A Comparative Study to Predict Student’s Performance Using Educational Data Mining Techniques

Annisa Uswatun Khasanah1, Harwati 2)
1,2) Department of Industrial Engineering, Faculty of Industrial Technology, Islamic University of Indonesia, Jl. Kaliurang Km 14.5 Yogyakarta, Indonesia
145220102@uii.ac.id1, harwati@uii.ac.id2)

Abstract. Student’s performance prediction is essential to be conducted for a university to prevent student fail. Number of student drop out is one of parameter that can be used to measure student performance and one important point that must be evaluated in Indonesia university accreditation. Data Mining has been widely used to predict student’s performance, and data mining that applied in this field usually called as Educational Data Mining. This study conducted Feature Selection to select high influence attributes with student performance in Department of Industrial Engineering Universitas Islam Indonesia. Then, two popular classification algorithm, Bayesian Network and Decision Tree, were implemented and compared to know the best prediction result. The outcome showed that student’s attendance and GPA in the first semester were in the top rank from all Feature Selection methods, and Bayesian Network is outperforming Decision Tree since it has higher accuracy rate.

Keywords: student performance, prediction, feature selection, educational data mining

1. Introduction
Data mining is the process of automatically discovering useful information in large data repositories [1]. Data Mining techniques are employed to discover meaningful pattern and rules from large quantities of data that unknown. There are many techniques, methods and rules in Data Mining to extract particular database. Recently, data mining has been widely applied in many applications, such as customer relationship, engineering, medical, marketing, education, etc. There are increasing number of research that used data mining in education, and data mining technique that applied in this filed can be usually called as Education Data Mining. Education Data Mining can be conducted with many methods such as Decision Tree, Neural Network, Naïve Bayes, K-Nearest neighbour, and many other [2].

Education Data Mining has been commonly applied in many universities. Implementing Education Data Mining can help universities to predict their student’s performance. Universities today are operating in high competitive environment and have very complex system. It is very important for modern universities to deeply analyse their performance in order to develop strategy and future actions. University management must consider their student’s performance as one of important key of the marketing campaign and approach the promising potential of the university [3]. Predicting student’s performance can help to predict the weak student and help the university management to make strategy and decision making related to student’s performance improvement. The university management can take early step to prevent student fail or student drop out. After knowing the prediction, it is also expected to the weak student that they can improve their performance and achieve better score. Student’s performance can be applied based on several factors like personal,
psychological, social, and other environment [4][5]. This kind of analysis can help to predict the weak student and help the university management to make strategy and decision making related to student’s performance improvement. The university management can take early step to prevent student fail or student drop out. After knowing the prediction, it is also expected to the weak student that they can improve their performance and achieve better score.

Among the major data mining techniques, one of the oldest and most useful technique is classification [6][7]. Tan et al [1] define classification as the task of assigning objects to one of several predefined categories. It uses a set of pre-classified example to build a model that can classify the large record population [8]. Classification methods has been commonly use to predict student’s performance [2][3][5][6]. Bayesian Network and Decision Tree are two popular classification methods that have been used by many scholars to conduct the student’s performance predictions [3][9][10][11][12][13][14]. Those algorithms are powerful classification algorithm and the rules are easy to be interpret.

This current study will compare Bayesian Network and Decision Tree to whether student will drop out (DO) or not in Department of Industrial Engineering Universitas Islam Indonesia based on several attributes. Feature selection was conducted first before implementing the classification algorithm to select high influence attributes. This study aims to know the best scenario to predict student’s performance in Department of Industrial Engineering Universitas Islam Indonesia by conducting feature section and comparing Bayesian Network and Decision Tree. The rest of this paper is organized as follows, Section 2 present the literature review, Section 3 presents the methods research related to this study, Section 4 shows the results and discussion. The concluding is finally made in Section 5.

2. Literature Review

2.1. Bayesian Network

Bayesian Network is as a statistical principle to combine prior knowledge of the classes with new evidence accumulated from the data [1]. The mathematical equation for this method is shown in this equation (1).

\[
P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}
\]  

(1)

X and Y are a pair random attributes. The join probability of X and Y, P(X = x, Y = y), is the probability that variable X will admit a particular value x, while variable Y will admit the value y. A conditional probability is the probability that a random variable will take on particular value given that the outcome for another random variable is known.

There have been many studies that implement Bayesian Network as classification method. Bharadwaj and Pal [5] worked on a study to predict performance of students. The object in this study is 600 students from five different degree colleges in Computer Application course for Bachelor degree, Dr. R. M. L. Awadh University Faizabad, India. This study use Bayesian Network as the classification method and involve 17 attributes. The results show that student’s grade in senior secondary exam, living location, teaching media, mother’s qualifications, student’s other habit, annual income of the family and family status of the student were highly correlated with academic performance of the student.

As well as Bharadwaj and Pal [5], Pandey and Pal [15] also conduct a study using Bayesian Network on category, language, and background. This study concluded that whether new comer students will perform or not. Millian et al [12] also try to use Bayesian Network for student model engineering. Ramaswami & Rathinasabapathy [16] also used Bayesian Network approach in the field of education to predict student’s performance. They conduct an analysis on student’s data of the higher secondary schools including 5650 records with 35 attributes, and the class data were varying in five cases. In this study, they implement feature selection before applying the prediction models. The data analysis for
this study was conducted using software WEKA. This study used a simple select estimator algorithm, like Hill Climbing, K2, LAGD Hill Climbing, Repeated Hill Climbing, Tabu Search and Network Augmented with Tree, to find the conditional probability tables of the Bayes Network. From the result, it can be concluded that Bayesian Network models with Network Augmented with Tree search algorithm was outperform over others types of search algorithms for all class values types. The result also revealed that the class variable HScGrade highly related with the grade that gained at secondary level, type of transportation to school, medium of instruction, sibling structure (number of brothers and sisters) and economic status of the family.

2.2. Decision Tree
Decision Tree is another classification popular method. Decision tree, as shown in Fig. 1, is as a flow-like tree structure. The tree consists of root node (the first node denoted by oval), internal nodes which can have two or more kinds (the middle node denoted by oval), and leaf nodes (denoted by rectangle). All internal nodes contain splits, which test the value of an expression of the attributes.

Yadav & Pal [2] implemented Educational Data Mining Techniques to make a prediction model of engineering student’s performance. The CART, ID3, and C4.5 decision tree algorithms were implemented to predict their final exam performance. The result of the decision tree can be used to predict whether the students will be likely to pass, failed or promoted to next year. The dataset in this study was collected for session 2010 from VBS Purvanchal University in Jaunpur (Uttar Pradesh), Institute of Engineering and Technology. Initially data consists of 90 students. The variables used in this study were 13 and for the class attributes final exam results was used. The data was performed in software WEKA. The result showed even though CART, ID3, and C4.5 algorithms gave an acceptable level of accuracy, C4.5 technique outperformed the other with accuracy of 67.7778% compared to other methods.

Yadav et al [12] also implemented decision tree (CART, ID3, and C4.5) to predict academic performance of students in Computer Applications department of course Master of Computer Applications from session 2008 to 2011 in VBS Purvanchal University Jaunpur (Uttar Pradesh), India. Initially size consists of 48 records. The variable that used were marks of previous semester, class test grade, seminar performance, assignment, attendance, lab work, and for the class variable, marks of end semester (first, second, third, fail) were used. The result indicated that a CART was the best with
accuracy of 56.25% compared to other techniques, however ID3 algorithm was also showed an acceptable level of accuracy [2][5][14].

As well as previous author, Ahmed & Elaraby [8] also tried to implement Educational Data Mining to conduct student’s performance analysis. The decision tree (using ID3 algorithm) method was applied in this study. The dataset was collected from Department of Information System student's database from session 2005 to 2010, and the initial data consists of 1547 records. The variables that included in this study were department of the students, high school degree, midterm marks, lab test grade, seminar performance, assignment, measure of student participate, homework, and the target class is final grade mark. Midterm mark has the biggest gain; then, it was used as the root node. This study concluded that by conducting student performance analysis can help the students to improve performance of the student. In addition, it can identify students who need a special attention to minimize failing ration and take suitable action at right time.

2.3. Bayesian Network compare with Decision Tree

There have been many scholars that try to compare the performance of Bayesian Network and Decision Tree in student’s performance prediction. Al-Radaideh et al [9] tried to predict the performance of Technology & Computer Science Faculty, Yarmouk University, Jordan students who took the Programming I course (C++). This paper also tried to compare ID3 algorithm, C4.5 algorithms and Naïve Bayes. There were 12 attributes, included gender, age, department, high school major, high school grade, study type, funding, place of residency, lecturer degree, lecturer gender, lecturer department, number of repetition and the grade of the C++ course (A, B, C, D). WEKA was used to analyse the data. The high school grade had the highest gain ration, then this attribute was considered as the root node of the decision tree. From the result it can be concluded that the classification accuracy for the three different classification algorithms was not so high. It indicated that the samples and attributes that used in this research were not enough to generate a high quality classification model.

Kabakchieva [3] presented the initial results of a research which was conducted in Bulgarian university. The main goal of the research conducted by Kabakchieva [3] was aimed to know if there are any patterns in the collected data which sufficient to predict performance of student at the university based on student’s personal characteristics and pre-university characteristics. Nearest Neighbour classifier, Decision tree classifier, and Bayes classifiers were used to analyse the data. The analysis was performed in software WEKA because it is widely implemented for research purposes in the data mining and it is freely available to the public. The initial data included 10330 students and 20 attributes. The final results showed that the decision tree classifier (J48) was outperform other classifier because it has the highest accuracy, then followed by the rule learner (JRip) and the kNN classifier, while the Bayes classifiers are less accurate than the others. Nevertheless, all of the classifiers were showed an overall accuracy below 70 % which means that the error rate was high and the predictions were not able to be trusted.

3. Research Method

This study consists of three steps. The first step is data cleaning, the second step is feature selection and the last step is classification analysis. Data cleaning is important step that must be conducted before analysing the data to identify missing value, noisy and inconsistent data. The second step is Feature Selection. Feature selection algorithm can determine as the combination of a search technique to propose new feature subsets, along with an evaluation measure which scores the different feature subsets [17]. The main objective of feature selection is to select a subset of input attributes by eliminating features, which are irrelevant or give less predictive information [16]. This step is necessary to be applied in order to define the high influencing variables in student’s performance prediction. Anuradha and Velmurugan [17] as well as Ramaswami and Bhaskaran [16] used several methods to conduct the feature selection in their study, and the five popular filter feature subset methods with Rank search or Greedy search method are Correlation-based Attribute evaluation, Gain-
Ratio Attribute evaluation, Information-Gain Attribute evaluation, Relief Attribute evaluation, and Symmetrical Uncertainty Attribute. Those methods were also used in this current study. The last step was conducting the classification analysis using Bayesian Network and Decision Tree. The student’s performance parameter that is used in this study is whether the student will be DO or not. The performance parameter to compare both algorithm that used in this study is Accuracy Rate. For the purpose of conducting the experiments, WEKA which is freely open source data mining software was used to implement the algorithm.

4. Result and Discussion

4.1 Data Cleaning
Data in this research were collected from student database that can be accessed from Universitas Islam Indonesia’s information system (UNISYS). The initial data consisted of 178 data of student in academic year 2007 and after cleaning the data, it is only 104 data set available with 12 attributes (including the class attribute, drop out or not) where 13 data (12.5%) classified as drop out, and 91 data (87.5%) classified as not. The attributes are divided into four categories and shown in Table 1.

| Category              | Attributes                  |
|-----------------------|-----------------------------|
| Personal Information  | Gender (GE)                 |
|                       | Origin (OR)                 |
| Family information    | Father education (FE)       |
|                       | Father occupation (FO)      |
|                       | Mother education (ME)       |
|                       | Mother occupation (MO)      |
| Pre-university         | Senior high school type (ST)|
| characteristic        | Senior high school department (SD)|
| University            | first semester Attendance (AT)|
| features              | first semester GPA (GPA)    |
|                       | Drop out (DO) or Not (as the Class attribute) |

4.2 Feature Selection
Feature Selection revealed the 12 attributes that have high influence with student’s performance. The Feature Selection results from the five methods are shown in Table 2.

| Feature Selection Method                      | Selected Attributes |
|-----------------------------------------------|---------------------|
| Correlation-based Attribute evaluation        | MO, AT, GPA         |
| Gain-Ratio Attribute evaluation               | GPA, AT, SD, GE, FO, MO, ME, OR, ST, FE, SF |
| Information-Gain Attribute evaluation         | GPA, AT, SD, GE, FO, MO, ME, OR, ST, FE, SF |
| Relief Attribute evaluation                   | AT, GPA, FO, SD, FE, MO, GE, ME, OR, ST, SF |
| Symmetrical Uncertainty Attribute             | GPA, AT, SD, GE, FO, MO, ME, OR, ST, FE, SF |

Correlation-based Attribute evaluation is a Greedy search method that only select the attribute that have high influence with the prediction attribute. Based on Correlation-based Attribute evaluation, only three attributes (MO, AT, GPA) that have high influence with student’s performance. The other selection methods are Rank search method that rank the attributes from the highest influencing until the less influencing attributes. Gain-Ratio Attribute evaluation, Information-Gain Attribute evaluation, and Symmetrical Uncertainty Attribute have exactly the same results, while Relief Attribute evaluation have slightly different result. From the whole result it can concluded that two attributes that...
appear in the first list in almost all selection methods are GPA and AT. It means that first semester GPA and attendance have high influence whether the student will DO or not. For the less high influence attribute the four Rank search method show identical result. ST, FE, SF are in the bottom three of the rank from Gain-Ratio Attribute evaluation, Information-Gain Attribute evaluation, and Symmetrical Uncertainty Attribute, while in Relief Attribute evaluation FE is in the middle rank. Based on this result, the attributes that would be used in further classification analysis are GPA, AT, SD, GE, FO, MO, ME, OR because most of the methods show the same result. This classification analysis would be compare with the analysis that included FE attributes as shown from Relief Attribute evaluation result.

4.3 Classification Analysis
In the Classification analysis three scenarios to compare the performance between Bayesian Network and Decision Tree algorithm to predict whether student DO or Not were conducted. The first scenario used all the attributes, the second scenario compared both algorithms using selected attributes from the Feature Selection result included FE (GPA, AT, SD, GE, FO, MO, ME, OR, FE), and the last scenario only used the most selected attributes (GPA, AT, SD, GE, FO, MO, ME, OR). Table 3 shows the Classification result from each scenarios.

| Scenario | Attributes Used                  | Classification Algorithm | Accuracy Rate | Number of Incorrectly Classified Instance |
|----------|----------------------------------|--------------------------|---------------|------------------------------------------|
| 1        | All                              | Bayesian Network         | 95.19%        | 5                                        |
| 2        | GPA, AT, SD, GE, FO, MO, ME, OR  | Bayesian Network         | 97.11%        | 3                                        |
| 3        | GPA, AT, SD, GE, FO, MO, ME, OR  | Bayesian Network         | 98.08%        | 2                                        |

From the experimental result, it can be seen that within the same algorithm, scenario 1 that include all the attributes has lower accuracy rate than scenario 2 and 3 that only include selected attributes from Feature Selection. Then, it can be proved that Feature Selection can improve the accuracy of prediction model. It also can be seen that scenario 2 has lower accuracy rate that scenario 3. It means that it is better to exclude attribute FE from the analysis. For the classification algorithm comparison, it can be seen that from all scenarios Bayesian Network indicate outperform Decision Tree, and from the whole experiment, prediction with Bayesian Network using selected attributes GPA, AT, SD, GE, FO, MO, ME, OR is the best among the others.

5. Conclusion and Suggestion
Student performance prediction is essential to be conducted for a university to prevent student fail. Number of student drop out is one of parameter that can be used to measure student performance and one important point that must be evaluated in university accreditation in Indonesia. Of course this number must be minimized. Some high influence attributes in student prediction that can be considered by university management to minimized student drop out based on the result of this study are First Semester GPA, First Semester Attendance, Senior High School Department, Gender, Father Occupation, Mother Occupation, Mother Education, and Origin. It also can be proved from this study that feature selection can increase the accuracy rate of the prediction, and Bayesian Network is outperforming Decision Tree since it has higher accuracy rate. Then, the best scenario to predict whether the student will drop out or not in Department of Industrial Engineering Universitas Islam Indonesia is using the high influence selecting attributes from feature selection and by implementing Bayesian Network as the classification algorithm.
For future study, it is better if attribute Gender and Origin can be evaluated based on social aspect to make sure that these attributes are not bias and not containing discrimination aspect.

6. Reference

[1] Tan, P. N., Steinbach, M, Kumar, V., (2006) Introduction to Data mining, Introduction to Data mining. MA: Pearson Education.

[2] Yadav, S. K. and Pal, S., (2012), Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification, World of Computer Science and Information Technology Journal WCSIT, 2(2), 51–56.

[3] Kabakchieva, D., (2013), Predicting student performance by using data mining methods for classification, Cybernetics and Information Technologies, 13(1), 61–72.

[4] Hijazi, S. and Raza-Naqvi, S. M., (2006), Factors Affecting Students’ Performance, 3(1), 1–10.

[5] Bhardwaj, B. K. and Pal, S., (2011). Data Mining: A prediction for performance improvement using classification, (IJCSIS) International Journal of Computer Science and Information Security, 9(4).

[6] Romero, C., Ventura, S., Espejo, P. G. and Hervás, C., (2008), Data Mining Algorithms to Classify Students, EDM, 8–17.

[7] López, M. and Luna, J. M., (2012), Classification via clustering for predicting final marks based on student participation in forums, Proceedings of the 5th International Conference on Educational Data Mining, pp. 4–7.

[8] Ahmed, A. B. E. D. and Elaraby, I. S., (2014), Data Mining: A prediction for Student’s Performance Using Classification Method, World Journal of Computer Application and Technology, 2(2), 43–47.

[9] Al-Radaideh, Q. A., Al-Shawakfa, E. M. and Al-Najjar, M. I., (2006), Mining Student Data Using Decision Trees, International Arab Conference on Information Technology.

[10] García, P., Amandi, A., Schiaffino, S. and Campo, M., (2007), Evaluating Bayesian networks’ precision for detecting students’ learning styles, Computers and Education, 49(3), 794–808.

[11] Dekker, G. W., Pechenizkiy, M. and Vleeshouwers, J. M., (2009), Predicting students drop out: A case study, EDM’09 - Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, 41–50.

[12] Millán, E., Loboda, T. and Pérez-De-La-Cruz, J. L., (2010), Bayesian networks for student model engineering, Computers and Education. Elsevier Ltd, 55(4), 1663–1683.

[13] Kumar, S. A. and Vijayalakshmi, M. D., (2011), Efficiency of Decision Trees in Predicting Student’ Academic Performance, © Cs & It-Cscp 2011, 335–343.

[14] Yadav, S., Bharadwaj, B. and Pal, S., (2012), Data mining applications: A comparative study for predicting student’s performance, International Journal of Innovative Technology & Creative Engineering, 1(12), 13–19.

[15] Pandey, U.K. and Pal, S., (2011), Data Mining: A prediction of performer or underperformer using classification, International Journal of Computer Science and Information Technologies, 2(2), 686–690

[16] Ramaswami, M. and Bhaskaran, R., (2009), A Study on Feature Selection Techniques in Educational Data Mining, 1(1), 7–11.

[17] Anuradha, C. and Velmurugan, T., (2016), Feature Selection Techniques To Analyse Student Academic Performance Using Naïve, The 3rd International Conference on Small & Medium Business, 345–350.