Instructions

- This homework assignment is worth 70 points.
- Please submit a **.ipynb** file to Blackboard.
- **Please strive for clarity and organization.**
- **Due Date: February 24, 2023 by 11:59 pm.**

For this homework assignment and for future one, we will work on the challenge presented in the `data mining cup 2019`. Please read the task and get familiar with the data. For this week homework assignment, answer the following:

Exercise 1

(5 points) Using the bucket, that you create in the last homework assignment, and the pandas library, read the `train.csv` and `test.csv` data files and create two data-frames called `train` and `test`, respectively.

Exercise 2

(65 points) Using the `train` data-frame (including the features that were engineered in homework assignments 4), do the following:

(i) Split the `train` data-frame intro `training` (80%) and `testing` (20%) (taking into account the proportions of 0s and 1s).

- Run RFE with logistic regression as a base algorithm (with `n_features_to_select = 5`). Extract and store the support of each of the features.
- Run RFE with random forest (with 500 trees and max depth equal to 3) as a base algorithm (with `n_features_to_select = 5`). Extract and store the support of each of the features.
- Run RFE with AdaBoost (with 500 trees, max depth equal to 3, and learning rate equal to 0.01) as a base algorithm (with `n_features_to_select = 5`). Extract and store the support of each of the features.

Repeat (i) 100 times. Combine the results and rank the features. If you run into converge issues with logistic regression, try to transform the input features to 0-1 scale.