University of Padova

Master Degree in *Computer Science*

a.y. 2017/2018

Data Mining

Teacher: Annamaria Guolo

Example of practical assessment

INSTRUCTIONS: The examination takes 2 hours and 30 minutes. Write a brief report explaining the data analysis performed and the results. Provide a printed copy of the report. Remember to write your i) name and surname, ii) enrolment number, iii) date. Reports without the last three information will not be evaluated.

In the following you can find a description of the data and a brief guide for the analysis to be carried out. The use of the material of the course (slides, notes) is allowed. The use of internet is not allowed.

Dataset `movie`: data refer to the characteristics of some movies at their first run in US.

- `box`: box office takings at the first run in US (in dollars)
- `mprating`: rating of the movie, 1: general audience, 2: parents guidance, 3: not recommended for a younger audience, 4: parental accompaniment required for younger audiences
- `budget`: budget (in thousands of dollars)
- `starpower`: effect of celebrities in the movie
- `sequel`: is the movie a sequel? TRUE/FALSE
- `action`: is it an action movie? TRUE/FALSE
- `comedy`: is it a comedy? TRUE/FALSE
- `animated`: is it an animated movie? TRUE/FALSE
- `horror`: is it a horror movie? TRUE/FALSE
- `addict`: trailer views
- `cmngsoon`: comments on comingsoon.net
- `fandango`: attention at fandango.com
- `cntwait`: percentage of fandango votes that can't wait to see

The aim of the analysis is the evaluation of the variables associated to the box office takings.

1. Construct a new variable distinguishing box office takings larger than 20,000,000 dollars (value =1) or equal to/smaller than 20,000,000 dollars (value=0). Consider the dataset composed by `budget`, `action`, `cmngsoon`. Construct the most appropriate model for the purpose of the analysis. Insert in the report the outputs from `R` and the graphical evaluation of the model/models that are considered most useful in order to explain the analysis and the results.

2. Consider all the variables in the dataset. Construct the most appropriate model for the purpose of the analysis. Insert in the report the outputs from `R` and the graphical evaluation of the model/models that are considered most useful in order to explain the analysis and the results.

If needed, please report the seed used in your analyses.