# Introduction to Machine Learning

**Data Preparation and Results' Presentation**

*These exercises can be solved in any programming language of your choice. It is assumed you are proficient with programming. It will be helpful if the language has a library to plot graphics.*

1. Import the credit-approval data file from the UCI Machine Learning Repository
https://archive.ics.uci.edu/ml/datasets/credit+approval

2. Discretize the 2nd row in two ways:

- Equal number of examples per class (choose an appropriate number of exemples per class)
- Equal-sized class borders (choose an appropriate number of exemples per class)

Calculate the information gain obtainable by each of the divisions when using the 2nd characteristic to divide the set into multiple sets. What can you conclude?
Verify graphically if the number of positive and negative elements is similar in all classes. Draw your conclusions from the graphs.

3. Select only the numerical columns of the set. Normalize each column using two different methods Min-max and Z-Score. Use the K-NN algorithm implemented on the last exercise sheet (Supervised Learning, Part 2) on all three sets (non-normalized, min-max normalized, Z-Score normalized) and compare the results (use 10-fold cross validation to compare the results).

4. Using this bank data set, plan two preparations (only a description is necessary, an accurate one, not an implementation):

A. If you were going use the K-NN algorithm for classification
B. If you were using the ID3 algorithm for classification

Assume you were going to use the algorithms made in the last exercise, Supervised Learning, Part 2, Ex 1 and 2, as they are, adapting only the examples' dimensions if necessary.

5. Assume you had two datasets for a given problem where the number of positive examples is very low compared to the number of negative example. In this case assume 500 positive examples and 9500 negative examples.
a) Calculate precision, accuracy, recall and F1-Score for each of the tables below
b) Given the following results what is your interpretation on the quality of the models? How do you think each would behave when classifying future data?
c) How do you build your training and test sets for this data set?

| | Real classification | | |
|---|---|---|---|
| Model's output | | Positive | Negative |
| | Positive | 0 | 0 |
| | Negative | 500 | 9500 |

Tab. 1 Results of model trained with data subset A (5% positive examples, 95% negative examples, 10000 examples)

| | Real classification | | |
|---|---|---|---|
| Model's output | | Positive | Negative |
| | Positive | 300 | 50 |
| | Negative | 200 | 450 |

Tab. 2 Results of model trained with data subset A (50% positive examples, 50% negative examples, 1000 examples)