Implementation Naïve Bayes Algorithm for Student Classification Based on Graduation Status

Ayundyah Kesumawati[1)], Din Waikabu[2)]

Statistics Department, Faculty of Mathematics and Natural Sciences,
Islamic University of Indonesia (UII), 55584 Sleman, Yogyakarta, Indonesia)
[1]ayundyah.k@uii.ac.id, [2]12611071@students.uii.ac.id

ARTICLE INFO

Article history:
Received April 12, 2017
Revised June 04, 2017
Accepted June 20, 2017

Keywords:
Naive Bayes
Graduation Status
Classification

ABSTRACT

Length of study in collage is a time it takes a student to have completed the study in college. Bachelor degree in achieving normal that it takes time for four years, but still there are students who completed their studies beyond normal limits (over four years). This such as influence on the value of accreditation of institution. In this paper we used five variables: grade point average (GPA), Concentration in High School, Sex, participation in assistance and city of residence, which are classified by the Graduation Status students over four years and less than equal to four years. The method used for the classification of a student's study time is Naive Bayes algorithm. This study investigated classification student based on Graduation Status in Department Statistics of Islamic University of Indonesia. From the result, Naïve Bayes algorithm classification is quite good with accuracy value for Naïve Bayes is 81,18%.

I. Introduction

Higher education in Indonesia are under Directorate of Higher Education and accredited by the National Accreditation Board for Higher Education. The higher education institution is categorized into two types: public and private higher education [1]. There are four types of higher education institutions: universities, institutes, academies and polytechnics. Universities in Indonesia are largely offered by the private sector. Out of around 3,500 institutions, only around 150 institutions are public (established and operated by the government) [2]. Islamic University of Indonesia is a private university in Indonesia. There are 9 faculties, and 46 program under faculty. Department of statistics is one of the program under Mathematics and Natural Science Faculty.

One of the standard quality from higher education is based on the student body it means the comparison student and lecturer. Expectation for any higher education is afford alumni with high quality. One of criteria for a high quality can describe by graduation status of the student in a institution. The graduation status can affect the quality of alumni especially for Statistics Department. This situation will affect the quality of the Department of statistics, especially courses in Indonesia measured by the accreditation conducted by the National Accreditation Board of Higher Education (BAN PT).

Quality is measured by seven major standards, is one of its students. Especially with regard to the evaluation of the student standard component of the assessment is the grade point average and the duration of the study [3]. Based on the assessment matrix instrument of accreditation that the percentage of students who graduate on time is one element of the accreditation assessment [4]. This issue is extremely important for the management of department considering the percentage of students graduating on time is one of the elements of accreditation set by the National Accreditation Board.

One of method that can used to solve this problem is detected factor that implied Graduation Status. In this paper used five variable that supposed to be a factor that implied Graduation Status of student which are grade point average (GPA), Concentration in High School, Sex, participation in assistance and city of residence.

Classification method used in this research method is Naive Bayes Algorithm and Naïve Bayes is the best classifier against several common classifiers in term of accuracy and computational efficiency [5].

The remaining parts of this paper are organized in the following structures. In section 2, some backgrounds of Naïve Bayes Algorithm and accuracy classification will be reviewed. The methodology and result which are used accuracy for classification are presented in section 3. In section 4, numerical experimental Naïve Bayes algorithm is presented. Finally, discussions and conclusions are presented in section 5.

II. Methodology

This section discusses on techniques that have been performed in this study.

A. Data

This research was conducted in the Department of Statistics Faculty of Mathematics and Natural Sciences, Islamic University of Indonesia, the data used in this research is registration data from student in department of statistics Islamic University of Indonesia and alumni data since 1997 - 2012. The number of data used is 404 alumni since 1997 - 2011 and 116 active student class of year 2012. In this paper used 5 variable which is graduation status, gender, city of residence, concentration in high school, assistances, and Grade Point Average (GPA). The category for each variable as follow :

Table 1. Variabel Category

Variable	Description	Category	Description
Variable Dependent	Graduation Status	On Time	The student that complete their study for 4 years (8 semester)
		Not On Time	The student that complete their study beyond 4 years
Variable Independent	Sex	Male	
		Female	
	City of Recidence	Yogyakarta and Central Java	
		Outside Yogyakarta and Central Java	
	Concentration in High School	Science	
		Social	
		Other	
	Assistance	Yes	
		No	
	GPA	Plenty	< 2,75
		Satisfy	2,75 – 3,00
		Very Satisfy	3,01 – 3,51
		Cumlaude	> 3,51

2.2. Naïve Bayes Algorithm

One of the statistical classifier are Bayesian Classifier. They can predict class membership probabilities, such as the probability that a given training sample belongs to a particular class [6]. So many studies comparing by performance decision tree and selected neural network classifier with A simple Bayesian classifier known as the Naïve Bayes Classifier. Attribute value on a class is assumed independent in Naïve Bayes Classifier or it called class conditional independence. The Naïve Bayes. Naive Bayes is based on Bayes theorem which has a similar classification capability with decision tree and neural network. For large data implementation Naive Bayes proved to have high accuracy and speed.

Bayes Theorem has the following general form:

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

with :

X	:	Data with unknown class
H	:	Hypothesis X data is a specific class
P(H|X)	:	The probability of the hypothesis H is based on the condition X
P(H)	:	The probability of the hypothesis H (prior prob.)
P (X|H)	:	The probability of X on the condition
P(X)	:	The probability of X

Bayes formula can be expressed informally by saying that

$$Posterior = \frac{likelihood \times prior}{evidence}$$

We know how frequently some particular evidence is observed, given a known outcome. We can use this known fact to compute the reverse, to compute the chance of that outcome happening, given the evidence

The Naïve Bayes classifier, works as follows [5]:

Given a set of labelled training samples and their associated class label. As usual, each training sample is represented by an n-dimensional attribute vector, $X = (x_1, x_2, ..., x_n)$, depicting n measurements made on the training sample from n attributes, respectively, $A_1, A_2, ..., A_n$.

Suppose that there are m classes, $C_1, C_2, ..., C_m$. Given a training sample, *X*, the classifier will predict that *X* belongs to the class having the highest posterior probability, conditioned on *X*. That is, the Naïve Bayes Classifier predicts that training sample X belongs to the class C_i if and only if

$P(C_i \mid X) > P(C_j \mid X)$ for $1 \le j \le m, j \ne i$

Thus we maximize $P(C_i \mid X)$. The class C_i for which $P(C_i \mid X)$ is maximized is called the maximum posteriori hypothesis. By Bayes theorem (equation (1)),

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \quad (2)$$

Suppose P(X) is equal for all classes, only $P(X \mid C_i)P(C_i)$ maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = ... = P(C_m)$, and we would therefore maximize $P(X \mid C_i)$. Otherwise, we maximize $P(X \mid C_i)P(C_i)$. Note that the class prior probabilities may be estimated by $P(C_i) = |C_{i,D}| / |D|$, where $|C_{i,D}|$ is the number of training samples of class C_i in *D*.

If data sets with many attributes given, it would be extremely computationally hard to compute $P(X \mid C_i)$. In case to reduce computation for evaluating $P(X \mid C_i)$, the Naïve Bayes assumption of class conditional is independence. To make presumes that the values of the attributes are conditionally independent of one another, given the class label of the training sample. Thus,

$$P(X \mid C_i) = P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times ... \times P(x_n \mid C_i) \quad (3)$$

We can easily estimate the probabilities $P(x_1 \mid C_i),..,P(x_n \mid C_i)$ can be estimated from the training samples, where

If Ak is categorical, then $P(x_k \mid C_i) = \mathrm{S}_{ik} / S_i$, where S_{ik} is the number of training samples of class Ci having value x_k and S_i is the number of training samples belonging to C_i.

If Ak is continuous-valued, then the attribute is assumed to have a Gaussian distribution so that $P(x_k \mid C_i) = g(x_k, \mu c_i, \sigma c_i)$ where $g(x_k, \mu c_i, \sigma c_i)$ is Normal density function for attribute Ak, while μc_i and σc_i are the mean and standard deviation, respectively.

In order to classify an unknown data input X, $P(\mathrm{X} \mid C_i)\,\mathrm{P}(\mathrm{C}_i)$ is evaluated for each class Ci.

2.3. Classification Accuracy

The classification accuracy A_i of an individual program *i* depends on the number of samples correctly classified (true positives plus true negatives) and is evaluated by the formula. [6]

$$A_i = \frac{true\,classification\,statement}{number\,of\ statement} \times 100\% \quad (4)$$

III. Numerical Experimental Result

In this section describe the results of classification for Statistics Department Student based on Graduation Status including Naive Bayes models, and the results of the classification. First of all, a datasets with 404 alumni classified in two different categories is used for evaluation.

A. Descritive Statistics

In this section describe the descriptive statistics from the data. Alumni in Statistics Department are spread throughout Indonesia as follows :

Fig. 1. Distribution Alumni of Statistics Department

Based on Figure 1., the highest city residence of alumni is from Central Java and Yogyakarta. In this paper, for city residence divided by two region are city residence from Central Java and Yogyakarta, and outside Central Java and Yogyakarta.

The percentage for the graduation status of a student department of Statistics as follows :

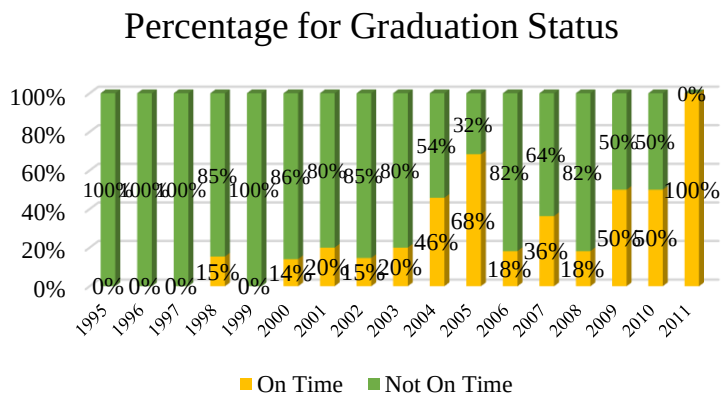

Fig. 2. Percentage for Graduation Status

From Figure 2, it can be conclude that from 1995 until 2011 there are several data that showed a 100% that the student did not completed their study in a less than equal to four years.

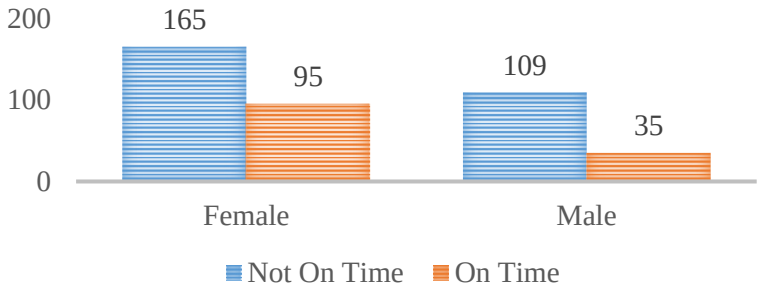

Fig. 3. Graduation Status Based on Sex

Based on figure 3, it can be conclude the student that female student has an higher number for Graduation status not on time and on time than a male student. It's possible happen, because there are many female student than male student so it implied this condition.

B. Naïve Bayes Classifier

In applying the Naive Bayes Classifier the selected dataset contains two categories of data : Graduation on Time and Graduation Not On Time. 30% data (404 data) are used to builds the training dataset for the classifier. The other (116 data) data are used as the testing dataset to test the classifier. The data used in this paper as follow :

Table 2. Data Training

No	Sex	City of Residence	Concentration High School	GPA	Graduation Status
1	Male	Yogyakarta and Central Java	Science	Very Satisfy	Not On Time
:					
404	Female	Outside Yogyakarta and Central Java	Science	Cumlaude	On Time

The data in table 2 proceed by using R software and Weka 3.8 to solve the problem. Then, in this section the prior probability of each class in graduation status (Y) and the conditional probability for each independent variable (except : Graduation Status) to Y showed in output R program. The output program from R as follow :

Table 3. Prior Probability for Graduation Status

Prior Probability	
Not On Time	**On Time**
0.6782178	0.3217822

The prior probability for graduation status in table 3 show that the biggest probability is graduation not on time for this case. It implied the conditional probability for each variable as follow :

Table 4. Conditional Probability for Graduation Status

Variable (X)	**Categories**	**Probability of Graduation Status (Y)**	
		Not on Time	**On Time**
Sex	Male	0.6021898	0.7307692
	Female	0.3978102	0.2692308
GPA	Cumlaude	0.1277372	0.6384615
	Very Satisfy	0.5364964	0.3615385
	Satisfy	0.1934307	0
	Plenty	0.1423358	0
City of Residence	Outside Yogyakarta and Central Java	0.5182482	0.5
	Yogyakarta and Central Java	0.4817518	0.5
Concentration in High School	Science	0.72262774	0.95384615
	Social	0.13868613	0.02307692
	Other	0.13868613	0.02307692

Based on table 4, it can be conclude that the biggest probability in each variable there are male for on time status, cumlaude GPA in on time status, city of residence that outside Yogyakarta and Central Java in not on time status, and science concentration in on time status. That probability will implied the accuracy of the classification that showed in the next section as follow.

C. Accuracy of Classification

According to the table 5, it can be seen from the 302 Graduation Student with "On Time" status there are 250 students classified correctly, and 52 other students are not appropriate.

Table 5. Accuracy of Classification

Graduation Status	**On Time**	**Not On Time**	***Percentage Correct***
On Time	250	24	81,18%
Not On Time	52	78	

Thus, it can be said that the Naive Bayes algorithm successfully classifies Graduation status student in Statistics Department UII with a percentage of 81,18% accuracy.

IV. Conclusion

In this paper, Naïve Bayes Classifier has been discussed as the best classifier in this problem. Thus, the level of accuracy in classification models Naive Bayes algorithm 81,18%.

Acknowledgment

The authors acknowledgment with thanks the Statistics Department, Islamic University of Indonesia for providing the financial assistance in the research activity.

References

[1] Indonesian Government Regulation No. 66 (2010). *Management and Delivery of Education.*

[2] Moeliodihardjo, B. Y., (2014). Higher Education Sector in Indonesia. International Seminar on Massification of Higher Education in Large Academic Systems. British Council. New Delhi.

[3] Han, J. and Kamber, M. (2006). Data Mining : Concepts and Techniques. Elsevier. San Fransisco.

[4] BAN PT - National Accreditation Board of Higher Education. (2008). Book VI Matrix Instrument Rating Program Accreditation.

[5] S.L. Ting, W.H. Ip, Albert H.C. Tsang. (2011). Is Naïve Bayes a Good Classifier for Document Classification ?. International Journal of Software Engineering and Its Application, 5(3), Hongkong.

[6] BAN PT - National Accreditation Board of Higher Education. (2011). Accreditation of Higher Education Institutions - Book III Guidelines for Preparation of Form, pp.4.

[7] Rachimawan, A. F., & Ulama, B. S. S. (2016). Ads Filtering Mengunakan Jaringan Syaraf Tiruan Perceptron, Naïve Bayes Classifier, dan Regresi logistik. Jurnal Sains dan Seni ITS, 5(1), D83-D89.