Second lab assignment WUM (Ver. Beta, subject to small changes)
starting date: May 20th
submission date: June 3rd

Dataset: You will work with an artificially generated dataset, consisting of 2000 samples, 400 input variables and two output variables called *class* and *output*. All input variables are standard-scaled real numbers. The class variable is discrete and has two possible values - 0 and 1. The other output variable is also a real number. Both output variables have non-trivial dependency on some of the input variables. Your task is to build ML models predicting the output variables based on the input variables and analyze the dataset, your models and the nature of the dependencies you uncovered.

Desired output: You are supposed to submit Jupyter notebook with the solutions, commentary, and results by Moodle. Please make sure your notebook opens and works in Google Colab, it will not be graded otherwise. Your solutions will be graded by lab assistants of respective groups.**Your solution should include a function that can process a "validation_data.csv" file with the same structure as your training data and compute the classification accuracies and R^2 values on these data for both your baseline and best models (in total 4 numbers).**

Specific tasks to perform:
Task 1. Building baseline models (6 points: 3 for regression + 3 for classification) Given the dataset, you should build baseline models using **linear regression and logistic regression** for output and class variables respectively using all input variables as predictors, without any pre-processing (such as dimensionality reduction). Assess their **performance on the training data and estimate their ability to generalize beyond training data** by one of the methods you have learned in our lectures. Comment on the results.

Task 2. More advanced classification (**12 points: 3 for using a more advanced method, 3 for parameter optimization, 3 for feature selection, 3 for the result analysis**) Using some of the classification methods you have learned in the lecture try to create an improved model that performs better than your baseline model. You can use more complex strategies using resampling, cross-validation and/or model selection. Utilize the knowledge that the class variable depends on some of the predictors, but not necessarily all. You can also use pre-processing of the input variables, such as dimensionality reduction. If your model is based on some features selected from the original data, you should provide the list of features you have selected. Comment on any choices you make in the process and what is the relative improvement over the baseline model.

Task 3. More advanced Regression (**12 points: 3 for using a more advanced method, 3 for parameter optimization, 3 for feature selection, 3 for the result analysis**) Using one of the approaches to model selection in classification that you have learned in the lecture, try to create an improved model that performs better than your baseline regression model. You can use more complex strategies using resampling, cross-validation and/or model selection. Utilize the knowledge that the output variable depends on some of the predictors, but not necessarily all of them. You can also use pre-processing of the input variables, such as dimensionality reduction. If your model is based on some features selected from the original data, you should provide the list of features you have selected. Comment on any choices you make in the process and what is the relative improvement over the baseline model.

Hint: It is possible to obtain classification accuracy >0.8 and regression with r^2 coefficient >0.5 in the data provided