Design and application data mining in academic fields

R Widyati*, F H Indiyah and M Mulyono

Computer Science Department, Faculty of Mathematics and Natural Science, Universitas Negeri Jakarta Jl. Rawamangun Muka No 1, Jakarta Timur, Indonesia

*ratna.widyati@unj.ac.id

Abstract. Classification is a data mining technique for a process of finding a model that explains or distinguishes a concept or class of data. The goal is to be able to estimate the class of an object whose class is unknown. Seeing the importance of education quality that must increase continuously, the study was conducted to evaluate the educational process in the department based on academic historical academic data which is used as input for which new and unknown classes of objects will be obtained. The level of competitiveness in entering a department in higher education is one of the things that becomes an assessment of department accreditation. Thus it is necessary to analyze how many students accepted into a department do not move to other department the following year. The analysis used C4.5 methods and in the process used Rapid miner software to make decision trees. The results obtained an accuracy rate of around 61.2% with confusion matrix.

1. Introduction

The academic field is a field that includes academic activities such as teaching, research, community service, administrative evaluation and academic development in the higher education environment. Achievement or academic performance is a term to show an achievement level of success about a goal because a learning effort has been carried out by someone optimally. Learning efforts can be done in universities, because higher education is a place to get knowledge. Academic performance of an academic community can be seen from the results of the learning process in higher education so that the achievement of the level of academic success can be influenced by the quality of college education.

In improving the quality of data mining education which is a semi-automatic process to gain knowledge from a set of data can be used to evaluate academic performance of the academic community. Classification is a data mining technique to process placing an object or concept into a set of categories based on the object or concept concerned and for find a model or function that describes and distinguishes a data class or concept in order to use a model to make predictions of object classes where the label class is unknown [1,2]. Seeing the importance of the quality of education that must continue to increase, research is conducted to evaluate the educational process in the study program based on academic data which is used as input for which new classes of unknown objects will be obtained.

Achievement or academic performance is a term to show an achievement level of success about a goal because a learning effort has been carried out by someone optimally. Learning efforts can be done in universities, because higher education is a place to get knowledge. Academic performance can be seen from the results of the learning process in higher education so that the achievement of the academic success level of the academic community can be influenced by the quality of college education.
This research broadly covers several core activities, namely making proposals, collecting data, processing data, implementing methods, testing, and analyzing results. At the processing stage there are several activities in accordance with the stages that exist in data mining, namely data cleaning, data integration, data selection, data transformation, and the formation of datasets which will be used as training data and testing data.

One example of the application of data mining in the academic field is Sivasakthi's research that predicts student performance can be more challenging because of the large volume of data in the education database [3]. Other research related to the academic field is research on the relationship between students entering university the results of the examination and its success were studied using cluster analysis and the K-Means algorithm technique, predicting the academic performance of students with the application of K-Means clustering algorithm [4,5].

There is also research focused on developing data mining models for predict student performance, based on their personal, pre university and university performance characteristics [6], how to predict academic performance of the new students so that the lecturers will know the level of the new students’ preparedness at admission [7]. Several studies have been carried out using techniques data mining to explore various information from a database college student [8,9].

2. Material and methods

2.1. Data
In this study the data is used as follows:
- Data training and testing data
- Target data
- History data

2.2. Research stage
The plan of this study, the stages of research carried out are as in Figure 1. This study broadly covers several core activities, namely making proposals, collecting data, processing data, implementing methods, testing, and analyzing results. At the processing stage there are several activities in accordance with the stages that exist in data mining, namely data cleaning, data integration, data selection, data transformation, and the formation of datasets which will be used as training data and testing data.

![Figure 1. Stage of research.](image)
2.3. Design system
The following is the system design used in this study:

2.3.1. System architecture. In the system components are divided into four environments, namely the database, engine, knowledge base, and user interface.

2.3.2. System modeling. In this study the system is modeled using use case diagrams to model the behavior of the system to be created.

![System model](image)

**Figure 2.** System model.

2.3.3. Algorithms. At this stage the algorithm will be used.

2.3.4. Database design. In database design some tables, attributes, and relationships are described which will be used as storage of training data, testing data, and target data in this study.

3. Results and discussion

3.1. Analysis
In this chapter is a description of the process of analyzing a problem and a description of the application of methods or algorithms to be able to solve the problem at hand. To support data analysis in knowledge search, a manual data transformation is found in the State University of Jakarta Computer Science Study Program. In this case the system design is a database designer using the WPS Spreadsheet where the application program is very helpful and supports and verifies the problem solving analysis.

3.2. System architecture
In data mining, there are several data processing techniques so that the data is more useful and valuable. The C4.5 Algorithm method is one of the techniques that can be used. The database used is a data collection, namely:

- Data of Student UKT
- Student scholarship data
- Data on student GPA
With the same data, then the data will be processed and analyzed using the C4.5 methods. For the use of the C4.5 Algorithm it is also done in several stages, namely as follows:

- Determine the Root Attribute
- Making branches for each value
- Distribution of Cases in each branch
- Establish a Decision Tree
- General Rules

3.3. Data collection
The method of data collection is carried out with data contained in the UNJ Pustikom. Student data needed is data on students entering whether or not bidik misi, scholarship data for active students, active student GPA data and student UKT data.

3.4. Pre-processing data
Pre-processing Data is one of the steps used to validate a data that will be tested. In pre-processing one of the steps used is to transform each value of the same attribute into a numerical form so that it is easy to do for the problem solving process and the formation of sample data.

3.5. Designing
Existing applications are Rapid Miner, so in this design process there are 2 (two) things that will be discussed, namely:

- Database Design
- System Modeling

4. Conclusion
The design of data mining applications that are made using Rapid Miner, so there are two things, namely database design and system modeling. For database design created using WPS Spreadsheet Tools while for modeling this system used is UML (Unified Modeling Language) and Flowchart. From 143 data, the value with a good predicate is 114, less predicate is 29, with the entropy value is -0.203694657. The data used are data from Computer Science Study Program students from the year of 2013 to 2017. The highest gain values for the UKT criteria are 0.051331889. The results obtained for the case studies discussed are the highest criteria for grouping UKTs that still produce a GPA in the high group. The criteria for entry through the Bidik Misi path also produce a GPA in the high group as well. The results obtained an accuracy rate of around 61.2% with confusion matrix.

Acknowledgment
The author would like to thank Lembaga Penelitian dan Pengabdian Masyarakat State University of Jakarta for funding under the schema research Faculty of Mathematics and Science.

References
[1] Han J and M Kamber 2006 Data Mining: Concepts and Techniques, Second Edition (San Fransisco: Morgan Kaufmann Publishers)
[2] Larose D T 2005 Discovering Knowledge in Data: An Introduction to Data Mining (John Willey & Sons, Inc)
[3] Sivasakthi M 2018 Classification and prediction based data mining algorithms to predict students' introductory programming performance International Conference on Inventive Computing and Informatics (ICICI)
[4] Erdogan S Z and Timor M 2005 A Data Mining Application In A Student Database Journal of Aeronautics And Space Technologies 2(2) 53-57
[5] Shovon H I and Haque M 2012 Prediction of Student Academic Performance by an Application
of K-Means Clustering Algorithm *IJARCSSE* 2(7) 353-355

[6] Kabakchieva D 2012 Student Performance Prediction by Using Data Mining Classification Algorithms *IJCSMR* 1(4) 686-690

[7] Garcia E P I and Mora P M 2011 Model Prediction of Academic Performance for First Year Students *IEEE Computer Society*

[8] Hien N T N and Haddawy P 2007 A Decision Support System for Evaluating International Student Applications *IEEE Frontiers in Education Conference*

[9] Ogor E N 2007 Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques *IEEE Computer Society*