Dr. Eick

COSC 4335 *“Data Mining”* Fall 2018

Assignment1: (Exploratory) Data Analysis

for a Portuguese School Performance Dataset

Individual Assignment

Third Draft



Due: Tuesday, September 18, 11p (electronic Submission)

Last Updated: Sept. 5, 10a

Tentative weight of Assignment1: 18-20% of the points allocated to the four assignments

**Learning Objectives**:

1. Learn how to manage and preprocess datasets and how to compute basic statistics and to create basic data visualizations (using R)
2. Learn how to interpret popular displays, such as histograms, scatter plots, box plots, density plots,…
3. Get some practical experience in exploratory data analysis
4. Learn how to create background knowledge for a dataset
5. Learn how to deal with nominal attributes in conjuction with methods that learn numerical models
6. Learn to distinguish expected from unexpected results in data analysis and data mining—in general, this task is quite challenging, as it requires background knowledge with respect to the employed data mining technique, domain knowledge, and also practical experience.

Download *the School Performance[[1]](#footnote-1)* dataset from <https://archive.ics.uci.edu/ml/datasets/Student+Performance>; however, we will only use the following attributes of the Dataset and only will analyze the *student-por.csv* dataset at: [http://*www2.cs.uh.edu/~ceick/UDM/student-por.csv*](http://www2.cs.uh.edu/~ceick/UDM/student-por.csv) We call this Reduced School Performance Dataset, RSP for short:

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')   
11 **reason** - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')   
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)   
26 **goout** - going out with friends (numeric: from 1 - very low to 5 - very high)   
30 **absences** - number of school absences (numeric: from 0 to 93)   
# these grades are related with the course subject 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)

The first 5 Examples in the RSP dataset:

4;"at\_home";"course";2;0;4;4;0;11;11

1;"at\_home";"course";2;0;3;2;9;11;11

1;"at\_home";"other";2;0;2;6;12;13;12

2;"health";"home";3;0;2;0;14;14;14

3;"other";"home";2;0;2;0;11;13;13

Assignment1 Tasks:

Apply the following exploratory data analysis techniques **using R** to the dataset:

1. Compute the mean value and standard deviation for attributes G1, G2, G3**.** Remove 0’s that do not make sense prior to computing these statistics. **1 point**
2. Compute the covariance matrix for attributes G1, G2, and G3 next, compute the correlations for each of the 3 pairs of the 4attributes. Interpret the statistical findings! Remove 0’s that do not make sense prior to computing the covariance matrix and correlations. **6 points**
3. Create a scatter plot for attributes G1 and G3 of your dataset and a second scatter plot for attributes G1 and G2. Interpret the two scatter plots**! 6 points**
4. Create histograms for attributes Mjob. Fedu, studytime, failures, and goout. Then create the same histograms for the 5 attributes for students with G3>12 and for students with G3≤12; interpret the obtained 15 histograms. **9 points**
5. Create box plots for the attributes absences. G1, G2, G3. Remove null values that do not make sense prior to computing the box plots.Interpret and compare the obtained 4 boxplots! **6 points**
6. Create scatter plots with the dependent variable G3 and the attributes studytime, failures, gooout, and absences. Interpret the obtained 4 plots for the 4 pairs of attributes—G3 and each of the four attributes . **4 points**
7. Create 2-D Histograms for Fedu and Mjob, for gout and failures and for studytime and G3 and study time. Interpret the obtained displays **6 points**
8. Fit a linear model that predicts the dependent variable G3 using the 7 numerical attributes as independent variables, and transform the two nominal attribute Mjob and reason into 9 binary attributes[[2]](#footnote-2) which are also considered to be independent variables. That is you have 16 independent variables total to prediction G3. Next, z-score the obtained dataset, and fit a linear model to the z-scored dataset. Report the R2 of the obtained linear model and the coefficients of each attribute in the obtained regression function. Next,drop G2 and 5 more attributes, whose coefficients are the closest to 0, and obtain a linear model using the ten remaining attributes as independent variables. Next, interpret the results! What do the coefficients tell you about the importance of the attribute for final grade prediction? What about negative and positive coefficients? Also compare the two regression functions and their degree of fit they accomplished. **15 points**
9. Add a column Grade to the RSP dataset (A; G3>15, B: G3=14or15, C: G3=12or13, D: G3=10or11, F: G3<10). Next. create 3 different decision tree models that predicts the Grade attribute and has 20 or less nodes[[3]](#footnote-3) using the modified RSV dataset with attributes G2 and G3 excluded from training sets(total number of nodes should be less than 21 do not submit models with more than 20 nodes!); Explain how the 3 decision tree models were obtained. Report the training accuracy and the testing accuracy for each decision tree; interpret the learnt decision tree—what do they tell you about the importance of the 8 attributes in the used dataset for the classification problem? **14 points**
10. Write a conclusion (at most 18 sentences!) summarizing the most important findings of the assignment—what did we learn about the dataset? In particular, address the findings obtained related to predicting the variable G3. **6 points (and up to 4 extra points)**

Remarks:

* About 27-36% of the Assignment1 points will be allocated to interpreting statistical findings and visualizations!
* The dataset contains some missing attribute values (values of 0 that do not make sense); therefore, some kind of data cleaning has to be done.

**Submission Guidelines**:

Create a folder and name it as LastName\_StudentId\_HW1. HW1 folder should include:

* R code for the tasks.
* The data files needed to run the R codes.
* The assignment report containing all the plots and results along with the interpretations for them.

Submit the LastName\_StudentId\_HW1 folder in a zipped file through Blackboard.

**Points will be deducted for incomplete submission.**

1. We will also use a “cleaned version” of this dataset, for some tasks of Assignment1 (see below!). [↑](#footnote-ref-1)
2. For the nominal attribute Mjob create 5 binary attributes named 'Mteacher', 'Mhealth', Mservices' 'Mat\_home' and 'Mother' and for the reasons attribute create 4 binary attributes named

   ‘Rhome', ‘Rreputation', 'Rcourse' 'Rother'. [↑](#footnote-ref-2)
3. Intermediate nodes count! [↑](#footnote-ref-3)