Prediction Of Student Performance Using Decision Tree C 4.5 Algorithm

Rames Krisnan Kuntoro[1], Rukin Sudarwanto[2], Sriyanto[3]

Institute Informatics and Business Darmajaya , University Teknikal Malaysia

sriyanto@darmajaya.ac.id[3]

ABSTRACT

This paper aims to make predictions of student achievement based on socioeconomic status of parents, student discipline and student achievement using data mining method with algorithm decision tree classifier C 4.5. For comparison, the research data were analyzed also with CHAID (Chi Squared Automatic Interaction Detection) and multiple regression. The research approach used is quantitative. The subject of this research is the elementary school students in SD Negeri 4 Trimulyo. Data collection techniques used are documented. The results of this study are very helpful for educational institutions to monitor the early improvement of student academic achievement, so that can be accompanied the learning process in order to achieve the expected performance

Keywords: Data Mining, Classifier, Decision Tree, C45, CHAID, and Multiple Regression.

1. Introduction

The learning achievements of students are influenced by many factors. One important factor in education in order to increase student learning achievement is management of learning at school. The better the learning management offered to school students' increasingly greater likelihood of student learning achievement will be good.

Aware of the importance of the quality of learning in the framework of the intellectual life of the nation, then the Government together with the private circle of same has been and continues to strive to realize the mandate through a variety of educational development efforts more quality, among others, through the development and improvement of curriculum and evaluation system, improvement of the means of education, development and procurement of learning materials, as well as training for teachers and other educational personnel. But in fact the Government's efforts have not been enough means in enhancing the quality of education[1]. The process of teaching and learning is one of the determinants of the success of the education at the school. The low qualities of education are a result of the poor quality of the learning process that is conducted in the school.

1. One thing that hasn't been much done by institutions or elementary school is doing anticipation towards learners who potentially experienced barriers or underachievement in education. This is considered important because the early institutions of elementary school or know of any potential students who are likely to experience barriers in education, then the institutions or schools can do anticipatory measures.

2. Based on the description above, this research aims to make predictions early on against students who are potentially not doing or experiencing barriers in their learning, so that it can be done from the anticipatory measures for school prevent yourself from the possibility of not going up grade even the releasing students from school. Steps that can be taken after the school aware of any students who are potentially not doing is to do a special accompaniment towards the students. Hope the end is all the students from different backgrounds each factor can be maximized in the learning achievements of their students.

Some of the factors that influence the learning achievements of elementary school students, among others, parents, socioeconomic learned the school facilities, student discipline and student achievement. This is evidenced by the large numbers of studies that have been done before.

Socioeconomic parents status factors who realize financial ability. Different financial capabilities a little lot will have an effect on the learning achievements of students. With the financial ability of the elderly, will certainly affect the learning facilities provided by the parents against the infrastructure and facilities needed by a student to improve their learning achievements.

Factors of discipline in the management of teaching are a very important thing. Without any awareness of the necessity of carrying out predetermined rules that teaching is not possible reach maximum target. A student need to have the attitude of discipline by doing exercises that strengthen itself to always accustomed to obey and heightens the power of control of the self. The attitude of discipline arising from his own would be more stimulating and durable compared to the attitude of discipline arising due to the scrutiny of others.

Discipline can be grown and nurtured through practice, education or planting a habit that must be started in the family environment, starting at infancy and continues to grow so that it becomes an increasingly strong discipline. Just as mentioned by Tulus[2] that with a discipline that emerged because self awareness, students conditioned results in education, without the discipline of a good atmosphere of the school and also became less conducive to learning activities in a positive environment that support discipline calm and orderly atmosphere for learning, discipline is a way for students to be successful in learning and later while working due to the awareness of the importance of norms, rules, compliance and obedience is the success of a person.

Based on the explanation about the factors that influence student learning achievements above, this research was conducted with the aim to find out how big the contributions of the various factors against the achievements of the student learning and factors which are the most dominant, its contribution to the achievement of student learning. Expectations from the results of this research is the result of the formulation factors affecting learning achievement of students. Based on those factors, students are predicted to have strong factors are experiencing barriers to their learning achievements could be made in steps of anticipation early on against the students.

Studies on the factors affecting students prior learning achievements, many of which use data processing statistics. In this study, researchers using data mining. Data mining is a new branch of science in the field of computers, quite a lot of applications can do. It supported the richness and diversity of disciplines (artificial intelligence, database, statistics, mathematical modeling and image processing) to make the application of the data mining became more widespread. The main reason why data mining is very interesting information in the industry's attention in recent years is due to the availability of large amounts of data and the greater the need to transform the data into information and useful knowledge. Data mining is an activity extract or mine the knowledge from the data size/large, this information is very useful for later development.

This research uses quantitative approach, and conducted in public elementary school 4 Sub Trimulyo Sekampung, East Lampung regency, Lampung Province.

Related Work

Education in the life of a country holds a very important role to ensure the survival of the State and the nation. Based on the research that has been done by Paulo Cortez and Silvia that over the past decade educational level in Portuguese has increased. However, statistics show that the level of education the Portuguese ranked low due to the large number of students dropping out of school. Causes of student dropouts in Portugal because of the failure of students completing a few fields of study, i.e. majors in mathematics and Portuguese[3].

There are many ways that you can use to analyze the ability of middle school students, one of them that is data mining. Data mining can also be used as a prediction to estimate the value of the future. By applying these techniques will be built decision tree (decision tree) to the possibility that students can complete the study well. One of the techniques of data mining and decision tree can be used as a predictive algorithm C is 4.5. Which algorithm is a classification algorithm 4.5 C data with a decision tree technique that can manipulate numerical data (continuous) and discrete, can handle a missing attribute value, resulting in rules that are easily interpreted and fast among the other algorithms-algorithms. Previous research has already been done by Paulo Cortez and Alice Silva (2008)[3].

Such research does prediction of students in Portugal who majors in mathematics and Portuguese with how to determine the factors that may affect the value of the students. In the study, there are three methods of data mining are used to predict the ability of students with different accuracy, value, i.e. Naïve Predictor (60.5%-78.5%), Decision tree (depth 62.9% 76.1%), Random Forest (33.5%-36.7%).

The author chose to use the algorithm C 4.5 because it can perform predictions by providing an ideal level of accuracy values to predict students ' ability.

3. Research Method

Research procedure using stages KDD (Knowledge Data Discovery). Stages-stages are: (1) the Data Selection, i.e. stages was carried out to select the data that correspond to the required variable in research; (2) Preprocessing/ Cleaning, before the process of data mining can be carried out, needs to be done on the data cleaning process became the focus of KDD. The process of cleaning includes, among others, throw out the duplication of data, examine the data that were strongly inconsistent, and correct errors in the data, such as printing errors (typographical); (3) Transformation, coding is the process of transformation of data that has been selected, so that the data is suitable to process data mining; (4) The analysis of data; (5) Pattern Evaluation, the evaluation stage is to identify a pattern that is really interesting that represents knowledge based on an existing data source. The framework of thought that are used in this research can be seen in Figure 1.

The study obtained data from the documentation. Documentation technique used to take socioeconomic data parents, discipline of students and student performance.

This research uses the performance of Decision Tree C4.5 to perform predictive learning performance of students public elementary school 4 Sub Trimulyo a number of 390 students will be used to make predictions of the model Decision Tree. Models that have been created will then be calculated the level of accuracy of the predictions.

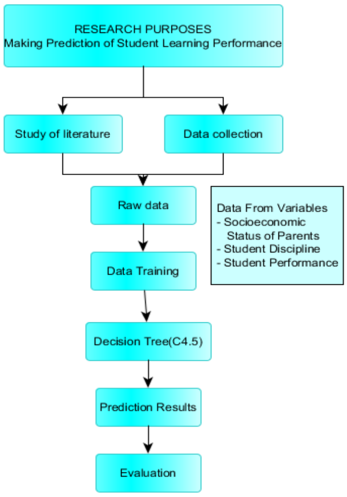


Figure 1. Data Analysis Techniques

1. Data Mining

According to Larose[4], data mining is defined as a process to discover relationships, patterns and trends of new meaning with a very large filtering data using pattern recognition techniques such as statistics and mathematical techniques. Into the tasks in data mining is generally divided into two main categories:

a) Predictive The goal of the task is to predict the predictive value of a specific attribute based on the values of other attributes. Predictable attribute that is commonly known as the target or the variable is not free, whereas attributes are used to make predictions is known as explanatory or the free variable.
b) Descriptive The purpose of the descriptive task is to lower the patterns (of correlation, trends, clusters, the trajectory and anomalies) that summarizes the principal relationships in the data. Deskripstif data mining tasks have been often a probing and often require postprocessing techniques for validation and explanation of the results.

2. Decision Tree

Decision tree model prediction is a technique that can be used for classification and prediction tasks. Decision tree using the technique of "divide and conquer" to divide the search space of the problem into a set of problems [5].

Process on the decision tree is changing the shape of the data table into a model tree. The model tree will produce a rule and simplified [6]. The concept of the decision tree in Figure 2.

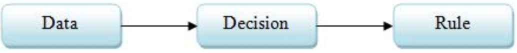


Figure 2. The flow of decision tree

Decision tree classification technique is one against an object or record. This technique consists of a set of decision nodes, and linked by branches, move down from the root node until it ends at leaf nodes [7].

3. Algoritma *C4.5*

There are several stages in making a decision tree algorithm in C 4.5[4], namely:

a) Prepare the training data. Training data are usually taken from the historical data that never happened before or referred to past data and are already grouped in specific classes.

b) Calculate the roots of a tree. The root will be taken from the attributes that will be chosen, by calculating the value of the gain of each attribute, the value of the gain of the most high that will become the root of the first. Before calculating the gains from the first count attribute, the value of entropy. To calculate the value of entropy used formula:

$$Entropy\ (S) = \sum_{i=1}^{n} -p_i \log_2(p_i) \qquad (1)$$

Description:
S = the set of cases
n = number of partition S
PI = The Proportion against S

c) Calculate the value of the Gain using Equation 2

$$Gain(S, A) = entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} Entropy\ (Si) \qquad (2)$$

Description:
S = the set of cases
A = Feature
n = number of partitions attribute of A
| SI = Proportion of Si against S
| S | = the number of cases in the S

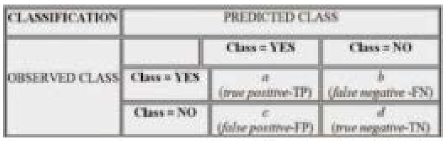
d) Repeat step 2 and step 3 until all of the records divided.

e) Process the partition decision tree will stop when:
- All records in node N got the same class.
- There are no attributes in the record which Partitioned again.

4. Evaluation Technic

a) Confusion Matrix

To do the evaluation of classification model based on calculation of object testing where the predicted correct and incorrect. This calculation is tabulated into a table called the confusion matrix. Confusion matrix is a data set has only two classes, one class as a class are positive and others as negative. Consisting of four cells i.e. True Positives (TP), False Positives (FP), the True Negatives (TN) and False Negatives (FN).



CLASSIFICATION	PREDICTED CLASS		
		Class = YES	Class = NO
OBSERVED CLASS	Class = YES	*a* *(true positive-TP)*	*b* *(false negative -FN)*
	Class = NO	*c* *(false positive-FP)*	*d* *(true negative-TN)*

Figure 3. Confusion matrix for class 2 models
To calculate the accuracy using the formula:

$$Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+FP+TN+FN}$$

b) Area Under Curve (AUC)

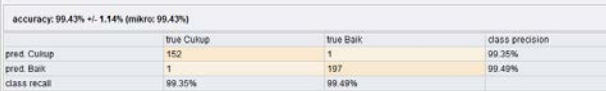
The AUC is widely recognized as the measure of a diagnostic test's discriminatory power. The AUC value measures discriminative performance by estimating output probability selected randomly from the positive or negative population. The larger the AUC, the stronger the classification used. AUC values ranged between 0.0 and 1.0.

4. Discussion

Research results based on the stages of the process in KDD (Knowledge Discovery Data) as follows: (1) the Data Selection, this step is done to select the data that correspond to the required variable in the study. The trick is to select or define an attribute-attribute data which will be used in the study of a group of operational data. One of them is determining the attributes of social, economic variables for Parents culled from operational data, i.e. Personal Data of students; (2) Pre-processing/ Cleaning, the cleaning process is done to the overall data examined that add up to 390 students. After making the process of cleaning the data a number of 390, produced data is clean as much as 352 record data that is used to process the following analysis; (3) Transformation, this stage yields one data recordset that is ready for data analysis; (4) data analysis.

Data analysis using decision tree algorithm C 4.5. The software used is Rapidminer 5. The results obtained are the prediction accuracy rate of 99.43%, as shown in table 1.

Table 1. The value of Data Accuracy Training Model Algorithm C 4.5



accuracy: 99.43% +/- 1.14% (mikro: 99.43%)

	true Cukup	true Baik	class precision
pred. Cukup	152	1	99.35%
pred. Baik	1	197	99.49%
class recall	99.35%	99.49%	

$$Accuracy = \frac{(TN + TP)}{(TN + FN + TP + FP)}$$

$$= \frac{(197 + 152)}{(197 + 1 + 152 + 1)}$$

$$= 0{,}99430$$

$$= 0{,}9943$$

In the meantime the results obtained from the processing of the AUC for algorithm C 4.5 using the training data of 0.550 where horizontal line is false positives and vertical line true positive, can be seen in Figure 4.

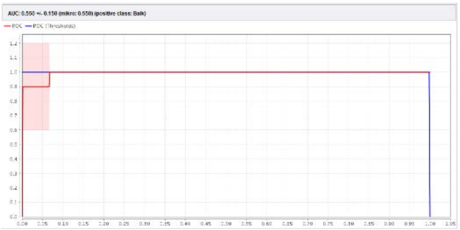


Figure 4. The curve AUC Algorithm C 4.5

5. Conclusion

Based on data analysis using decision tree, data mining to predict student learning achievement based on socioeconomic status of parents, discipline of students and student achievement using data mining methods obtained the following results: economic variables are the variables that determine the potential of a student success or not learning achievements in the future. This is evidenced by the existence of variables being the root node in the decision tree is formed. Variable student achievement, then was the second important variable in student success attended his studies. This suggests that aspects of knowledge or intelligence students very influential towards the success of their learning. In contrast, although students are less predictable knowledge, but with a high willpower can still be accomplished on at least category B or c. Average success algorithm c 4.5 in conducting classification data reached 99.43% in accuracy. This shows that this algorithm has a reliable performance in conducting classification.

Bibliographies

[1] Basuki, Ahmad dan Syarif, Iwan, 2003. “Decision Tree”. Surabaya: Politeknik Electronika Negeri.

[2] Cortez, Paulo & Silva, Alice Maria Gonçalves, 2008 “Using Data mining to Predict Secondary School Student Performance”. Portugal: University of Minho

[3] Dunham, Margareth H., 2003. “Data Mining Introductory and Advanced Topics”. New Jersey:Prentice Hall.

[4] Larose, & Daniel T. (2005). Discovering knowledge in data: an introduction to data mining. USA: John Wiley and Sons

[5] Tulus. (2004). Peran disiplin pada perilaku dan prestasi siswa. Jakarta: Grasindo

[6] Umaedi. (2001). Manajemen peningkatan mutu berbasis sekolah. Jakarta: Departemen Pendidikan Nasional Direktorat Jendral Pendidikan Dasar dan Menengah Direk-torat Sekolah Lanjutan Tingkat Pertama

[7] Yusuf W, Yogi, 2007. “Perbandingan Performasi Algortima Decision Tree C5.0, CART, dan CHAD: Kasus Prediksi Status Resiko Kredit di Bank X”. Bandung: Universitas Katolik Parahyangan.