DATA MINING LAB COURSE
SOSE 2016

PRESENTED BY

SOHEL MAHMUD

Masters in Informatics

Matriculation Number : 03663613

2

Proposed Dataset
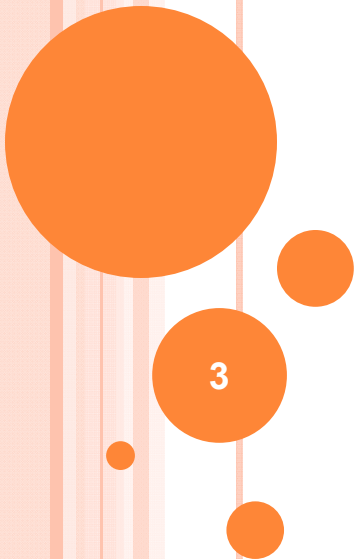
STUDENT ALCOHOL CONSUMPTION DATA SET

Source:
Fabio Pagnotta, Hossain Mohammad Amran.
University Of Camerino, Italy

Collected by:
Paulo Cortez and Alice Silva
University of Minho

3

Repository: UCI Machine Learning Repository

University of California, Irvine

WHY???

- The data illustrates the social impact of alcohol on the academic life of students in Portugal.

- The dataset is interesting enough for doing research in Social Computing as it provides the cause and effect relationship.

- Using this data, we can determine the effects of family life of a student along with the social one.

4

WHAT???

- ***Data Definition***: The dataset is about Portuguese student of secondary school on two courses(Mathematics and Portuguese) regarding the addiction on alcohol.

- ***Data Selection Area:*** **Social**

- ***Data Set Characteristics:*** **Multivariate**

- ***Associated Tasks***: **Classification**

- The dataset has been divided into two parts:
  - *student-mat.csv (for the student of Mathematics)*
  - *student-por.csv (for the students of Portuguese)*

- The Number of attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets is **32**

DATA FILE DESCRIPTION

- school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)

- sex - student's sex (binary: "F" - female or "M" - male)

- age - student's age (numeric: from 15 to 22)

- address - student's home address type (binary: "U" - urban or "R" - rural)

- famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)

6

DATA FILE DESCRIPTION

- Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)

- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

- Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")

7

DATA FILE DESCRIPTION

- Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")

- reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")

- guardian - student's guardian (nominal: "mother", "father" or "other")

- 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)

- studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

8

DATA FILE DESCRIPTION

- failures - number of past class failures (numeric: n if 1<=n<3, else 4)

- schoolsup - extra educational support (binary: yes or no)

- famsup - family educational support (binary: yes or no)

- paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

9

DATA FILE DESCRIPTION

- activities - extra-curricular activities (binary: yes or no)

- nursery - attended nursery school (binary: yes or no)

- higher - wants to take higher education (binary: yes or no)

- internet - Internet access at home (binary: yes or no)

DATA FILE DESCRIPTION

- romantic - with a romantic relationship (binary: yes or no)

- famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

- freetime - free time after school (numeric: from 1 - very low to 5 - very high)

- goout - going out with friends (numeric: from 1 - very low to 5 - very high)

DATA FILE DESCRIPTION

- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

- Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

- health - current health status (numeric: from 1 - very bad to 5 - very good)

- absences - number of school absences (numeric: from 0 to 93)

DATA FILE DESCRIPTION

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

GOAL

- Prediction of secondary student grades of two core classes (Mathematics and Portuguese) by using the variables:
    - past school grades (first and second periods)
    - Demographic (Student's age, parent's job and education)
    - Social (going out with friend's, alcohol consumption)
    - Other school related data (number of absences, reason to choose school, extra educational school support)
- Data Mining Goal:
    - Classification
    - Regression

14

THANK YOU

15