Data Mining (CS524)

Indian Institute of Technology Ropar

Lab Assignment 1 **Release Date: 27/01/2019**
Maximum Marks: 50 **Due Date: 06/02/2019**

Download the Student Performance dataset from `https://archive.ics.uci.edu/ml/datasets/Student+Performance`, and extract the file "student-por.csv" in the dataset. Create a new dataset D from the extracted file (can be done manually), by selecting all the data objects (rows), but only the following 10 attributes (columns).

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')

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

14: studytime - weekly study time (numeric: 1 - less than 2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - greater than 10 hours)

15: failures - number of past class failures (numeric: n if $1 \leq n < 3$, else 3)

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)

The following grades are related to the course subject Portuguese:

31: G1 - first period grade (numeric: from 0 to 20)

32: G2 - second period grade (numeric: from 0 to 20)

33: G3 - final grade (numeric: from 0 to 20, output target)

The first 5 examples of the newly created dataset D are as follows.

- 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

Perform the following data analysis tasks on the dataset D.

T 1. Compute the mean value and standard deviation for the attributes G1, G2, G3. Document the results. [*5 marks*] ■

T 2. Compute the covariance matrix for attributes G1, G2, and G3. Compute the correlations for each of the 3 pairs of attributes. Interpret the statistical findings, and document them. [*5 marks*] ■

T 3. Create scatter plots for the pairs of attributes: *a*) G1 and G2, *b*) G2 and G3, and *c*) G1 and G3. Interpret the three scatter plots, and document them. [*5 marks*] ■

T 4. Create histograms for the attributes: *a*) Fedu, *b*) Mjob, *c*) studytime, *d*) failures, and *e*) goout. Then create the same histograms for the 5 attributes for students with $G3 > 12$ and for students with $G3 \leq 12$. Interpret the obtained 15 histograms, and document them. [*5 marks*] ■

T 5. Create box plots for the attributes: *a*) absences, *b*) G1, *c*) G2, *d*) G3. Interpret and compare the obtained 4 boxplots. Document them. [*5 marks*] ■

T 6. Create scatter plots for the pairs of attributes: *a*) G3 and studytime, *b*) G3 and failures, *c*) G3 and goout, and *d*) G3 and absences. Interpret the obtained 4 plots, and document them. [*10 marks*] ■

T 7. Compute the triangular dissimilarity matrix for the first 10 data objects (rows), based on all the 10 attributes in the dataset D. Interpret the obtained results, and document them. [*15 marks*] ■

End of the Assignment

Submission Guidelines

Create a folder and name it as `entrynumber_A1`, where `entrynumber` is your own entry number. This folder should include the following.

(*i*) Properly named and organised implementation source code files (with minimum self-explanatory comments) in Python for the tasks. For a common standard of implementation, we will use only Python in the lab assignments. Use of existing libraries is allowed only for reading the input files, generating the graphical plots, and writing into output files. For all other computations, like computing statistical measures, distances, etc., detailed codes are to be written (without using any libraries or inbuilt-functions).

(*ii*) A properly named electronic file in pdf format containing documentations of all the tasks. It can be prepared using any text editor or latex, but needs to be finally converted into pdf for submission.

(*iii*) Properly named and organised data files, needed to run the Python codes.

(*iv*) A `ReadMe` text file containing details about executing the source code to obtain the desired results. Any additional information or note may also be included in this file.

Submit the `entrynumber_A1` folder in a zipped file through Moodle.

Note:

(*i*) *This is an individual assignment.*

(*ii*) *Marks distribution within each task is as follows. a) 10% for minimum comments in the source code files, b) 50% for implementation, c) 30% for interpretation of results, and d) 10% for proper documentation and organisation of the submission.*

(*iii*) *Late submissions will face a penalty of 10% (of the full marks) for each day of delay.*

(*iv*) *Presenting some other person's work as your own without proper citation of the source is an act of plagiarism. It is a serious offence and will be treated strictly.*

(*v*) *Queries, if any, can be directed to our TA Aroof Aimen (`2018csz0001@iitrpr.ac.in`) through email.*