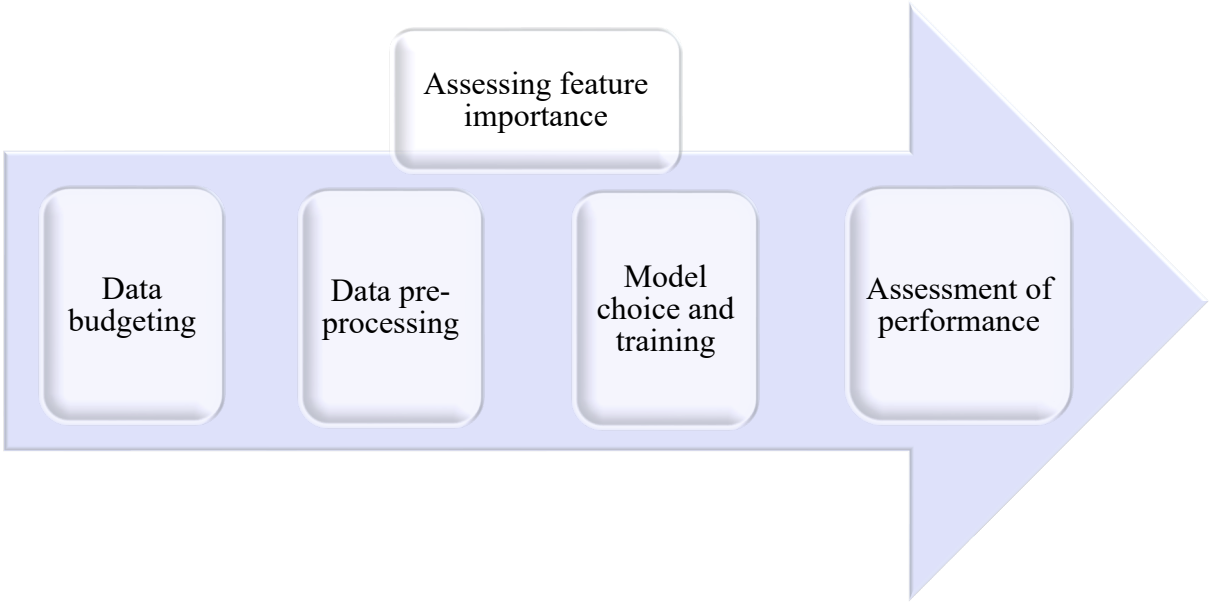
Justina Razanauskaite
202109601

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.*

Question 1



Data budgeting

The data is split into two parts: training (80% of the data) and testing (20 %) data. The split should preserve the structure of the data, meaning that participants from training data shouldn't appear in the testing data, there should be approximately equal amount of control and schizophrenic people (in this case) in the training data in order for the algorithm to learn from both categories of participants equally. While the training set of data will be used for learning, the test set will be used to "verify" whether the algorithm could learn and infer the patterns.

Data pre-processing

The data (in our case, the simulated and empirical data) has to be scaled. The scaling of the data is performed on the training dataset, as it would make sure that the information of

heterogeneity of population would not affect the performance of the algorithm on the test set. The same recipe (i.e., the mean and standard deviation) is then used on the test set.

Model choice and training

In order to find the model that might work best with our data, I chose three models to begin with: model with fixed effects, varying intercepts, varying slopes. To assess the quality of each model (how well the model can classify a person belonging to the group of control or schizophrenia), the algorithm of logistic regression is used. Furthermore, I assessed how priors affect the models' performance (sensitivity analysis of accuracy) to see whether the priors should be more conservative or looser.

Assessment of performance

How well the model performs the classification is assessed by looking at the accuracy estimate of classification, and seeing what kind of errors does the model make. It might be that the model classifies more "controls" as "schizophrenics", which might seem to be less significant error than as classifying "schizophrenics" as "controls".

Assessing feature importance

The feature importance is assessed by looking at the coefficients of each of the model, and by carrying out the analysis of global feature importance. The results indicate which features are used by the model a lot, and are most significant when classifying the sample into the groups of control and schizophrenia.

Justina Razanauskaite
202109601

Question 2

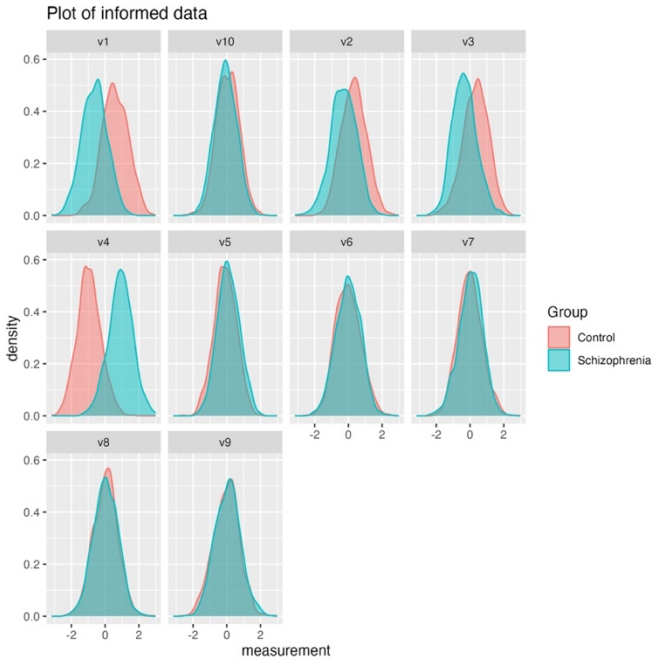
Data simulation will allow to understand the problem of classification more, as I know what simulated data consists of and which information, encoded in the data simulation, is useful when separating people into control and schizophrenic. Later, the machine learning pipeline is used as "marker" to see whether the data patterns are inferred and whether, according to specific measures, participants can be classified correctly. Moreover, it will help to figure out which features impact the classification algorithm the most. The results of simulation will allow to understand the general outcome of the empirical data and judge it's results in a better way.

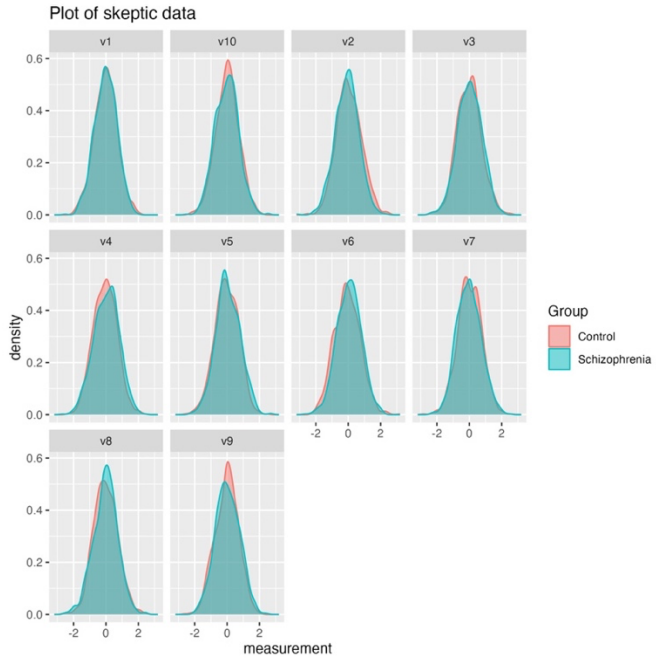
Generating data (simulation)

Two datasets with 100 matched pairs of controls and schizophrenic patients were simulated. One dataset consists of 10 acoustic measures – noise variables ("skeptic" data), another one ("informed") includes 6 measures from the meta-analysis and 4 of random noise. The meaning of the measures from the meta-analysis is indicated below:

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

The plots of informed and skeptic data are shown below. The biggest difference can be seen in measures v1 – v4, as these have the biggest effect sizes, whereas measures v5 – v6 are closer to 0 and seem to overlap more. As effect sizes in v7 – v10 are just additional noise and are equal across both data sets, there is no clear difference in the plots as well.





Justina Razanauskaite
202109601

Application of ML pipeline on simulated data sets

The data budgeting is performed on both informed and skeptic data sets. 80% of the data from each data set is used as a training data, the rest 20% as a test data. Moreover, it was made sure that the same participant would not appear in both training and testing 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.

In order to check which features in classification problem matter the most, I set up three different models for informed and skeptic data sets separately: one consisting of fixed effects, one of varying intercepts and the last one of varying 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: $Group \sim 1 + v2 + v1 + v3 + v4 + v5 + v6 + v7 + v8 + v9 + v10 + (1 + v2 + v1 + v3 + v4 + v5 + v6 + v7 + v8 + v9 + v10|ID)$

	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.

Justina Razanauskaite
202109601

Informed training data:

It seems like both – fixed effects and varying intercept – models on average perform similarly when classifying the patients into controls and schizophrenics (also have the same estimate of accuracy). The model with varying slope performs categorization more accurately, with accuracy of 0.995.

Informed test data:

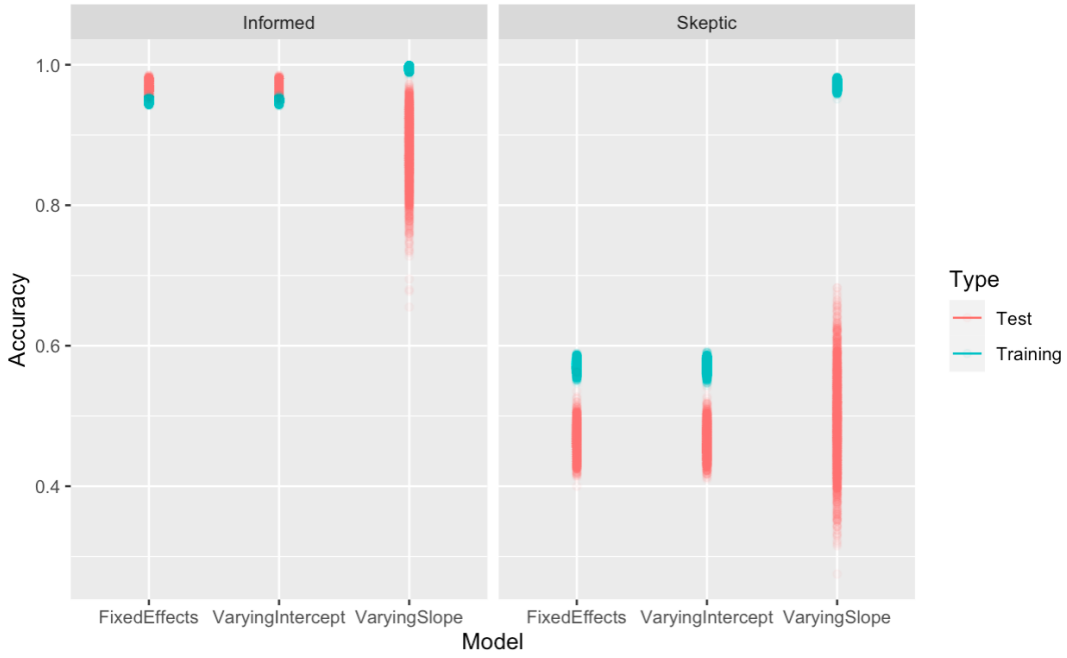
The first two models performed in the same way (the same accuracy estimate). However, the model with varying intercepts classified only 4 schizophrenics as controls, compared to the model with fixed effects (classified 7 schizophrenics as controls), making the second model better than the first. The accuracy of the model with varying slopes decreased even more.

Skeptic training data:

The model with varying slopes performs the best, although not as accurately as the same model with informed data set. The accuracy among the first two models is approximately the same.

Skeptic test data:

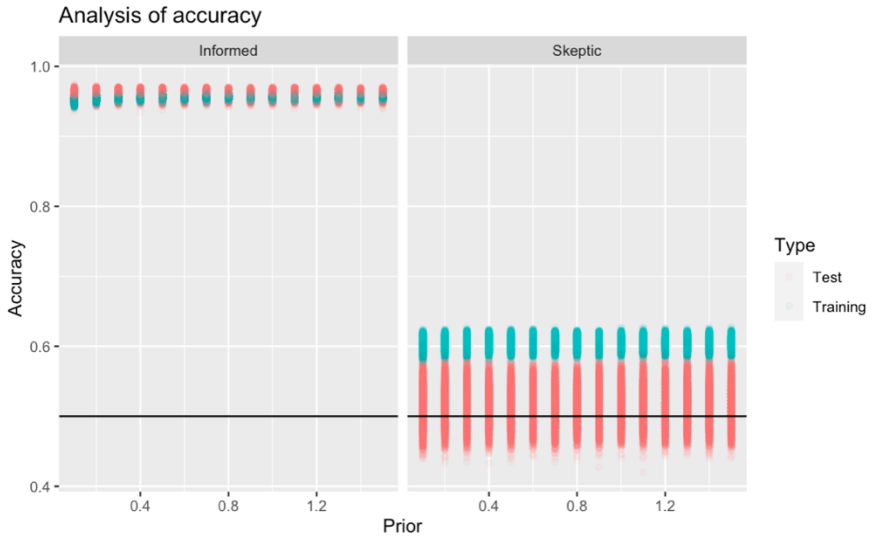
The accuracy is just above the chance level, showing that models with skeptic data cannot successfully classify the diagnosis.



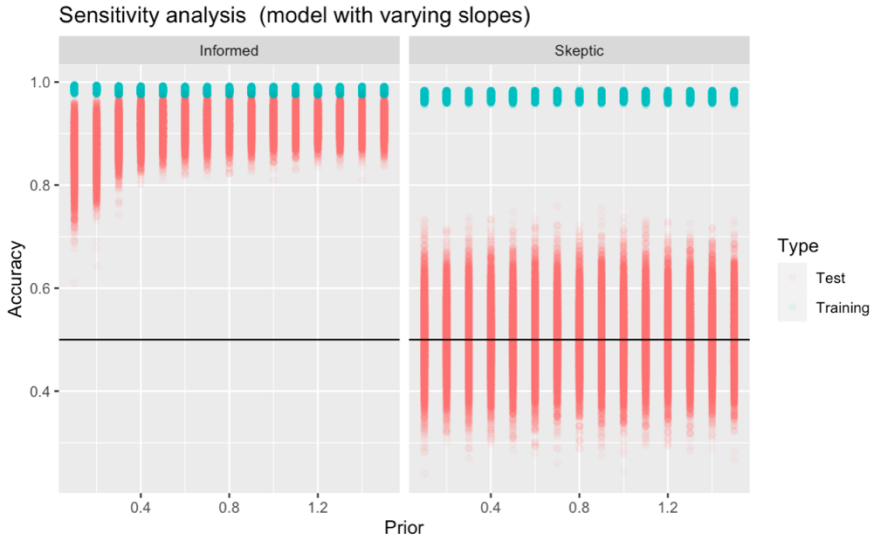
The figure above summarizes the results of accuracy when fitting different models.

Justina Razanauskaite
202109601

Sensitivity analysis of accuracy (assessing the impact of priors)



The figure above indicates how priors impact the accuracy of the model when it classifies participants into the groups of diagnosis. Here, the performance of fixed effects model is captured. It seems that the prior does not really affect the accuracy of classification, as the uncertainty for both informed and skeptic data does not change across all of the values of the prior.



However, the results are a bit different when it comes to setting priors for the model with varying slopes (above). Here, less conservative priors reduce the uncertainty (in the test set) of classification (with informed data set), nevertheless, do not affect the uncertainty of the model with skeptic data set.

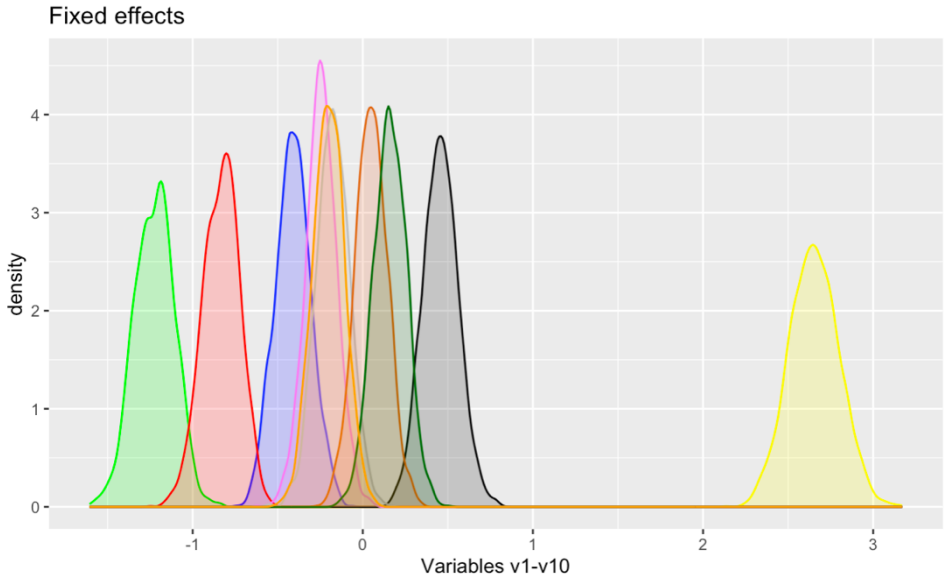
Justina Razanauskaite
202109601

Feature importance

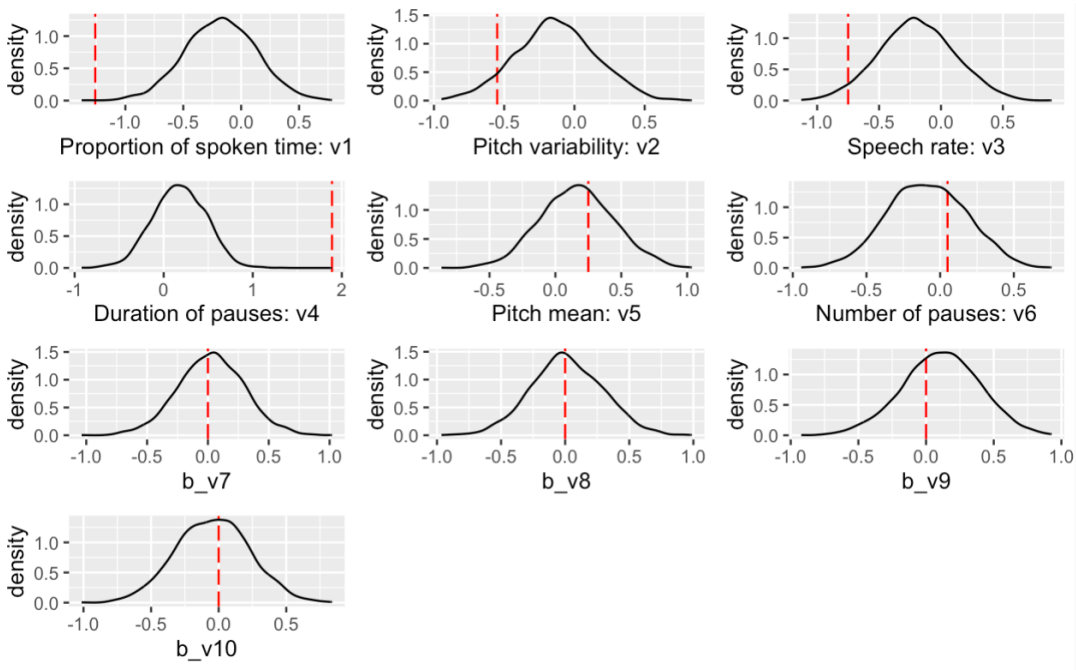
I have assessed the feature importance in all of the models with informed data set. The results of each model are summarized below.

Fixed effects model:

By looking at the posterior distributions of each of the variable, the duration of pauses (yellow), the proportion of spoken time (green), and the speech rate (red) are used a lot by the model.

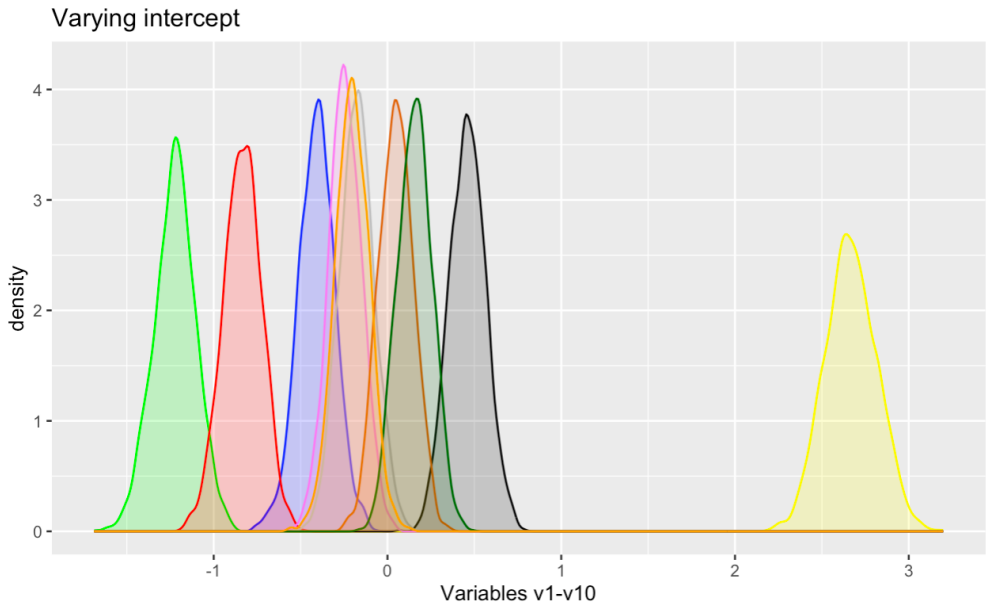


The density plots above illustrate the posterior distributions and whether they capture the true effect.

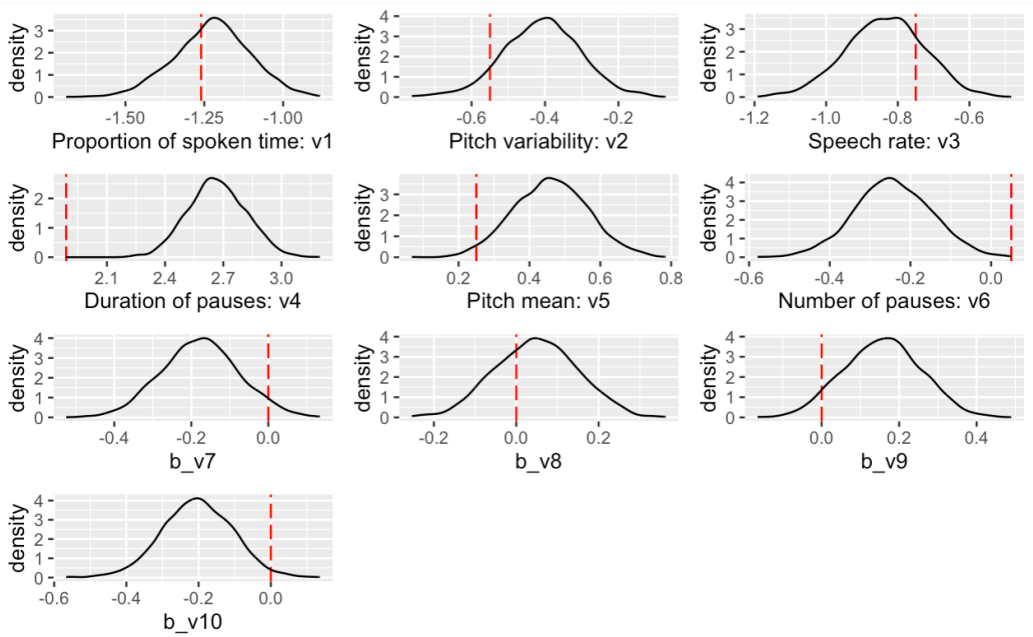


Justina Razanauskaite
202109601

Varying intercepts model:



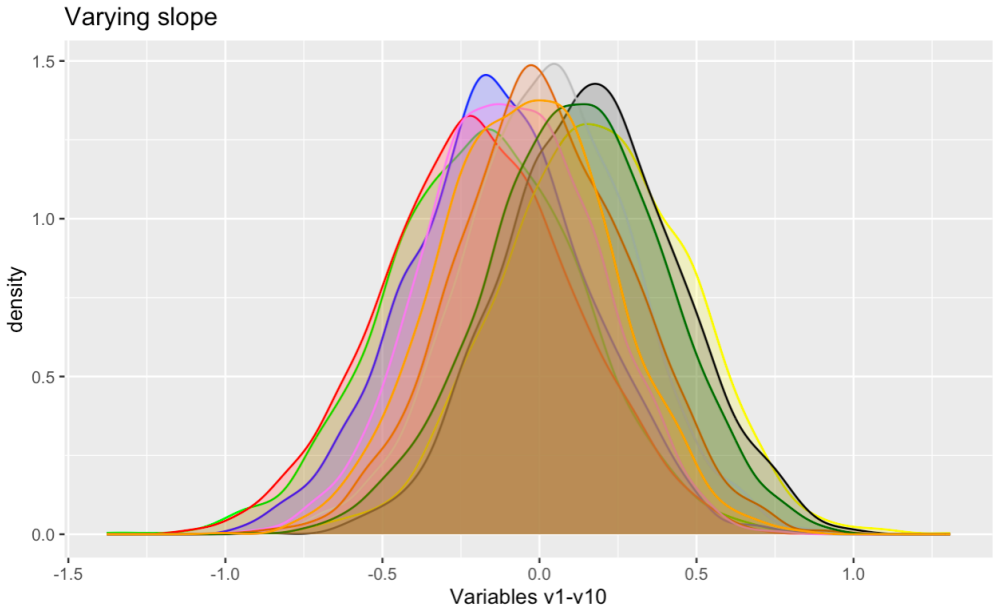
The output seems to be very similar compared to the model of fixed effects. The importance of the features is the same.



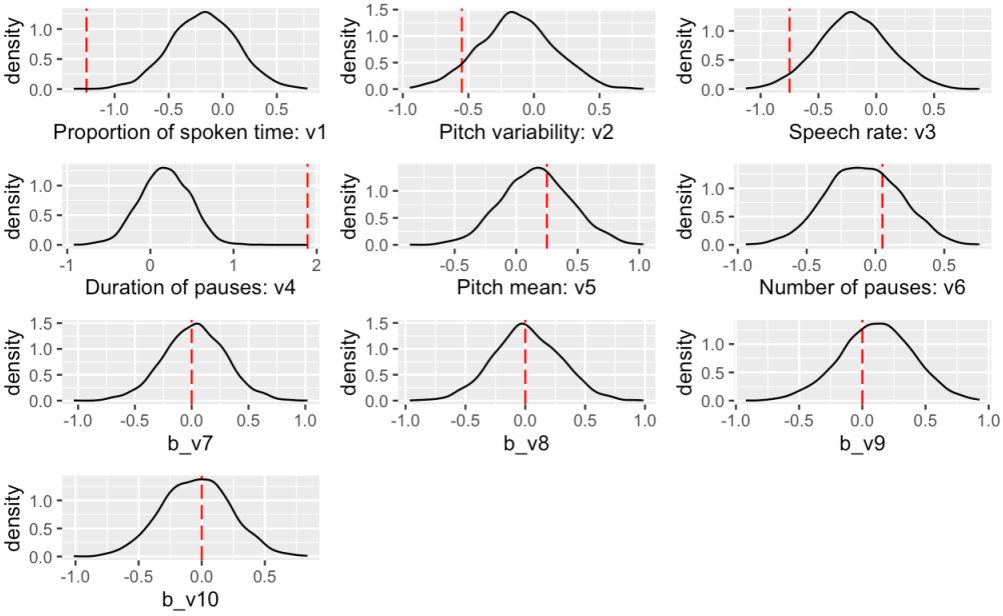
The density plots are also pretty similar to the ones shown above, however, in this case, the model with varying intercepts captures the true effect of v1 and v3 more accurately.

Justina Razanauskaite
202109601

Varying slopes model:

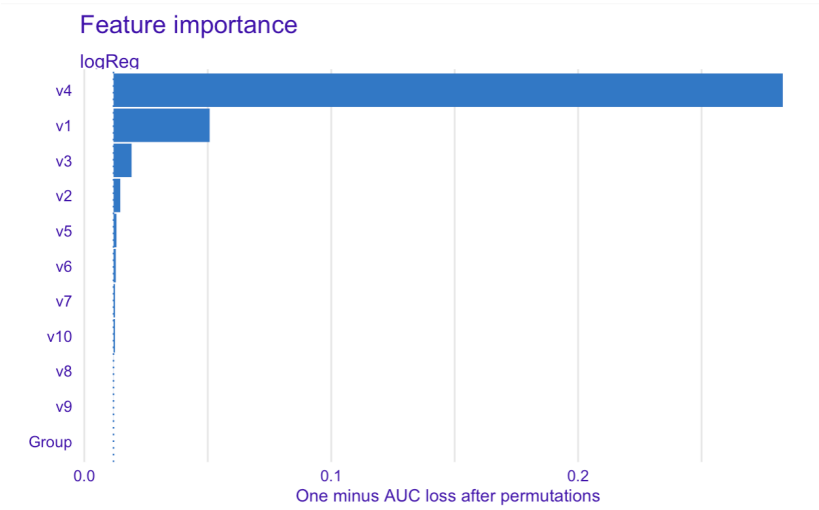


The distributions are overlapping a lot, the duration of pauses (v4 – yellow) becomes the important feature in this model (see the distribution of v4 below).

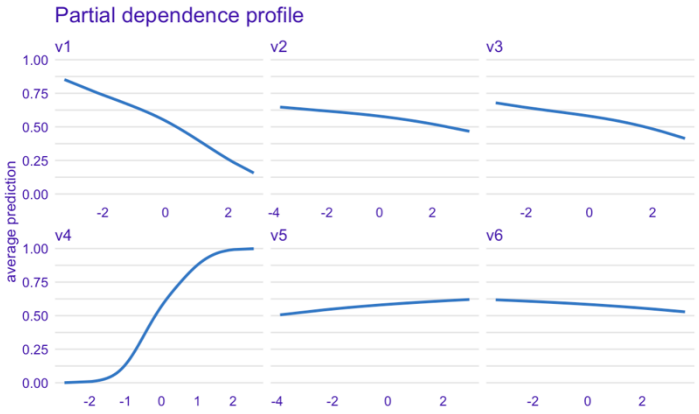


Justina Razanauskaite
202109601

Global feature importance:



By using logistic regression algorithm to assess the feature importance, the results indicate that again, v4 is the most significant feature that is used, followed by the importance of v1 and v3. The general pattern of importance is similar to the one with models' coefficients described above.



The importance of the features that the model is estimating is plotted above. Therefore, as the duration of pauses increases (v4), the probability of being diagnosed as schizophrenic increases as well. Moreover, as the proportion of spoken time decreases (v1), the probability of being diagnosed as schizophrenic decreases too. The measure of speech rate (v3) follows the same pattern as the proportion of spoken time; however, it is not as significant measure as the latter one is. From this plot, the "noise" measures v7-v10 might seem to show the importance to the model, but as the effect of these measures is 0, and also, plots are generated on a smaller scale compared to the measures v1-v6, they are just pure noise which predict the diagnosis at the chance level only.

Justina Razanauskaite
202109601

Question 3

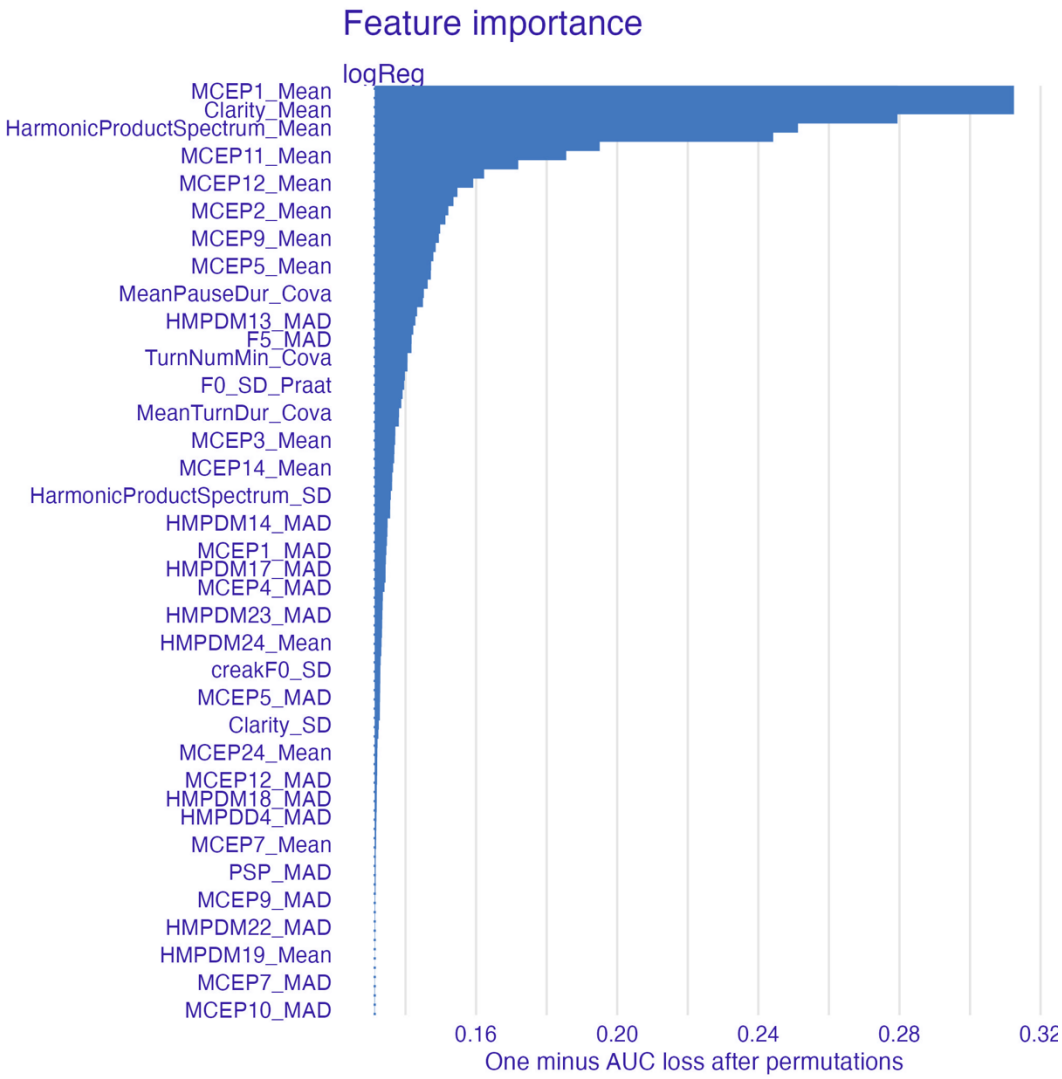
The machine learning pipeline is applied to empirical data. The data was split into two sets, the training set (80% of the data) and the test set (20 %). I have also attempted to balance the groups so that the same participant wouldn't appear in both sets, moreover, I attempted to roughly balance the gender and diagnosis among the test and training sets. Training data set consists of 654 females and 868 males, out of which 801 are diagnosed as controls and 721 as schizophrenics. Test data set contains 154 females and 213 males, out of which 188 are controls and 179 – schizophrenics. Therefore, it seems like the groups are balanced, both consisting of higher number of males/controls (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.

Empirical data has way more predictors than the simulated data, therefore I chose to fit the model using "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. In case of the random forest model, the accuracy is 0.73 on the test set, and 1 on the training set, therefore, the classification using random forest model is more accurate than logistic regression.

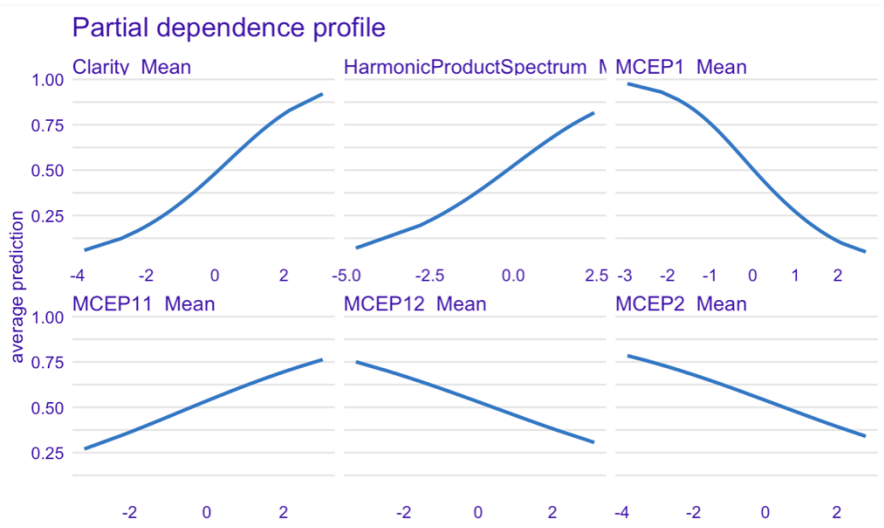
In case of the logistic regression model, I decided to perform a global feature importance, in order to see which features are the most important for classification. To do that, I have removed non-acoustic features, such as IDs, gender, language, etc. Moreover, the features, that are highly correlated (correlation > 0.7) were also removed from the dataset.

Justina Razanauskaite
202109601

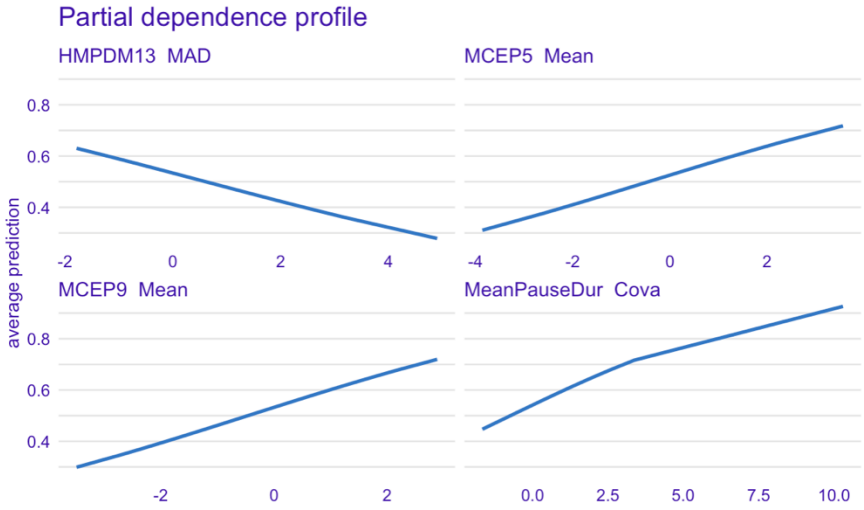


Global feature importance - empirical data (logistic regression model)

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.



Justina Razanauskaite
202109601



The importance that the model is estimating is also indicated in the profile plots above. 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. In all other cases shown above, the decrease in the predictors reduces the probability of being classified as SCZ.

The results of machine learning pipeline provide a lot of information about the empirical data at hand. First of all, when doing the data budgeting, we should be aware of the predictors that the data consists of, and how they might affect the results if the balancing is not applied. 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. Having done 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.

The analysis of global feature importance is beneficial when one wants to analyze which measures play the greatest role in classification problems. The algorithms, that might successfully predict the diagnosis (with as less errors as possible) might also be used in real life settings.