# ML2_assignment1

March 12, 2024

## 1 Assignment 1

### 1.1 General information

You are required to submit two files to Moodle: an `.ipynb` file and the rendered `.pdf` file with your solutions. Do not zip them together so I will be able to annotate the `.pdf` directly.

Please give short (2-3 sentences) interpretations / explanations to your answers, not only the program code and outputs. Be concise and focused (less could be more ;)).

Grades will be distributed with the following rule: from the points you earn, you get 100% if you submit until the due date (`2024-03-22 20:00`), 50% within 24 hours past due date, and 0% after that.

### 1.2 Predict real estate value (20 points)

In this exercise you will predict property prices in New Taipei City, Taiwan using this dataset. (I have uploaded the data to the repo for you with cleaned up variable names. You can find it in the `real_estate` folder, here.) Let's say you want to build a simple web app where potential buyers and sellers could rate their homes, and the provided `.csv` contains the data you have collected.

Similarly to what we did in the class, let's just work with a 20% subsample of the original data first. Put aside 30% of that sample for the test set. (*Hint:* Extend the snippet below.)

[ ]:
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split

prng = np.random.RandomState(20240311)

real_estate_data = pd.read_csv("https://raw.githubusercontent.com/divenyijanos/
 ↪ceu-ml/2023/data/real_estate/real_estate.csv")
real_estate_sample = real_estate_data.sample(frac=0.2)

outcome = real_estate_sample["house_price_of_unit_area"]
features = #TODO
X_train, X_test, y_train, y_test = train_test_split(features, outcome,␣
 ↪test_size=0.3, random_state=prng)

print(f"Size of the training set: {#TODO}, size of the test set: {#TODO}")
```

- (2 points) Think about an appropriate loss function you can use to evaluate your predictive models. What is the risk (from a business perspective) that you would have to take by making a wrong prediction?
- (2 points) Build a simple benchmark model and evaluate its performance on the hold-out set (using your chosen loss function).
- (2 points) Build a simple linear regression model using a chosen feature and evaluate its performance. Would you launch your evaluator web app using this model?
- (2 points) Build a multivariate linear model with all the meaningful variables available. Did it improve the predictive power?
- (6 points) Try to make your model (even) better. Document your process and its success while taking two approaches:
    1. Feature engineering - e.g. including squares and interactions or making sense of latitude&longitude by calculating the distance from the city center, etc.
    2. Training more flexible models - e.g. random forest or gradient boosting
- (2 points) Would you launch your web app now? What options you might have to further improve the prediction performance?
- (4 points) Rerun three of your previous models (including both flexible and less flexible ones) on the full train set. Ensure that your test result remains comparable by keeping that dataset intact. (*Hint:* extend the code snippet below.) Did it improve the predictive power of your models? Where do you observe the biggest improvement? Would you launch your web app now?

[ ]:
```
real_estate_full = real_estate_data.loc[~real_estate_data.index.isin(X_test.
↪index)]
print(f"Size of the full training set: {#TODO}")
```