ICT-6522 Data Warehousing and Mining
Term Project
Marks: 15
Deadline: 20/11/2021

- Download the 'vote' dataset containing information about 1984 United States Congressional Voting Records. The output attribute called class, has two values: 'democrat', and 'republican' showing the vote of a voter.
- Using Weka/PyLib Data Mining software, you are required to:

1. Pre-process the dataset in order to select the 12 best attributes. Include in your report screenshots showing the algorithm you have applied for pre-processing (include the chosen parameter values if any).
2. On the dataset obtained at point (1), apply precisely 4 different classification algorithms in order to produce 8 models (2 models per algorithm), that can be used to automatically predict future votes.
    a. At least one of the produced models has to be a decision tree.
    b. All the models will be learned and tested by splitting the dataset in a training and a test dataset, each of which consisting in 80% and 20% of instances, respectively.
    c. For each built model, report the algorithm name and the parameter values chosen for its application, and the confusion matrix and the measures of performance of the model (in particular the accuracy). Include a screenshot with one decision tree that you obtained.
3. Calculate for each model the accuracy, precision, recall, sensitivity, and specificity, for the class 'democrat'.
4. Choose the best, the second best and the third best model from step (2). Justify your answer.
5. Mention three characteristics of 'democrat' voter based on the decision tree built and displayed in (2).
    a. List the production rule(s) that you have used to mention these characteristics.

- Submission must be in MS Teams.
- **COPYING CODE AND REPORT FROM ANYONE ELSE IS STRICTLY PROHIBITED. IF PLAGIARISM IS DETECTED, BOTH STUDENTS WILL GET ZERO.**