Classification of Subject Concentration using Algorithm C4.5

A P Fadillah* and B Hardiyana
Universitas Komputer Indonesia, Jl. Jalan Dipatiukur No. 112-116 Indonesia

*annisa@email.unikom.ac.id

Abstract. Universitas Komputer Indonesia (UNIKOM) is one of developing a private university in Bandung. Information system department is one of the big department at UNIKOM. Due to the applicable curriculum at information system department, the students who want to continue their study to 6th semester, they have to choose their subjects concentration. There is the unknown pattern classification of the subject selections, because of there is not standard to manage the selection of subject concentration at information system department of UNIKOM. This is because of the less of processing data maximally. It is necessary to do the classification pattern analyzing of subject’s concentration selection deeply, therefore the hidden pattern can be known and found out the things that affected to the selection of subjects. The resulting pattern can be used to design the standard selection of concentration subject at information system department. The analysis in this research used Algorithm C4.5. The selection of Algorithm C4.5 as an algorithm to facilitate the formation of a decision tree that able to create the selection pattern of subject concentration. The resulting pattern can be a guide to design a recommendation for selecting the subject concentration of students at information system of UNIKOM.

1. Introduction
UNIKOM is one of the universities that use technology development in the form of information system services. The main information system is one of the largest majors in UNIKOM, with a student population of over 200 students in each generation. In accordance with the curriculum applicable to the current major information system, students who will continue their studies in the 6th semester should make a choice of courses based on their interests. Students in 6th semester have difficulty in choosing subject concentration, because there is no pattern or standard that become reference in choosing subject concentration. So that lecturers also have difficulty to give recommendation of subject concentration election. To know the pattern or standard of subject selection, then used an algorithm to do the process of classification by using algorithm C4.5 with decision tree.

In this research entitled "Classification of Selection of Core Subject Concentration Using Algorithm C4.5" apply algorithm C4.5 as pattern forming to know the concentration of course, beside research conducted there is other research applying algorithm C4.5 as algorithm which assist in making decision tree, the title of the research is "Implementation of Algorithm C4.5 in Application Creation to Analyze Possible Resignation of New Student Candidates". This research discusses the making of application to analyze the possibility of resignation of new student candidate. The analysis is conducted by using case-based reasoning theory, by comparing new prospective cases with previous cases [1].

Another study used the C4.5 algorithm with the title "Application of CRISP-DM Method for Graduation Prediction of Student Study Taking Courses (Case Study of XYZ University)". This study discusses the predictions of students' graduation courses in taking courses. Differences between research
conducted with the research, namely in research with the title "Application of CRISP-DM Method for Prediction of Graduation of Student Study Taking Courses (XYZ University Case Study)" data used only focus on one course, while research is done determine the pattern, for the selection of subject concentration in the sixth semester students with research data of several courses related in the selection of the concentration of the course [2]. The research that will be carried out refers to a study entitled "Data Clustering using CRISP-DM methodology for pattern recognition of Tridharma implementation proportions", but the theme raised discusses the classification of selection of concentration courses using C4.5 algorithm, the difference is the object of research where the research will be conducted in Information Systems Study Program in UNIKOM, and using k-means algorithm, but using C4.5 algorithm used to build decision tree [3]. Another study using a C4.5 algorithm titled ‘predicting students’ performance using ID3 AND C4.5 CLASSIFICATION ALGORITHMS’. This study discusses the performance prediction of new students enrolled in the future, using a classification algorithm ID3 (Iterative Dichotomiser 3) and C4.5 [4]. Other studies related to the student's learning process is research with the title ‘Mining Student Behavior Models in Learning by Teaching Environments'. This paper discusses the approach to model building and analysis student behavior in different versions of learning with teaching environment where students learn by teaching a computer agent named Betty using visuals concept map representation [5]. Other research that discuss about the data mining that is ‘Data Mining Algorithms to Classify Students’. This study compares various methods of data mining and techniques for classifying students based on their Moodle usage data and final marks are obtained in each course [6].

2. Method
Data collection techniques can be done by observation, interviews, and using questionnaires. Data Preparation, Data that has been collected in the previous stage, then the data preparation process will be done, by collecting the data of the students as a whole.

Modelling is a direct step involving data mining. Selection of data mining techniques, algorithms and determining parameters with optimal values. In the modelling stage, there are several things done, among others selecting modelling techniques, modelling, and modelling [3].

Draft Recommendation, From the design pattern that has been done in the previous stage, then at this stage can generate recommendation from result of pattern analysis that exist, and hopefully this recommendation can become reference or standard for student subject selection and expected to support decision in choosing of subject concentration.

3. Results and discussion
The process of conducting the research is based on the method described earlier.

3.1. Data collection
From the results of data collection obtained sample data of 40 students. The data retrieval process is done by using samples in the 2013 student class that are available in the department of information systems. The data used is the data of students actively doing learning activities. From the data is known:
- There are 22 students who choose the concentration of Information Systems Engineering subjects
- Visible 15 Students who choose the concentration of Information Technology subjects
- While there are 3 Students who have not chosen the subjects.

3.2. Data preparation
The data used for the research is the student's transcript data.

3.3. Modelling
The initial stage is done before modelling data using data mining, that is knowing the pattern of curriculum used so that it can be known subject related to the selection of concentration on the sixth
semester students. From the subject curriculum pattern can be known subject which became the initial standard for the choice of subjects, which can be seen in the Table 1.

| No | Information Technology | Information Systems Engineering |
|----|------------------------|---------------------------------|
| 1  | Programming Laboratory I | Information System Concepts     |
| 2  | Programming Laboratory II | Business Process Analysis       |
| 3  | Programming Laboratory III | E-Business Concepts             |
| 4  | Web Programming         | Information Systems Management  |
| 5  | Computer network        |                                 |

From the data that has been collected above then the next step is to separate the sample data value of students based on the concentration of subjects, in order to calculate patterns with data mining algorithms can be done.

The collected value data is then converted into a predicate form, in accordance with the rules applicable at the university, which can be seen in the Table 2.

| VALUE       | INDEX | PREDICATE          |
|-------------|-------|---------------------|
| 80 ≤ NA ≤ 100 | A     | PASS, Very Good    |
| 68 ≤ NA ≤ 79  | B     | PASS, Good         |
| 56 ≤ NA ≤ 67  | C     | PASS, Enough       |
| 45 ≤ NA ≤ 55  | D     | PASS, Less         |
| 0 ≤ NA ≤ 44   | E     | NOT PASS           |

After the sample data is converted then calculated using data mining algorithm C4.5, from the calculation then produced a pattern From the pattern that has been formed in Figure 1 (in Appendix), from the sample data, obtained some criteria to make the process of choosing subject concentration.

4. Conclusion
Based on the results of research conducted, it can be concluded as follows:
- Algorithm C4.5 is used to perform decision tree formation. Decision tree that formed to produce pattern of course election in study program of information system with conclusion result of concentration of course of Information System Engineering and Information Technology.
- The decision tree pattern resulting from Algorithm C4.5 in the department of information systems is highly dependent on the assessment of each course given by each lecturer in the course.
- Recommendations produced can be used as a standard for the selection of subject concentration for students of UNIKOM information system study program.

Acknowledgements
Acknowledgments are addressed to RISTEKDIKTI who has funded this research, and to Rector UNIKOM, INCITEST team and lecturer staff and secretariat staff majoring in information system.

References
[1] R Gelar G 2011 "elib.unikom.ac.id," [Online]. Available at http://elib.unikom.ac.id/gdl.php?mod=browse&op=read&id=jbptunikompp-gdl-ranggagela-26808. [Accessed 24 Mei 2015].
[2] Fadillah A P 2015 "Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ)," JUTISI 1 260-270.
[3] Budiman I, Prahasto T and Christyono Y 2011 "Data Clustering menggunakan metodologi CRISP
–DM untuk pengenalan pola proporsi pelaksanaan Tridharma," Jurnal Sistem Informasi Bisnis 1 (3) 129-134.

[4] Adhatrao K, Gaykar A, Dhawan A, Jha R and Honrao V 2013 "PREDICTING STUDENTS’ PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS," International Journal of Data Mining & Knowledge Management Process (IJDKP) 3 (15) 39-52.

[5] Jeong H and Biswas G 2008 "Mining Student Behavior Models in Learning-by-Teaching Environments," in The 1st International Conference on Educational Data Mining, Montréal, Québec, Canada.

[6] Romero C, Ventura S, Espejo P G and Hervás C 2008 "Data Mining Algorithms to Classify Students," in The 1st International Conference on Educational Data Mining, Montréal, Québec, Canada.
Figure 1. The decision tree pattern generated from the calculation of the c4.5 algorithm.