Application Of Decision Tree Approach To Student Selection Model– A Case Study

Harwati¹ and Amby Sudiya²

¹Industrial Engineering, Islamic University of Indonesia, Yogyakarta, Indonesia
E-mail: harwati82@gmail.com

²Industrial Engineering, Islamic University of Indonesia, Yogyakarta, Indonesia
E-mail: ambysudiya@gmail.com

Abstract. The main purpose of the institution is to provide quality education to the students and to improve the quality of managerial decisions. One of the ways to improve the quality of students is to arrange the selection of new students with a more selective. This research takes the case in the selection of new students at Islamic University of Indonesia, Yogyakarta, Indonesia. One of the university’s selection is through filtering administrative selection based on the records of prospective students at the high school without paper testing. Currently, that kind of selection does not yet has a standard model and criteria. Selection is only done by comparing candidate application file, so the subjectivity of assessment is very possible to happen because of the lack standard criteria that can differentiate the quality of students from one another. By applying data mining techniques classification, can be built a model selection for new students which includes criteria to certain standards such as the area of origin, the status of the school, the average value and so on. These criteria are determined by using rules that appear based on the classification of the academic achievement (GPA) of the students in previous years who entered the university through the same way. The decision tree method with C4.5 algorithm is used here. The results show that students are given priority for admission is that meet the following criteria: came from the island of Java, public school, majoring in science, an average value above 75, and have at least one achievement during their study in high school.

Keywords: Classification, Decision tree, Selection, Criteria, Student

1. Introduction

Data mining is the process of searching for a pattern or interesting information in the selected data by using techniques or methods. With the increasing sophistication of technology, databases are now able to store large capacity data. Data mining is useful whenever the system is dealing with large data sets (Agarwal, et al, 2012). In a data set which is very much present, hidden information stored is an important strategy. Rich data without analysis just makes the database only as a repository for data

¹ To whom any correspondence should be addressed.
that rarely visited. Consequently, important decisions are often made not by the rich data information stored in the database but only based on the intuition of decision makers. Data mining has been widely applied in various fields such as healthcare, manufacturing, banking, marketing, business, security, entertainment and so on. Today has emerged a renewed interest in research that uses data mining in the field of education. This is a relatively new field appears, called Educational Data Mining (EDM). EDM is emerging discipline, concerned with developing methods to explore patterns in the data that is unique to find the knowledge derived from the education domain (Abu Tair and El Halees, 2012). That knowledge is then used to better understand the students and the learning process. In the education system, the amount of hidden information in data stored in the database of institution increased rapidly. This database can contain student records such as registration details, profiles of students and academic achievement.

One data mining techniques that most useful in the field of education is a classification. Classification is the process of finding a model that describes and differentiates the data classes or concepts, in order to able to use the model to predict the class of the object whose class label is unknown (Umamaheswari and Niraimathi, 2013). Not only in education, but the classification has also been widely applied in various fields such as healthcare, manufacturing, CRM, even in the field of intelligence (Padhy, et al, 2012). The use of this technique on several studies in the field of education have provided benefits to educational institutions. The main objective of the institution is to prepare a quality education for students and to improve the quality of managerial decisions. One way to improve the quality of students is by select the new students more selective. At the Islamic University of Indonesia, admission new students consists of a wide variety of selection paths, one of them is Searching Student with Achievement (SSA). SSA is a pattern of admissions by selection for Academic achievement and interest talent. All the department admit students through the SSA by setting quotas and the selection is determined by the Faculty. Students for Academic Achievement, Student Achievement Division of Sports and Arts, Seed Excellence Scholarship, and Hafiz / Hafidzah Qur’an (www.pmb.ui.ac.id). Industrial Engineering Department at this time, especially not having a standard model selection criteria for SSA. Selection is only done by comparing the application file of prospective students in every registration schedule by management assessment, so that subjectivity is possible to occur because of the lack comparison criteria that can discriminate the quality of prospective students with each other. Certainly this issue demands to be addressed immediately considering SSA selection takes place every year in eight period of admission. If there is no standard model selection, it is highly possible to the injustice in accepting or rejecting new students where prospective students have the same relative performance.

By utilizing the technique of classification, it can be constructed a model selection for new SSA students that contains criteria to certain standards such as the area of origin, the status of the school, the average value of high school grade and so on. Such criteria are determined by using classification rules that appear based on academic achievement, in this case represented by GPA of SSA student in previous years. Classification will be implemented using Decision Tree algorithm C.4.5 with consideration of decision tree models preferred among others because the classification algorithm is an algorithm that easily developed and implemented (Anyanwu and Shiva, 2010). Selection models built is expected to become the standard for the selection of new students in the future, especially for the SSA track. Thus, subjectivity in determining the acceptance of SSA could be avoided and could ultimately homogenize the quality of new students accepted in the Industrial Engineering Department.

2. Aim of the Research

Referring to the problem above, the purpose of this study was to determine the criteria in the admission model of new student from SSA track based on classification academic achievement previous SSA student with decision tree technique.
3. Research Methodology

Classification is a process to find a model that describes or differentiates the concept or class of data, in order to be able to predict the class of an object whose class is not known (Umamaheswari and Niraimathi, 2013). There are several classification techniques that can be used, among others: Decision Tree, rule-based, neural networks, support vector machine, Naive Bayes and nearest neighbour. One method commonly used Data Mining is a decision tree. Decision tree is a flowchart structure that resembles a tree (the tree), where each internal node denotes a test in an attribute, each branch represents a test result, and a leaf node represents a class or class distribution (Sivaram and Ramar 2010). Groove on the decision tree traced from the root node to the leaf node that holds the class prediction for the example. Decision tree is easy to convert to classification rules. The concept of Decision Tree is essentially converts the data into a decision tree and decision rules.

Much of the research on the application of data mining in particular classification techniques in various fields such as the medical field, marketing, manufacturing and in the academic field. Chaurasia and Pal (2013) predict the possibility of cardiac patients by using three classification techniques well as the classification algorithm Naïve Bayes, J48 Decision Tree, and Bagging algorithm. By using 11 attributes such as gender, age, blood pressure, heart rate and other physical conditions, this study managed to classify patients with heart disease by about 85% and the accuracy of the pre-diction time is only about 0.05 seconds. It is very useful to predict the possibility that a person will develop heart dis-ease simply by looking at the previous data record in a short time and accuracy are relatively high.

Classification can also be used to predict the possibility of a customer of a bank, insurance and retail (Karim and Rahman, 2013). By using individual data such as status, occupation, education, home ownership status of these studies can predict a person's tendency to be open a savings account, a credit card or become a member in a place of shop-ping. The algorithm used is a C4.5 decision tree algorithm and Naïve Bayes classification as easily applied and under-stood by the relatively high degree of accuracy. Manufacturing also take advantage of this classification technique in decision making. As well as a study in Taiwan conducted on 66 of data to predict the failure of where 43 the data used as the training data set and the data 23 is used as a data testing (Yeh, et al, 2011). With the level of accuracy of 97.6% for the training set and 86.9% for data testing, the study resulted in rules of classification failure moulding process is based on several attributes that occur during the manufacturing process are: Temperature of injection, velocity, time packaging, injection pressure, injection time and so on.

In the academic field, Thomas and Galambos (2004) using a decision tree classification techniques to predict the satisfaction of students in the school. It is done to predict whether the student concerned shall continue education to the next level at the same school or not, as well as to determine what are the attributes that make the students decide to continue their education in the same school. Decision tree can also predict and evaluate the performance of students in a university (Baradwaj and Pal, 2011). The results can help university earlier to identify students were threatened by dropouts so that it can be given special treatment and provide an opportunity for the supervisor to provide guidance and counseling. Osmanbegović and Suljić (2012) have utilized the technique decision tree is compared with the two techniques other classification that Naive Bayes (NB) and the Multilayer Perceptron (MLP) for predicting the performance of students based on 12 attributes were: gender, origin of high school, the number of families, distance home, scholarship, learning time, the value of entrance exams and so on.
In this study classification techniques will be used to predict the performance of students who entered the university through the SSA where the rules obtained can be input for decision-makers in determining the criteria for acceptance of next new student who entering through the same track.

3.1 Attribute

Attribute functions as the object to be observed and the effect of the changes. In the new student admission system SSA track, some attributes are taken from the database that was available so it is not necessary anymore distributing questionnaires to obtain data. Some of these attributes include:

- Origin
- The origin of school (public or private)
- Department of high school (science or social)
- Average value
- Achievement
- GPA

Data cleansing is done to remove data that has entries imperfect as missing data, invalid data or also just a typo. Cleaning this data will affect the performance of the system because the data will decrease in amount and complexity.

3.2 Profile of Respondent

Data were collected from 179 secondary data of active students who entered through SSA track in the space of 3 years. The profile of the respondents is presented in the following table:

| No | Attributes                  | Profile                                                   |
|----|-----------------------------|-----------------------------------------------------------|
| 1  | Gender                      | Men 61%, Women 39%                                        |
| 2  | Origin                      | West Java 22%, Middle Java 40%, East Java 7.9%, Non Java 29% |
| 3  | High School Status          | Public 72%, Private 27%, Islamic 6.2%, others 5%          |
| 4  | Concentration/Majoring      | Science 97%, Social 3%                                   |
| 5  | Parents Occupation          | Gouverment Staff 39%, Private Staff 15.6%, Entrepreneur 30%, Others 15% |
| 6  | Average Value in High School| Excellent (above 75) 51%, Good 49%                       |

Meanwhile, distribution of achievements obtained by the SSA student during his education at the university are presented in the following figure:
4. Result and Discussion
In the process of data classification, C4.5 Decision Tree algorithm is used. This technique is very commonly used to construct a decision tree because it is easy to understand and flexible compared with other algorithms such as Naïve Bayes, CART, J48, and so on. The stages in the classification consists of: Develop tree, tree transform into rules, simplify and test rules, determines the final rules, counting accuracy and accuracy level

4.1. Information gain Calculation
Information gain (IG) is used as a reference for the formation of a branch in the decision tree. IG value indicates the effectiveness of an attribute in classifying data. The value of the highest IG will serve as the root node in the decision tree. IG is calculated using the following formula:

$$IG(S,A) = Entropy(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v)$$  \hspace{1cm} (1)

Information Gain calculation for all attributes used is shown in the following table:

| No | Attributes                  | Information Gain |
|----|-----------------------------|------------------|
| 1  | Gender                      | 0.001308678      |
| 2  | Origin                      | 0.016357926      |
| 3  | High School Status          | 0.06366159       |
| 4  | Concentration/Majoring      | 0.004976685      |
| 5  | Parents Occupation          | 0.003094282      |
| 6  | Average Value in High School| 0.005293175      |
| 7  | Achievement                 | 0.004571812      |
| 8  | Type of SSA (Academic/Art)  | 0.025250454      |

From the result above, it was found that the greatest value of information gain is school status. It means that attribute is the root node that is elected and for other attributes automatically become a certain level of internal node according with the calculation some iterations

4.2. Decision Tree Model
Classification calculations is assisted by application software Rapid Miner. One part of the resulting decision tree model presented in the figure below:
4.3. Rules Construction

There are 50 rules that are formed from the whole decision tree model. Some of the main rules involving all the attributes and characteristics of students with satisfaction GPA are:

- IF the status of the school is the State, the achievement is yes, origin is from Central Java, majoring in school is science: THEN GPA is Satisfactory
- IF the status of the school is the State, the achievements are not, origin is from Central Java, Value SSA is Good, the work of parents are civil servants, majoring in school is science, SSA is from Academic Track, gender is Female: THEN GPA is Satisfactory
- IF the status of the school is the State, the achievements are not, origin is from outside Java, the job of parents is the business man, gender is Male, majoring in school is science, Line SSA is Academic: THEN GPA is Not Satisfactory
- IF status is a State school, the achievement is yes, origin is from outside Java, the work of parents are civil servants, Value SSA is Good: THEN GPA is Satisfactory
- IF the status of the school is private, gender is Male, origin regions are East Java, Value SSA is Very Good parents work is businessman, majoring in school is science, Line SSA is Academic achievements are not: THEN THEN GPA is satisfactory
- IF is a private school status, gender is female, the region of origin is Central Java, the job of parents is businessman: THEN GPA is satisfactory

Table 1. Attributes for Satisfactory GPA.

| No | Attributes                        | Category     |
|----|-----------------------------------|--------------|
| 1  | Gender                            | Men, Women   |
| 2  | Origin                            | Java         |
| 3  | High School Status                | Public       |
| 4  | Concentration/Majoring            | Science      |
5. Parents Occupation  
6. Average Value in High School > 75  
7. Achievement  
8. Type of SSA (Academic/Art) Academic

### 4.4. Rules Construction

Accuracy level the proximity of measurement of the quantity of the actual values. In decision tree accuracy level is calculated by this following formula:

\[
\text{Accuracy Level} = \frac{\text{Number of correct Prediction}}{\text{Total Prediction}} \times 100\%
\]

Accuracy Level = \[\frac{15}{90}\] \times 100\%

Accuracy Level = 22%

The result shows that the prediction accuracy rate is above 22%. This means the possibility of failure prediction is 78%. Low levels of accuracy can be caused by several things, including the amount of data that is too few, too many variables, or even may be due to the random pattern of the respondents that are difficult to predict based on existing data.

### 5. Conclusion

From all the analysis above, it can be concluded that although the accuracy rate is very low but it can be predicted the characteristics of students form SSA admission that may be able to produce a good performance. These characteristics are came from the island of Java, public school, majoring in science, an average value above 75, and have at least one achievement during their study in high school.

### 6. Acknowledgement

This research was supported by a grant of Industrial Engineering Department of Universitas Islam Indonesia. The data, statement, and the view expressed are solely the responsibility of the authors.

### References

Abu Tair Mohammed M., El-Halees M Alaa, 2012, Mining Educational Data to Improve Students’ Performance: A Case Study , ISSN 2223-4985 Volume 2 No. 2.

Agarwal Sonali, Pandey G. N., Tiwari M. D, 2012, Data Mining in Education: Data Classification and Decision Tree Approach International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2.

Anyanwu N Matthew, Shiva G Sajjan G, 2010, Comparative Analysis of Serial Decision Tree Classification Algorithm International Journal of Computer Science and Security, (IJCSS) Volume (3): Issue (3)

Ayinde A., Dr Adetunji., Bello and Odeniyi, 2013 Performance Evaluation of Naive Bayes and Decision Stump Algo-rithms in Mining Students’ Educational Data, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 4, No 1
Baradwaj K Brijesh, Pal Saurabh, 2011, Mining Educational Data to Analyze Students’ Performance, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, page 63

Baradwaj K Brijesh, Pal Saurabh, 2011, Data Mining: A Prediction For Performance Improvement Using Classification, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4

Chaurasia Vikas, Pal Saurabh, 2013, Data Mining Approach to Detect Heart Dieses, International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol. 2, No. 4, Page: 56-66, ISSN: 2296-1739

Gubosh Soumi, Kumal K Sanjay, 2013, Comparative Analysis of KMeans and Fuzzy C Means Algorithms, IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No. 4

Kalyani M Raval, 2012, Data Mining Techniques, International Journal Of Advanced Research In Computer Science And Software Engineering Volume 2, Issue 10.

Karim Masud, Rahman M Rashedur, 2013, Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing, Journal of Software Engineering and Applications, 6, 196-206.

Keim A Daniel, 2002, Information Visualization and Visual Data Mining, IEEE Transactions On Visualization And Computer Graphics, Vol. 7, No. 1

Kotsiantis Sotiris, Kanellopoulos Dimitris, Association Rules Mining: A Recent Overview, GESTS International Trans-actions on Computer Science and Engineering, Vol.32 (1), pp. 71-82

N Sivaram, K Ramar, 2010, Applicability of Clustering and Classification Algorithms for Recruitment Data Mining, International Journal of Computer Applications (0975 – 8887) Volume 4 – No.5.

Osmanbegović Edin, Suljić Mirza, 2012, Data Mining Approach For Predicting Student Performance, Economic Review – Journal of Economics and Business, Vol. X, Issue 1.

Padhy Neelamadhab, Mishra Pragnyaban, Panigrahi Rasmita, 2012, The Survey of Data Mining Applications And Feature Scope, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3

Pujari Pushpalata, Gupta Jyoti Bala, 2012, Exploiting Data Mining Techniques For Improving The Efficiency Of Time Series Data Using Spss-Clementine, Journal of Arts, Science & Commerce E-ISSN 2229-4686ISSN 2231-4172, ol.– III, Issue 2(3).

Sahu Hemlata, Shrama Shalini, Gondhalakar Seema, 2012, A Brief Overview on Data Mining Survey, International Journal of Computer Technology and Electronics Engineering(IJCTEE) Volume 1, Issue 3:114

Thomas H Emily, Galambos Nora, 2004, What Satisfies Students? Mining Student-Opinion Data with Regression and Decision Tree Analysis Research in Higher Education, Vol. 45, No. 3

Umamaheswari K, S. Niraimathi, 2013, A Study on Student Data Analysis Using Data Mining Techniques, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, ISSN: 2277 128X Issue 8.

Yeh Duen-Yian, Cheng Ching-Hsue, Hsiao Shih-Chuan, 2011, Classification Knowledge Discovery In Mold Tooling Test Using Decision Tree Algorithm,) 22:585–595 DOI 10.1007/s10845-009-0321-7