# Classification on the ‘credit-g’ dataset

You can download the dataset with ‘fetch_openml(‘credit_g’)’ and see it’s description at https://www.openml.org/d/31

1. Determine which features are continuous and which are categorical.
2. Visualize the univariate distribution of each continuous feature, and the distribution of the target.
3. Split data into training and test set. Do not use the test set until a final evaluation. Preprocess the data (such as treatment of categorical variables) without using a pipeline and evaluate an initial LogisticRegression model with a training/validation split.
4. Use ColumnTransformer and pipeline to encode categorical variables (your choice of OneHotEncoder or another one from the categorical_encoder package, or both). Evaluate Logistic Regression, linear support vector machines and nearest neighbors using cross-validation. How different are the results? How does scaling the continuous features with StandardScaler influence the results?
5. Tune the parameters using GridSearchCV. Do the results improve? Evaluate only the be model on the test set. Visualize the performance as function of the parameters for all three models.
6. Change the cross-validation strategy from ‘stratified k-fold’ to ‘kfold’ with shuffling. Do the parameters that are found change? Do they change if you change the random seed of the shuffling? Or if you change the random state of the split into training and test data?
7. Visualize the 20 most important coefficients for LogisticRegression and Linear Support Vector Machines using hyper-parameters that performed well in the grid-search.