DATA 201 – Assignment 4

Total marks: **20**

Due date: **11:59 p.m., Friday, September 29, 2023**.

Submit **code** and **outputs** in a **single Jupyter notebook file**. *Do not expect the marker to rerun your code in order to get the outputs.*

The aim of this assignment is to develop a machine learning model to predict the house prices using information in file **data.csv**. The description about the data is given in file **description.pdf**.

Requirements:

- Use root mean square error (RMSE) as the evaluation metric. [**2 marks**]
- Load the dataset, determine the target column, remove irrelevant variables (if any), and use function `train_test_split` with `random_state=1` to split the data into two sets: a training set (80%) and a test set (20%). [**3 marks**]
- Explore the training set to gain insights. [**2 marks**]
- Select one machine learning model, train it, optimise it (e.g., add pre-processing transformers, perform hyper-parameter tuning, etc.), and estimate the performance of the model. [**9 marks**]
- Test the final model on the test set and report the RMSE and at least two other evaluation metrics (e.g., mean absolute percentage error (MAPE), R^2-score, etc.). [**3 marks**]
- Include a discussion at the end of your notebook (about what you have learned, difficulties, what has worked and not worked, future directions, etc.). [**1 mark**]

Notes:

- Write **your name and student ID** at the beginning of your notebook. After completing your work, use menu item ***Kernel => Restart & Run All*** in Jupyter, then submit your notebook file.
- You can use any public Python package.
- The requirements above have no order that you have to follow.
- Use your own assumptions and judgement if you are unsure about any information in the dataset. However, remember to mention it in the discussion.
- Try to write functions for all data transformations you apply, try feature engineering (e.g., creating new features), and try to automate all the steps as much as possible (e.g., using pipeline and data transformers, etc.). You may have **bonus mark**s for this; however, your total mark will not exceed **20**.