TDS 3301

Data Mining (Phase 1)

Group Leader: Benjamin Au Wen Wei 1141327105

Group Member 1: Mah Siew Chin 1132702455

Group Member 2: Oh Soon Kit 1132702874

Group Member 3: Tan Chong Raen 1121115725

In Part 1, the group needs to identify and characterize a data set. The datasets can be acquired from resources such as Kaggle, CrowdFlower, and UCI Machine Learning Repository etc.

Prepare a report on the chosen dataset by completing the following tasks:

A. Describe the dataset in your own words.

The particular dataset chosen was from Paulo Cortez and Alice Silva's work, "Using Data Mining To Predict Secondary School Student Performance". Generally this dataset describes the poorly performed Portuguese students on Mathematics and unfortunately their own native language, Portuguese. There are various reasons to be considered in affecting the student's performance such as travelling time to school, parents' cohabitation status.

However, the statistics made Portugal less efficient than Europe end due to its high student failure and dropping out rates. In fact it is very a serious problem for Portuguese students not able to master Mathematics and their native language Portuguese.

The recent world data such as student grades and school related features was collected with school reports and questionnaires, and integrated into 2 datasets namely Mathematics with 395 records, and Portuguese language with 649 records. The classes contain 33 variables. The 2 datasets used are not merged, also training sets and test sets cannot be used due to their differences in both datasets.

Attribute Information:

1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - student's sex (binary: 'F' - female or 'M' - male)
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: 'U' - urban or 'R' - rural)
5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. guardian - student's guardian (nominal: 'mother', 'father' or 'other')

13. traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic - with a romantic relationship (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)
30. absences - number of school absences (numeric: from 0 to 93)

These grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

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

- **Improvement on accuracy**

The chosen datasets that are extracted to mine are ranged from the year of 2005 and 2006. It is not convincing enough to believe in the accuracy of the possible mining result for only two years. The range of the year should be extended in order to acquire better accuracy of the mining result.

- **Dataset lifespan should be extended**

The datasets used in this research only lasted for 2 years from 2005 – 2006. To provide better results, in our opinion it should have extended longer time frame perhaps for another 3 – 5 years to collect more data for a more accurate and precise prediction.

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.

One of the data mining technique used is binary and five-level classification, which intends to forecast student performances along with factors and key variables that may affect their outcome. One of the usage of binary classification is to determine pass or fail of a student. There are 3 types of grades namely G1, G2, and G3. Regardless of the grades, a student must be able to achieve at least 10 ranging from 0 – 20 in order to pass. Whereas the final grade (G3) will be indicated as the target variable to evaluate the academic performance of the students who are taking Mathematic subjects and Portuguese subjects.

On the other hand, five level classification is based on Eramus grade conversion system, whereby is a European exchange programme that enables student exchange in 31 countries. This classification system ranging from 0 – 20 has 5 grading systems namely I (excellent / very good), II (good), III (satisfactory), IV (sufficient), and lastly V (fail). The score benchmark ranges from 16 – 20, 14 – 15, 12 – 13, 10 – 11, and 0 – 9 respectively for Portugal and France; grades A, B, C, D, and F respectively for Ireland.

Both classifications use and classify the students' performances on the subjects Portuguese and Mathematics.

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

Currently the datasets used have no missing values and no noise. It is considered as untidy data which requires only tidying up data. Melting is not required as all the attributes and objects are in proper manner. One of the important step to be taken is to ensure both datasets are merged in order for pre-processing to be done. All values in the datasets are unique as in no duplication in values. Outliers do exist however the range is within the box-plot. Examples of variables used would be absences and age. Both datasets Mathematics and Portuguese language are used as training sets.

E. Perform a pre-processing task on the dataset chosen.

You may refer the attachments uploaded along with this report for this section.

Reference:

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.