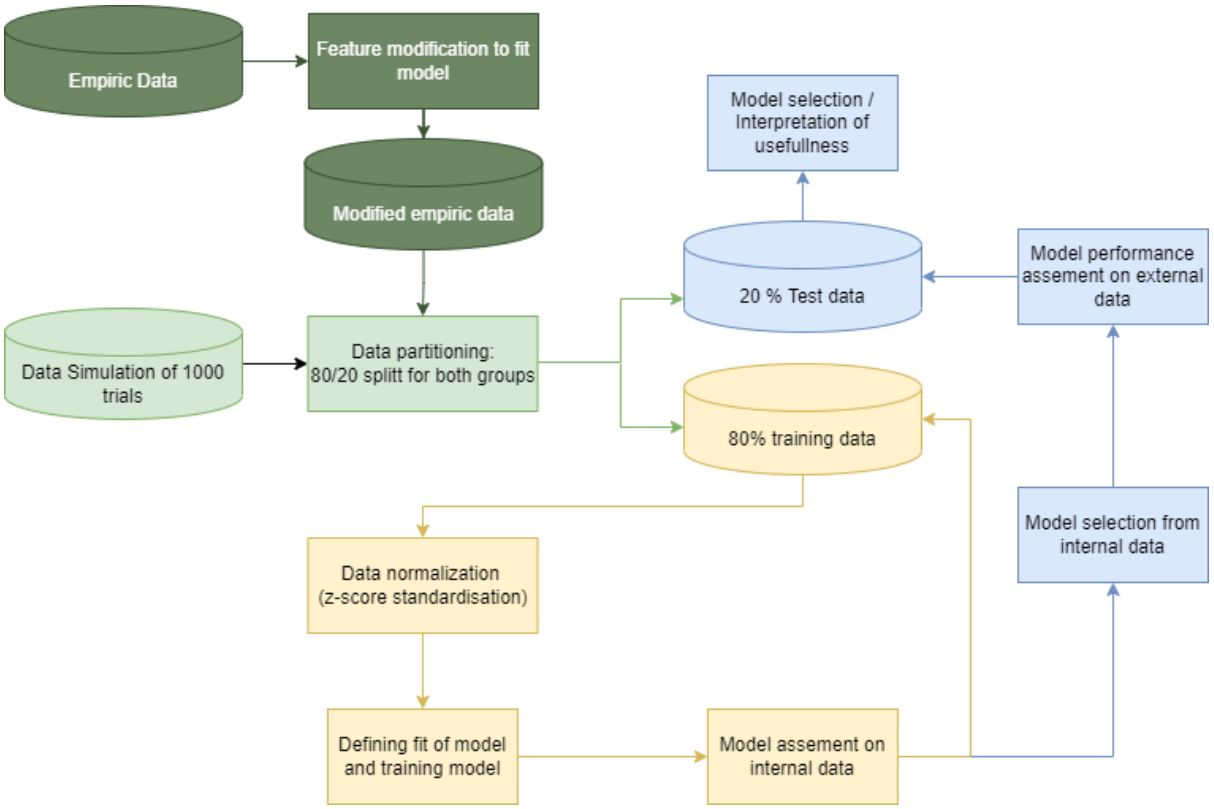
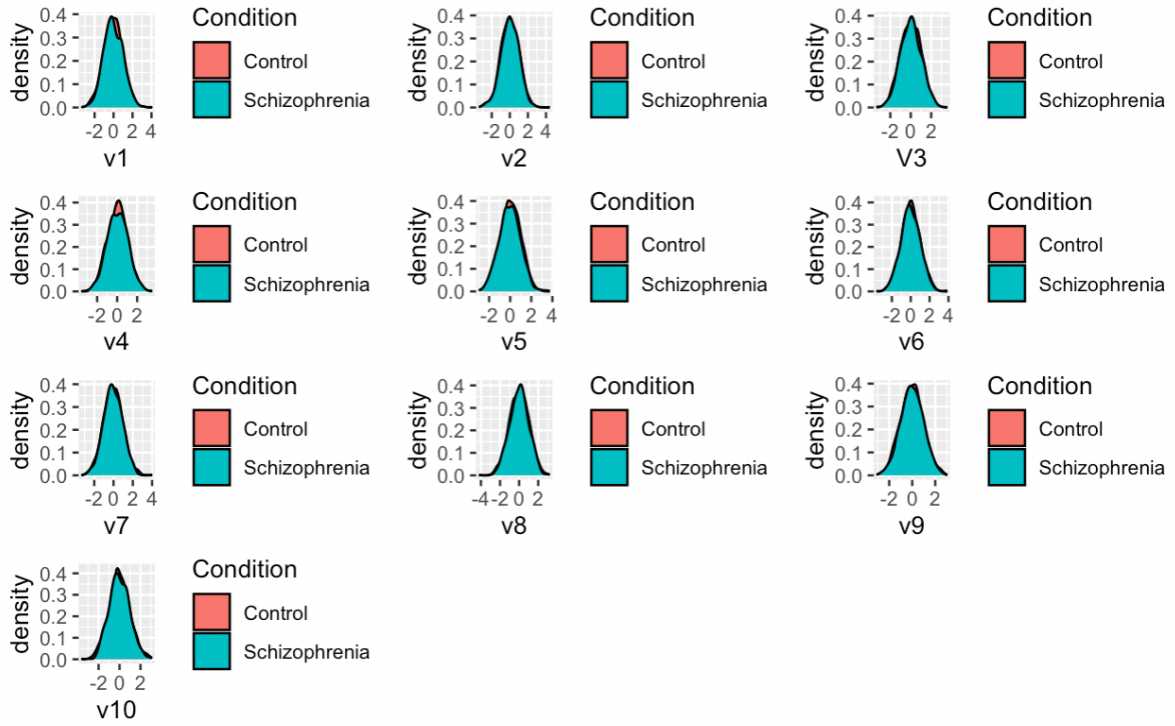
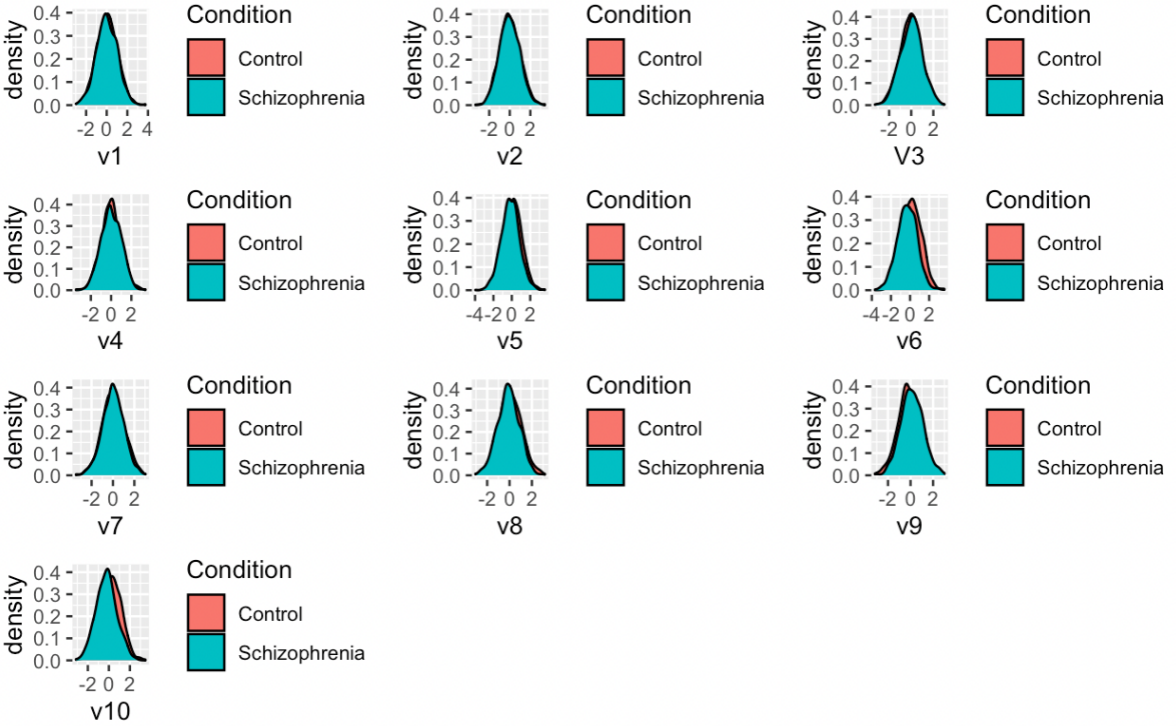
Pipeline for Assignment 3:



The pipeline shows two different starting points: Simulated data and Empirical data.
Both data-sets follow the same pipeline after modification.
The data is partitioned into 80% training data and 20 % test data.
We then normalize the training data, and save the transformation-scale.
Once the data is transformed a model is fitted onto the data and trained.
An assessment is run on the data's performance on the training data.
If the training datasets assessment is satisfying, the trained model is fitted onto the untouched normalized test data.
The model performance on the test data is then used to interpret usefulness and select the correct model.



This plot shows the 10 simulated variables for both control and schizophrenia for the sceptic data set. As expected there doesn't seem to be a difference between the two categories.

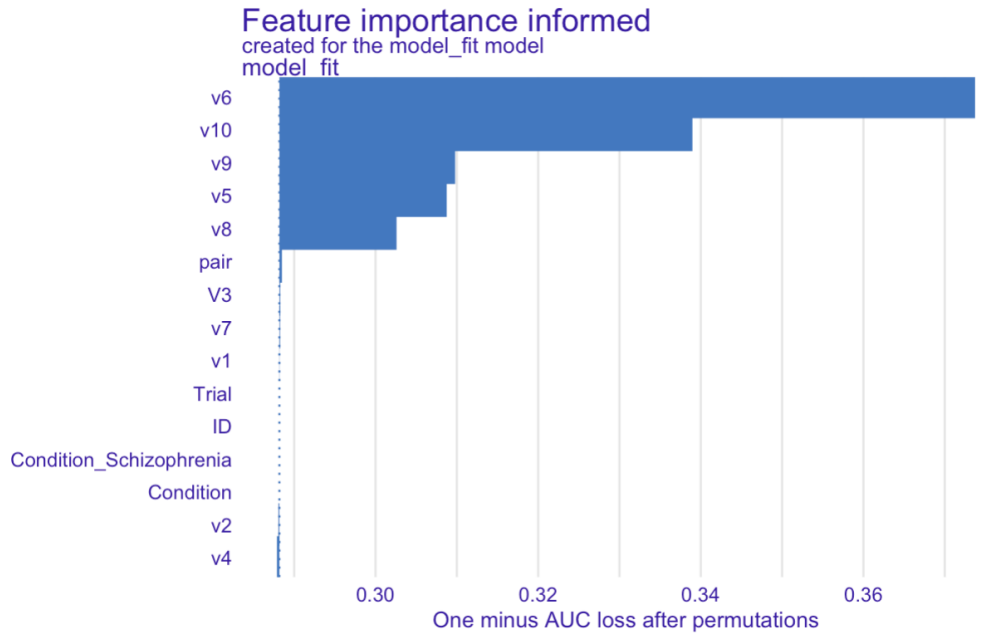


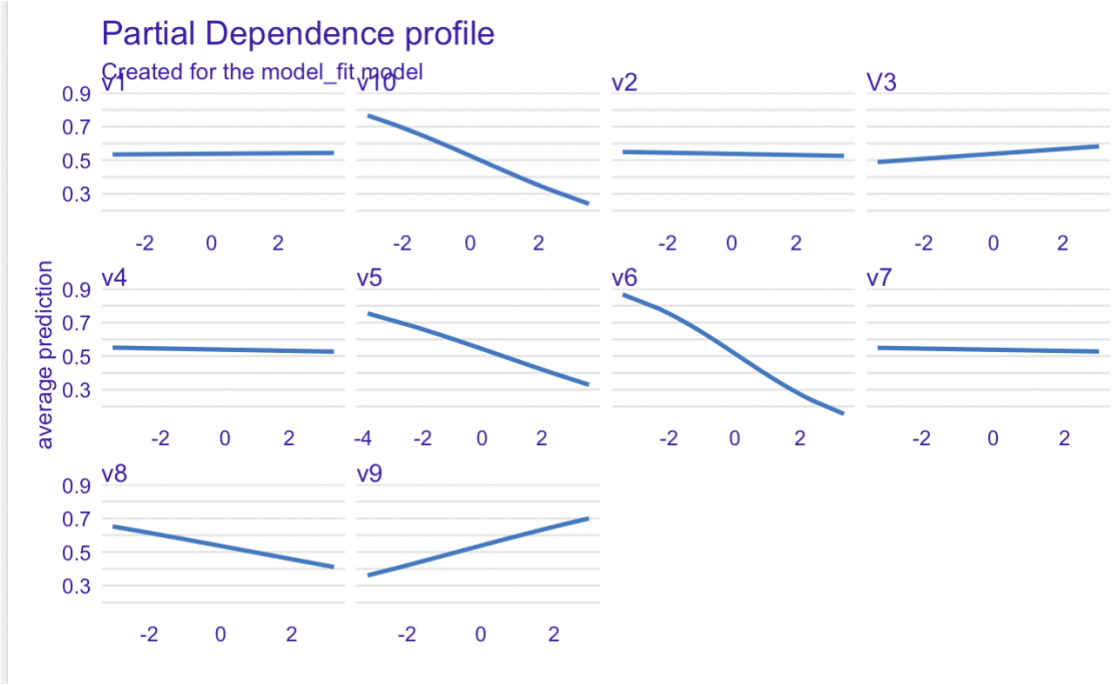
This plot shows the 10 simulated variables for both control and schizophrenia for the informed data set. As expected there is some difference for some of the variables, since we simulated a difference for 6/10 of the variables.

```
Confusion Matrix and Statistics

              Reference
Prediction      Control Schizophrenia
  Control           123            60
  Schizophrenia      77           140

               Accuracy : 0.6575
                 95% CI : (0.6087, 0.7039)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 1.475e-10

                  Kappa : 0.315

 Mcnemar's Test P-Value : 0.1716

            Sensitivity : 0.6150
            Specificity : 0.7000
         Pos Pred Value : 0.6721
         Neg Pred Value : 0.6452
             Prevalence : 0.5000
         Detection Rate : 0.3075
   Detection Prevalence : 0.4575
      Balanced Accuracy : 0.6575

       'Positive' Class : Control
```

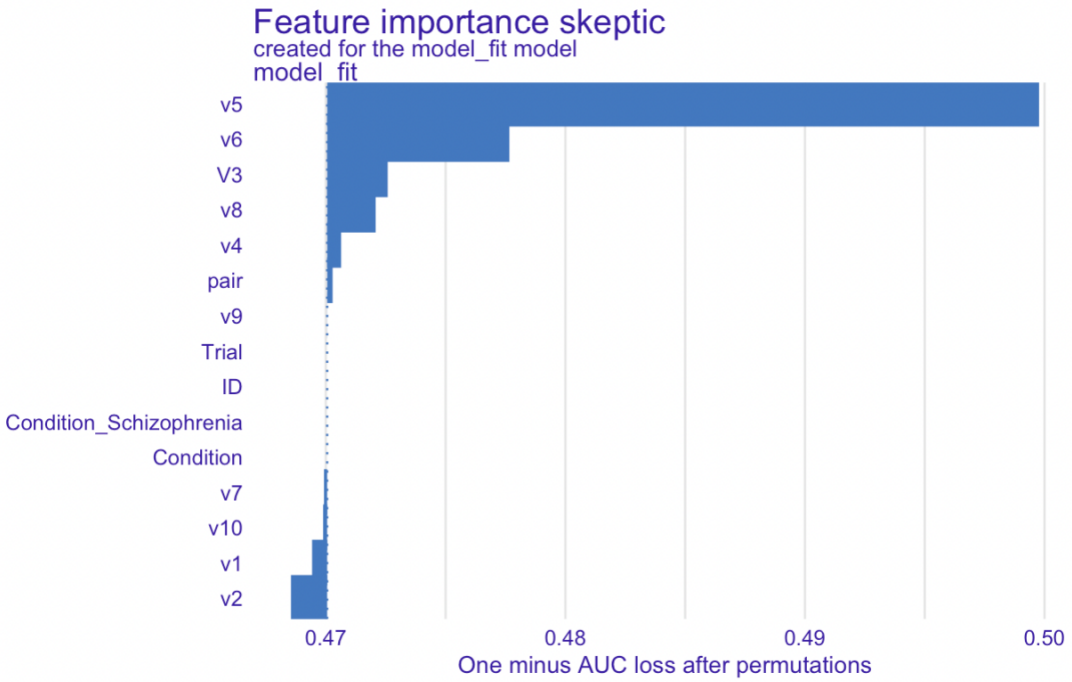
```
Confusion Matrix and Statistics

              Reference
Prediction      Control Schizophrenia
  Control           103           103
  Schizophrenia      97            97

               Accuracy : 0.5
                 95% CI : (0.4499, 0.5501)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.5199

                  Kappa : 0

 Mcnemar's Test P-Value : 0.7237

            Sensitivity : 0.5150
            Specificity : 0.4850
         Pos Pred Value : 0.5000
         Neg Pred Value : 0.5000
             Prevalence : 0.5000
         Detection Rate : 0.2575
   Detection Prevalence : 0.5150
      Balanced Accuracy : 0.5000

       'Positive' Class : Control
```

We split the data set into a training and test set to calculate prediction accuracy based on the 10 predictors. The accuracy for the informed model was 65.6 % and the accuracy for the skeptic model was 50 % (chance level). This makes sense since there is no information in the skeptic data set that predicts the condition.





Feature importance plot. As expected there seems to be 6 variables that predicts Schizophrenia for the informed data set.



For the skeptic data set we would expect that none of the variables predicted Schizophrenia or that they all predicted Schizophrenia equally. However the variation in how much each variable predicts varies quite a lot.

The empirical data set contained 398 variables. Therefore, in order to avoid multicollinearity in the model, PCA was performed on the data set. Using the same model pipeline as in the simulation task, different numbers of PCAs were tested. Firstly, models containing 3, 6 and 9 PCAs were evaluated using cross-validation. The models yielded 52%, 54% and 56% accuracy respectively. Another approach was tested in which the number of PCAs were set to capture 90% of the variance within the empirical data, which came out to be 55 PCAs. The accuracy using these as predictors yielded an accuracy of 58%. However, we deemed that the tradeoff between interpretability and the small increase in accuracy was not worth it. Therefore, we ended up using 9 PCAs in the final model which yielded 56% accuracy on the unseen data. The chosen PCAs are shown in the plot above.