DEVELOPMENT OF MODEL FOR PROVIDING FEASIBLE SCHOLARSHIP

Harry Dhika
Department of Science, School of Information Systems
University Indraprasta PGRI, Jakarta 13760 Indonesia
Email: dhikatr@yahoo.com

***Abstract*—The current work focuses on the development of a model to determine a feasible scholarship recipient on the basis of the naivë Bayes' method using very simple and limited attributes. Those attributes are the applicants academic year, represented by their semester, academic performance, represented by their GPa, socio-economic ability, which represented the economic capability to attend a higher education institution, and their level of social involvement. To establish and evaluate the model performance, empirical data are collected, and the data of 100 students are divided into 80 student data for the model training and the remaining of 20 student data are for the model testing. The results suggest that the model is capable to provide recommendations for the potential scholarship recipient at the level of accuracy of 95%.**

Keywords: **Scholarship; Data Mining; Naive Bayes; Knowledge Discovery in Databases.**

I. INTRODUCTION

The Republic of Indonesia legislation No. 20, 2003, regarding the national education system states, in Chapter 2 and Section 3, that

> National education serves to develop the ability and character development as well as a dignified civilization in the context of educating the nation, aims at developing student's potentials in order to become a man of faith and fear of God Almighty, noble, healthy, knowledgeable, skilled, creative, independent, and become citizens of a democratic and responsible.

Furthermore, the first paragraph of Article 5 of the law reads

> Every citizen has the same rights to acquire quality education.

On the legal basis, every citizen is entitled to an education regardless to race, religion, or social class. However, in the reality, many Indonesia citizens are not able to afford to have an education particularly higher education.

Received: March 22, 2016; received in revised form: April 06, 2016; accepted: April 28, 2016; available online: May 20, 2016.

Article 11 of the law states that the central and local governments have an obligation to ensure that funding for education is available for every citizen within the age of 7–11 year old. Thus, elementary, middle, and senior high schools are available freely to anybody. Meanwhile, for higher educations, Indonesia government provides a competitive scholarship via Council of Education Funding Management (*Lembaga Pengelola Dana Pendidikan*, LPDP). Due to the large number of applicants, there is the need of system that capable to provide recommendation of applicants who have great potential to use the scholarship wisely.

The current issue is rather similar to the problems of recommendation systems such as those discussed in Ref. [1–5]. However, the current work strives on adopting a much more simple approach that is on the basis of naivë Bayes's method.

II. METHODS

A. Data Mining

The Knowledge Discovery in Database (KDD) or data mining is essential in the modern life and is designed to extract important information from a database, which often is huge in size. This field is very important and has a huge potential to empower any company [6].

The KDD process is schematically shown in Fig. 1. The detail of each process is of the following [7].

The first is the formation of the understanding of the application domain. This stage determines the purpose of the end-user and related parts where KDD is applied. Develop an understanding of the application domain is the initial preparatory steps. It prepares the scene to understand what to do with a lot of decisions (about transformation, algorithms, representation, etc.). The people in charge of the project KDD need to understand and define the purpose of the end user and the environment in which the knowledge discovery process

will take place (including relevant prior knowledge). As a result KDD process, there may be a revision and improvement of this step. Having understood the purpose KDD, preprocessing of data begins, as defined in the following three steps (note that some of the methods here are similar to Data Mining algorithms, but is used in the context of preprocessing).

The second is choosing and creating a set of data to support the knowledge discovery process will be conducted. Determining the data to be used for the KDD process is done at this stage. Looking data are available, obtain additional data needs, integrating all data for KDD into a set of data, including attributes required in the KDD process. There are interactive and iterative of the KDD. Starting with the data provided both set up and then expands and observe its effect in KDD.

The third is preprocessing and cleansing. In this stage, enhanced data reliability. Including data clearing, such as dealing with incomplete data, eliminating distractions or outliers. Including using complex statistical methods, or specific data mining algorithms in KDD.

The fourth is data transformation. At this stage, the generation of better data for data mining prepared and developed, making better use of data into a dimension reduction methods and transformation attributes. For example, in the medical examination, the results for the attributes may often be the most important factor, and not one by one. In marketing, we may need to consider effects beyond our control and efforts and temporal issues (such as studying the effect of the accumulation of the ad). However, even if we do not use that right at the beginning of the transformation, we can obtain the surprising effect that instructs us about the transformation needed (In the next iteration). Thus, the KDD process reflects upon itself and causes the understanding of the transformation needed (such as quick knowledge of an expert in a particular field of the key leading indicator).

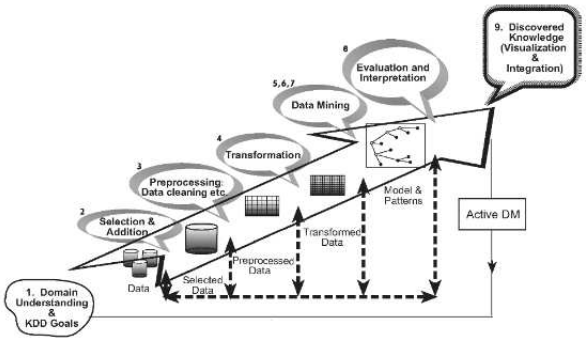

Fig. 1. Knowledge Discovery in Database.

The fifth is choosing a suitable data mining tasks. At this stage determined the type of data mining that will be used, whether the classification, regression, or clustering, depending on the purpose of KDD and the previous stage.

The sixth is selecting the data mining algorithms. Selection of the most appropriate algorithms to find patterns at this stage. There are two main objectives in Data Mining: prediction and description. Prediction is often referred to as Supervised Data Mining while Unsupervised Descriptive Data Mining covers aspects of visualization and Data Mining. Most data mining techniques based on inductive learning, where the model is constructed explicitly or implicitly by generalizing from an adequate number of the training data. The underlying assumption of the inductive approach is that the trained model is valid for future cases. This strategy also takes into account the level of meta-learning for a particular set of data is available.

The seventh is the use of data mining algorithms. At this stage, the implementation of data mining algorithms that have been determined in the previous stage. For example, by setting up the parameters of the control algorithms, such as the minimum number of cases in a single leaf of the decision tree.

The eigth is evaluation. At this stage, the evaluation and interpretation of the patterns obtained, with respect to the goals set in the first step. This step focuses on comprehensibility and usefulness of the model induction. In this step, the knowledge found also documented for further use. The last step is the use and the overall feedback on patterns and findings obtained by data mining.

Finally, the ninth is the use of knowledge is found that incorporate knowledge into other systems for further action. Knowledge becomes active in the sense that we can make changes to the system and measure the impact of the success of this step actually determine the effectiveness of the overall KDD process. There are many challenges in this step, such as the loss of "laboratory conditions". For example, it found knowledge of a certain static snapshot (typically samples) of the data, but now the data into a dynamic.

The study of data mining is also consistent with data processing in the Business Intelligence (BI). BI data includes the acquisition of data and information from various sources is varied and processes them into decision-making. BI can be used to support the company in achieving the success criteria such as [8]: to help to make decisions with speed and better quality, to accelerate operations, to shorten the product devel-

opment cycle, to maximize the value of the products available and anticipating new opportunities, and to create a better market and focused, as well as improving relationships with customers and suppliers.

Information technology is useful for automating business processes involving sizeable transaction data in every day. Currently, the company had to overcome a great deal of data that are difficult to be handled manually. It is very difficult for a person to manually extract useful information from large data sets despite the fact that the information may be useful in the decision-making process [9].

B. Naive Bayes Algorithm

Naive Bayes, which is also called an idiot Bayes, Bayes sample, or independence Bayes, is a good method because it is easily performed and does not require complicated and looping parameter estimation scheme. This means it can be applied to large size sets of data [10]. Bayes classification with the Naive Bayes is also known to have the capability comparable to the decision trees and neural networks [11]. Easily interpreted so that users who do not have expertise in the field of classification technology can understand. The Bayes theorem is written:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \tag{1}$$

where y is the data with unknown class, x is the y data that hypothesized in a certain class, $P(x|y)$ is the probability of the hypothezing x based on the conditions y (posteriori probability), $P(x)$ is the probability of hypothezing x (prior probability), $P(y|x)$ is the probability of y given x, and finally, $P(y)$ denotes the probability of y.

Naive Bayes is a simplified method of Bayes. Bayes theorem simplifies to:

$$P(x|y) = P(y|x)P(x) \tag{2}$$

The use of the algorithm Naive Bayes theorem Bayes which combine prior probability and the conditional probability in a formula that could be used to calculate the probability of each possible classification [12].

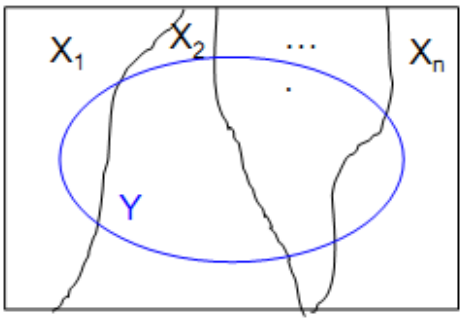

Fig. 2. The graphical presentations of the posterior probability of the random variable X_i in Y and the prior probability of the random variable Y in X_i.

C. Evaluation

This method uses a matrix table as in Table 1. If the data set consists of only two classes; one class is regarded as positive and the other negative [12].

True positives are positive that the number of records classified as positive, false positives is the number of records that are classified as positive of negative, false negatives is a positive number of records that are classified as negative, true negative negatives is the number of records that are classified as negative, then enter test data. After the test data is inserted into confusion matrix, calculate the values that have been entered to calculate the amount of sensitivity (recall), specificity, precision, and accuracy. Sensitivity is used to compare the amount of TP to record a positive number while specificity is the ratio of the number of TN against the record number of negative

III. RESULTS AND DISCUSSION

The attribute data for classification are student semester, grade-point average (GPa), capability, and the level of organizational activity. The variable student semester denotes the current academic term of the student, which for most Indonesia university is from Semester I to Semester VIII. Regarding the GPa, only those who have GPas higher than 2.75 qualify for the consideration. The GPa is categorized into four: 'Grade 1', 'Grade 2', 'Grade 3', and 'Grade 4' depending on the student GPa. The student GPa is within the range of four and is divided into four categories. The attribute capability denotes the capability from the economic perspective and is categorized into capable, quite capable, and incapable. Consideration is only provided to those in the last two categories. Regarding the level of organization activity, only those who are active in university organizations would receive higher

TABLE I
CONFUSION MATRIX MODEL

Correct classification	Classified as	
	+	-
+	true positives	false negatives
-	false positives	true negatives

TABLE II
THE DATA ATTRIBUTES AND THEIR CATEGORIES.

Attribute	Category
Semester	I, II, ..., VII
GPa	Grade I–IV
Ability	Capable, Quite Capable, Incapable
Organization	Active, Non active

TABLE III
THE STATISTICS OF THE DATA.

```
=== Classifier model (full training set) ===
Naive Bayes Classifier
                        Class
Attribute         Rejected   Approved
                   (0.21)     (0.79)
===================================
Semester
  mean              5.7        4.9125
  std. dev.         1.7635     1.9506
  weight sum       20         80
  precision         1          1

GPA
  mean              2.4        3.475
  std. dev.         1.2806     0.4994
  weight sum       20         80
  precision         1          1

Ability
  Capable          10.0        1.0
  Quite Capable    11.0        2.0
  Unable            2.0       80.0
  [total]          23.0       83.0

Organization
  Active            9.0       81.0
  Non Active       13.0        1.0
  [total]          22.0       82.0
```

consideration. These attributes and their categories are presented in Table II. The outcome of the model is the student status of receiving the scholarship; thus, the available categories are 'Rejected' or 'Approved'.

The above classification problem is solved using Weka software [13]. The results are depicted in Table III for the descriptive statistics of the data, in Table IV for the performance of the classifier, and in Table V of the level of the accuracy of the current classifier.

The results in the confusion matrix of Table V indicate that the current classifier is able to achieve the level of the accuracy of 95%. For this level of the accuracy, following the categorization provided by Ref. [14], the current classifier can be said to perform extremely well.

Finally, the study concludes the proposed scholarship assessment model can provide recommendation by considering the aspects of the student academic term, GPa, economic ability, and organizational involvement up to a very high level of the accuracy.

TABLE IV
THE PERFORMANCE OF THE CLASSIFIER.

```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          19   95
Incorrectly Classified Instances         1    5
Kappa statistic                          0.7727
Mean absolute error                      0.0337
Root mean squared error                  0.1275
Relative absolute error                 11.1085
Root relative squared error             35.0564
Total Number of Instances               20
```

TABLE V
THE ACCURACY OF THE MODEL IN THE FORM OF THE CONFUSION MATRIX.

		Classification	
		Rejected	Approved
Actual	Rejected	2	1
	Approved	0	17

IV. CONCLUSION

This study uses the methods of knowledge discovery in databases (KDD). The result using naive Bayes algorithm, the data of 100 scholarship recipients, the resulting level of accuracy in the model of scholarships is 95% or by 0.95 so that classification in the category very well with a percentage of 0.90 - 1.00. The accuracy level results should be improved in order to minimize error in awarding scholarships one of which can be done by comparison with other algorithms.

REFERENCES

[1] Y. Cao and Y. Li, "An intelligent fuzzy-based recommendation system for consumer electronic products," *Expert Systems with Applications*, vol. 33, no. 1, pp. 230–240, 2007.

[2] M.-H. Park, J.-H. Hong, and S.-B. Cho, "Location-based recommendation system using bayesian users preference model in mobile devices," in *Ubiquitous Intelligence and Computing*. Springer, 2007, pp. 1130–1139.

[3] F. E. Walter, S. Battiston, and F. Schweitzer, "A model of a trust-based recommendation system on a social network," *Autonomous Agents and*

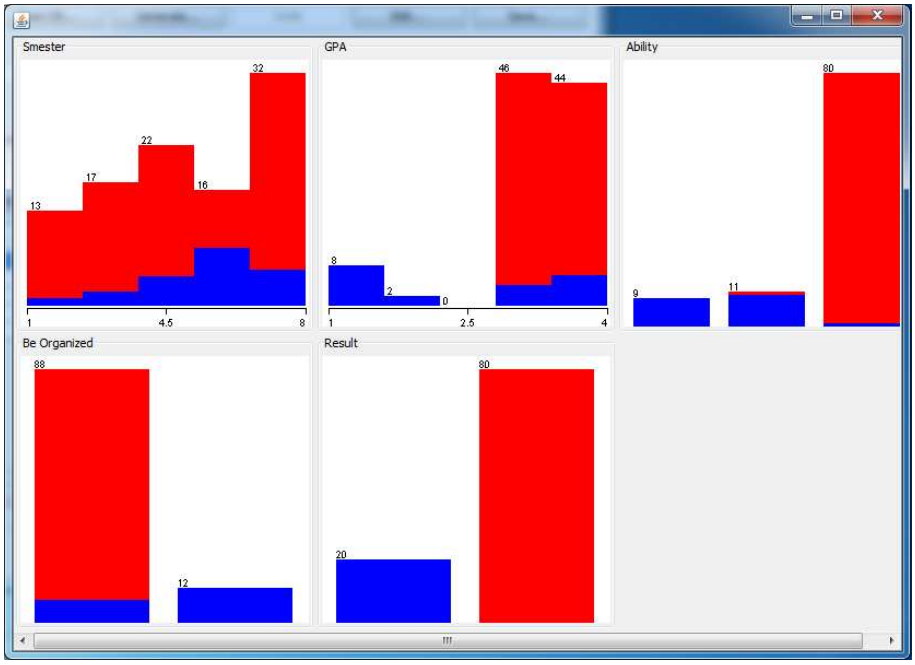

Fig. 3. The Distribution of the Data Used in the Analysis.

Multi-Agent Systems, vol. 16, no. 1, pp. 57–74, 2008.

[4] S. Debnath, N. Ganguly, and P. Mitra, "Feature weighting in content based recommendation system using social network analysis," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 1041–1042.

[5] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston *et al.*, "The youtube video recommendation system," in *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 2010, pp. 293–296.

[6] F. Sulianta and D. Juju, *Business forecasting data mining companies*. PT Elex Media Komputindo, 2010.

[7] O. Maimon and L. Rokach, *Data mining and knowledge discovery handbook*. Springer, 2005, vol. 2.

[8] S. Darudiato, S. W. Santoso, and S. Wiguna, "Business intelligence: konsep dan metode," *Communication and Information Technology Journal*, vol. 4, no. 1, pp. 63–67, 2010. [Online]. Available: http://journal.binus.ac.id/index.php/commit/article/view/537/515

[9] A. M. Sundjaja, "Implementation of business intelligence on banking, retail, and educational industry," *Communication and Information Technology Journal*, vol. 7, no. 2, pp. 65–70, 2013.

[10] X. Wu and V. Kumar, *The top ten algorithms in data mining*. CRC Press, 2009.

[11] J. Han and M. Kamber, *Data mining concepts and techniques*. San Fransisco: Morgan Kauffman, 2006.

[12] M. Bramer, *Principles of data mining*. Springer, 2007, vol. 131.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, *The WEKA Data Mining Software: An Update*, 2009, vol. 11.

[14] F. Gorunescu, *Data mining: concepts, models and techniques*. Springer Science & Business Media, 2011, vol. 12.