A project,

that consists in exercises that require the use of data mining tools for analysis of data. Exercises include: sequential patterns, time series, classification (alternative methods and validation), outlier detection. The project has to be performed by max 3 people. It has to be performed by using Knime, Python, other software or a combination of them. The results of the different tasks must reported in a unique paper. The total length of this paper must be max 20 pages of text including figures. The project must be delivered at least 2 days before the oral exam.

- **Dataset**: the data is a time series dataset on air quality, which can be downloaded here: Dataset.

https://data.world/uci/air-quality

- **Task 1: Time series**: Consider only attribute “PT08.S1(CO)” and split the corresponding time series into daily series, deleting those with too many missing values (value = -200) and fixing the others in some way. Make also sure that all time series have 24 values. Compute clustering (with an algorithm of your choice) based on DTW and Euclidean distances and compare the results.

- **Task 2: Sequential patterns**: discover contiguous sequential patterns of at least length 4. Before that, time series should be discretized in some way.

- **Task 3:Classification methods**: define a target variable “WE” for the time series data set to “true” for weekend days, and “false” for the others. Test the K-NN classification method using DTW as distance measure, and at least another classification method using the 24 values as separate variables.

- **Task 4: Outlier detection**: from the original dataset (i.e. the raw records with all attributes, not the time series built only on the “PT08.S1(CO)” attribute), identify the top 1% outliers. Adopt at least two different methods belonging to different families (i.e. model-based, distance-based, density-based, angle-based, …) to identify the 1% of input records with the highest likelihood of being outliers, and compare the results. Before doing the analysis, the records containing missing values should be deleted to avoid trivial results.