# Supervised learning competition

Juan Moreno Díez (1840886), Leticia Marcal Russo (0664618), Luc Lubbers (1558501)

**Assignment:** Predict student's scores on math tests in secondary schools in Portugal.

**Data preparation:** We did not change the original dataset. We only explored it to make sure there were no missing values or some anomaly in some attribute and run correlation measures in order to understand the relationship among the variables. We used Pearson's correlation for the continuous variables and Cramer's V coefficient for categorical attributes.

**Supervised learning method chosen:** We have used Lasso to predict the student's scores on math tests. This is done with the function GLM() from the package glmnet.

As our goal is to make good predictions and not inference, we chose a regularization method (Hastie *et al.*, 2021, p. 244).

**Validation strategy:** First, we splitted the data into train and validation sets . We set validation aside and moved on to train the model. Important to note that we couldn't use the test set to look for performance, as we don't have the outcome variable (score). In our case, validation will work as a test set (to check performance through the MSE) and in the training set we will use cross validation to calibrate the hyperparameter.

So after splitting the data, we used cross validation -cv.glmnet ()- to choose the best lambda for the model. We had two options for the hyperparameter: *lambda.1se and lambda.min. The first one would give us* a slightly higher MSE, but less variance because there are less predictors in the model.

We went for *lambda.min* as our main goal is to make a more accurate prediction and get a lowest MSE.

Next step, after having the best model tuned, it was to check its performance in our validation/ test set. In the training set we got a MSE of 0.8451153; in the validation, MSE was 0.726242. We are aware that the splits in the data can make this result vary.

Hastie, T., James, G., Tibshirani, R. & Witten, D. (2021). *An Introduction to Statistical Learning with Applications in R*. (2nd edition). Springer. Retrieved from https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

**Files**

Our group tried different algorithms and we chose the one with the lowest MSE (Lasso, described above). The R script for that is ***assigment5_lassoRegression.R.***

We also uploaded the R script with linear regression model and random forest: ***Assignment 5 Random Forest + Linear Regression.R*** just in case you would like to see our work.