# Portfolio 2

## Experimental Methods 3: Multilevel models and machine learning

### E22.147201U013.A

Submission: XX. YY 2022

Student: Sara Kjær Kristensen

Email: [au702259@uni.au.dk](mailto:au702259@uni.au.dk)

Student no: 202105320

School of Communication and Cognition, University of Aarhus

Nordre Ringgade 104 4 th, 8200 Aarhus N, Denmark

Parts of this assignment have been made in a group consisting of Freddy Wulf (FW), Ida Møller (IM), Maria Mujemula (MM), Sabrina Zaki (SZ) & Sara Kjær Kristensen (SK). Each section has a highlighted initial of the responsible writer with editors following.

E.g. Introduction (MM, SK)

Maria wrote the main part and Sara adapted it.

The main parts of the assignment have been discussed in the mentioned group, assessed and written by SK. The group discussions will be addressed as ‘we’ or ‘the group’ hereafter.

Character count with spaces w/o plots:

# Table of content

[1. Portfolio 3](#_Toc116331246)

[2. Portfolio 10](#_Toc116331248)

[3. Portfolio 11](#_Toc116331250)

# The assignment

The Machine Learning assignment has 3 main parts: First we create a skeptical and an informed simulation, based on the meta-analysis. Second we build and test our machine learning pipeline on the simulated data. Second we apply the pipeline to the empirical data.

The report for the exam, thus, consists of the answer to all the following prompts:  
- Describe your machine learning pipeline. Produce a diagram of it to guide the reader (e.g. see Rybner et al 2022 Vocal markers of autism: Assessing the generalizability of ML models), and describe the different parts: data budgeting, data preprocessing, model choice and training, assessment of performance.  
- Briefly justify and describe your use of simulated data, and results from the pipeline on them.  
- Describe results from applying the ML pipeline to the empirical data and what can we learn from them.

Remember: plots are very very important to communicate your process and results.

## Part I - Simulating data

Use the meta-analysis reported in Parola et al (2020), create a simulated dataset with 100 matched pairs of schizophrenia and controls, each participant producing 10 repeated measures (10 trials with their speech recorded). for each of these "recordings" (data points) produce 10 acoustic measures: 6 from the meta-analysis, 4 with just random noise. Do the same for a baseline dataset including only 10 noise variables. Tip: see the slides for the code.

## Part II - ML pipeline on simulated data

On the two simulated datasets (separately) build a machine learning pipeline: i) create a data budget (e.g. balanced training and test sets); ii) pre-process the data (e.g. scaling the features); iii) fit and assess a classification algorithm on the training data (e.g. Bayesian multilevel logistic regression); iv) assess performance on the test set; v) discuss whether performance is as expected and feature importance is as expected.

Bonus question: replace the bayesian multilevel regression with a different algorithm, e.g. SVM or random forest (but really, anything you'd like to try).

## Part III - Applying the ML pipeline to empirical data

Download the empirical dataset from brightspace and apply your ML pipeline to the new data, adjusting where needed. Warning: in the simulated dataset we only had 10 features, now you have many more! Such is the life of the ML practitioner. Consider the impact a higher number of features will have on your ML inference, and decide whether you need to cut down the number of features before running the pipeline (or alternatively expand the pipeline to add feature selection).

Data: <https://www.dropbox.com/s/7ky1axvea33lgye/Ass3_empiricalData1.csv?dl=0>