Alina Maria Nechita
Student number: 202104831

Assignment 3

Machine learning

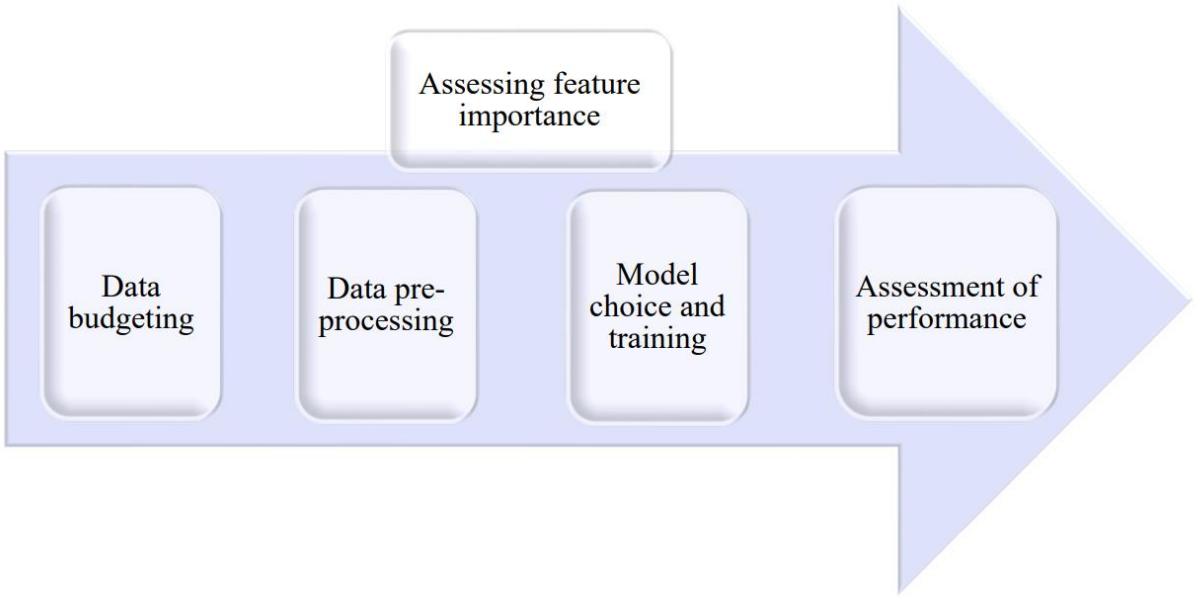
---

1. 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.

2. Briefly justify and describe your use of simulated data, and results from the pipeline on them.

3. *Describe results from applying the ML pipeline to the empirical data and what we can learn from them.*

Part 1



Alina Maria Nechita
Student number: 202104831

Data budgeting

Training data make up 80% of the total data, while testing data make up the remaining 20%. In order for the algorithm to learn equally from both categories of participants, the split should preserve the structure of the data, which means that participants from training data shouldn't appear in the testing data and there should be roughly equal numbers of controls and schizophrenic people (in this case) in the training data. The test set will be used to "check" if the algorithm could learn and infer the patterns, whereas the training set of data will be utilized for learning.

Data pre-processing

Scaling is necessary for the data (in our instance, the simulated and empirical data). Data scaling is done on the training dataset to ensure that population heterogeneity information won't have an impact on how well the algorithm performs on the test set. The test set is subsequently subjected to the same procedure (i.e., the mean and standard deviation).

Model choice and training

I first picked three models: model with fixed effects, varied intercepts, and different slopes in an effort to identify the model that could perform the best with our data. The logistic regression model is used to evaluate each model's quality (how effectively it can categorize an individual as belonging to the control or schizophrenia groups). In order to determine whether the priors should be more cautious or looser, I also evaluated how the priors impact the models' performance (sensitivity analysis of accuracy).

Assessment of performance

Examining the classification accuracy estimate and the model's type of errors allows one to judge how well the model performs classification. It's possible that the model incorrectly labels more "controls" as "schizophrenics," which could seem like a less serious error than labeling "schizophrenics" as "controls."

Assessing feature importance

The coefficients of each model are examined, and the analysis of the overall feature importance. The findings show which traits the model relies on often and which are most important for sorting the sample into the control and schizophrenia groups.

Alina Maria Nechita
Student number: 202104831

Part 2

Data simulation will help me better comprehend the classification issue since I'll be able to distinguish between control and schizophrenic individuals by knowing what makes up simulated data and which information it contains. Later, the machine learning process is utilized as a "marker" to determine whether the inferred data patterns and participant classification accuracy can be verified. It will also assist in determining which features have the greatest influence on the categorization method. The outcomes of the simulation will help us better grasp how the actual data will generally turn out.

Simulation part

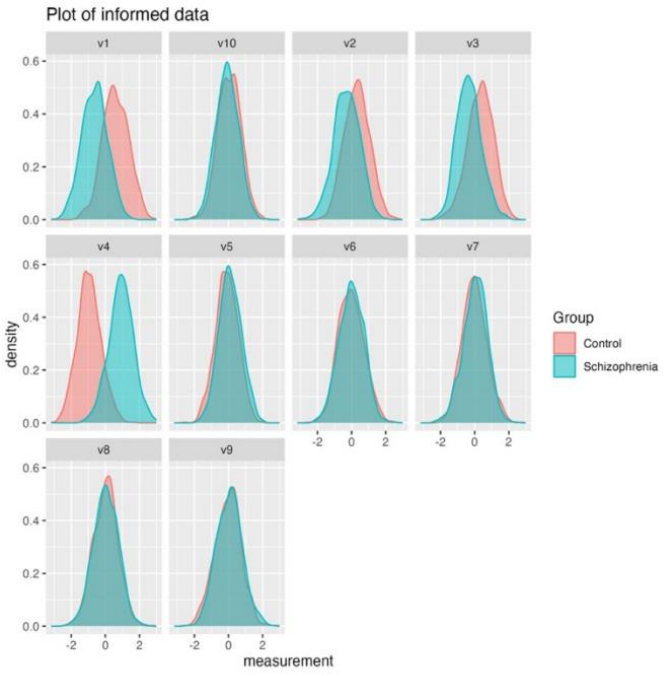
100 matched pairs of controls and schizophrenia patients were simulated in two datasets.

One dataset ("skeptic") has 10 acoustic measurements and noise variables, whereas the other ("informed") contains 6 meta-analysis measures and 4 random noise variables.

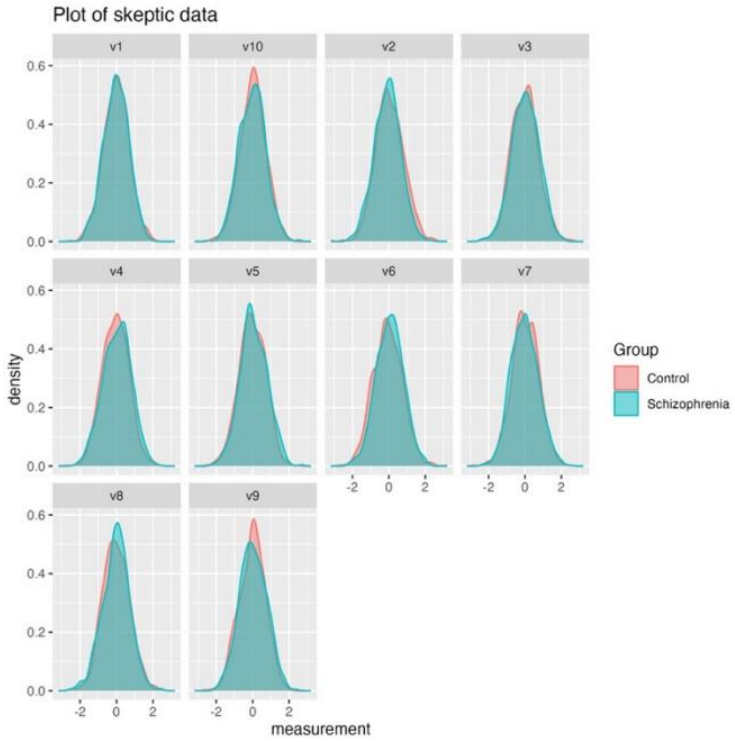
Below is a description of the significance of the meta-metrics: analysis's.

Acoustic measure	Proportion of spoken time	Pitch vari-ability	Speech rate	Duration of pauses	Pitch mean	Number of pauses
Effect size	- 1.26	- 0.55	- 0.75	1.89	0.25	0.05

Below are charts of data from knowledgeable and sceptical sources. Measures v1–v4 show the biggest differences because they have the largest impact sizes, whereas measures v5–v6 are closer to 0 and appear to overlap more. There is no discernible change in the graphs either because the effect sizes in v7-v10 are merely more noise and are the same for both data sets.



Alina Maria Nechita
Student number: 202104831



Application of ML pipeline on simulated data sets

Both informed and sceptical data sets are used for the data budgeting. Each data set's remaining 20% of data is utilized as test data, with the remaining 80% being used as training data. Additionally, it was ensured that different participants wouldn't show up both in the testing and training data sets. The measures in the training data sets of both skeptic and informed were scaled by using the mean and standard deviation of each feature (v1 – v10). The same recipe was applied on the test data.

I created three distinct models for informed and skeptic data sets independently in order to test which characteristics of the classification issue matter the most: one with fixed effects, one with variable intercepts, and the last one with variable slopes.

Fixed effects: $Group \sim 1 + v2 + v1 + v3 + v4 + v5 + v6 + v7 + v8 + v9 + v10$

Varying intercepts: $Group \sim 1 + v2 + v1 + v3 + v4 + v5 + v6 + v7 + v8 + v9 + v10 + (1|ID)$

Varying slopes: $Group \sim 1 + v2 + v1 + v3 + v4 + v5 + v6 + v7 + v8 + v9 + v10 + (1 + v2 + v1 + v3 + v4 + v5 + v6 + v7 + v8 + v9 + v10|ID)$

Alina Maria Nechita
Student number: 202104831

	Informed: training	Informed: test	Skeptic: training	Skeptic: test
Fixed effects	**Truth** **Prediction** CT SCZ CT 758 36 SCZ 42 764	**Truth** **Prediction** CT SCZ CT 196 7 SCZ 4 193	**Truth** **Prediction** CT SCZ CT 448 332 SCZ 352 468	**Truth** **Prediction** CT SCZ CT 92 107 SCZ 108 93
Varying intercepts	**Truth** **Prediction** CT SCZ CT 755 33 SCZ 45 767	**Truth** **Prediction** CT SCZ CT 193 4 SCZ 7 196	**Truth** **Prediction** CT SCZ CT 437 317 SCZ 363 483	**Truth** **Prediction** CT SCZ CT 87 104 SCZ 113 96
Varying slopes	**Truth** **Prediction** CT SCZ CT 798 6 SCZ 2 794	**Truth** **Prediction** CT SCZ CT 183 18 SCZ 17 182	**Truth** **Prediction** CT SCZ CT 771 31 SCZ 29 769	**Truth** **Prediction** CT SCZ CT 81 128 SCZ 119 72

The table above shows the results of classification of each of the model with both: informed and skeptic data sets.

Informed training data

When categorizing the patients into controls and schizophrenics, it appears that both fixed effects and varying intercept models perform equally on average (also have the same estimate of accuracy). With an accuracy of 0.995, the model with varying slope performs classification better.

Informed test data

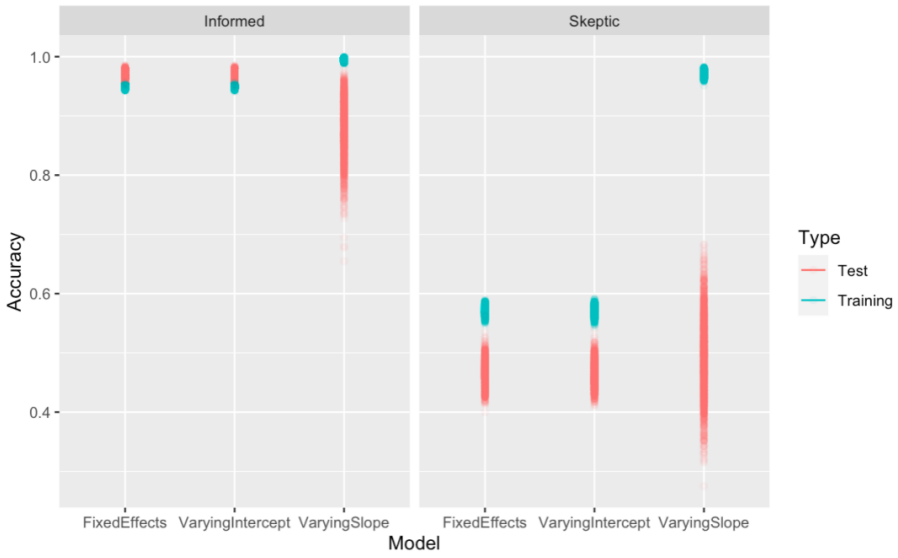
The first two models behaved similarly (the same accuracy estimate). The model with fixed effects, which categorized 7 schizophrenics as controls, performed better than the model with varying intercepts, which only classified 4 schizophrenics as controls. With varying slopes, the model's accuracy fell even further.

Skeptic training data

Although it does not perform as well as the same model with informed data set, the model with varying slopes performs the best. The first two models have about equal levels of accuracy.

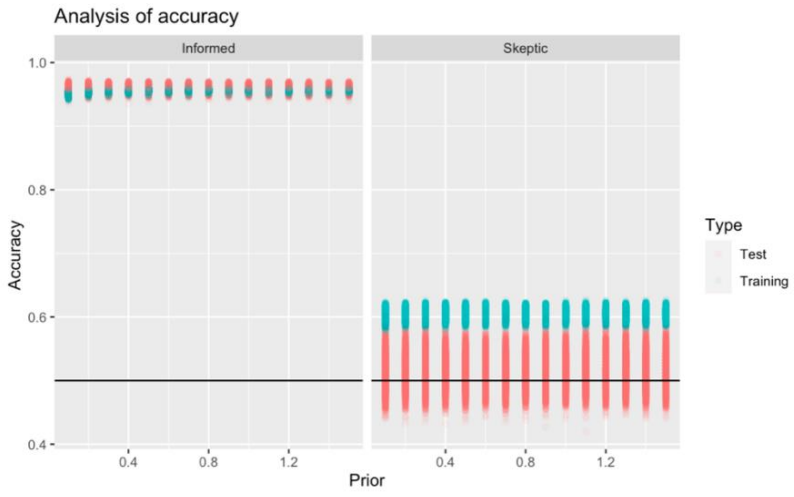
Alina Maria Nechita
Student number: 202104831

Skeptic data



The figure above summarizes the results of accuracy when fitting different models. The accuracy is just above the chance level, showing that models with skeptic data cannot successfully classify the diagnosis.

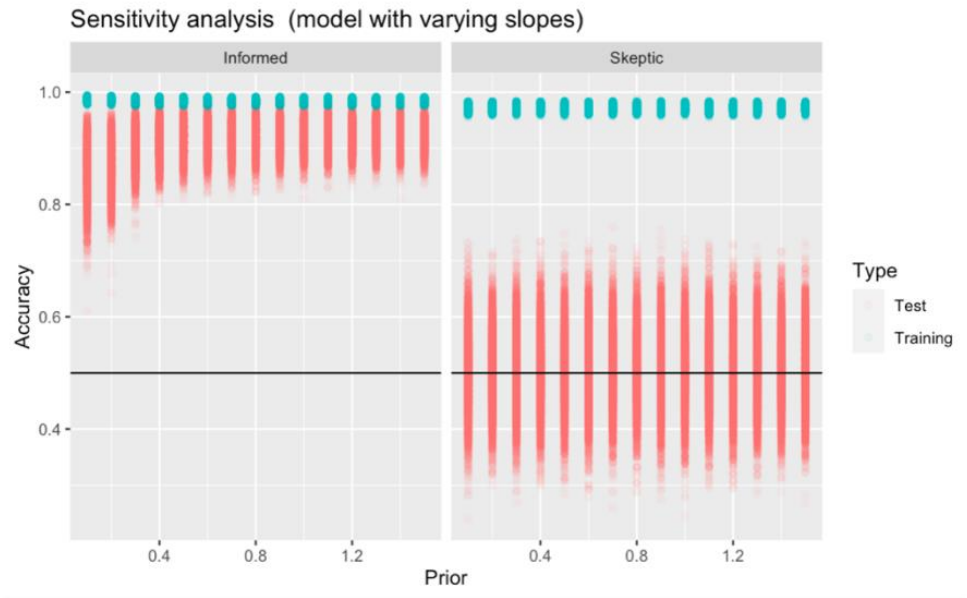
Sensitivity analysis of accuracy (assessing the impact of priors)



The graphic above shows how priors affect the model's ability to accurately divide participants into groups according to diagnoses. Here, the fixed effects model's performance is recorded.

Alina Maria Nechita
Student number: 202104831

Given that the uncertainty for both informed and skeptic data remains constant across all prior values, it appears that the prior has no effect on the accuracy of categorization.

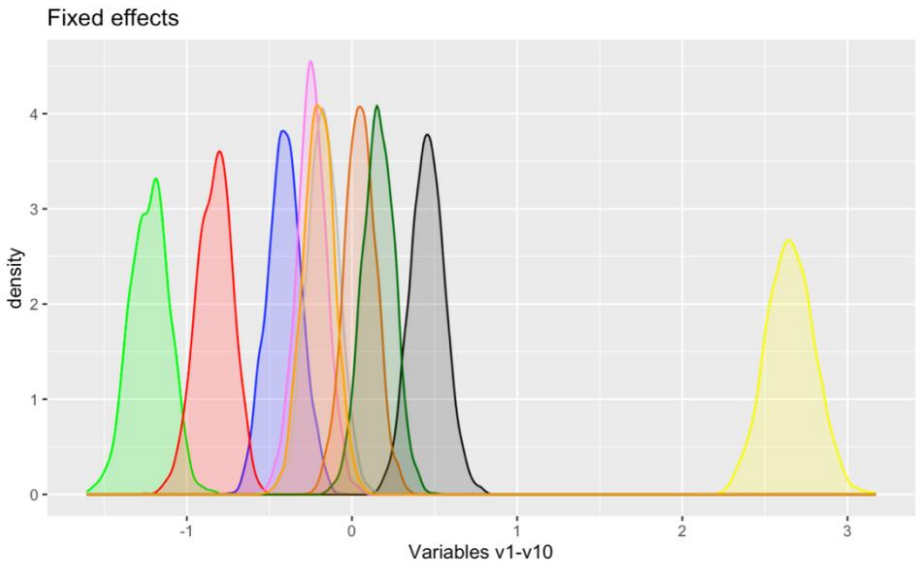


The graph above indicated that when priors are set for the model with varying slopes, the outcomes are somewhat different. Less conservative priors in this case decrease the classification's level of uncertainty in the test set with the available information, but have no impact on the model's level of uncertainty with the available skeptic data.

Feature importance

Using a thorough data set, I evaluated the feature importance in each model. Below is a summary of each model's findings.

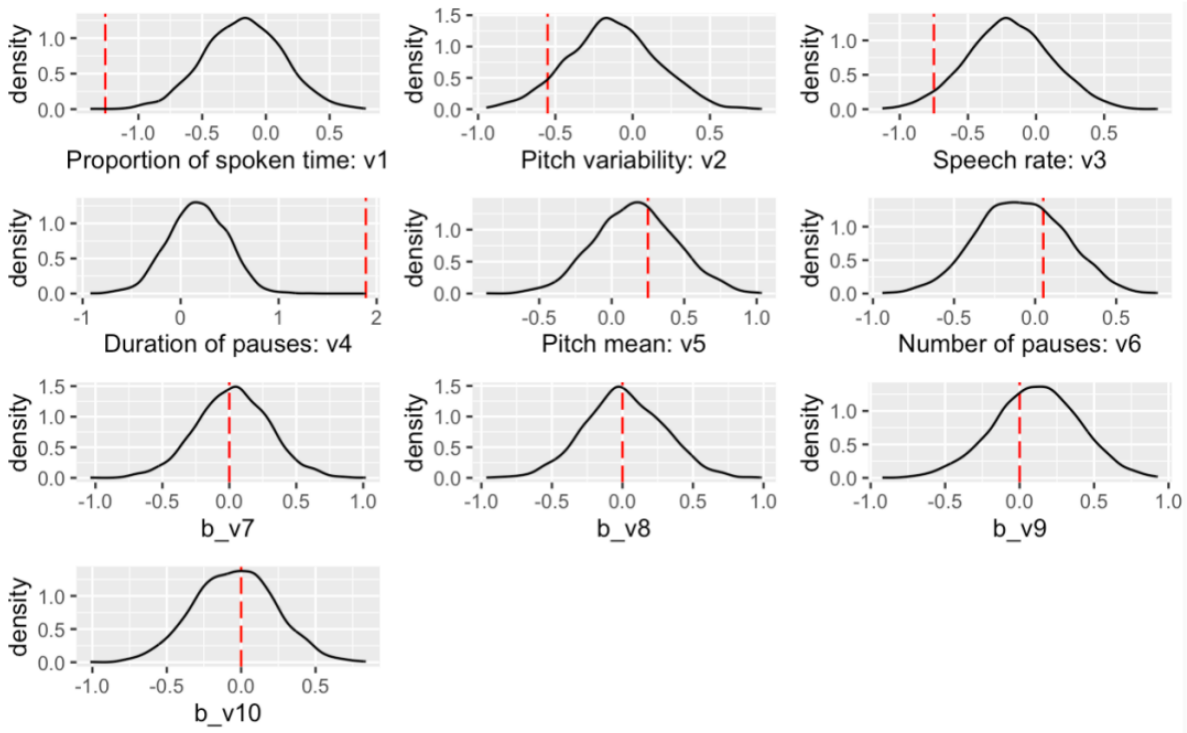
Fixed effect model



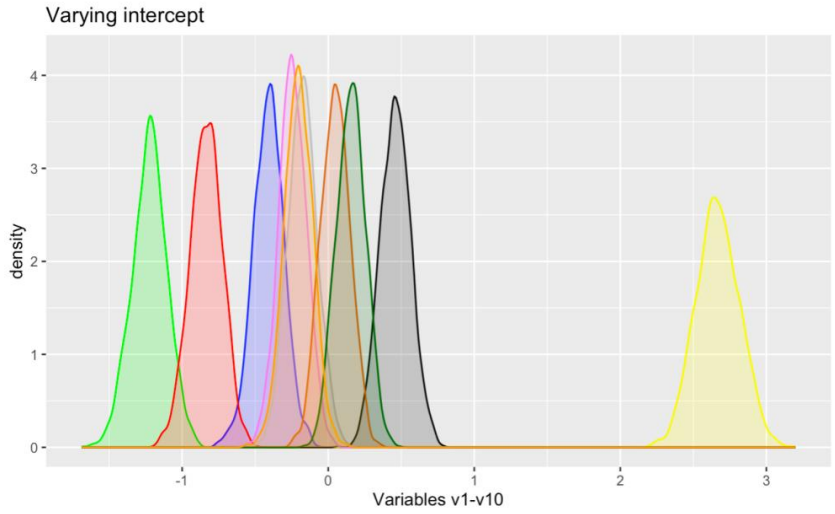
Alina Maria Nechita
Student number: 202104831

By examining the posterior distributions of each variable, it can be seen that the model heavily relies on the length of pauses (yellow), the percentage of spoken time (green), and the speech tempo (red).

The density graphs show the posterior distributions and their accuracy.

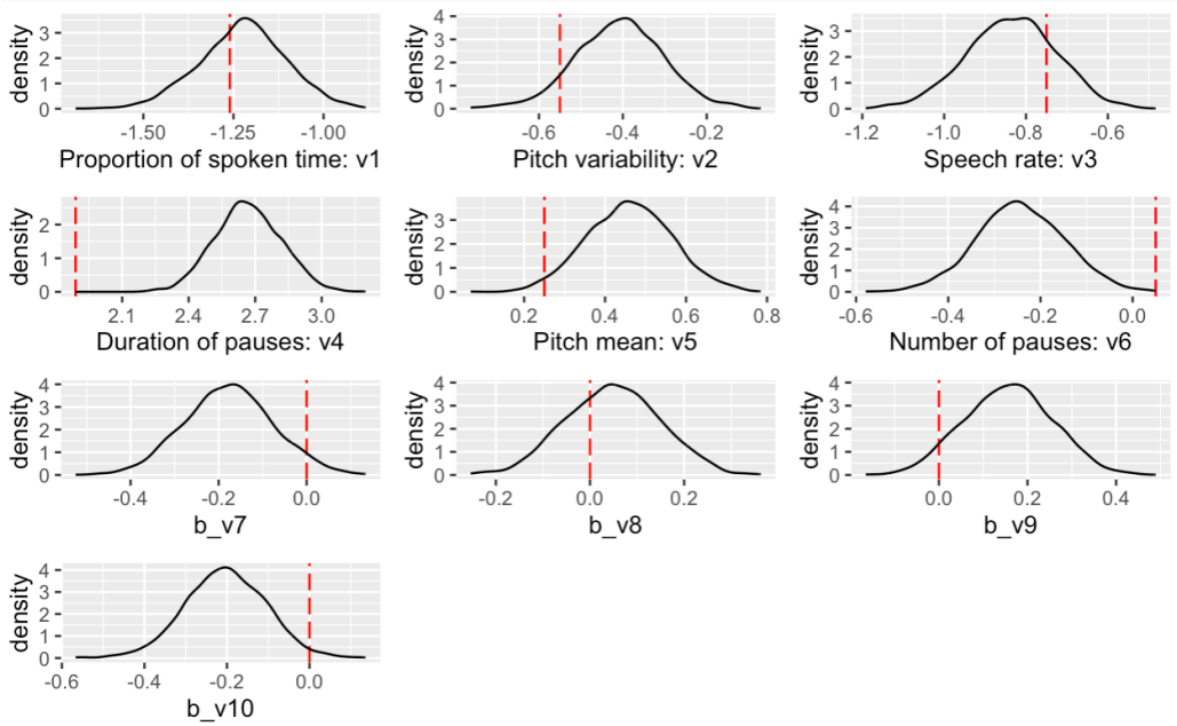


Varying intercept model



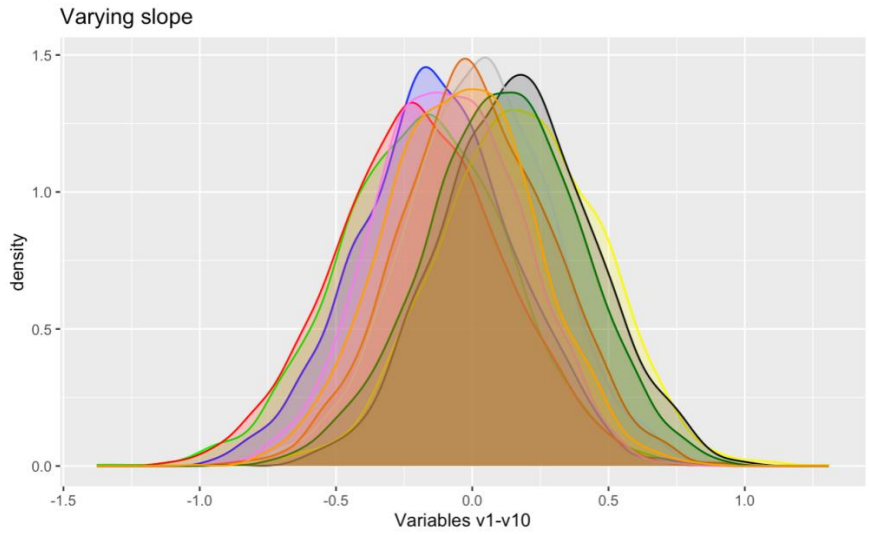
Alina Maria Nechita
Student number: 202104831

The output seems to be very similar compared to the model of fixed effects. The importance of the features is the same.



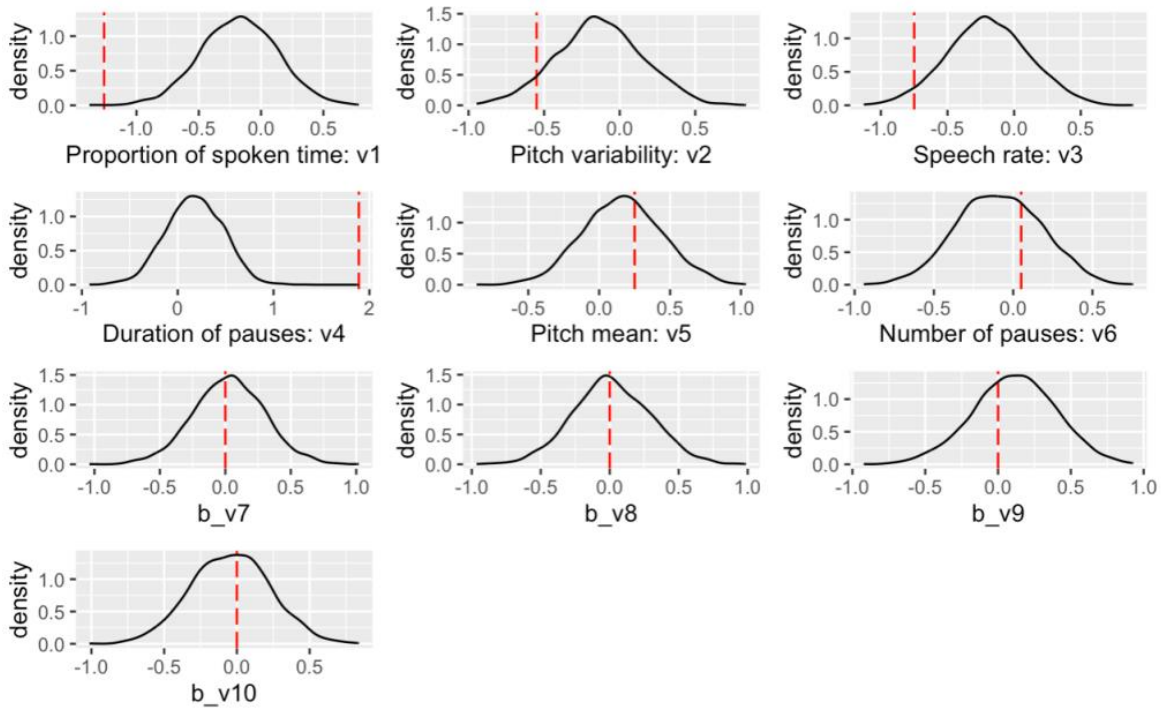
The density plots resemble those displayed previously in terms of similarity, but in this instance, the model with variable intercepts better represents the true effect of v1 and v3.

Varying slopes model

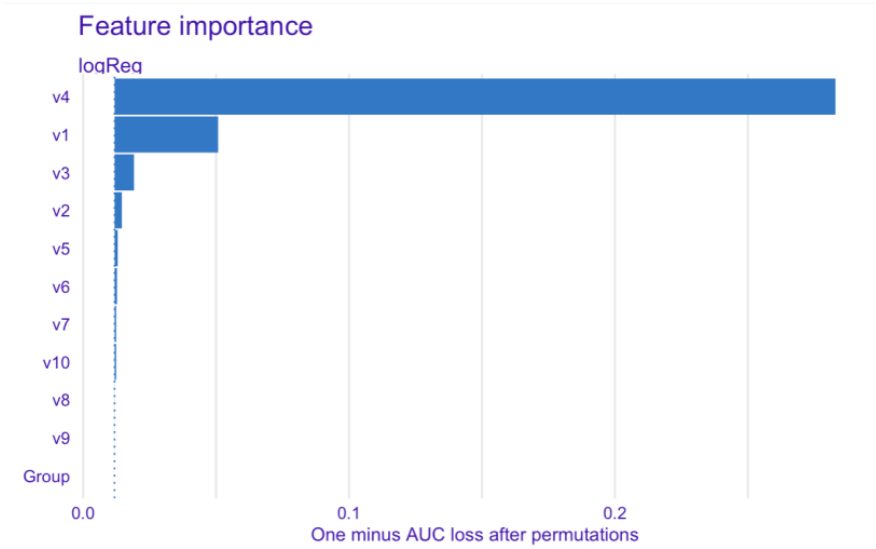


Alina Maria Nechita
Student number: 202104831

The distributions are overlapping a lot, the duration of pauses (v4 – yellow) becomes the important feature in this model (see the distribution of v4 below).



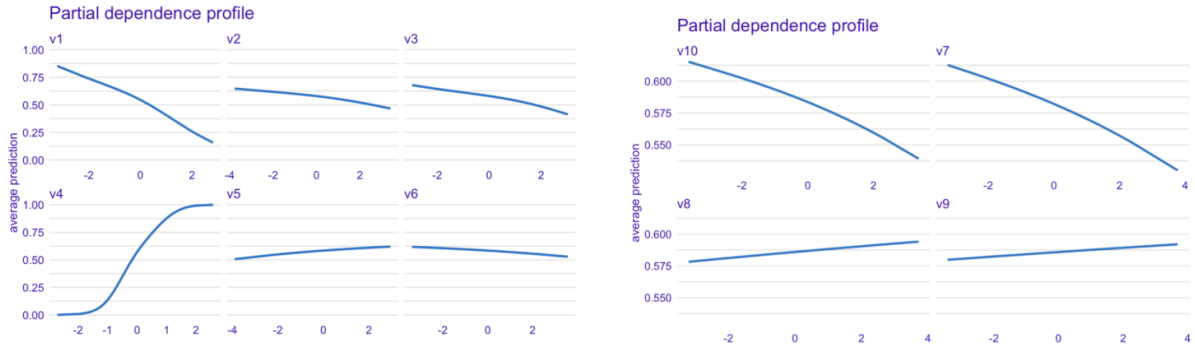
Global feature importance



The findings show that, once again, v4 is the most significant feature that is used, followed by the relevance of v1 and v3, when the logistic regression algorithm is used to evaluate the feature importance.

Alina Maria Nechita
Student number: 202104831

Similar to the pattern with model coefficients previously discussed, the overall pattern of significance is present.



The figure above shows the relevance of the characteristics that the model is estimating. Therefore, the likelihood of being identified as having schizophrenia increases as the length of pauses increases (v4). Additionally, as the percentage of spoken time declines (v1), so does the likelihood of receiving a schizophrenia diagnosis. The proportion of spoken time and the measure of speech rate (v3) have similar patterns, although the latter is a more important metric. The "noise" measurements v7-v10 would appear to be significant to the model from this plot, but since their impact on the model is zero and their plot sizes are less than those of the measures v1-v6, they are just random noise that can only accurately predict the diagnosis at the level of chance.

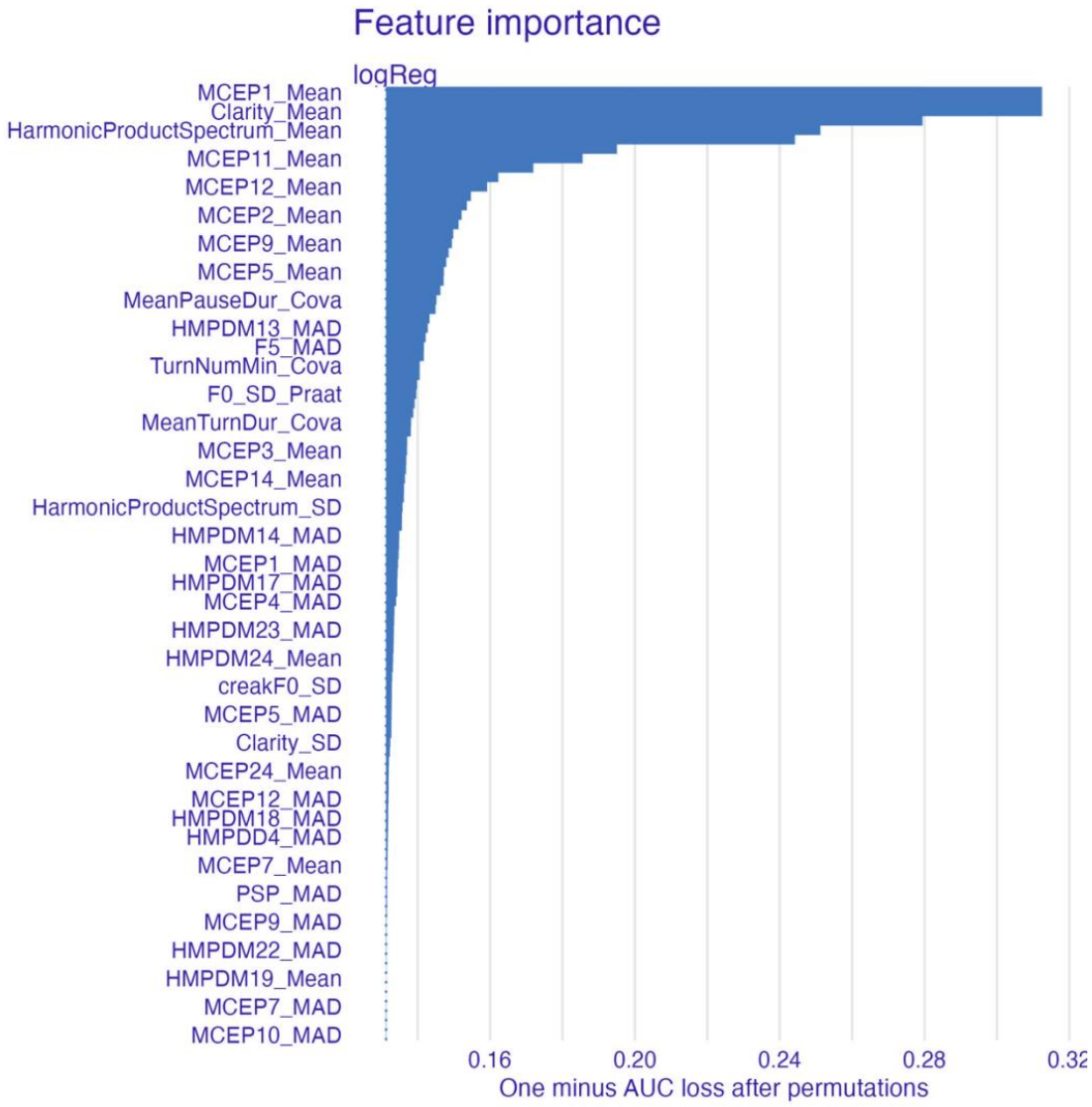
Part 3

The machine learning pipeline is used on the empirical data. The training set had 80% of the data, and the test set contained 20%. Additionally, I tried to fairly distribute the gender and diagnoses throughout the test and training sets, as well as to balance the groups such that the same person wouldn't appear in both sets. There are 654 females and 868 males in the training data set, of whom 721 have schizophrenia and 801 have been classified as controls. s. Test data set contains 154 females and 213 males, out of which 188 are controls and 179 – schizophrenics. Therefore, it seems that the groups are balanced, with a larger proportion of men and controls in each (as it is this way in the whole empirical data set). The training data is scaled in the same way as the simulated data was. The recipe of scaling the training set was applied to the test set as well.

Since empirical data contains a much higher number of predictors than simulated data, I decided to train the model using the "tidymodels" package. Here, I have chosen two models to set up the algorithm for classification: the logistic regression and random forest. In case of the logistic regression model, the accuracy of predicting the diagnosis on the test set is 0.664, on the training set – 0.787. For the classification procedure, I have built up two models: random forest and logistic regression. For the logistic regression model, the test set's accuracy in predicting the diagnosis was 0.664, whereas the training set's accuracy was 0.787.

Alina Maria Nechita
Student number: 202104831

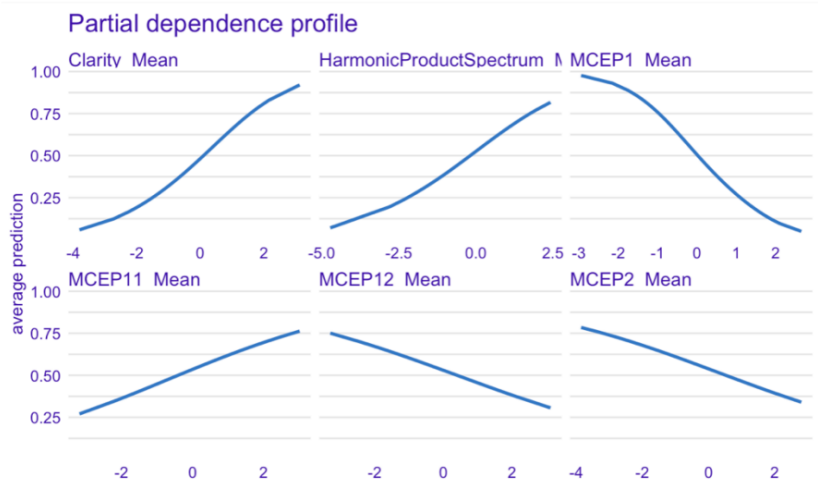
I choose to do a global feature importance analysis on the logistic regression model to determine which characteristics are most crucial for classification. In order to achieve that , I have taken off non-acoustic data like IDs, gender, and language. Additionally, the highly correlated features (correlation > 0.7) were also taken out of the dataset.

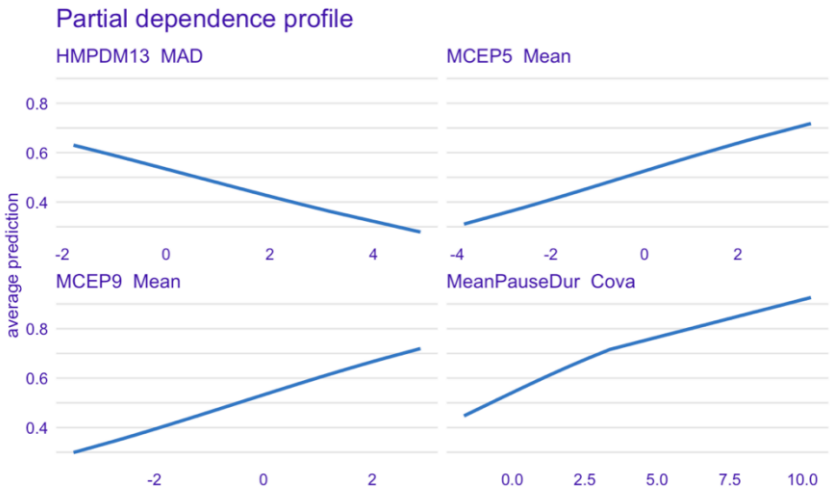


Global feature importance - empirical data (logistic regression model)

From the results of global feature importance, it is clear that the predictors MCEP1_Mean, Clarity_Mean and HarmonicProductSpectrum_Mean are 3 variables that are used the most when predicting the diagnosis.

Alina Maria Nechita
Student number: 202104831





The profile graphs above also show the estimated relevance that the model is predicting.

According to the algorithm of logistic regression, as mean of Clarity, Harmonic Product Spectrum, MCEP11, MCEP5, MCEP9 and MeanPauseDur_Cova increases, the probability of being classified as SCZ (schizophrenic) increases. The chance of being labeled as SCZ decreases in all other scenarios depicted above as a result of the drop in predictors.

The outcomes of the machine learning process reveal a great deal about the available empirical data. First and foremost, while performing data budgeting, it is important to understand the variables that make up the data and how they could influence the outcomes if balancing is not performed. In this case, by balancing the training and test data by gender and diagnosis, we make sure that there are no significant differences between the two data sets. By doing that, we get additional information about the proportion of males and females in the data set as a whole (the same is with diagnosis), which might help us to interpret the prediction results in a clearer way

(in this case, it's clear that the empirical data consists of more males, and people that are classified as "controls", therefore the algorithm might get more accurate when predicting diagnosis for these types of people. When determining which measures are most important in classification issues, it is useful to examine the value of global features. The algorithms that can accurately anticipate the diagnosis (with the fewest mistakes) may also be applied in real-world situations.