KLU
KÜHNE LOGISTICS UNIVERSITY

# DATA SCIENCE

## Group Assignment

**Prof. Dr. Henrik Leopold**

Kühne Logistics University

# GOAL

The goal of this assignment is to:

- Give you the opportunity to apply the acquired knowledge of this course in a setting of your choice.
- Explore what is possible with data mining techniques and what problems you may face.
- Benefit from the exchange with your fellow students.
- Obtain a grade for applying your knowledge rather than obtaining a grade for correctly calculating details in a written exam.

# GENERAL SETTING

- The general idea is to work in groups of 3-4 students:
  - Exceptions are possible if, for instance, there a problems because of different time zones.
  - However: Nobody should work alone.
  - I can help with the group allocation (if desired).
- You can pick both the problem as well as the dataset you would like to work on:
  - I would like to give you the opportunity to work on something you like.
  - You can employ supervised learning, unsupervised learning, text mining, or a combination of two or more.
  - Data sets can be found everywhere on the web. One of the most prominent sources is: https://www.kaggle.com/.

# TASKS (1/2)

- I essentially would like you to walk through the stages of the data science life cycle. This translates into four main tasks.

1. **Business / data understanding**: After picking a dataset, obtain useful knowledge about the domain and the data set (also consider consulting other websites / sources). Conduct exploratory analyses to understand what kind of data set you are dealing with and, e.g., how feature varaiables are related with the target variables (in case of supervised learning).

2. **Data preparation**: Prepare your data for building a model. This may include selecting feature variables, transforming feature variables, and also adding feature variables from other sources.

# TASKS (2/2)

3. **Modeling**: Build a (potentially) well-performing model. The goal is clearly not to build a model with perfect performance but rather to explore what is possible to improve a first (or really simple) model. Build several models and compare them.

4. **Evaluation**: Evaluate your model using appropriate mechanisms. Again, use different mechanisms and compare the impact on the results.

5. **Documentation:** While many of the tasks above will result in Python code, don‘t forget to also document your strategy, ideas, thoughts, considerations, etc. using text in an exerternal document (the structure of this document will be discussed on the deliverable slide).

# ADDITIONAL REQUIREMENTS

- It is clear your data set / task ***should not be too easy***. If you are in doubt, contact me and brieflt explain your plans. I will provide you with feedback.
- As orienation: You data set / problem should be at least as complex as the ones in the exercises.
- Make sure you ***exploit the different options*** we discussed for obtaining better results:
  - Different types of models.
  - Varying parameters.
  - Combining different techniques (e.g. unsupervised and supervised learning).
  - Testing different evealuation methods.

# DELIVERABLES

1. **Presentation**:
   - Present your results to the group in about 5 minutes.
   - Explain the problem and the data set.
   - Explain the (prelimanary) solution and results.
2. **Python implementation**:
   - Submit your (executable) code incl. all data files.
   - Don‘t forget to use comments in your code to explain the different steps.
3. **Report:**
   - Create a report (with about 3 pages / group member).
   - Explain each task / step in detail.
   - Include plots / graphs to illustrate the data set, analyses, and results.
   - Add a section in which you reflect about the assignment: How good is the final model? What else could be done? What you have you learned?

# ASSESSMENT CRITERIA

- Is the problem conceptualized appropriately from a technical and from a domain perspective?
- Has the data set been appropriately prepared for further processing (e.g. introduction of dummies, balancing, removing rows with NaNs, etc.)?
- Are the chosen techniques (supervised, unsupervised, text mining/ etc.) appropriate to solve the problem at hand?
- Are the chosen evaluation metrics and methods suitable for evaluating the chosen techniques?
- Are the choices (from above) properly explained in the report?
- Does the report show that the students have developed a solid understanding of the problem and the solution?
- Is the report in a good shape (structure, flow, quality of text, etc)?

# DEADLINE

- **Presentation**:
  - To be held on May 11, 2022.
  - Please submit slides to henrik.leopold@the-klu.org with your group members in CC.

- **Report / Python Code**:
  - To be submitted by Sunday, May 31 (Midnight, AOE).
  - Please submit a single PDF for the report and single Python file to henrik.leopold@the-klu.org with your group members in CC.