# 1. Project assignment

Each project defines the dataset or other potential requirements, all the other decisions are up to you.
Typically, the work on the project consists of the following:

1. Exploratory data analysis
2. Data preprocessing
3. Choosing one specific model for machine learning
4. A brief general explanation of how the model works
5. Training the model on preprocessed dataset
6. Interpreting the trained model
7. Evaluating the trained model on the dataset
8. Writing a brief summary about the results of the comparison

# 2. Description of individual tasks

2.1. **Exploratory analysis**: Explore the dataset, i.e. see how much data is in the dataset and how much and what type of data is in the values of the individual columns, how the individual items correlate with each other. Each dataset includes an explanatory commentary that summarizes what is in the data and, if applicable, how the data was collected. It's better to read this commentary before starting the analysis. Output of this analysis will typically be tables and graphs. Comment on your observations with a few sentences.

2.2. **Preprocessing**: Prepare the dataset so that individual models can be run on it. This step includes all manipulation with the data. For example, data type conversion (e.g. to datetime), dealing with missing values, scaling and normalization, feature selection, feature extraction, splitting into train and test sets, resampling, adding more external data and more. Not all the things listed need to be done, depending on the data and models. When you decide to do some preprocessing, briefly comment on what are you doing and why.

2.3. **Model selection**: The choice of model is limited only by the type of data (regression, classification, clustering, anomaly detection) of your specific task, otherwise you are free to choose. You can choose a model from the scikit-learn library, but you can also use another library or even implement your own model.

2.4. **Model explanation**: The explanation of the model will be around two paragraphs long. The aim is to present the technique and to give a brief and accurate description of how it works. You may assume that the reader has basic understanding of machine learning but does not know your specific model. In the explanation, focus on the description of the learning model (what happens during the learning process, what is the result of the learning process, ...).

2.5. **Training the model**: Train your selected model on the preprocessed dataset. A more precise way of training is up to you, but should be comparable across other models. Thus, it is not good for each model to be trained on different subsets of the data. The model does not have to be run on all columns; you can choose only their subset. The tuning of hyperparameters will also be a part of the training.

2.6. **Model interpretation**: Interpret your selected and trained model. The specific interpretation will depend on the chosen model, but the goal is to explain what the model has learned and why it has learned this. Interpretation of the model may include an analysis of the parameters learned, the meaning and effect of the individual columns on the predictions, and analysis of the sensitivity on values in the individual columns.

2.7. **Model evaluation**: Evaluate the model on your dataset using appropriately chosen measures depending on task solved (regression, classification, clustering, anomaly detection). Explain your choice of the evaluation method and describe what is measures. The evaluation will also include a comparison with the 'baseline' model which will solve the problem in a naive way (e.g. constant value, simple statistics, randomness, ...).

2.8. **Summarize the results**:
Compare the results of the different models and summarize the results of the evaluation in a few sentences. Particularly interesting are findings about which model works best and why. Also, whether some of the model is more robust when selecting the different divisions of the data into training and the test set.

<u>The project will be rated by the following requirements:</u>

- Exploratory data analysis
- Appropriate data processing according to the type of data and selected models
- Advanced preprocessing such as feature extraction or external data
- Justification and commentary of the techniques used for data analysis
- Description of the functioning for each selected model
- Training of all models on the dataset
- Appropriate choice of model parameters and their tuning
- Interpretation of the learned models
- Evaluation of the models using several appropriately chosen measures/metrics
- Comparison of the models with a naive "baseline" model/approach
- Brief summary of the results and observations

- An explanatory commentary documenting individual project decisions
- Correct methodology of learning and evaluating the model