isep
Institut Superieur
De'Electronique D'Paris

PROJECT ON MACHINE LEARNING

Dr. Patricia CONDE-CESPEDES

I. INSTRUCTIONS

This project has to be made by groups of 2, 3 or 4 students. There is no possibility to work alone.

OBJECTIVE

The main objective is to use supervised and unsupervised learning methods treated in the lectures and tutorial courses in a real data.

EVALUATION CRITERIA

The project notation is **30%** of the final mark. The presentation and quality of the final report as well as the oral defense will be taken into account.

THE FINAL REPORT

Particular attention will be paid to the presentation of the final report. The language of the report can be either English or French. In any case, the quality of writing will be appreciated. The sentences must be clear, explicit and well understandable.

The final report must contain a cover page, a table of contents, an introduction, the body of the report explained in the following sections (results, figures, tables, etc.), the conclusion and the appendix for the code.
The number of pages should not exceed 25 pages and 10 pages for the appendix.

The cover page must contain the first name, last name and the student identification number of all the authors.

THE DEFENSE

An oral defense will be held in **January 19th 2018**.

The defense will last about 15 minutes per group and it will consist in about 10 minutes of presentation plus 5 minutes of questions.

DELIVERY OF THE REPORT

In moodle, you will find a deposit box named Project-Reports-box. The final report must be uploaded in this deposit box not later than **January 12th 2018**. If you want to upload more than 2 files compress all of them in only one file in zip format. The final report must be sent in pdf format. The file names must be as follows:

LastNameStudent1_LastNameStudent2_ LastNameStudent3.pdf.

Just one deliver per group must be done.

II. PROJECT DESCRIPTION

2.1 Data collection

The site web http://archive.ics.uci.edu/ml/index.php is a Machine Learning repository created by the Center for Machine Learning and Intelligent Systems in the University of California, Irvine). The dataset "Las Vegas" was taken from this repository. This dataset contains 504 records and 20 quantitative and categorical variables from online reviews from 21 hotels located in Las Vegas Strip, extracted from TripAdvisor .

2.2 Descriptive Statistics

In this part, you are going to make an exploratory analysis of the data set. Calculate descriptive statistics on all the variables. The purpose is for you to get familiarized with this dataset.

2.3 Supervised Learning

In this part, you are going to use at least three supervised learning methods studied in the course. You are going to perform Classification and Regression. The target variable is the Hotel Score and the predictors will be the other variables or a subset of them that you consider pertinent for this analysis.

- For regression, you can consider the "Hotel Score" as a continuous variable
- For classification, you will create two classes: *High* (if the score is 4 and 5) and *Low* (otherwise).

Do not forget to evaluate your model performance. For example, by calculating the confusion matrix, accuracy, curve ROC, etc.

2.4 Unsupervised Learning

In this part, you are going to apply all the unsupervised learning methods that you learned during the lectures. This time, the variable *Score* will be part of your analysis.

IMPORTANT: Your creativity, your contributions, results and comments will be taken into account for the final mark. Any figure, output code, table or other result with no interpretation will not be considered for the mark.

2.5 BIBLIOGRAPHY

Do not forget to include the bibliography in your report

III. DATA SOURCE

- UC, Irvine machine learning repository http://archive.ics.uci.edu/ml/index.php
- Moro, S., Rita, P., & Coelho, J. (2017). Stripping customers' feedback on hotels through data mining: The case of Las Vegas Strip. Tourism Management Perspectives, 23, 41-52.