# Case 1

02582 Computational Data Analysis

February 2025

## Case 1

The data for this exercise consist of 100 observations $(y, x)$, of response $Y$ (vector), features $X$ (100-dimensional feature matrix). Further, we have 1000 additional observations, here denoted $x_{new}$. Data is presented in the .csv files

`case1Data.csv`

and

`case1Data_Xnew.csv`

which are found on the course page on DTU learn, under Assignments. You can use any programming language you prefer e.g. R, Python or matlab. You can choose the methods you find suited to solve the case, please argue for your choices in the report. You should work in groups of **no more than two people**. In short your task is to build a predictive model of $Y$ based on $X$. Argue your choices and assess the quality of the chosen model. Apart from your predictions, $\hat{y}_{new}$, you should also estimate your prediction error. To complete this case you have to hand in three documents.

- A report on the case (max 5 pages, all included).
- Your predictions $\hat{y}_{new}$ (in a file called

  `predictions_YourStudentNos.csv`

  ; please insert your student numbers as a replacement for *YourStudentNos*.
- Your estimated prediction error RMSE (in a file called

  `estimatedRMSE_YourStudentNos.csv`

The requirements for the documents are described in greater detail in the following sections.

## The report

Your report should be short (**No more than max 5 pages**), in pdf format. Please use the provided latex template

`case1reportTemplate.pdf`

-Confusion Matrix - Have it for every Model
- AUC - for every model

**without** modifying the margin and dimensions, for a fair comparison. The report should answer to the following items:

- Describe your model and method (including model selection and validation).
- Argue for your choices of model, model selection and validation.

  - Ridge regression [OLS]
  - Lasso regression
  - Splines???
  - Regression trees / + maybe using bagging
- Describe how you handled missing data. Mean for regression. Use missing values as a new category for tree
- Describe how you handled factors in the features (catergorical variables). --> Treat the variables as ordinal and purely categorical and see what performs best
- Estimate the predictive performance of your model on $x_{new}$. We are interested in the root mean squared error $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$. As you do not know the true values $y_{new}$, you cannot just calculate the error, you need to estimate it. Your RMSE estimate will be denoted $R\hat{M}SE$. Describe what you did.

## The predictions and estimated prediction error

Your predictions $\hat{y}_{new}$ and your estimated prediction error $R\hat{M}SE$ should be uploaded to DTU inside in two text files. $\hat{y}_{new}$ in a file named

`predictions_YourStudentNos.csv`

and

`estimatedRMSE_YourStudentNos.csv`

The formats are illustrated in

`sample_predictions_YourStudentNo.csv`

and

`sample_estimatedRMSE_YourStudentNo.csv`

Please do not include headers in the file. Your predictions $\hat{y}_{new}$ and $R\hat{M}SE$ will be evaluated by the teachers.

## The competition

There is no case study without a great competition - actually we have two. There will be a prize for the group who submits the best predictions $\hat{y}_{new}$ in terms of their $RMSE$ (calculated by the teacher). The other prize goes to the group who gives the closest estimate $R\hat{M}SE$ to their actual $RMSE$ (measured in percent deviation and again calculated by the teacher). The winner will be announced at the lectures.