MULTIMEDIA UNIVERSITY ®

TDS3301 Data Mining

Group Assignment (Part 1)

Exploratory Data Analysis

Prepared by:

Position	Student ID	Name	E-mail
Leader	1142701655	Nicholas Tan Yu Zhe	nicholas.290696@gmail.com
Member	1142700814	Choo Jia Sheng	jason952002@gmail.com
Member	1142700808	Liew Soon Pang	liewsoopang@gmail.com
Member	1142701084	Chow Chan Kit	chankitchow@gmail.com

Name: Students' Academic Performance

URL: https://www.kaggle.com/aljarah/xAPI-Edu-Data

A. Describe the dataset in your own words

This dataset is record all about student's academic performance. It contains 478 record of student with 17 variables. The dataset describe the gender, nationality, birth place, educational stage, grade level, classroom section, field of study, semester, responsible parent, number of time hand raised in the class, number of time course content visited, number of time discussion group participated, did parent answered the survey given by school, did parent satisfy the school, the absence day of student and class of student based on their mark. This dataset has inconsistence naming of their column name. For example, "NationalITy" have inconsistence naming across other column name. Other than that, we have found out some of the column name is not relevant to the column data. For example, "Discussion" is not relevant to the data whereas "participate_in_discussion_group" is more understandable and relevant to the data itself.

B. What possible insights can be obtained from mining the chosen dataset?

The insight we want to obtain from mining this dataset is will the students' academic performance remain or change when they go to the next semester. We want to find ways to increase their academic performance to at least a medium.

C. What type of data mining technique (association rule mining, classification or clustering) would be relevant? Give an example, for example, if you think classification is suitable, describe what will be classified and what the possible classes are.

Association rule mining will be the best technique for the dataset that we have chosen. By using this technique, we are able to see that some of the data might look like they have no direct relationships but actually they are connected. It will analyze the data and show the patterns with the help from criteria support and confidence. For example, in this dataset if the parents are satisfied with the school then most of the student's absence days will be under 7. It will also affect the performance of the students in the near future. For those students whose absence days are under 7, they will normally get an academic performance of medium (M) or high (H).

D. Describe data quality issues, and be specific. Identify which attribute (column) has issues, or if the structure of the data has problems.

When comparing our dataset with the six core dimensions of data quality consisting of accuracy, completeness, consistency, timeliness, uniqueness and validity, there were some issues displaying the content which is caused by not having a suitable column name:

- NationalITy – Inconsistence naming
- class – the column name doesn't imply the meaning of the column
- Discussion – the column name and its value is not relevant to each other