Frederieke Wullf
202105369

30.11.2022

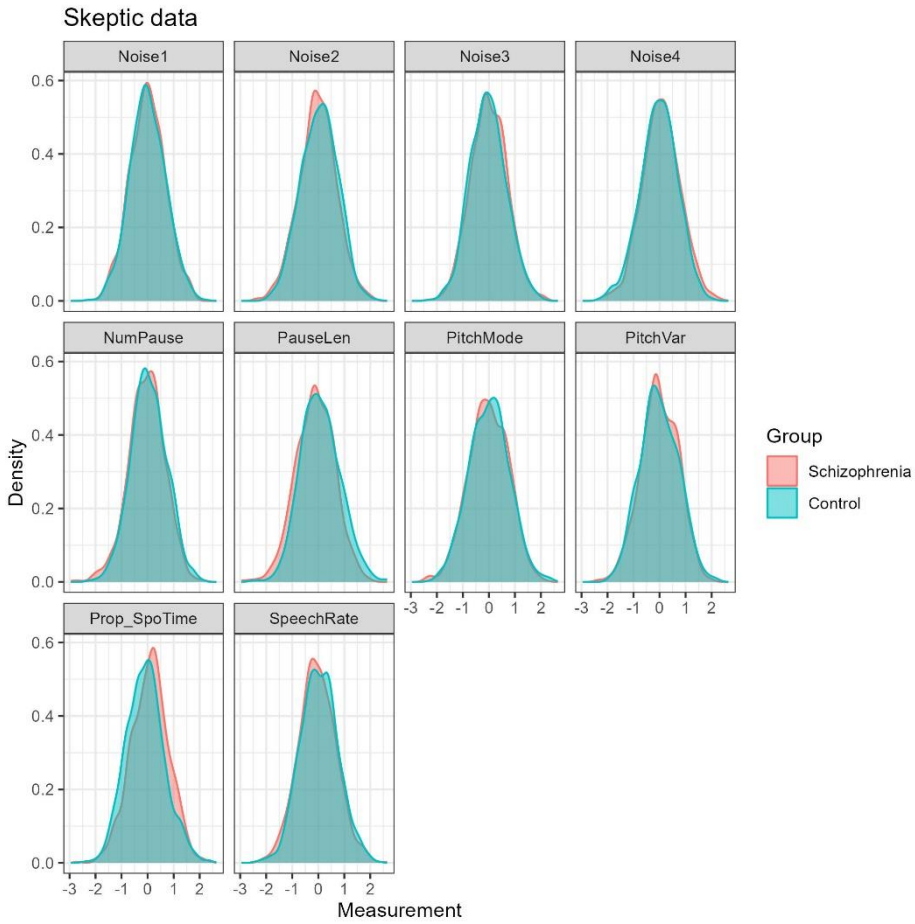
Portfolio 3

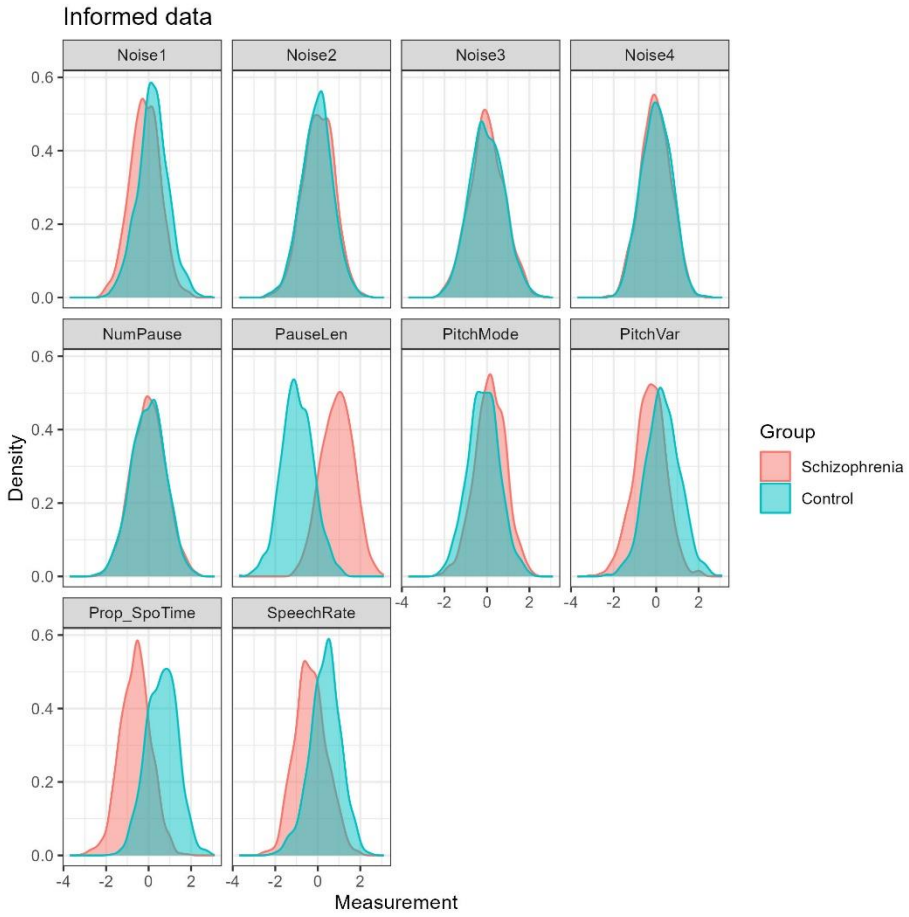
Part I

The goal of the assignment was to establish a machine learning pipeline that could assess whether someone is diagnosed with schizophrenia based on different vocal markers. The assignment was based on the meta-analysis by Parola et al. (2020).

Initially, some data was simulated to assess, which model to use. We created two different dataset, one that was more skeptical and one that was more informed, which was based on some parameters from the aforementioned meta-analysis. A total of 100 matched datapoints, one for a schizophrenic person and one for a control, were created. For each participant 10 repeated measures were created, each one a different acoustic measure. 6 out of 10 measures were taken from the meta-analysis and 4 were just random noise for the informed simulation. For the skeptical simulation, 0 was used for each of the 10 measures.

After creating the parameters, the error (set to 0.2) and the sd for the trials (0.5) and the individual sd's (1) were defined. The variables were then looped creating a tibble for each of the simulated datasets. To visualize the data, two plots were created that depict the difference of the individual variables between the schizophrenic group and the control group.





Part II

After simulating the two datasets, we then built a machine learning pipeline for each of the set respectively (*see Appendix*)

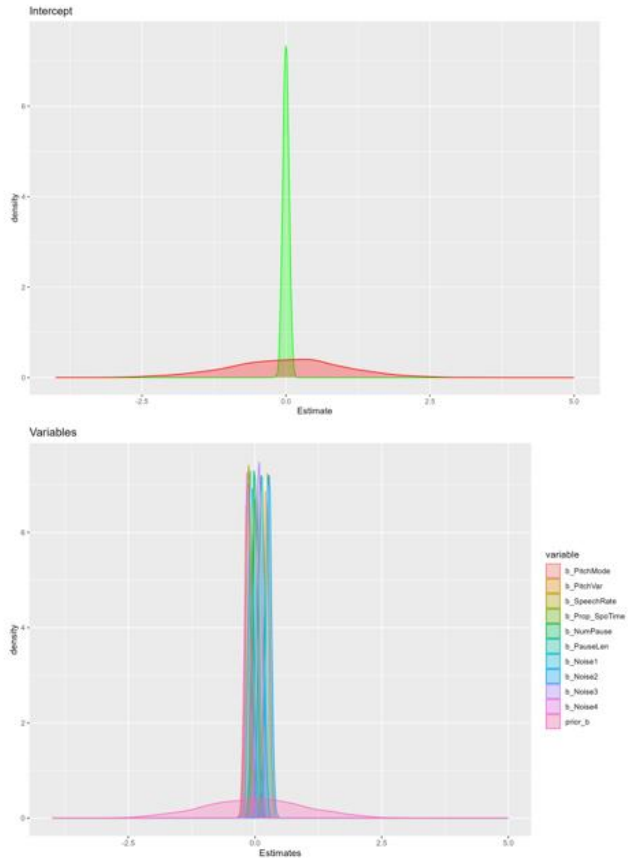
Firstly, we created a data budget, with 80% of the simulated data being split into the training set and the remaining 20% of the data being split into the test. This was again done for both simulated datasets and as will the following steps.

The next step was to pre-process the data. All four dataframes were standardized. Finally we defined the model parameters. Three different models were considered; a baseline model, a model with varying intercepts and a model with varying intercepts and slopes. It was chosen to only continue on with the varying intercepts- and the varying intercepts and slopes model, as it would be very unlikely that the baseline model would be successful at predicting as there is a lot of individual difference in between participants.

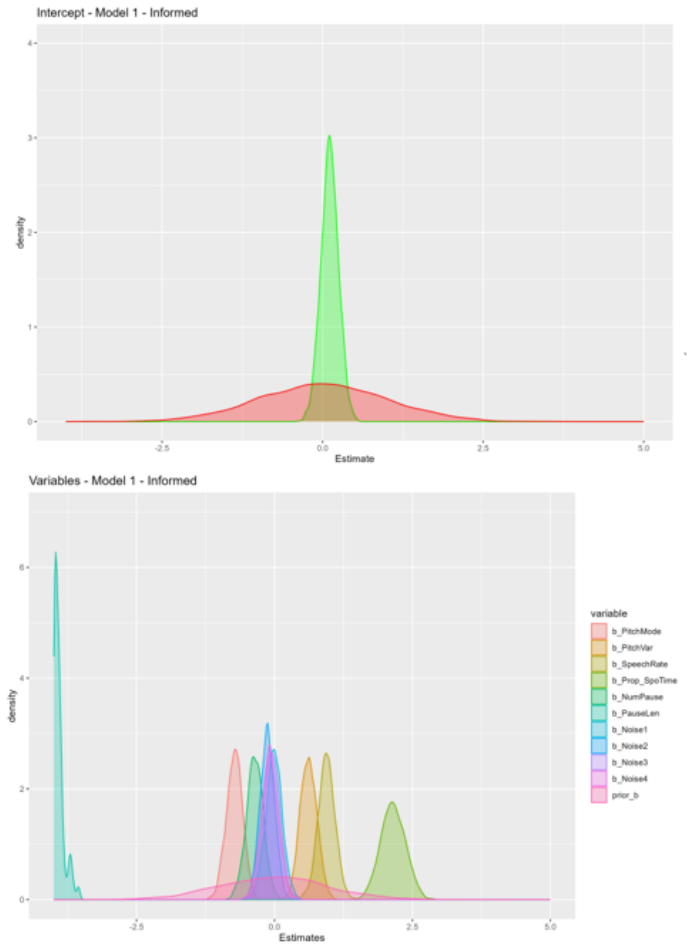
Model 1	PitchR_b1 ← bf(Group ~ (1 + PitchMode + PitchVar + SpeechRate + Prop_SpoTime + NumPause + PauseLen + Noise1 + Noise2 + Noise3 + Noise4)

Model 2	PitchR_b1 ← bf(Group ~ (1 + PitchMode + PitchVar + SpeechRate + Prop_SpoTime + NumPause + PauseLen + Noise1 + Noise2 + Noise3 + Noise4) + (1 | ID))

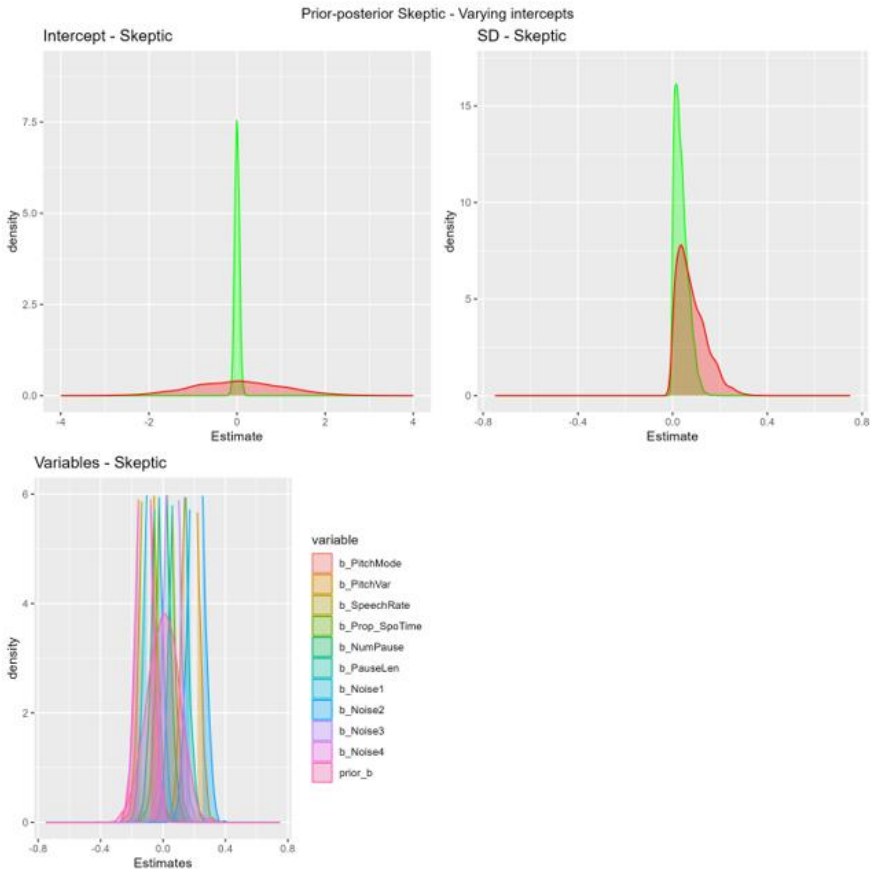
Afterwards, we defined normally distributed priors, which where then used to run the model with only priors. Following modelling the priors, we fitted the model and then conducted some prior-posterior update checks and visualized the data.



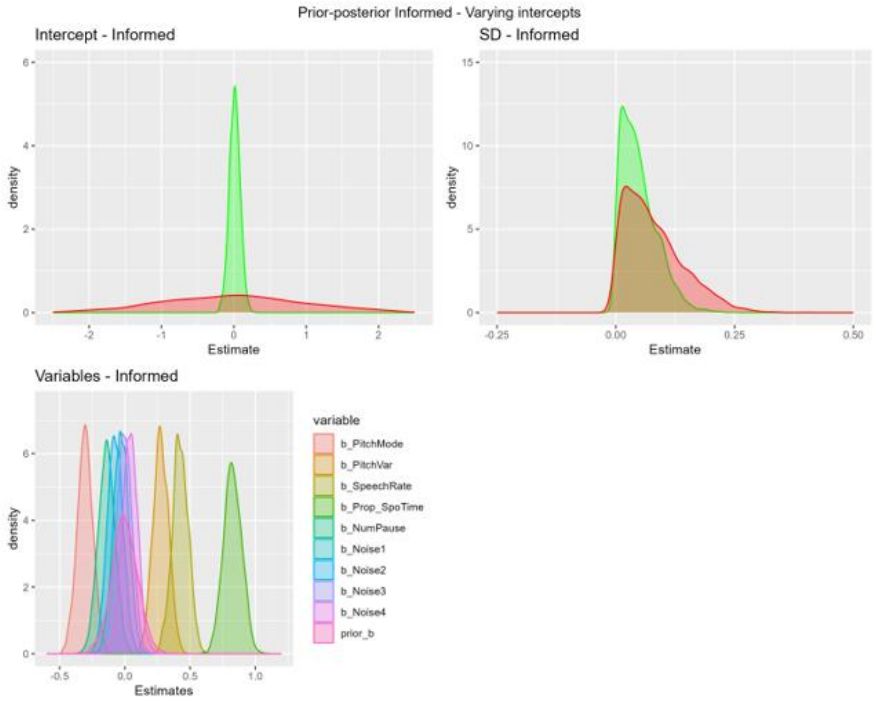
Model 1 - skeptical



Model 1 - informed



Model 2 - Skeptical



Model 2 - Informed

Looking at the prior-posterior updates, we decided to accept the priors. Some posterior distributions clearly show that they are bordering on the prior distribution, but as most of the posterior variables are well within the prior distribution, we settled on those priors.

Following the prior-posterior update checks, we were evaluating the two models using the leave one out (loo) weighted method.

Loo - Informed models

	Elp_diff	**SE_diff**	**Weight**
Model 1 (Baseline)	0.0	0.0	1.000
Model 2 (Varying Intercepts)	-166.1	9.3	0.0

Loo – Skeptic models

	Elp_diff	**SE_diff**	**Weight**
Model 1 (Baseline)	0.0	0.0	0.625
Model 2 (Varying Intercepts)	-0.4	1.8	0.375

Based one the leave one out comparison, the baseline model seems to be performing the best for both the skeptical and the informed model comparison. To further compare the models, we looked at the summaries of the models:

Model 1 – Informed

```
Population-Level Effects:
             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept        0.11      0.14    -0.16     0.37 1.00     4228     2147
PitchMode       -0.72      0.15    -1.01    -0.45 1.00     4860     2263
PitchVar         0.62      0.15     0.33     0.91 1.00     4879     2300
SpeechRate       0.94      0.15     0.63     1.25 1.00     4642     2275
Prop_SpoTime     2.15      0.22     1.72     2.59 1.00     4204     2305
NumPause        -0.36      0.15    -0.67    -0.06 1.00     4056     1866
PauseLen        -4.65      0.32    -5.30    -4.02 1.00     3824     2395
Noise1          -0.01      0.15    -0.31     0.29 1.00     4447     1930
Noise2          -0.13      0.13    -0.39     0.13 1.00     5255     2059
Noise3          -0.05      0.15    -0.34     0.23 1.00     4956     1809
Noise4          -0.09      0.15    -0.38     0.21 1.00     4007     2376
```

Model 1 – Skeptical

```
Population-Level Effects:
             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept       -0.00      0.05    -0.10     0.10 1.00     3725     2242
PitchMode       -0.15      0.05    -0.25    -0.05 1.00     3913     2068
PitchVar         0.23      0.05     0.12     0.33 1.00     3800     1760
SpeechRate      -0.12      0.05    -0.22    -0.01 1.00     3428     2085
Prop_SpoTime     0.02      0.05    -0.08     0.13 1.00     3383     2024
NumPause        -0.01      0.05    -0.11     0.09 1.00     3429     2136
PauseLen         0.13      0.05     0.03     0.23 1.00     3188     2167
Noise1          -0.08      0.05    -0.18     0.03 1.00     3262     2126
Noise2           0.27      0.05     0.16     0.37 1.00     3218     2225
Noise3           0.08      0.05    -0.02     0.18 1.00     3682     2564
Noise4          -0.14      0.05    -0.24    -0.04 1.00     3737     2246
```

Model 2 – Informed

```
Group-Level Effects:
~ID (Number of levels: 80)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.05      0.04     0.00     0.14 1.00     2064      990

Population-Level Effects:
             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept        0.01      0.07    -0.13     0.16 1.00     3884     1924
PitchMode       -0.30      0.06    -0.42    -0.19 1.00     4481     2195
PitchVar         0.28      0.06     0.16     0.39 1.00     4092     2384
SpeechRate       0.43      0.06     0.30     0.55 1.00     3662     2232
Prop_SpoTime     0.83      0.07     0.69     0.96 1.00     3575     2259
NumPause        -0.14      0.06    -0.26    -0.02 1.00     4454     2098
PauseLen        -1.34      0.07    -1.48    -1.21 1.00     4848     2008
Noise1          -0.03      0.06    -0.14     0.08 1.00     3695     2270
Noise2          -0.07      0.06    -0.19     0.04 1.00     3549     2300
Noise3          -0.00      0.06    -0.12     0.11 1.00     4390     2471
Noise4           0.04      0.06    -0.08     0.16 1.00     3771     2171
```

Model 2 - Skeptical

```
Group-Level Effects:
~ID (Number of levels: 80)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.04      0.03     0.00     0.10 1.00     2585     1038

Population-Level Effects:
            Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept       0.00      0.05    -0.10     0.11 1.00     5453     2046
PitchMode      -0.12      0.05    -0.21    -0.03 1.00     5827     2033
PitchVar        0.18      0.04     0.09     0.26 1.00     5791     2451
SpeechRate     -0.10      0.04    -0.18    -0.01 1.00     5350     2311
Prop_SpoTime    0.02      0.05    -0.07     0.11 1.00     5807     2283
NumPause       -0.01      0.05    -0.10     0.08 1.00     4832     1976
PauseLen        0.10      0.05     0.01     0.20 1.00     4488     2339
Noise1         -0.06      0.05    -0.15     0.03 1.00     5459     2254
Noise2          0.21      0.05     0.12     0.31 1.00     3943     1975
Noise3          0.06      0.05    -0.03     0.15 1.00     5326     1968
Noise4         -0.12      0.05    -0.21    -0.03 1.00     4135     2056
```

The output clearly depicts that all 4 models seem to perform quite well. The convergence diagnostic (Rhat) is 1.00 for all paramters in all of the four models. This generally means that the Markov Chains have mixed well during the sampling process. Furthmore, looking at the bulk- and tail-values, we can also see that the sampling efficiency seems to be very reliable.

Next, we calculated the accuracy of the predictions of the model, inlcluding various confusion matrices of average predictions of the different models with the test- and training data:

Skeptical – Training data

Truth	**Model 1**		**Model 2**	
Prediction	Schizophrenia	Control	Schizophrenia	Control
Schizophrenia	466	334	462	333
Control	334	466	338	467

Informed – Training data

Truth	**Model 1**		**Model 2**	
Prediction	Schizophrenia	Control	Schizophrenia	Control
Schizophrenia	774	33	768	39
Control	26	767	32	761

Skeptical – Test data

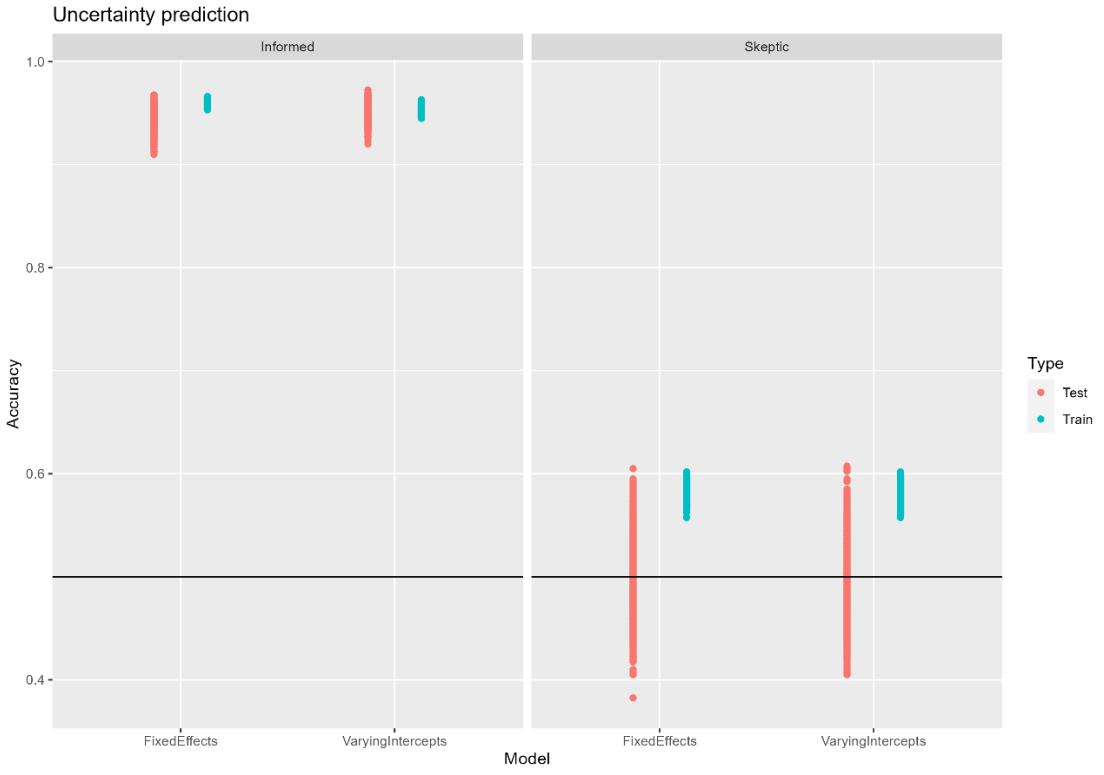
Truth	**Model 1**		**Model 2**	
Prediction	Schizophrenia	Control	Schizophrenia	Control
Schizophrenia	93	101	88	102
Control	107	99	112	98

Informed – Test data

Truth	**Model 1**		**Model 2**	
Prediction	Schizophrenia	Control	Schizophrenia	Control
Schizophrenia	189	10	192	7
Control	11	190	8	193

Based on the matrices, we can clearly see a trend in that the informed models seem to perform much better, while model 1 (Fixed Effects/baseline) and model 2 (Varying Intercepts) both seem to be quite good at predicting whether someone has schizophrenia or not.

Lastly, we calculated the uncertainty and plotted it together with the accuracy:

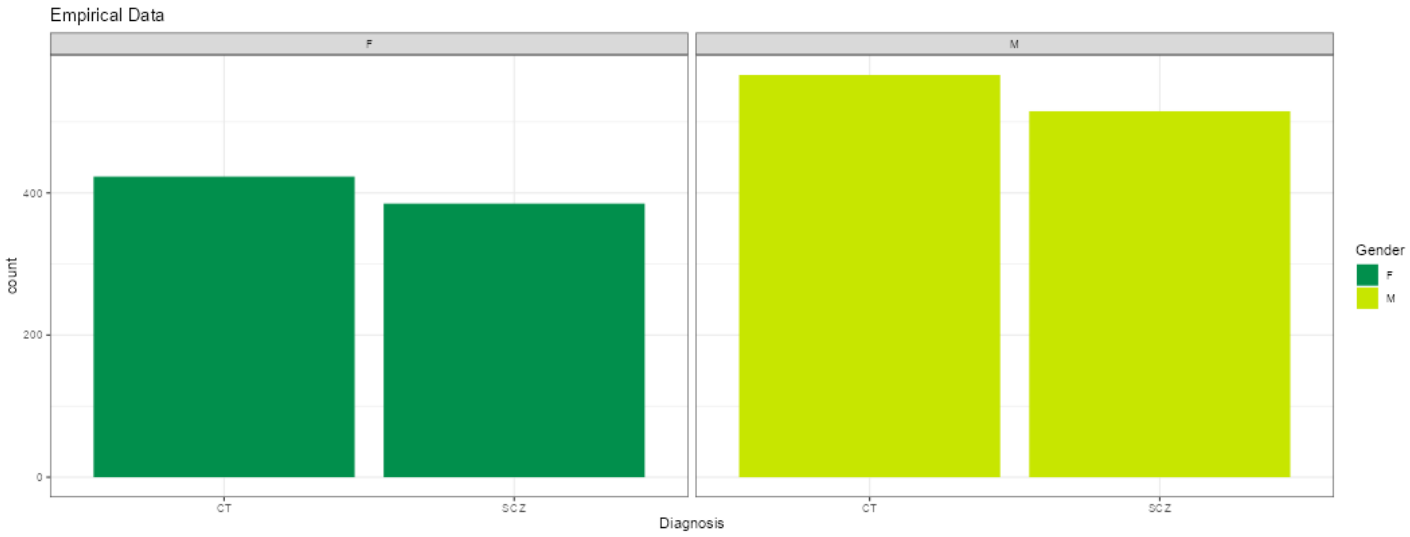


Based on the uncertainty it is obvious that the skeptical data would carry more uncertainty. Furthermore, the accuracy of the informed model also obviously is higher than the accuracy of the skeptical model. Only looking at the two different models, the accuracy and the uncertainty does not seem to change lot, based on the model.

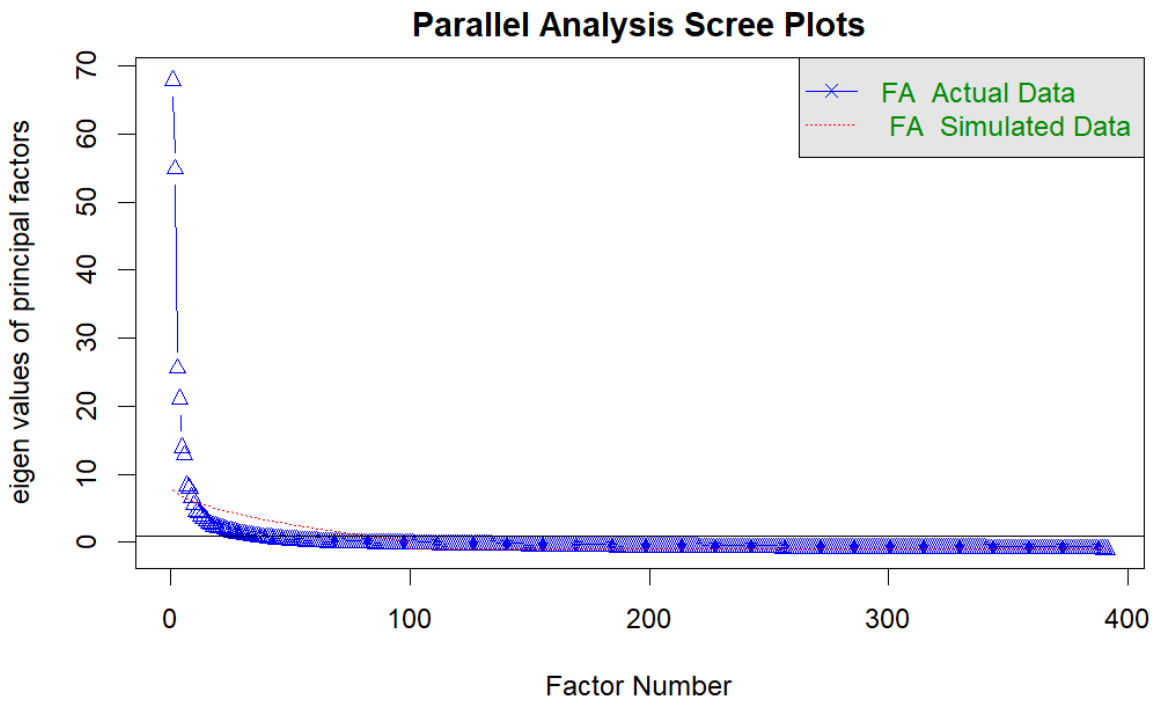
All in all, assessing the model brought the conclusion that model 1 *(Group ~ 1 + PitchMode + PitchVar + SpeechRate + Prop_SpoTime + NumPause + PauseLen + Noise1 + Noise2 + Noise3 + Noise4*), would be the best model for the Machine Learning Pipeline in part 3.

Part III

In part 3, we began with looking at the data that we were going to use for the ML pipeline. The data comes from the Meta-analysis of Parola et al. (2020). The data includes 1889 observations of 398 variables. Some of the variables include number of pauses, percentage of spoken time and the mean pitch (including gender differences). All in all, the data includes 221 participants of which 105 are schizophrenic. Of those 105 schizophrenic participants, 60 are male and 45 are female. The other 116 participants are heathy controls, of which 66 are male and 50 are female. The groups seem to be balanced fairly well based on gender and diagnosis differences.



After getting an overview of the data, we began with the preparation of the data. As the data contains almost 400 variables, it firstly makes sense to distinguish between how much influence and importance each variable/measurement has. We made a parallel analysis scree plot, which is based on a principal component analysis. The pca indicates that 9 factors were important to our investigation.



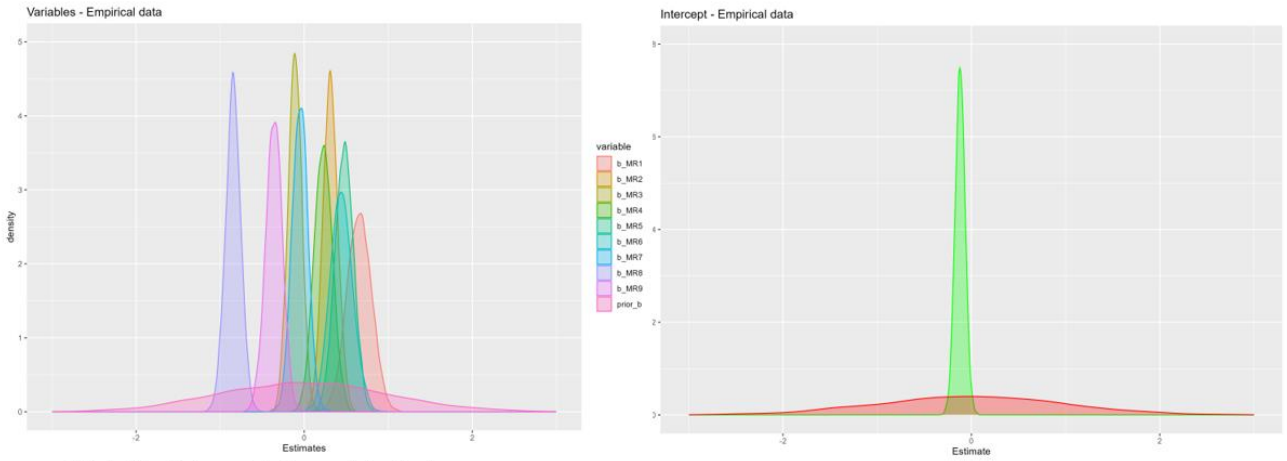
We then decided to follow with a factor analysis that would reduce the larger number of variables into 9 factors. Those factors were named *MR1, MR2, …, MR9*.

For the data pre-processing, the dataset was split into a training (0.8) and a test set (0.2), using the function *sample.split()* from the package *caTools*. Afterwards, both the test and the training set were pre-processed, using the *tidymodels* package.

Next, we modelled the data with the model we chose in part II.

Diagnosis ~ (1 + MR1 + MR2 + MR3 + MR4 + MR5 + MR6 + MR7 + MR8 + MR9)

Letting the model run with only priors and then fitting it with the data, gives the following prior posterior update check:



Model 1 – Prior posterior update check

The model summary reveals that the model has been successful at sampling. The Rhat value is 1.00 all the way through and the bulk and tail values indicate that the sampling efficiency seems to be very reliable.

```
Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    -0.12      0.05    -0.22    -0.02 1.00     6113     4128
MR1           0.65      0.15     0.37     0.93 1.00     3666     3982
MR2           0.31      0.09     0.14     0.49 1.00     3951     4125
MR3          -0.11      0.08    -0.27     0.05 1.00     4438     4221
MR4           0.23      0.11     0.02     0.44 1.00     3427     4474
MR5           0.48      0.11     0.26     0.69 1.00     4295     4366
MR6           0.44      0.13     0.19     0.71 1.00     5058     4427
MR7          -0.05      0.09    -0.23     0.13 1.00     4714     4041
MR8          -0.85      0.09    -1.03    -0.67 1.00     4703     3806
MR9          -0.36      0.10    -0.56    -0.17 1.00     4194     4277
```

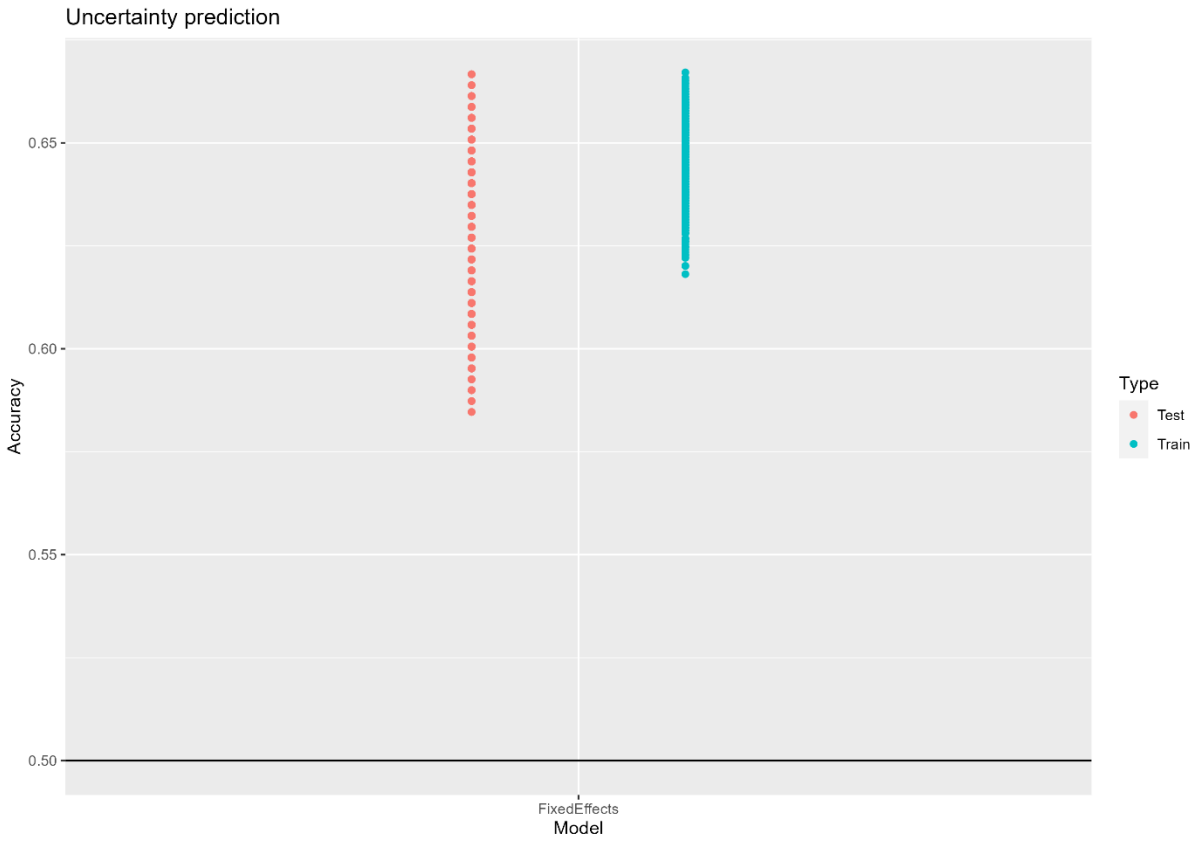
We then assessed the performance by calculating the average prediction.

Training & Test data

Truth	**Training**		**Test**	
Prediction	Schizophrenia	Control	Schizophrenia	Control
Schizophrenia	546	273	133	75
Control	245	447	65	105

Looking at the average prediction, it can be observed that the model definitely predicts the majority of the participants correctly, but there is still a lot of false positives and false negatives, which would be very problematic in a context like diagnosing an individual with a mental health condition.

Furthermore, we looked at the uncertainty and the accuracy of the model performance.



As one can see, the uncertainty is quite high and although the model is definitely performing above chance and it could even be considered good considering the sample size and the circumstances, there would still be a long way to go in order to have a model that could diagnose someone with schizophrenia.

Some suggestions for future improvements could be to consider using different parameters for assessing, whether someone has Schizophrenia based on voice-markers. Furthermore, an increase in sample size might increase the accuracy of the model. Another suggestion would be to conduct more feature selection/factor analysis/principal component analyses to maybe try to find other parameters, which might be important for the analysis. Lastly, one could increase the number of iterations of each model, but as we have very limited computational power, this was not possible.

Appendix:

Visualization of the machine learning pipeline