INTRODUCTION TO SPORT ANALYTICS (20630)
Exam May 23rd 2023
Allowed time 90 minutes

The dataset **olympics_100m.csv** contains all the relevant data for the exam.

Using this database generate a .R or .Rmd file containing all the relevant answers. Please name this file with your surname and the appropriate extension (.R or .Rmd)

Q1 (5 points)

1. (2 points) Clear the environment, set the correct default directory, import the data in R, install and run all the relevant packages for the programme.

2. (3 points) Check the data and their type, remove not available observations and create a balanced panel with data for the period 1928-2021. Indicate the number of total observations available and the dimensions of your panel.

Q2 (3 points) Create a table showing the number of 100m olympics titles won by each country for each gender.

Q3 (6 points, 1 for each answer) Create a graph showing as function of time:

1. the time of the winners of 100m F

2. the time of the winners of 100m M

3. use two different colours for the data points of the two different genders

4. the interpolant using the data points in the F competition

5. the interpolant using the data points in the M competition

6. the interpolant using all the available data without distinguishing by gender

Q4 (12 points) Estimate the following models and report summaries for each of them

1. (2 points) a regression on pooled data of Time on a constant and Year

2. (2 points) a regression on pooled data of Time a constant and Year, allowing the constant to be different for F and M

3. (2 points) a regression on pooled data of Time on a constant and Year, allowing the constant and the slopes to be different for F and M

4. (2 points) two separate regressions for F and M of Time on a constant and Year.

5. (4 points) Briefly comment on your results

Q5 (6 points) Using only one graphs please illustrate the results obtained in the answer to **Q4**

Q6 (4 points) Suppose you are given the observations without the gender label, indicate which type of machine learning technique you would use to group the data in the appropriate way by specifying if you would opt for a unsupervised or supervise machine learning.