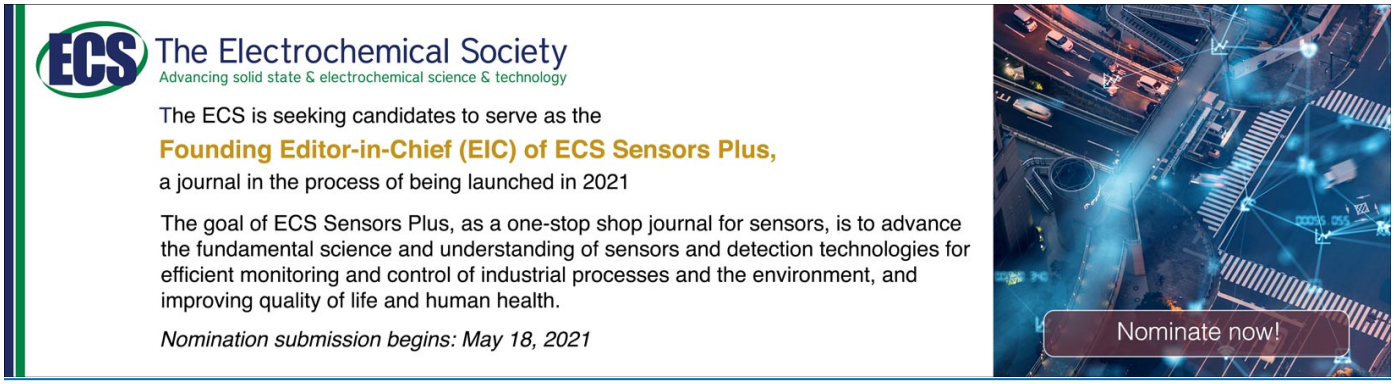
Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Academic achievement analysis of Universitas Negeri Semarang students using the naïve bayes classifier algorithm

To cite this article: A Purwinarko *et al* 2021 *J. Phys.: Conf. Ser.* **1918** 042130

View the article online for updates and enhancements.



This content was downloaded from IP address 36.80.199.194 on 15/06/2021 at 04:45

Academic achievement analysis of Universitas Negeri Semarang students using the naïve bayes classifier algorithm

A Purwinarko[1,*], W Hardyanto[2], and N P Aryani[2]

[1]Department of Computer Science, Universitas Negeri Semarang, Indonesia
[2]Department of Physics, Universitas Negeri Semarang, Indonesia

*Corresponding author: aji.purwinarko@mail.unnes.ac.id

Abstrak. This study aims to evaluate the academic data of students in third year and classified in the category of students who can graduate on time or not. This system uses students master data and student academic data as input data. Data of students who have graduated will be used as training and testing data while the data of students who have passed will be used as target data. Input data will be processed using the Naïve Bayes Classifier (NBC) data mining algorithm to form a probability table as the basis for the student graduation classification process. The output of this system is in the form of a classification of student academic performance that is predicted to pass and provides recommendations for the graduation process on time or graduates in the most appropriate time with optimal grades.

1. Introduction

The Monitoring student of ability, student achievement, graduation ratio, and graduate competencies should receive serious attention to gain stakeholder confidence in assessing and determining the use of graduates [1]. One indicator of success in a college can be seen from the many graduation rates of students each year [2].

The level of graduation of students is very likely to be predicted; one of the factors used is the value of student outcomes. The recapitulation process of students' academic scores has been routinely carried out in each semester [3]. So, to predict student graduation patterns, a technique called data mining is needed.

Data mining is the process of finding interesting patterns or information in selected data using specific techniques or methods [4]. Data mining in the world of education is known as Educational Data Mining [5]. One of the classification methods in data mining is the Naïve Bayes Classifier algorithm. Naïve Bayes Classifier (NBC) is a classification algorithm by calculating the probability value for each occurrence of the target attribute in each case [6]. Input data will be processed using the to form a probability table as the basis for the student graduation classification process. The output of an academic student's classification performance is predicted to pass and provide recommendations for the graduation process on time or graduates in the most appropriate time with optimal grades.

2. Method

Figure 1 shows the research flow chart. The study used data from Universitas Negeri Semarang (UNNES) students for the year 2016, amounting to 5,653 data.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

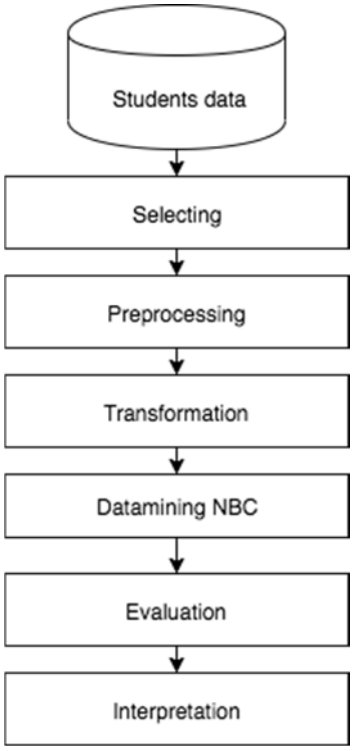


Figure 1. The research flow chart

2.1 Selecting

Data was collected from UNNES student academic system data. The data used for implementation has 16 attributes (Student Identification Number, Number of Subjects, GPA, GPA status, Academic Year, Student Name, Study Program Code, Study Program Name, Degree, Graduation Status, Semester 1, Semester 2, Semester 3, Semester 4, Semester 5, Gender).

2.2 Pre-processing

After selecting the data to be used, the next process is pre-processing. Pre-processing is a stage in the text classification process, and improving the effectiveness of classification is mostly determined by the selection of appropriate pre-processing techniques [7].

2.3 Transformation

The transformation process is a process to convert data into standard forms so that the data can be used for further processing [8].

2.4 Datamining NBC

Naïve Bayes Classifier algorithm is a classification algorithm that is simple, effective, and excellent performance [9, 10]. Equation 1 is a general form of the Bayes theorem [11].

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

Where:

X = Data with unknown classes
H = Data hypocryption X which is a specific class
$P(H|X)$ = The probability of H hypocryption based on the X condition
$P(H)$ = Probability of H hypocryption
$P(X|H)$ = Probability of X base on H condition
$P(X)$ = Probability of X

2.5 Evaluation

The evaluation is based on accuracy, precision, Area Under Curve (AUC), and Receiver Operating Characteristics (ROC).

2.6 Interpretation

Interpretation is a valuable asset in research [12]. Interpretation is the process of visualizing patterns, models, or data provided by such models, and, in the case, iteratively by reviewing the stages of the previous process [13].

3. Result and Discussion

This research was conducted using RapidMiner software. RapidMiner is a java-based software developed by the same company. RapidMiner provides integrated services for machine learning, data mining, text mining, predictive analytics, and business analytics [14]. The RapidMiner company developed this software.

In this research, a node utility is performed by discretizing data and calculating the estimated cross validation accuracy using Naive Bayes at each node use value of k = 2, 4, 6, 8, and 10 [15]. The performance results of this algorithm are in the form of classifying the testing data of students who fall into the class group "not on time" or "on time". The accuracy of the data predicted by using different k values in the Naïve Bayes algorithm is shown in Table 1.

Table 1. The accuracy of the data predicted by using different k values.

K-fold	Accuracy (%)	true "not on time"	true "on time"
2	94.32	1609	3723
4	94.27	1610	3719
6	94.22	1609	3717
8	94.30	1610	3721
10	94.23	1610	3717
20	94.23	1610	3717

Table 1 shows the results of research predictions with k = 2, having an accuracy rate of 94.32%. The results of forecasts that showed the right pass "not on time" amounted to 1609, while the correct predictions showed the pass on time amounted to 3723. The percentage of accuracy is more influenced by the predictive data that passed on time. The utilization of k-fold crossover validation is to reduce errors originating from random sampling and comparing it with the accuracy of several prediction models [16]. The precision of this calculation is 98.96%, as shown in Figure 2. Precision is the most critical characteristic of the classifier process [17].

precision: 98.96% +/- 0.06% (micro average: 98.96%) (positive class: TEPAT WAKTU)

	true TIDAK	true TEPAT WAKTU	class precision
pred. TIDAK	1609	282	85.09%
pred. TEPAT WAKTU	39	3723	98.96%
class recall	97.63%	92.96%	

Figure 2. Precision.

The Receiver Operating Characteristic (ROC) graph is an illustration of the relative tradeoff used to visualize the classifier's performance. ROC charts are very useful with unbalanced datasets between positive and false positive levels. One well-known method for classifying is to calculate the area under the ROC (AUC) curve. AUC always takes values between 0 (worst) and 1 (best). Figure 3 shows the AUC value of 0.992, and this shows that the classification results are better because it meets the AUC requirements greater than 0.5 [18].

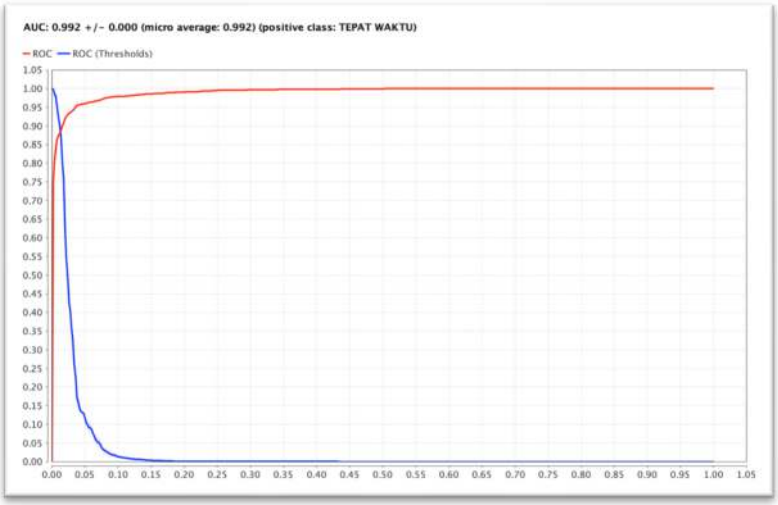


Figure 3. AUC graph

4. Conclusion

The testing process is used to predict the accuracy of student graduation rates. In this study, the selection of k-fold will determine the level of classification accuracy. The precision in this study is an essential part of the classification process because it illustrates the accuracy of predictions of passing on time correctly. From the ROC curve and the AUC value, it can be concluded that the performance of the Naïve Bayes Classifier algorithm for the case of this dataset can predict near perfect.

Reference

[1] Ridwan M, Suyono H & Sarosa M 2013 *EECCIS*, **7**(1) 59
[2] Nasution N, Djahara K & Zamsuri A 2015 *J Teknol. Inf. Komun.* **6**(2) 1
[3] Alfajri R M, Chrisnanto Y H & Yuniarti R 2016 *Proc. Seminar Nasional Sains dan Teknologi* (Semarang, 3 Agustus 2016), Universitas Wahid Hasyim Semarang **1**(1) 144
[4] Gunadi G, & Sensuse D I 2016 *J Telematika MKOM* **4**(1) 118
[5] Nugroho M F 2017 *J Inform. UPGRIS* **3**(1) 63
[6] Rahman A A & Kurniawan Y I 2018 *J Teknol. Manaj. Inform.* 4(1)

[7] Symeonidis S, Effrosynidis D & Arampatzis A 2018 *Expert Syst. Appl.* **110** 298
[8] Saputra M F A, Widiyaningtyas T & Wibawa A P 2018 *JOIV: Int. J. Inform. Vis.* **2**(3) 153
[9] Chandrasekar P & Qian K *2016 Proc. 40th IEEE Annual Computer Software and Applications Conference (COMPSAC)*(Atlanta, GA, USA, June 10-14, 2016) IEEE **2** 618
[10] Jabbar M A & Samreen S 2016 *2016 Proc. Int. Conf. on Circuits, Controls, Communications and Computing* (I4C)(Bangalore, India, Oct 4-6, 2016) IEEE **1** 1
[11] Gerhana Y A, Fallah I, Zulfikar W B, Maylawati D S & Ramdhani M A 2019 *J. Phys.: Conf. Ser.,* **1280** 022022
[12] Hameed M A & Meskele F 2019 *INFOCOMP J. Comput. Sci.* **18**(2) 1
[13] Gullo F 2015 *Phys. Procedia* **62** 18
[14] Arunadevi J, Ramya S & Raja M R 2018 *Int. J. Pure Appl. Math.* **119**(12) 15977
[15] Rakholia R M & Saini J R 2017 *Indian J. Sci. Technol.* **5** 1
[16] Hazra A, Mandal S K & Gupta A 2016 *Int. J. Comput. Appl.* **145**(2) 0975
[17] Granik M & Mesyura V 2017 *2017 Proc. IEEE First Ukraine Conference on Electrical and Computer Engineering* (UKRCON) (Kyiv, Ukraine, May 29, 2017) IEEE **1** 900
[18] Kim T, Do Chung B & Lee J S 2017*Computing* **99**(3) 203