Drop out Estimation Students based on the Study Period: Comparison between Naïve Bayes and Support Vector Machines Algorithm Methods

Harwati1, Riezky Ikha Virdyanawaty2 and Agus Mansur3
1,2,3 Industrial Engineering Department, Faculty of Industrial Technology
Universitas Islam Indonesia, Jl. Kaliurang km. 14,5 Yogyakarta, Indonesia

E-mail: 1Harwati82@gmail.com, 2ikha03oktober@gmail.com, 3agusmansur.am@gmail.com

Abstract. Industrial Engineering is one of the departments in Faculty of Industrial Technology. It has more than 300 freshmen in every academic year. However, many students are dropped out because they couldn’t complete their study in appropriate time. Variables that influence the drop out case are not yet studied. The objective of this paper is discovering the highest accuracy level between the two methods used, i.e. Naïve Bayes and Support Vector Machines algorithms. The method with the highest accuracy will be discovered from the patterns forms and parameters of every attribute which most influence the students’ length of study period. The result shows that the highest accuracy method is Naïve Bayes Algorithm with accuracy degree of 80.67%. Discussion of this paper emphasizes on the variables that influence the students’ study period.
Keyword: Drop Out, Naïve Bayes, SVM, Estimation, Student

6.1. Introduction
College is an institution of academic education for students. The role of students is an important aspect in the success of the program of study at the college [1]. [2] said that the department is obliged to monitor the students’ study progress and predict their study period. It is useful to anticipate students’ problems which will lead to the decreasing performance of department. Therefore, to improve the department’s performance, the students must have a good academic performance, as stated by [3] that the students’ academic performance are evaluated each semester to determine the learning outcomes that have been achieved. If a student cannot meet certain academic criteria to be declared worthy of continued study, the student is declared resign of the college or dropped out.

Industrial Engineering is one of the departments in Faculty of Industrial Technology. It has more than 300 freshmen in every academic year. There are many dropout cases in the department because the students couldn’t finish their study in 7 years. Meanwhile, the department is having difficulties in estimating the factors which influence the dropout cases. Whereas, by estimating the dropout factors, it can improve the department’s performance and help the academic system in giving early warning to the students by using a classification technique.
stated that the purpose of data mining is obtaining relations or patterns which may provide useful indications. Similarly, stated that data mining is a process of finding significant relations, patterns, and trends by examining a large group of data stores in storage using pattern recognition techniques such as statistical and mathematical techniques. A study by explained that one of the data mining techniques is classification technique which is a learning technique to predict the value of target’s category variables.

The definition of classification according to is a technique of seeing the behaviors and attributes of predefined groups. stated that this technique can classify new data by manipulating existing data which has been classified and by using the result to provide a number of rules. also states that prediction has similarities with classification and estimation, in prediction value of prediction result will be in the future. Several techniques used in classification and estimation can also be used for prediction.

A study by similar method was performed by using Naïve Bayes Classifier algorithm which showed that the most influential factor in determining the classification of students’ academic performance is Grade Point Average (GPA), Grade Point (GP) on the first semester, GP on the fourth semester, and gender. Similar method was also used by discussing Classification of Length of Study of FSM Students of Diponegoro University Using Binary Logistic Regression and Support Vector Machines (SVM) which showed that variables which influence Length of Study variable are Department and GPA variables.

The objective of this paper is discovering the highest accuracy level between the two methods used, i.e. Naïve Bayes and Support Vector Machines algorithms. The method with the highest accuracy will be discovered from the patterns formed and parameter of every attribute which most influence the students’ length of study period. It would help the Industrial Engineering Department in looking for solutions and policies to improve the students’ achievement to finish their studies on time.

6.2. Literature Review

2.1 Naïve Bayes Algorithm

Naïve Bayes is a prediction technique based on simple probabilistic based on the application of Bayes theorem (Bayes rules) with strong independence assumption [8]. In Naïve Bayes classification, X is input vector which contains features and Y is class label. Naïve Bayes is written as $P(Y|X)$. The notation means class label probability $Y$ is obtained after features of class $X$ are observed. This notation is also called posterior probability for $Y$, while $P(Y)$ is called prior probability of $Y$. During training process, posterior probability learning of $P(Y|X)$ must be performed on the model for every combination of $X$ and $Y$ based on information from training data. By developing the model, a test data $X'$ can be classified by discovering the value of $Y'$ by maximizing the value of $P(Y'|X')$ obtained. The Naïve Bayes formulation for classification is:

$$P(Y|X) = \frac{P(Y)\Pi_{i=1}^{q} P(X_i|Y)}{P(X)}$$

$P(Y|X)$ is data probability with vector $X$ in class $Y$. $P(Y)$ is the prior probability of class $Y$. $\Pi_{i=1}^{q} P(X_i|Y)$ is the independent probability of class $Y$ from every feature in vector $X$. The value of $P(X)$ is fixed, so in to calculate prediction, only $P(Y)\Pi_{i=1}^{q} P(X_i|Y)$ should be calculated by selecting the biggest as the result of the prediction. Meanwhile the independent probability $\Pi_{i=1}^{q} P(X_i|Y)$ is the influence of all features of data on every class $Y$, which is notated as $P(X|Y = y) = \Pi_{i=1}^{q} P(X_i|Y = y)$. Every feature set $X = \{X_1, X_2, X_3, ..., X_q\}$ consists of $q$ attributes (dimensions).
2.2 Support Vector Machines Algorithm

Support Vector Machines can be called a semi-eager learner classification technique because aside from requiring training process, SVM also stores a small part of training data to be used again in the prediction process. Some of the stored data is support vector [8]. SVM is actually a linear hyperplane which only works on data which can be separated linearly. Data with non-linear distribution usually use kernel approach on the initial data feature of data set. Kernel can be defined as a function which maps the data feature of initial dimension (low) to another feature with higher dimension (even much higher). Kernel mapping algorithm is shown below:

$$\phi : D^q \rightarrow D^r$$

$$x \rightarrow \phi(x)$$

$\phi$ is a kernel function used for mapping, $D$ is training data, $q$ is feature set in an old data, and $r$ is a new feature set as a mapping result for every training data. Meanwhile, $x$ is training data, where $x_1, x_2, ..., x_n \in D^q$ are features mapped to high dimension feature $r$, so data set used as training used algorithm from old feature dimension $D$ to new dimension $r$. For example for $n$ data sample:

$$(\phi(x_1), y_1, \phi(x_2), y_2, ..., \phi(x_n), y_n) \in D^r$$

Then the same training process as linear SVM was conducted. The mapping process in this phase requires dot-product calculation of two data in new feature space. Dot-product of both vectors $(x_i)$ and $(x_j)$ are noted as $\phi(x_i) \cdot \phi(x_j)$. The dot-product values of both vectors can be calculated indirectly, without knowing transformation function $\phi$. This computation technique is called kernel trick which is calculating the dot-product of two vectors in new dimensional space using the components of both vectors in original dimensional space, as follows:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

And prediction on data set with new feature dimension is formulated with:

$$f(\phi(x)) = \text{sign} (w \cdot \phi(z) + b) = \text{sign} \left( \sum_{i=1}^{N} \alpha_i y_i \phi(x_i) \cdot \phi(z) + b \right)$$

$N$ is total data which becomes support vector, $x_i$ is support vector, and $z$ is test data which will be predicted. For kernel function selections commonly used in application, see Table 1.

| Kernel Name     | Formula                                      |
|-----------------|----------------------------------------------|
| Linier          | $K(x, y) = x \cdot y$                        |
| Polinomial      | $K(x, y) = (x \cdot y + c)^d$                |
| Gaussian RBF    | $K(x, y) = \exp \left(\frac{-|x-y|^2}{2 \cdot \sigma^2}\right)$ |
| Sigmoid)        | $K(x, y) = \tanh \left(\sigma(x \cdot y) + c\right)$ |
| Invers          | $K(x, y) = \frac{1}{\sqrt{|x-y|^2 + c^2}}$   |
6.3. Research Objectives

3.1. Obtaining the value of comparison of the result prediction based on Naïve Bayes algorithm and support vector machine.

3.2. Obtaining a dropout estimation pattern of students based on the methods of classification with the highest value.

3.3. Obtaining an effective parameter on each attribute from students’ length of study period based on the method with the highest accuracy.

6.4. Research Methods

The focus of this research is to conduct drop out estimation based on the evaluation of academic achievement at the maximum limit of seven years study period which has been determined by the Industrial Engineering Department in Faculty of Industrial Technology, Universitas Islam Indonesia. The research’s subjects are the students from year 2003 to 2007 of the Industrial Engineering Department and the object is the duration of students who study for more than seven years and less than seven years. This research will estimate through classification technique using Naïve Bayes method and Support Vector Machines, then the data processing process assisted by Rapid Miner Software and Microsoft Office Excel. The processing results are in the form of a decision or rules and the amount of accuracy level of each methods degree. The chosen method is a method that has the highest accuracy percentage, and then the method is used as a basis for providing recommendations.

6.5. Results and Discussions

Data collection of this research was performed by conducting field study, which were direct observation and recording variables or quantitative attributes of study period, such as origin, how long they’ve studied in college, gender, high school type, department in high school, high school’s national final score, parents’ education, parents’ occupations, and GPA on the fourth semester. Data were collected in the Industrial Engineering Department on students from year 2003 to 2007 and the attributes used in the classification method referred to the Unisys service of Universitas Islam Indonesia.

The processing in the case of students’ length of study period used data mining software called Rapid Miner. Classification algorithm used in developing the model was Naïve Bayes algorithm and Support Vector Machine algorithm. The result of data processing based on the case of students’ length of study period using Rapid Miner software was the probability of each attribute by Naïve Bayes and weight of each attribute by Support Vector Machine algorithm. Below is the result of processing by Naïve Bayes and Support Vector Machine:
5.1. The Result of Naïve Bayes Processing

The result of processing by Naïve Bayes method was the instance probability value of each attribute and label probability value of overall test data used in Naïve Bayes processing.

Table 2. Parameter Probability Value of Each Attribute

| Attribute      | Parameter             | Do    | Not Do  |
|----------------|-----------------------|-------|---------|
| Origin         | Value=Outside Java    | 0.306 | 0.254   |
| Origin         | Value=Java            | 0.694 | 0.746   |
| Age            | Value=On Time         | 0.694 | 0.622   |
| Age            | Value=Passed Late     | 0.265 | 0.274   |
| Age            | Value=Pass Quickly    | 0.041 | 0.104   |
| Gender         | Value=Male            | 0.857 | 0.622   |
| Gender         | Value=Female          | 0.143 | 0.378   |
| High School    | Value=State           | 0.673 | 0.771   |
| High School    | Value=Private         | 0.327 | 0.229   |
| Majoring       | Value=Social Science  | 0.286 | 0.104   |
| Majoring       | Value=Natural Science | 0.714 | 0.895   |
| NEM            | Value=Upper           | 0.673 | 0.632   |
| NEM            | Value=Medium          | 0.306 | 0.368   |
| NEM            | Value=Low             | 0.020 | 0.000   |
| Father’s Education | Value=Junior High School | 0.082 | 0.055 |
| Father’s Education | Value=Senior High School | 0.408 | 0.458 |
| Father’s Education | Value=Bachelor Degree   | 0.224 | 0.249 |
| Father’s Education | Value=Elementary School | 0.082 | 0.035 |
| Father’s Education | Value=Diploma 4       | 0.041 | 0.080   |
| Father’s Education | Value=Master Degree   | 0.082 | 0.050   |
| Father’s Education | Value=Bachelor Degree | 0.020 | 0.000   |
| Father’s Education | Value=Diploma 3       | 0.020 | 0.070   |
| Father’s Education | Value=Doctorate Degree | 0.020 | 0.005   |
| Father’s Education | Value=Junior High School | 0.020 | 0.000   |
| Mother’s Education | Value=Senior High School | 0.408 | 0.448 |
| Mother’s Education | Value=Diploma 3       | 0.102 | 0.144   |
| Mother’s Education | Value=Junior High School | 0.143 | 0.149   |
### Table 3. Attribute Weights

| Attribute               | Parameter                  | Do  | Not Do |
|-------------------------|----------------------------|-----|--------|
| Mother’s Education      | Value=Bachelor Degree      | 0.143 | 0.104 |
| Mother’s Education      | Value=Elementary School    | 0.122 | 0.070 |
| Mother’s Education      | Value=Master Degree        | 0.020 | 0.010 |
| Mother’s Education      | Value=Diploma 4            | 0.061 | 0.070 |
| Father’s Occupation     | Value=Entrepreneur         | 0.224 | 0.249 |
| Father’s Occupation     | Value=Private Employee     | 0.326 | 0.264 |
| Father’s Occupation     | Value=Retired              | 0.102 | 0.040 |
| Father’s Occupation     | Value=Civil Servant        | 0.306 | 0.398 |
| Father’s Occupation     | Value=Farmer               | 0.041 | 0.025 |
| Father’s Occupation     | Value=Army/Police          | 0.000 | 0.015 |
| Mother’s Occupation     | Value=Entrepreneur         | 0.245 | 0.249 |
| Mother’s Occupation     | Value=Jobless              | 0.367 | 0.363 |
| Mother’s Occupation     | Value=Civil Servant        | 0.306 | 0.294 |
| Mother’s Occupation     | Value=Farmer               | 0.041 | 0.015 |
| Mother’s Occupation     | Value=Private Employee     | 0.020 | 0.065 |
| Father’s Occupation     | Value=Retired              | 0.020 | 0.015 |
| GPA of the Fourth Semesters | Value=Satisfy               | 0.571 | 0.488 |
| GPA of the Fourth Semesters | Value=Not Satisfactory    | 0.326 | 0.070 |
| GPA of the Fourth Semesters | Value=Very Satisfactory    | 0.102 | 0.433 |
| GPA of the Fourth Semesters | Value=Satisfy              | 0.000 | 0.005 |
| GPA of the Fourth Semesters | Value=Praiseworthy         | 0.000 | 0.005 |

5.2. The Result of Support Vector Machine Processing

The processing result by Support Vector Machine with the help of rapid software only produces weights which is shown in Table 3. To discover the decision on each existing data, manual calculation should be performed.
| Attribute                          | Weights |
|-----------------------------------|---------|
| Gender                            | 0.032   |
| High School                       | -0.005  |
| Majoring                          | 0.025   |
| NEM                               | -0.025  |
| Father’s Education                | 0.007   |
| Mother’s Education                | 0.021   |
| Father’s Occupation               | 0.020   |
| Mother’s Occupation               | 0.004   |
| GPA of the Fourth Semesters       | 0.150   |

The next calculation after obtaining the weight of rapid output was searching for score in every existing data by multiplying attribute in every data with the weight of each attribute. Below is formulation in searching for score of every data.

\[
Score = \sum_{i=1}^{A} N_i \cdot W_i
\]

\(N_i\) is attribute value and \(W_i\) is attribute weight. After finding score of each data type, all scores are totaled. The final step in the method of support vector machine is finding the median value from the smallest to the biggest score. After obtaining the median value, then performed grading, where data is entered in a class 1 if the score \(\leq 1.088\) and the data included in class 2 if the score \(> 1.088\).

5.3. Comparison Result Between Naïve Bayes Algorithm Method And Support Vector Machines Algorithm Method

The performances of the two methods generate different accuracy percentage. The accuracy of the results obtained based on the concept of each of the methods, i.e. Naïve Bayes algorithm and support vector machine algorithm. Whether the performance of the model is good or not, it depends on each case at hand. Below is the comparison of the accuracy based on a lot in common with the facts.

**Table 4. Comparison the Accuracy**

| Method                | Accuracy | Error  |
|-----------------------|----------|--------|
| Naïve Bayes           | 80.67%   | 19.33% |
| Support Vector Machine| 60%      | 40%    |

It can be found in a comparison table that outcome prediction accuracy rate is formed by using Naïve Bayes amounted to 80.67% and using Support Vector Machine has an accuracy rate of 60%. It proves that the Naïve Bayes method is more accurate to use for students drop out case.

6.6. Conclusions and Recommendations

6.7. Conclusions

- The level of accuracy based on the prediction that was formed by using Naïve Bayes is 80.67%, while accuracy rate by using Support Vector Machine is 60%. This shows that the Naïve Bayes method is more accurate in predicting the cases of drop out students based on a long period of study.
• The prediction pattern of drop out students that formed by using *Naïve Bayes* method generates data of the students origin area that come from Java, the age of students when entering the college punctually, most of drop out students are the male gender, private high school, they studied science in high school, high NEM category, their parents graduate of scholars, the father works as a private employee, the mother does not work, and most of the students who have been predicted to be drop out have unsatisfactory criteria or GPA <2.00 on the 4th semester.

• Parameters that affect every long period of study attribute based on *Naïve Bayes* method are the attribute of the island with Java parameter, the attribute of the age of students when entering the college punctually with the appropriate parameter, the attribute of gender with male parameter, the attribute type of school with domestic parameter, the attribute o what they studied in high school with science parameter, NEM category with high parameter, father's education with high school parameter, mother’s education with high school parameter, father's occupation with private employee parameter, mother’s occupation with doesn’t not work parameter, and the GPA category on the 4th semester with satisfactory parameter that is <2.75.

6.8. Recommendations

• Conduct a classification concept of comparison using more than researcher did, because the more concept is tried, the more open and widely knowledge is obtained. We can also determine a match between the case and the method that used on it.

• The classification method is not only *Naïve Bayes* and *Support Vector Machine*, so the further research can use the classification concept by using *Artificial Neural Network, Nearest Neighbour*, or *Decision Tree*.

• The results from *Naïve Bayes* obtained rules that being used in making a simple useful application as predictor of drop out students.

• Conduct a students drop out research based on a length of study period by inputting data that obtained qualitatively or by interview.

• The recommendations that given to the Industrial Engineering Department of Universitas Islam Indonesia is to provide a letter of warning to students with unsatisfactory status of GPA in 4th semester or GPA <2.00 and satisfactory status of GPA or GPA <2.75 by the time students completed the course in the 4th semester.

6.9. References

[1] Ridwan M, Suyono H and Sarosa M2013  *Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier* Eccis7(1) 59–64

[2] Jananto A2013  *Algoritma Naive Bayes Untuk Mencari Perkiraan Waktu Studi Mahasiswa* 18(1) 9–16

[3] Andriania A2012  *Penerapan Algoritma C4.5 Pada Program Klasifikasi Mahasiswa Dropout* Seminar Nasional Matematika 139–147

[4] Kusrini and E. T. Lutfi 2009  *Algoritma Data Mining* (Yogyakarta : Andi Publishing)

[5] Larose, Daniel. 2005  *Discovering Knowledge in Data - an Introduction to Data Mining* (New Jersey: John Wiley &Sons)

[6] Untari, D 2010  *Data Mining Untuk Menganalisa Prediksi Mahasiswa Berpotensi Non-Aktif Menggunakan Metode Decesion Tree C4.5*.

[7] Damanik S, Ispiryanti D, and Sugito 2015  *Klasifikasi Lama Studi Mahasiswa Fsm Universitas Diponegoro Menggunakan Regresi Logistik Biner Dan Support Vector Machine* 4(2000)123–132.

[8] Prasetyo, E2012  *Data Mining - Konsep dan Aplikasi Menggunakan MATLAB* (Yogyakarta: Andi Publishing)