Machine Learning (WS 24 / 25) – Project 1 **(Incl. Update after Lecture 5: Classification)**

HfT Stuttgart, Prof. Dr. Laura von Rueden

Deliverable:

- Python notebook (.ipynb) with code and text cells where you describe both your analysis steps and your results interpretation. An extra report is not needed when you describe everything in the notebook.
- Hand-in via Moodle until **27.11., 1 pm**
- Presentation of python notebooks in practice session on 28.11. and 5.12.

Goal:

- Select one of the datasets and train a model for regression (model 1) and for classification (model 2). Have fun ☺ !

Datasets:

- Bike sharing usage: https://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand
- Energy efficiency of buildings: https://archive.ics.uci.edu/dataset/242/energy+efficiency
- Wine quality: https://archive.ics.uci.edu/dataset/186/wine+quality

Tasks:

1. Business understanding and data collection
    a) Inform yourself about the listed datasets. What are they about? What are the analysis goals?
    b) Select the dataset that interests you the most. Create a python notebook and describe your understanding about the dataset.
    c) Download the data and save it in a pandas data frame.

2. Data exploration
    a) How many variables and instances does the dataset contain?
    b) Do the variables have understandable names? If not, think about renaming.
    c) Explore the data statistically and visually. How is the data distributed?
    d) Do you observe any correlations? If yes, between which variables?

3. Data preparation
    a) Is data cleaning needed?
    b) Is data encoding needed?
    c) Do you think any further feature engineering would be useful?
    d) Split the data into data subsets.
    e) Is feature scaling needed?

4. Modelling: Regression (= model 1)
    a) Define again the analysis goal. What is the target variable that you want to predict? (Remark - If you use the "energy" dataset: It is enough to predict only *one* target variable.) Which features do you want to use?
    b) Select a model, define a performance metric, select a learning algorithm.
    c) Run the learning algorithm. Monitor the learning curves. What do you observe? How is the model performance?
    d) Is fine-tuning needed? Try, e.g., other hyperparameters.
    e) Try regularization techniques. What is the effect?
    f) Save the final model, perform a final evaluation and demonstrate how it can be used for making predictions.

5. Modelling: Classification (= model 2)
    a) Define again the analysis goal. What is the target variable that you want to predict? Which features do you want to use?
        i. Remark - If you use the "bike" or the "energy" dataset: You first need to create a categorical class variable by binning the originally continuous target variable. You can use the code below for that.
    b) Train a logistic regression model for multinomial classification.
    c) Evaluate the performance on the train and the validation subset. Which performance measures do you use, and why? How good is the performance? Is the performance similar for all classes?

d) Do you have any idea, if something could be improved for the classification model?
    i. If yes, make those adaptations and train a new classification model. Afterwards measure the performance on the validation set. Has the performance improved?
e) Optional: Train another type of a classification model (e.g., decision tree, or SVM).
f) Select your best classification model (from b), d), and e)) and measure the performance on the test subset.

6. Comparison of regression (model 1) and classification (model 2)
    a) Describe advantages of the regression model, and advantages of the classification model.
    b) Which of the two models is more suited for the original analysis goal?

Code for creating a categorical cvariable (Task 5a.i):

For the energy dataset:

```
# Assuming your data is stored in a pandas dataframe called "df"
df['heating_class'], class_bins = pd.cut(df['Heating Load'],\
                                         bins = 5, labels = [0,1,2,3,4], retbins=True, right=False)
df['heating_class'] = df['heating_class'].astype('int64')
```

For the bike dataset:

```
# Assuming your data is stored in a pandas dataframe called "df"
df['bike_count_class'], class_bins = pd.cut(df['Rented Bike Count'],\
                                          bins = 5, labels = [0,1,2,3,4], retbins=True, right=False)
df['bike_count_class'] = df['bike_count_class'].astype('int64')
```