The Application of Data Mining in Determining Patterns of Interest of High School Graduates

Dedy Hartama¹*, Agus Perdana Windarto², Anjar Wanto³

¹,²,³STIKOM Tunas Bangsa, Pematangsiantar, North Sumatera - Indonesia

*dedyhartama@amiktunasbangsa.ac.id

Abstract. This study proposes an application of Data Mining in determining patterns of interest of high school graduates to attend AMIK Tunas Bangsa Pematangsiantar and the dominant factors which affect the interest of high school graduates. The data was picked up from the application forms completed by those enrolled in the AMIK. In this study, algorithm C 4.5 decision tree was applied in order to obtain the pattern of data classification, i.e. school, status, school location, and year of graduation. The interest pattern of these graduates demonstrated that the best variable of the predictor used is the school status which gives as much as 81.71% performance accuracy to registration.

1. Introduction
Pematangsiantar and Simalungun area is the second largest after Medan City. This region is cool and convenient as well as a tourist resource as it is immediately adjacent to Lake Toba, the largest in Indonesia. Pematangsiantar is also seen as a City of Education as the residents enthusiastically send their children to schools, even to the level of university. With the support of the students’ high learning-interest and parents’ spirit, the city and the regency have constructed a lot of schools, from kindergartens to High Schools. As a feature, 66 high schools, general and vocational, out of 70 available in the two areas have 60.837 attendants. The annual enrollment of new students is conducted by state-owned and private owned universities. For the Private institutes, this procedure is carried out before mid-October. For this purpose, a representative marketing management must be applied, and the classification of school data should be conducted to determine the interest pattern of the students. To maintain and to increase the number of the students, a certain method in the process of enrollment must be annually made especially for the period of admission, March to September every year. Without students, private schools’ management can never run.

The process of new student’s admission at AMIK Tunas Bangsa Pematangsiantar is divided into three phases: Information, Listing and Registration. Information is when prospective students come, fill out data-forms, without making any payment. Listing is when a students come, fill out data-forms, pay a listing-fee and take a test. While registration is when students come, fill out forms, pay for listing fee and take a test. And those who pass the test I obliged to pay for tuition fee and some other costs. The candidates who are high school graduates are mostly the residents of Pematangsiantar City and Simalungun regency. The data the students have to fill in forms are those of the previous
schools, addresses, geographic area of living, and source of information about the availability of AMIK Tunas Bangsa. From the detail, data clustering can be constructed, which will lead to the making of an accurate method. And this method, is not only used to predict the interest of prospective students, but also to accelerate target-achievement. To manage the data obtained, a method that can be utilized to dig up hidden information is required.

This method is called “data mining “. With the help of software, data mining supported by algorithm C 45 will run a data analyzing process to predict the interest of prospective students enrolling at AMIK Tunas Bangsa. By using Algorithm C 45[1], this study will offer a pattern model in the form of “decision tree” so that private institutes are able to find out the marketing model on the basis of school types, status and locations of those interested in attending AMIK Tunas Bangsa by choosing the programs available. This study is expected to provide some contribution to Private Institutes especially to AMIK Tunas Bangsa Pematangsiantar. Based on the background above, the formulation of the problems in the research can be formulated as follows:

a) How to determine the interest patterns of high school graduates using decision tree for modeling to search for a solution to problems of new-student admission.

b) How to classify the determination of the best marketing model in schools in deciding the interest of prospective high school graduates that are based on schools, status, location and year of graduation.

c) How to analyze the interest of prospective high school graduates that is carried out with the help of decision tree supported by algorithm C 4,5

The formulation of the problem above is limited by several things as follows:

a) Technical algorithm of data mining to seize decision tree is algorithm C4,5.

b) The data was obtained from the forms issued by the New Student Admission Committee of AMIK Tunas Bangsa.

c) In the making of rule or prediction, the writer used “rapid miner 5,0 software to analyze data.

The objectives to be achieved in this research are:

a) Top gain a prediction of the interest of prospective students going to AMIK Tunas Bangsa.

b) To help the management to take preventive action for the marketing process for the coming years.

2. Methodology

Data mining is the term used to dig up hidden knowledge in database. Data mining is semi automatic process using statistical technique, mathematics, artificial intelligence and machine learning to extract and to indentify the information of potential and useful knowledge saved in big database [3]. According to Gartner Group, data mining is a process of finding meaningful relation, pattern and tendency by examining a large cluster of data saved in a saving system using a pattern identifying technique like statistical technique and mathematics [4]. Data mining is an automatic analysis of large number of data that is used to discover some important patterns or tendency whose availability is usually ignored [6]. The term of data mining and knowledge discovery in database (KDD) is usually interchangeably used to describe the process of hidden-information digging in a big database. In fact, the two terms have different concepts, but related to one another. And one of the steps of the whole KDD process is data mining.

The KDD process can roughly be described as follows: [3]

a) Data Selection

b) Pre-processing Cleaning

c) Transformation

d) Data mining

e) Interpretation evaluation

This phase includes an examination to find out if the pattern of the information is contrary to the fact or hipo-research that was previously available. The description above can be represented by figure 1.
In CRISP-DM, a data mining project has a lifecycle that is divided into 6 phases. Figure 2. The whole sequential phases are adaptive. The next phases in the range depend on output of the previous phases. An important relation between the phases is represented by arrows. For example: If a process is in a modeling phase, in accordance with behavior and characteristic model, the process may return to data preparation phase for further recovery to the data or move forward to evaluation phase [7].

Algorithm C4.5 is a method of making a decision tree that is based on previously provided training data. Algorithm C4.5 is a form of development ID3. Some of the development conducted to C4.5 is, among others, potential to overcome continue data and pruning. In general, algorithm C4.5 is to construct a decision tree as follows:

1. Select attribute as roots
2. Make a branch for each score
3. Divide cases in the branches
4. Repeat process for each branch until the whole cases in the branches have the same grade

To select attributes as roots is based on the highest Gain score of the available attributes. To count the Gain, the formula as seen in Formula 1 [1]

\[
Gain(S,A) = \text{Entropy}(S) - \sum_{i=1}^{N} \left( \frac{|S_i|}{|S|} \times \text{Entropy}(S_i) \right)
\]  

With:
- \( S \): Set of cases
- \( A \): Attributes
- \( N \): Number of Partition Attribute A
- \( |S_i| \): Number of cases in partition 1
- \( |S| \): Number of cases in S

While the calculation of Entropy Scores can be seen in Formula 2 [1]:

\[
\text{Entropy}(A) = \sum_{i=1}^{N} p_i \log_2 p_i
\]

With:
- \( S \): Set of cases
- \( A \): Features
- \( N \): Number of partition S
- \( p_i \): Proposal of Si against S
3. Results and Discussion
Data Mining is In this case of study, for data set, the writer collected data from New Student Admission Committee database which consists of number, names, types, school locations and study program of applicants. The data set can be seen in table 1

| No | Role     | Name       | Type      |
|----|----------|------------|-----------|
| 1  | Label    | Description| nominal   |
| 2  | Regular  | school     | nominal   |
| 3  | Regular  | The status of | nominal  |
| 4  | Regular  | location   | nominal   |
| 5  | Regular  | year       | nominal   |

In table 1, information attribute as label is the destination of school type attribute, school status, location and study program, The data available in table 1 selected for the research are available in table 3 by adding remark field on the basis of test winning and tuition paying.

| No | fee       | remark |
|----|-----------|--------|
| 1  | nor pay   | information |
| 2  | registration | listing |
| 3  | enrollment | registration |

A. Decision Tree Experiment Output
Sample Model of Decision Tree rules with graphic model of software rapid miner to be used can be seen in figure 3

![Decision Tree Grafik](image)

From figure 3, it can be described that status factor has the greatest effect on this research data. After having conducted an experiment using rapid miner in which the status variable is located on the top node, then compared with year of graduation and location to obtain information of registration and listing. For the simplification of decision tree in the form of arrangement model of induction rule, see figure 4
Figure 4: Model Rules

Arrangement Model of Induction Rule

From the above figure, the best rule can be simplified on the basis of gain Ratio, as seen in table 2 with R = Registration and D = Listing.

Table 3: Simplification Model of Induction rule arrangement

| Rule | Description Rule | R | D |
|------|------------------|---|---|
| 1    | if Status = Negeri and year = 2014 then Registrasi | 26 | 0 |
|      | if Status = Negeri and School = SMK then Registrasi | 21 | 0 |
| 2    | if year = 2014 and location = LPS then list | 0 | 20 |
|      | if Status = Swasta and location = PS then list | 0 | 18 |

Table 3 describes that Registration Rule consists of “if the school status is state owned and year of graduation is 2014, so is that if the status is state owned and vocational. Listing Rule consists of “if year of graduation is 2014 and the school is located outside the city boundary and Simalungun Regency and so is that if the school status is private owned and located in the city of Pematangsiantar. The pattern of high school graduates earned interest shows that the best predictor variables used are school status which gives the performance accuracy of 81.71% on registration and listing as in the Figure 5 below:

Figure 5: Accuracy Performance

4. Conclusion

This research resulted in some conclusions as follows:

a) Obtained a high school graduates whose interest in pattern registration is the status of the public schools and by 2014 completed new students.

b) In the case study on Computer Informatics Management Academy (AMIK) Tunas Bangsa Pematangsiantar that most high school graduates who came from the town of pematangsiantar and simalungun
c) Research  This the best predictor variables used has shows that is the status of the school performance accuracy that gives registration 81.71% with respect to registration and the list of.

Advice the author of this research are as the following:

a) To get the results of this research may be markedly more can be developed with engineering Fuzzy data mining others such as genetic algorithm, Decision Tree Association rule Neighbor for and algorithms looking for KNearest or rules approach to the model to be achieved.

b) In further research, testing the model rules can use the SSVM method (Smooth Support Vector Machine) as a means of testing the accuracy of the model rules of the truth

References

[1] Craw, S. (2005). Case Based Reasoning : Lecture 3: CBR Case-Base Indexing www.comp.rgu.ac.uk/staff/smc/teaching/cm3016/Lecture-3-cbr-indexing .ppt,retrieved on 11 maret 2015

[2] CRISP (Cross Industry Standard Process for Data Mining http://www.crisp-dm.org/Process/index.htm, retrieved on 11 maret 2015

[3] Fayyad, U. M, 1996, Advances in Knowledge Discovery and Data Mining. Cambridge, MA: The MIT Press.

[4] Larose D, T., 2006, Data Mining Methods and Models, Jhon Wiley & Sons, Inc. Hoboken New Jersey

[5] Merceron, A and Yacef, K. 2005. Educational Data mining: A case study. In proceedings of the 12th International Conference on Artificial Intelligence in Education AIED 2005, Amsterdam, The Netherlands, IOS Press, Vol 5, pp 1-8

[6] Pramudiono, I, 2006, Apa itu Data Mining ? dalam http://datamining.japati.net/cgi-bin/indodm.cgi?bacaarsip%115552761&artikel, retrieved on 11 maret 2015

[7] SPSS, Clementine 7.0 User’s guide, 2004