**CSC105M: Dataset Writeups**

Ryan Austin Fernandez

CS-ST Student

2401 Taft Avenue,

Manila, Philippines 1004

(+63)917-621-6469

ryan\_fernandez@dlsu.edu.ph

Clarisse Felicia Poblete

CS-ST Student

2401 Taft Avenue,

Manila, Philippines 1004

(+63)915-624-5417

clarisse\_poblete@dlsu.edu.ph

**I. THE STUDENT ALCOHOL CONSUMPTION DATASET**

This dataset provides information about two sets of students, one from a math class and another from a Portuguese language class. It is a multivariate data set with 32 attributes and 1044 instances, donated on March 3, 2016. The attributes included are as follows:

**Table 1. Attributes of the Student Alcohol Consumption Data Set**

|  |  |  |
| --- | --- | --- |
| **Attr #** | **Attribute** | **Description** |
| 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 5th to 9th grade, 3 secondary education or 4 higher education) |
| 8 | Fedu | father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 5th to 9th grade, 3 secondary education or 4 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 | gout | 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) |
| 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) |

Tasks that can be performed are prediction of grades, social, gender, and study time attributes (attrs # 23,24,25,26,29,30,31,32) to determine what factors most affect these predictable variables.

**2. WINE QUALITY DATA SET**

This dataset includes two sets of data: one for red wine and one for white wine. It is multivariate data with real values, with 4989 instances and 12 attributes, donated on October 7, 2010.

The attributes are as follows:

**Table 2. Attributes of the Wine Quality Data Set**

|  |  |
| --- | --- |
| **Attr #** | **Attribute** |
| 1 | fixed acidity |
| 2 | volatile acidity |
| 3 | citric acid |
| 4 | residual sugar |
| 5 | chlorides |
| 6 | free sulfur dioxide |
| 7 | total sulfur dioxide |
| 8 | density |
| 9 | pH |
| 10 | sulphates |
| 11 | alcohol |
| 12 | quality (score between 0 and 10) |

Regression analysis to predict the quality of the wine can be performed to classify the wines and determine via regression coefficients which chemical attributes have the most effect on wine quality.

**3. BOOK-CROSSING DATASET**

This is a dataset that is made up of the following 3 tables:

* BX-Users, which contains a list of 276272 users with their locations and ages
* BX-Books, which contains a list of 113032 books with their titles, authors, years of publication, publishers and image URLs
* BX-Book-Ratings, which contains a list of 1000 ratings, with values from 0 to 10, with their corresponding user IDs and book ISBNs

The denormalized version of this dataset contains the following attributes:

**Table 3. Attributes of the Book-Crossing Dataset**

|  |  |  |
| --- | --- | --- |
| **Attr #** | **Attribute** | **Description** |
| 1 | ISBN | The book’s ISBN |
| 2 | User-ID | The user’s ID |
| 3 | Book-Title | The book’s title |
| 4 | Book-Author | The book’s author |
| 5 | Year-of-Publication | The book’s year of publication |
| 6 | Publisher | The book’s publisher |
| 7 | Location | The user’s location |
| 8 | Age | The user’s age |
| 9 | Book-Rating | The user’s rating of the book |

This dataset can be used to come up with recommendations. Recommendations can be based on a user’s age or location, on publishers, authors, or years of publication of books that were rated highly by the user, or on the user’s own ratings, e.g. through recommending books that were rated highly by users that highly rated books that you rated highly. This data can also be used to profile the types of readers that enjoy specific books, based on the age or location of the users rating these books highly.

**REFERENCES**

[1] A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal

[2] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, Georg Lausen; Proceedings of the 14th International World Wide Web Conference (WWW '05), May 10-14, 2005, Chiba, Japan. To appear.

[3] Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez

[4] 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.