Assignment 3

Course: *Big Data*
Due date: *April 7th, 2024*

Assignment

You will be working with Dask. Download the NYC Yellow Taxi 2015 sample data `https://www.kaggle.com/datasets/elemento/nyc-yellow-taxi-trip-data`. Load into Dask the data for January 2015 and 2016 (yellow_tripdata_201501.csv and yellow_tripdata_201601.csv) and concatenate them.

First perform some date preprocessing:

- Do an exploratory data analysis.
- Think about the following issues: Which columns will you use? How are you going to use dates and times? Would you encode some columns?
- Handle missing data.
- Clean your data. Describe what you did.

Be careful not to do feature selection on the entire dataset.

Think about standardizing your data before training your models. Describe the procedure.

Predict the *Trip_distance* using different machine learning models. Select models where you have to tune some parameters. Describe the procedure.

As you will be predictiing the distance you will be doing regression.

Evaluate the importance of the inputs in your model.

Test (use machine learning evaluation techniques like cross-validation) your model and report the results.

Submit a Jupyter notebook file with your results to Učilnica.

Be careful to print out only relevant results. Do not forget to include different visualizations.