS NOT ON TIME
TRUE ON TIME	398	37
TRUE NOT ON TIME	32	33

Information:

1) Number of data true on time with prediction on time = 408
2) Number of data true on time with predictions not on time = 42
3) The number of true data not on time with predictions not on time = 22
4) The number of true data not on time with predictions not on time = 48

The evaluation used in this method is the value of accuracy and error.

Accuracy results are:

$$Accuracy = \frac{398 + 37}{408 + 42 + 22 + 28} = 87\ \%$$

$$Error = \frac{32 + 37}{408 + 42 + 22 + 28} = 13{,}8\ \%$$

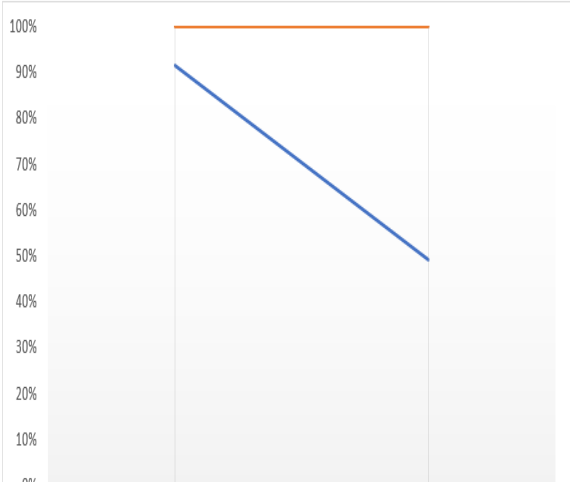


Fig. 7. Graph of the accuracy of the KNN algorithm.

V. CONCLUSION

The final results of each method, namely accuracy and error, will be compared so that it can be concluded which algorithm is better at predicting student graduation with a graduation time of 3 years and a GPA >= 3.00.

TABLE VI: COMPARISON OF NAÏVE BAYES AND KNN . CONCLUSIONS

Algorithm	Accuracy	Error
Naïve Bayes	87,2 %	12,8 %
KNN	87 %	13,8 %

The conclusions obtained from this research are:

1. All student graduation information is easily obtained when using data mining Naives Bayes and KNN algorithms are suitable for predicting student graduation.
2. 2 From the calculation results, it is obtained that a higher level of accuracy is used by using the Nave Bayes algorithm with an accuracy rate of 87.2% and an error of 12.8% smaller than the KNN algorithm.

REFERENCES

Azahari, A., Yulindawati, Y., Rosita, D., & Mallala, S. (2020). Komparasi Data Mining Naive Bayes dan Neural network memprediksi Masa Studi Mahasiswa S1. Jurnal Teknologi Informasi Dan Ilmu Komputer, 7(3), 443. https://doi.org/10.25126/jtiik.2020732093

Gorunescu, Florin. (2011). Data Mining Concept ,Model Technique.Springer-Verlag Berlin Heidelberg

Lizsara, P. A., Oyama, S., & Wardani, S. (2020). Implementasi Data Mining Menggunakan Metode Naive Bayes Untuk Memprediksi Ketepatan Waktu Tingkat Kelulusan Mahasiswa (Study Kasus: Program Studi Informatika

Universitas PGRI Yogyakarta). Seri Prosiding Seminar Nasional Dinamika Informatika, 4(1), 34–37.

Lizsara, P. A., Oyama, S., & Wardani, S. (2020). Implementasi Data Mining Menggunakan Metode Naive Bayes Untuk Memprediksi Ketepatan Waktu Tingkat Kelulusan Mahasiswa (Study Kasus: Program Studi Informatika Universitas PGRI Yogyakarta). Seri Prosiding Seminar Nasional Dinamika Informatika, 4(1), 34–37

Murtopo, A. A. (2015). Prediksi Kelulusan Tepat Waktu Mahasiswa STMIK YMI Tegal Menggunakan Algoritma Naive Bayes Time Graduation Prediction by Using Naive Bayes Algorithm at STMIK YMI Tegal. Vol.7 No.3, 145–154

Pratama, A., Wihandika, R. C., & Ratnawati, D. E. (2018). Implementasi Algoritme Support Vector Machine (SVM) untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa. 2(4), 1704–1708.

Resti Hutami1, E. Z. A. (2016). Implementasi Metode K-Nearest Neighbor

Sillueta, C.Y. (2016). Implementasi Data Mining Untuk Memprediksi Kelulusan Mahasiswa Dengan Metode Klasifikasi Dan Algoritma Knearest Neighbor Berbasis Desktop (Studi Kasus : Fakultas Teknologi Informasi, Program Studi Teknik Informatika, Tugas Akhir

Suyatno. (2017) Data Mining Untuk Klasifikasi dan Klasterisasi Data. Bandung: Informatika.