# Code implementation

The supplied python code includes 4 files.

The main is “data\_preparation.py”.

It loads the dataset, splits it into 6 datasets (train, validation, test) according to desired partitions set by user.

Then, some preliminary learning is done on the training set, e.g. the Principal Components from the PCA analysis. The number of principal components is, again, set by user. In our case, we chose 5.

Then, we created a custom Pipeline. It uses the basic Pipeline from the sklearn, but only applies the transformations (without ‘fit’ stage in the end). This way, it allows us to add stages to the Pipeline which modify each dataset. This Pipeline is applied on training, validation, and test set. Each stage in a pipeline makes some modification on some feature in the dataframe – it fills missing data, it changes some feature to categorical, or applies PCA transform, which drops the original features and leaves only the desired principal components. The final stage in a pipeline drops all the columns which have at least one NA value.

After the pipeline, each dataset is saved into a .csv file which is presented here.