DATA SCIENCE, BUSINESS ANALYTICS E INNOVAZIONE

METODI DI APPRENDIMENTO STATISTICO PER IL DATA SCIENCE

ClassWork 1

Si utilizzi il seguente formato, seguendo l'ordine delle istruzioni, per produrre uno script che svolga le funzionalità richieste. L'uso di ggplot2 è un plus.

```
########################  ClassWork 1  ######################################
# Nome:
# Cognome:
# Matricola:
# CdL:
#############################################################################
#######  1   #######
…your code here…

#######  2   #######
…your code here…
```

N.B. Il file dovrà essere rinominato nel seguente modo: **00000_Nome_Cognome.R**

Esempio: 01234_Mario_Rossi.R

1. Importare, via script, il dataset "student-mattt.csv" salvandolo in una variabile (**prestare attenzione ai tipi delle variabili!**).
2. Effettuare l'analisi esplorativa rappresentando graficamente **almeno**:
    a. la correlazione lineare (fra le opportune variabili);
    b. G1, G2, G3, studytime;
    c. Un Istogramma per *G3* ed uno per *studytime,* colorati entrambi in funzione di *sex.*

3. Si rimuovano le variabili *G1* e *G2*.
4. Produrre un opportuno modello di previsione per *G3*.
5. Trasformare la variabile *G3* secondo il seguente schema e chiamarla *G3_qual*:
    a. se *G3* < 12, “Insufficiente”
    b. se *G3* ≥ 12 e *G3* < 15, “Buono”
    c. se *G3* ≥ 15 e *G3* ≤ 20, “Ottimo”
6. Produrre un opportuno modello di previsione per *G3_qual*.
7. Usare la K-fold Cross-Validation (K=10) per osservare le performance del modello di previsione per *G3* creato al punto 4.

Data Set Information

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. This dataset is provided regarding the performance in Mathematics.

Attribute Information

Attributes for both student-mat.csv (Math course) datasets:

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

6. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
7. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
8. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
9. 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)
10. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
11. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
12. internet - Internet access at home (binary: yes or no)
13. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
14. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
15. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
16. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
17. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
18. health - current health status (numeric: from 1 - very bad to 5 - very good)
19. absences - number of school absences (numeric: from 0 to 93)

these grades are related with the course subject:

20. G1 - first period grade (numeric: from 0 to 20)
21. G2 - second period grade (numeric: from 0 to 20)
22. **G3 - final grade (numeric: from 0 to 20, output target)**