Machine Learning Project (3EC): *House Prices in Ames, Iowa*

1 Introduction

In this project you will solve a 'real' Machine Learning assignment by joining in on a 'Kaggle Competition'. The dataset on which you have to train your ML model and its benchmarking is made for you by Kaggle.

2 Description

You will have to predict the price of residential homes in Ames, Iowa in the Kaggle competition: "House Prices - Advanced Regression Techniques". You can test the accuracy of your model by uploading the price prediction on Kaggle. They have the answers, compute a Root-Mean-Square Error and send it back to you.

You start with four files:

- *data_description.txt* Contains a short description of the training data.
- *train.csv* Training dataset of 1460 houses including 79 attributes and their sales price.
- *test.csv* Test dataset of 1460 houses including 79 attributes, but now without their sales price.
- *sample_submission.csv* Example of the file and format of your predicted prices for the test data.

You will encounter many problems related to a real ML project: missing data, data that does not correlate with the training target, building 'one-hot' representations etc. The ML method that you are going to apply is up to you. However, you are required to:

1. Choose a method from the previous weeks and fit the code you developed to this problem.
2. Choose another method you are interested in, and use a library like Scikit-learn to implement an ML method.

3 Deliverables

- An ipython notebook with models according to 1. and 2. You must load train.csv, train the models on this data, and use the trained model to make predictions on test.csv. At the top of the file you have to write a short report on the solution that you have found. How did you solve missing data entries? Which attributes did you pick as features and how did you decide on which were not so important and could be dropped? Given more time, how could you further improve your model? Use the LaTeX environment in a cell.
- submission.csv containing the predictions on test.csv. You can try to upload this submission.csv to kaggle to check where you stand on the leaderboard.

Time-limit: 3 weeks (1EC = 28 hours (max), so roughly 10 hours per week)
Deadline: Hand in the ipython notebook with your project report before April 17th 2023 13:00 hours.
Presentation: On Wednesday 19th of April you will present your notebook in an oral exam.
Team: The project and its presentation are to be made by a pair of 2 students.

Acknowledgements

The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It's an alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.