EEG readings:
You sure it was only a glass?

Claudio Pella

INTRODUCTION

GOAL

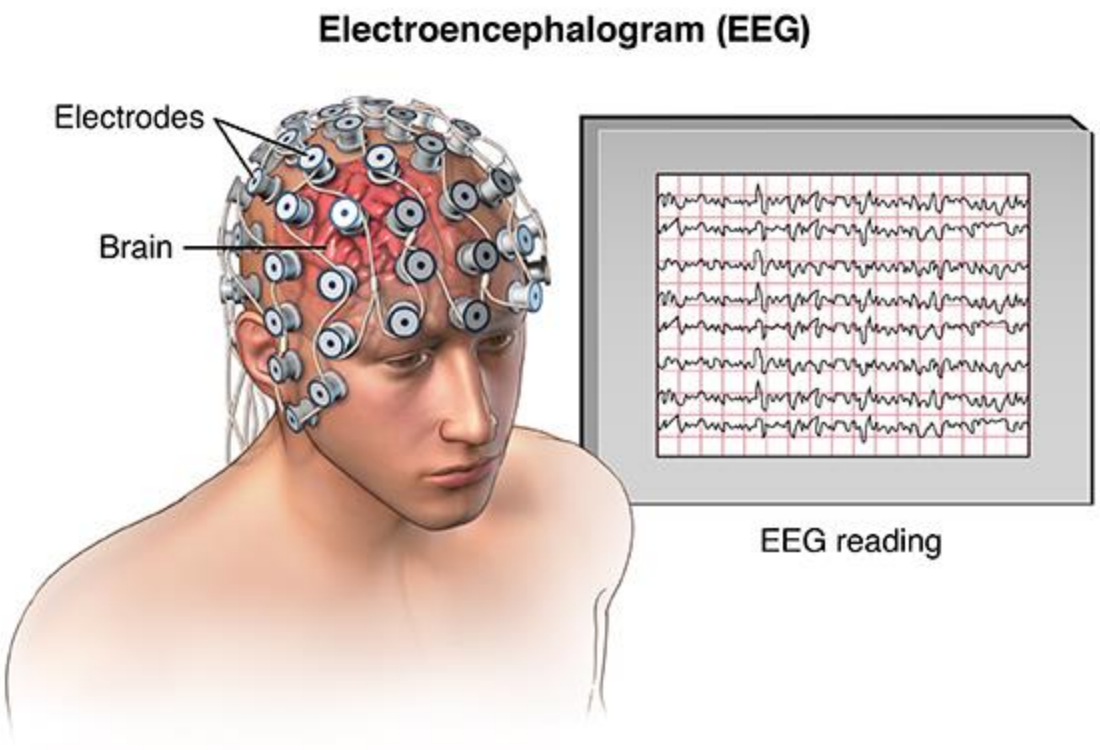
EXPLORATORY ANALYSIS

MODEL DEFINITION

RESULTS AND APPLICATIONS

INTRODUCTION – Dataset

https://archive.ics.uci.edu/ml/datasets/eeg+database



The dataset contains EEG readings of 16 people equally split between alcoholics and control.

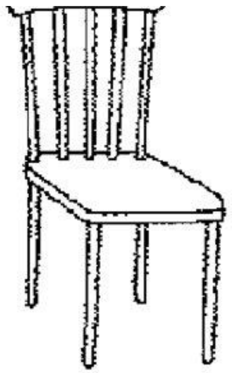
They were exposed to visual stimuli.

Overall, the dataset is **perfectly balanced**, containing the extact number of observations per each sub-group.

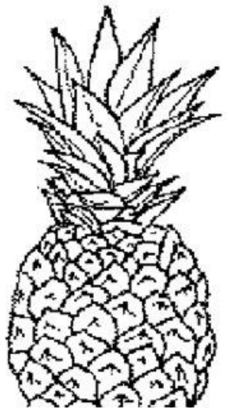
INTRODUCTION – Dataset

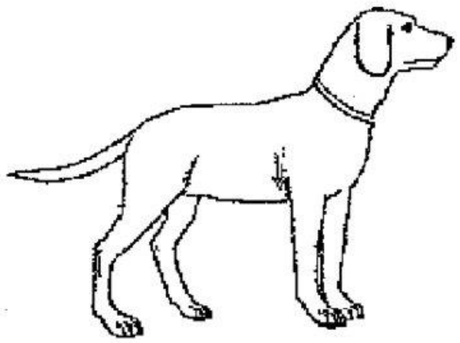
https://archive.ics.uci.edu/ml/datasets/eeg+database

Each subject was exposed to either a single stimulus (S1) or to two stimuli (S1 and S2) which were pictures of objects chosen from the 1980 Snodgrass and Vanderwart picture set.



When two stimuli were shown, either S1 was identical to S2 or different.





- INTRODUCTION
- GOAL
- EXPLORATORY ANALYSIS
- MODEL DEFINITION
- RESULTS AND APPLICATIONS

Goal

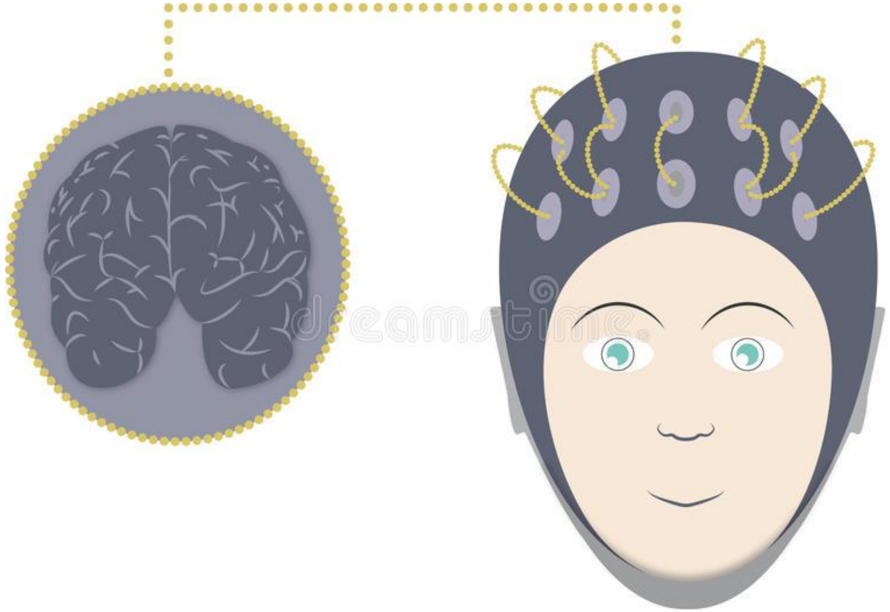
This project has a research goal:

We all know brain acts differently under the effect of alcohol, but can an EEG reading tell it?

Is there any practical use of this information?

Goal – Practical use

Imagine Government wants to introduce a new way to test if a driver is drunk or not.



The experiment is easy to replicate: people were shown one or two images for a few seconds. We could use an helmet containing electrods capable of reading an EEG.

INTRODUCTION

GOAL

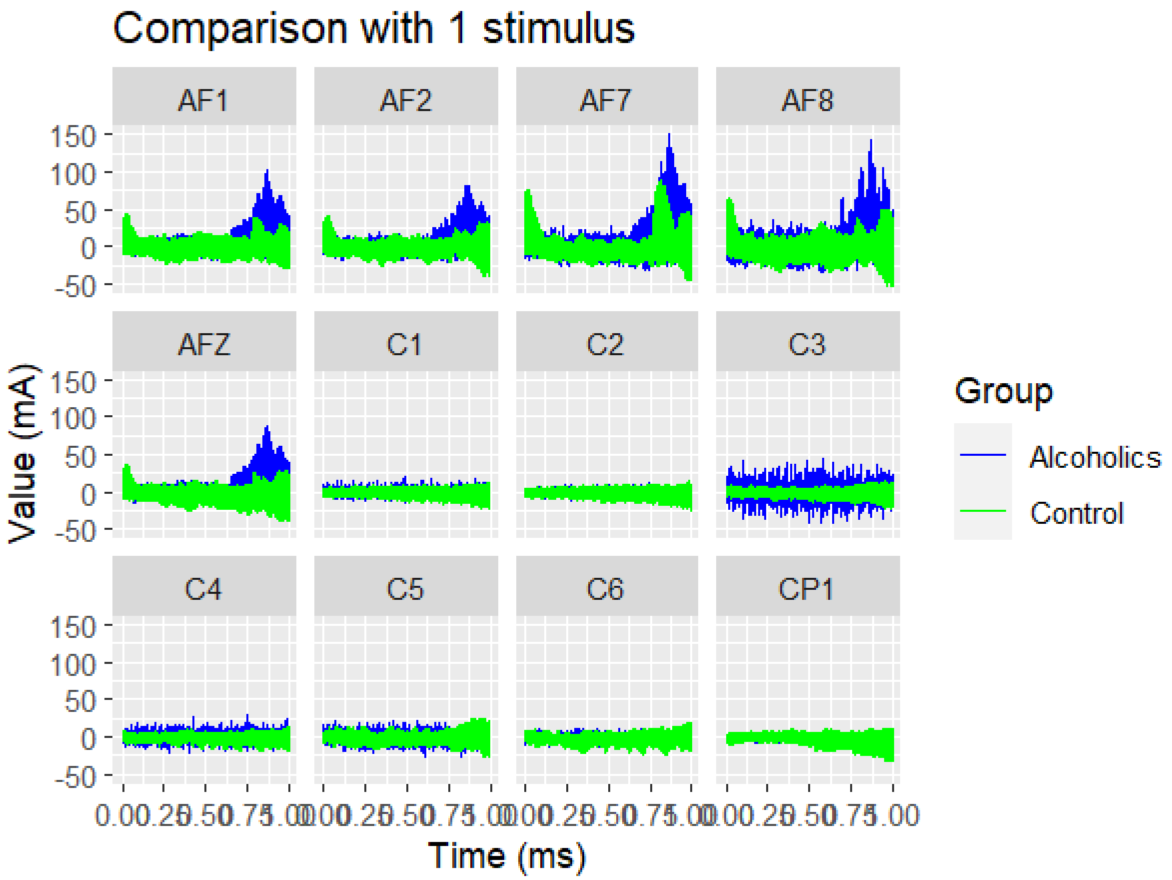
EXPLORATORY ANALYSIS

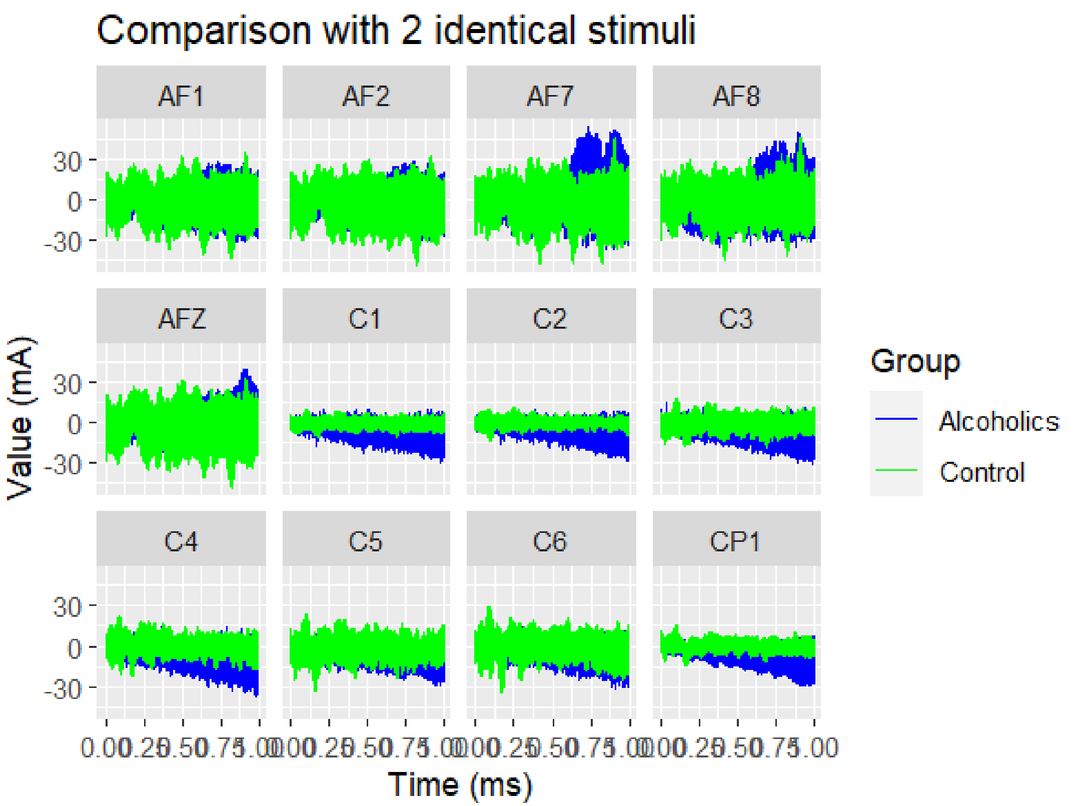
MODEL DEFINITION

RESULTS AND APPLICATIONS

EXPLORATORY ANALYSIS

First, let's see if a sample of electrodes has any visual information. Here we have the comparison between Alcoholics and Control with 1 and 2 stimuli: the two groups (a and c) have different behaviours.





PRE-PROCESSING

The original dataset contains 9 variables and 7 million rows. This is hardly usable. It is the sum of hundreds of .csv files (there was a .csv per electrode, per trial)

```
> str(brain)
'data.frame':	7308288 obs. of  9 variables:
 $ trial.number      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ sensor.position   : Factor w/ 64 levels "AF1","AF2","AF7",..: 35 35 35 35 35 35 35 35 35 35 ...
 $ sample.num        : int  0 1 2 3 4 5 6 7 8 9 ...
 $ sensor.value      : num  -8.92 -8.43 -2.57 5.24 11.59 ...
 $ subject.identifier: Factor w/ 2 levels "a","c": 1 1 1 1 1 1 1 1 1 1 ...
 $ matching.condition: Factor w/ 3 levels "S1 obj","S2 match",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ channel           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ name              : Factor w/ 16 levels "co2a0000364",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ time              : num  0 0.00391 0.00781 0.01172 0.01562 ...
> 
```

PRE-PROCESSING

To perform PCA I need to change the form of the dataset. Using the **reshape** library, I created one variable per each electrode.

```
> str(new.df)
'data.frame':   119808 obs. of  66 variables:
 $ trial.number      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ sample.num        : int  0 1 2 3 4 5 6 7 8 9 ...
 $ time              : num  0 0.00391 0.00781 0.01172 0.01562 ...
 $ matching.condition: Factor w/ 3 levels "S1 obj","S2 match",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ subject.identifier: Factor w/ 2 levels "a","c": 1 1 1 1 1 1 1 1 1 1 ...
 $ AF1               : num  -2.146 -2.146 -1.658 -0.682 2.248 ...
 $ AF2               : num  1.129 0.641 -0.336 -0.824 0.641 ...
 $ AF7               : num  -16.86 -7.09 7.56 19.28 23.18 ...
 $ AF8               : num  -10.02 -7.09 1.21 10.49 13.91 ...
 $ AFZ               : num  -0.987 -1.475 -0.987 -0.01 2.431 ...
```

INTRODUCTION

GOAL

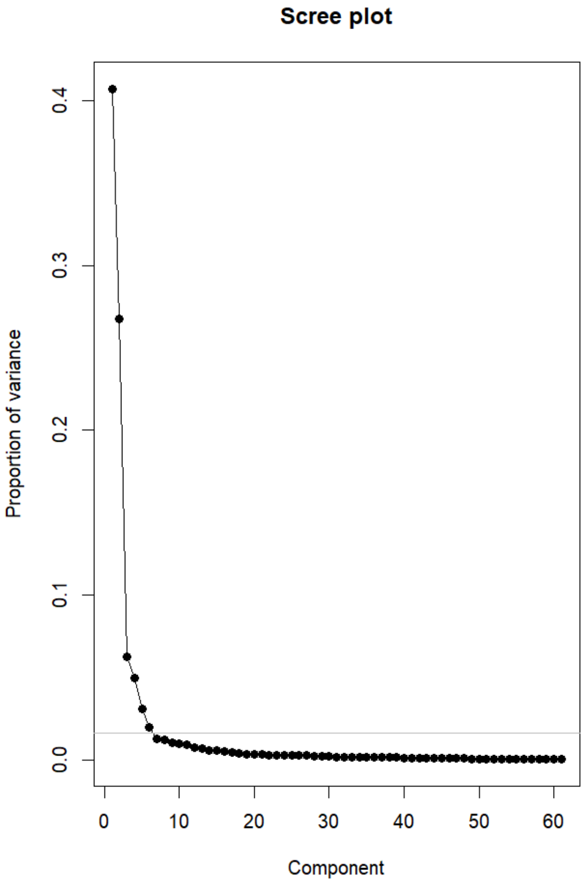
EXPLORATORY ANALYSIS

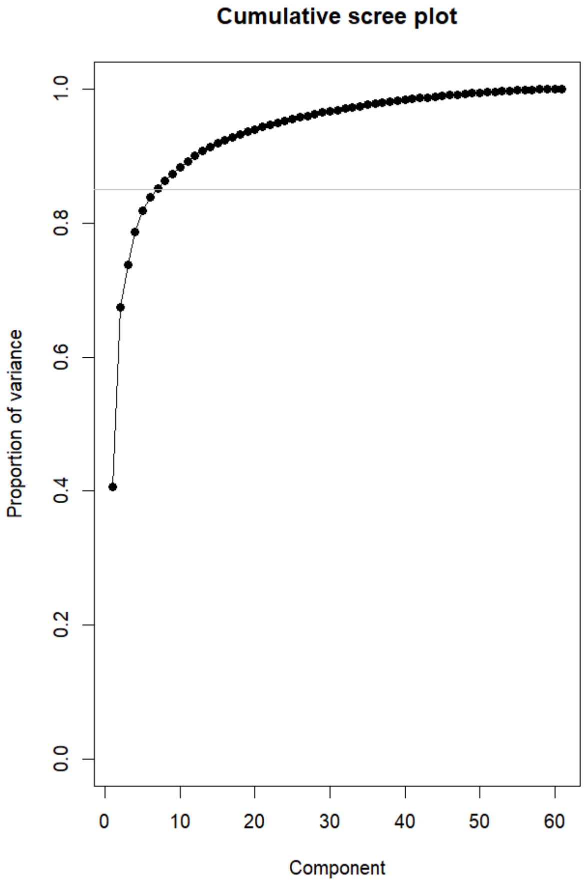
MODEL DEFINITION

RESULTS AND APPLICATIONS

PCA

Even though the goal is to reach maximum accuracy, the size of the dataset forces to use dimensionality reduction. I performed PCA:

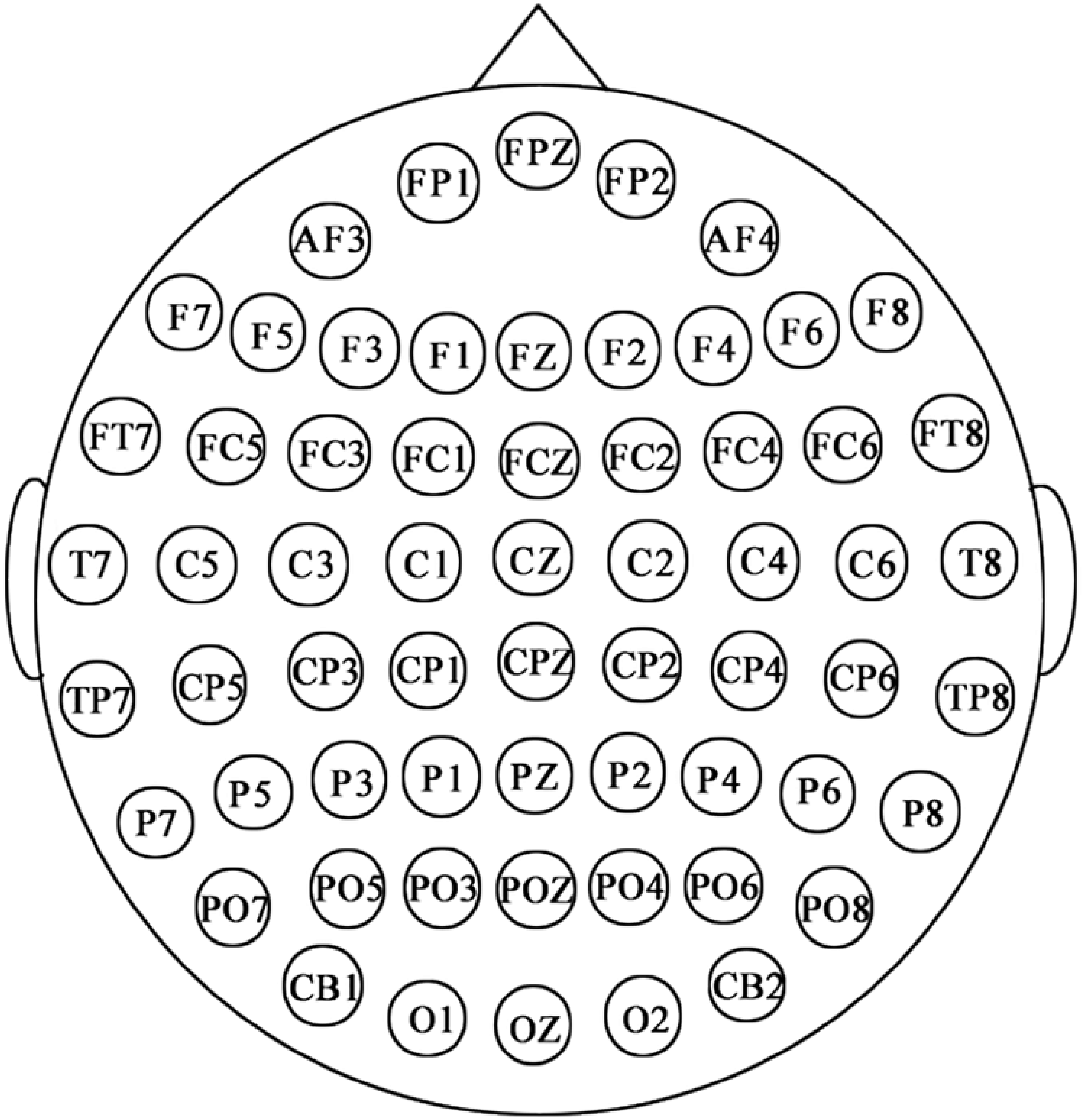




The results are good: I only need 6 principal components to explain 85% of the variability of the data, instead of 62.

The elbow rule suggests to stop at 6 or 7 PCAs.

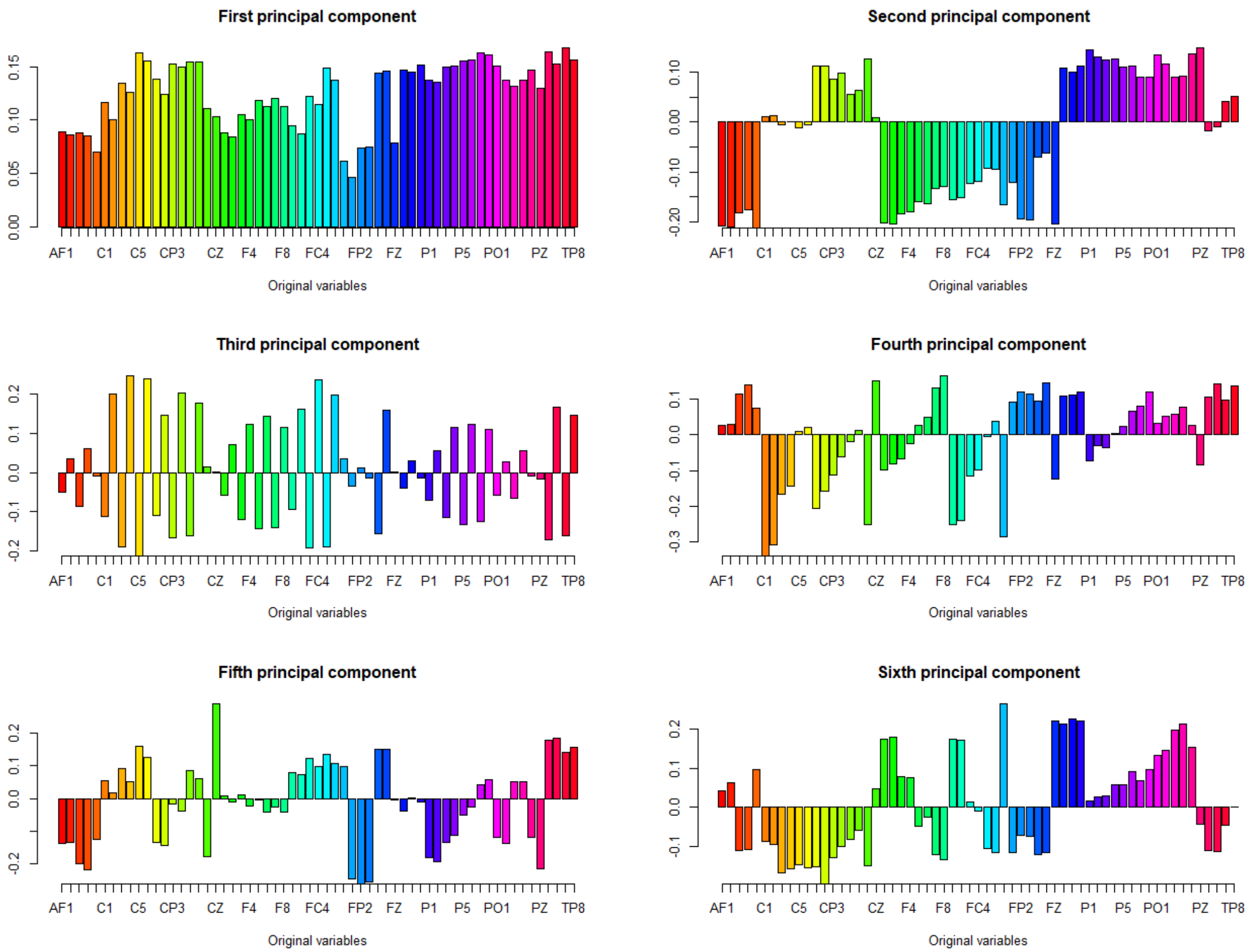
Variables names



Electrodes with the same initial are from the same group.

The triangle represents the nose, the two curves the ears.

PCA



They are difficult to interpret.

First principal component looks like a sort of weighted average of the values.

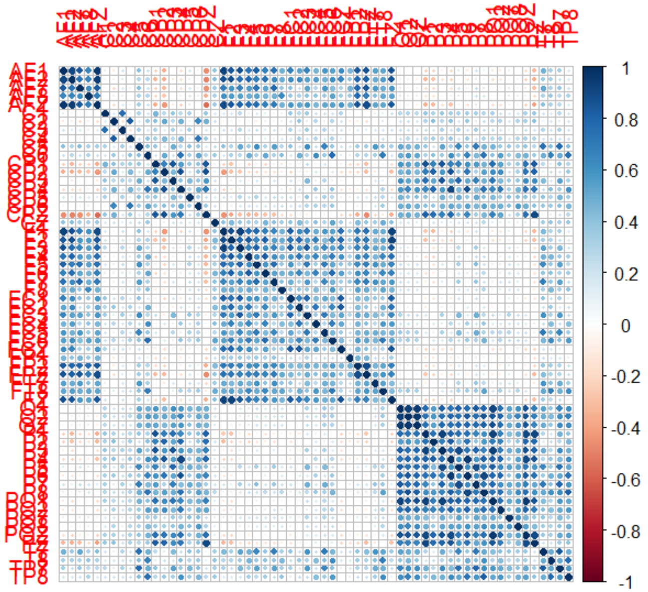
The second represents the contribution of the areas of the brain.

Correlation – 1 stimulus

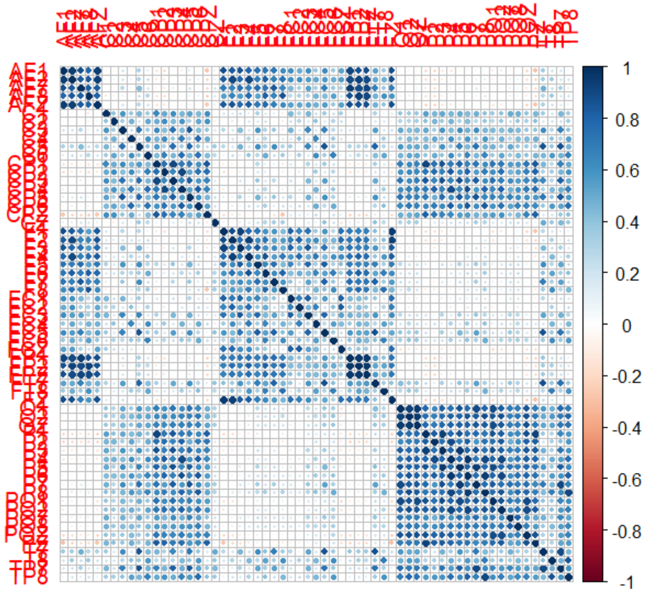
But what about correlation?

We can see there are differences:

- Close electrodes have a strong correlation
- For control, there are only positive correlation
- For Alcoholics, the C area of the brain has weaker correlations.
- Alcoholics has a few negatively correlated electrodes



Alcoholics

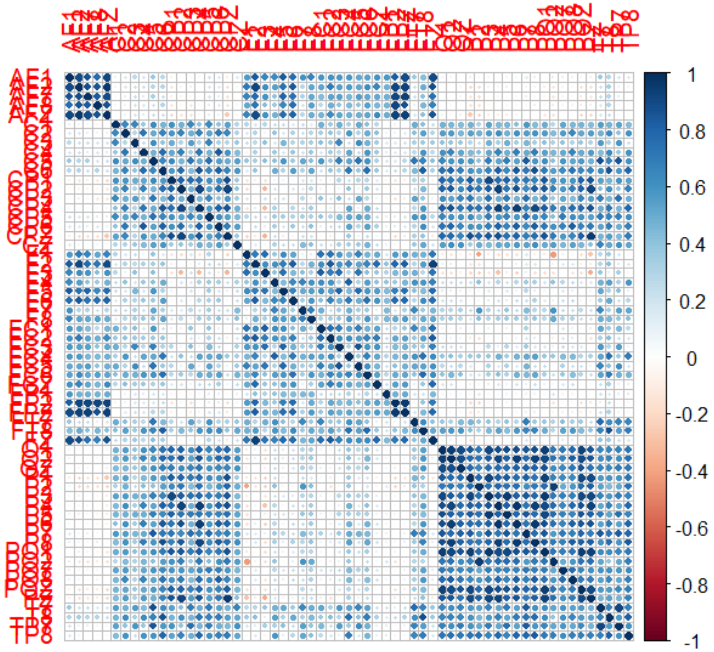


Control

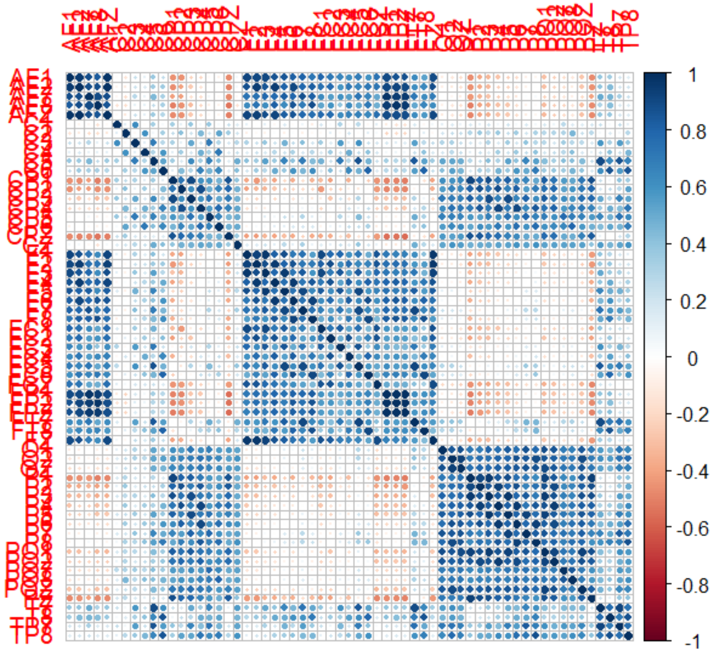
Correlation – 2 matching stimuli

But what about correlation?

The two groups have opposite reactions! When presented to two identical images, control group tend to have negative correlation between areas and weaker correlation within areas.
Viceversa for alcoholics.



Alcoholics

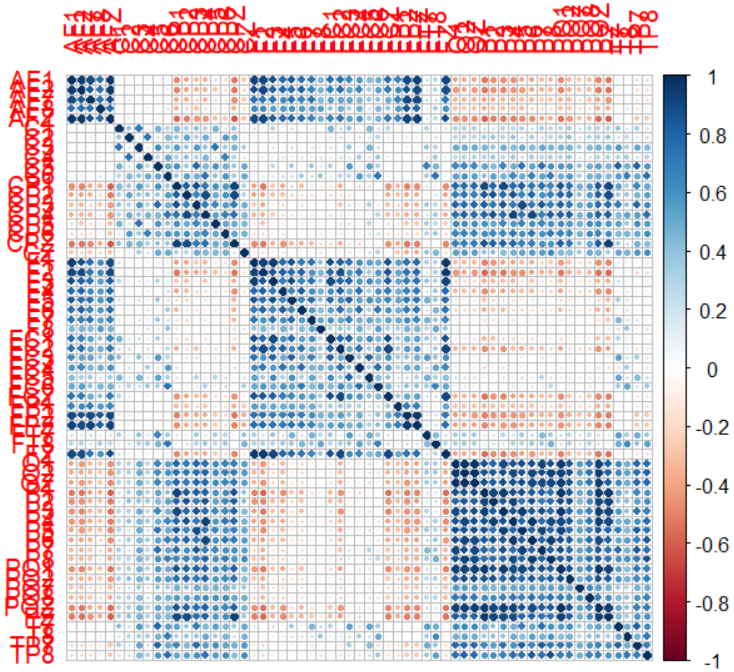


Control

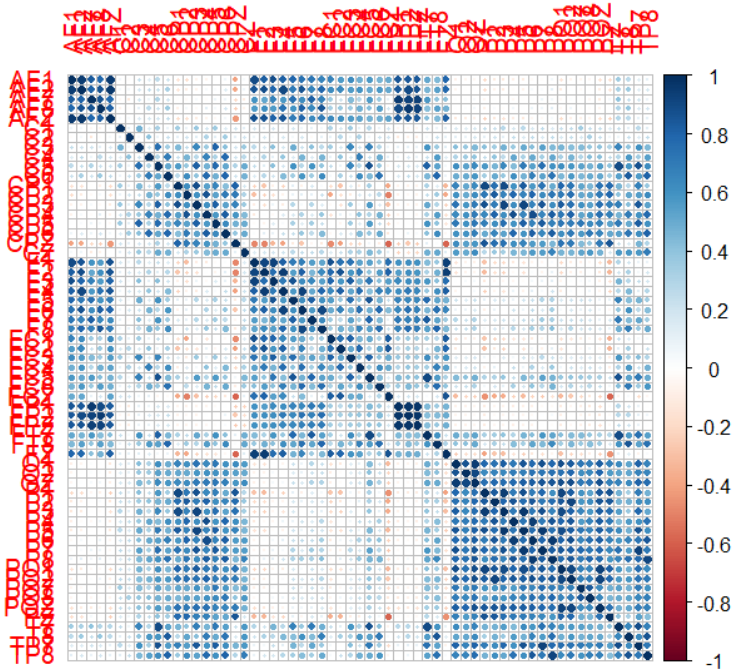
Correlation – 2 non matching stimuli

But what about correlation?

When the image changes, alcoholics reacts with negative correlation between areas



Alcoholics



Control

Classification

Now that we have reasons to suspect that EEG can distinguish between drunk and sober, we should try to implement a model able to classify our «drinking» status based on parameters.

PCA or Full Dataset?

This is a fake question. Even though Full would be preferred because of greater accuracy, my computer is not able to manage 7 **millions** rows.

Therefore, PCA will be preferred.

Classification - Logistic

Logistic regression is the only case in which I was able to run a full dataset model. We can make a comparison with a PC model:

```
Confusion Matrix and Statistics

          Reference
Prediction     a     c
         a 12972  7254
         c  7151 12559

               Accuracy : 0.6393
                 95% CI : (0.6346, 0.644)
    No Information Rate : 0.5039
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.2785

 Mcnemar's Test P-Value : 0.3954

            Sensitivity : 0.6446
            Specificity : 0.6339
```

Full

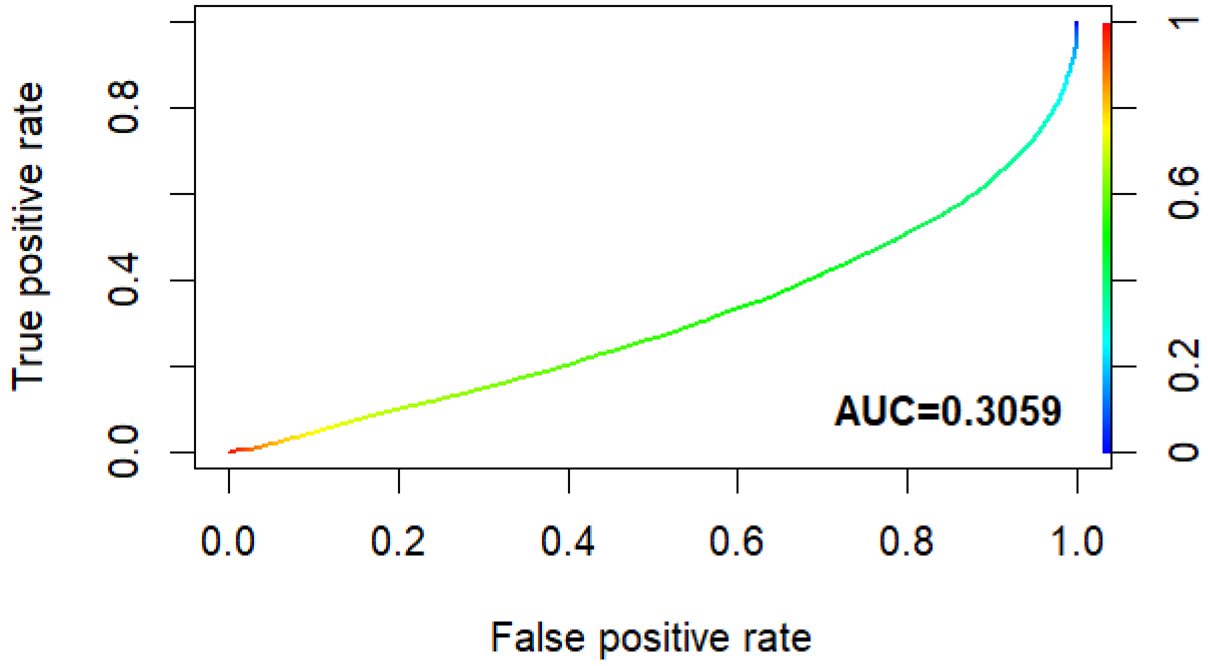
```
Confusion Matrix and Statistics

          Reference
Prediction     a     c
         a 10704  9322
         c  9419 10491

               Accuracy : 0.5307
                 95% CI : (0.5258, 0.5356)
    No Information Rate : 0.5039
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.0614

 Mcnemar's Test P-Value : 0.4831

            Sensitivity : 0.5319
            Specificity : 0.5295
```

PCA

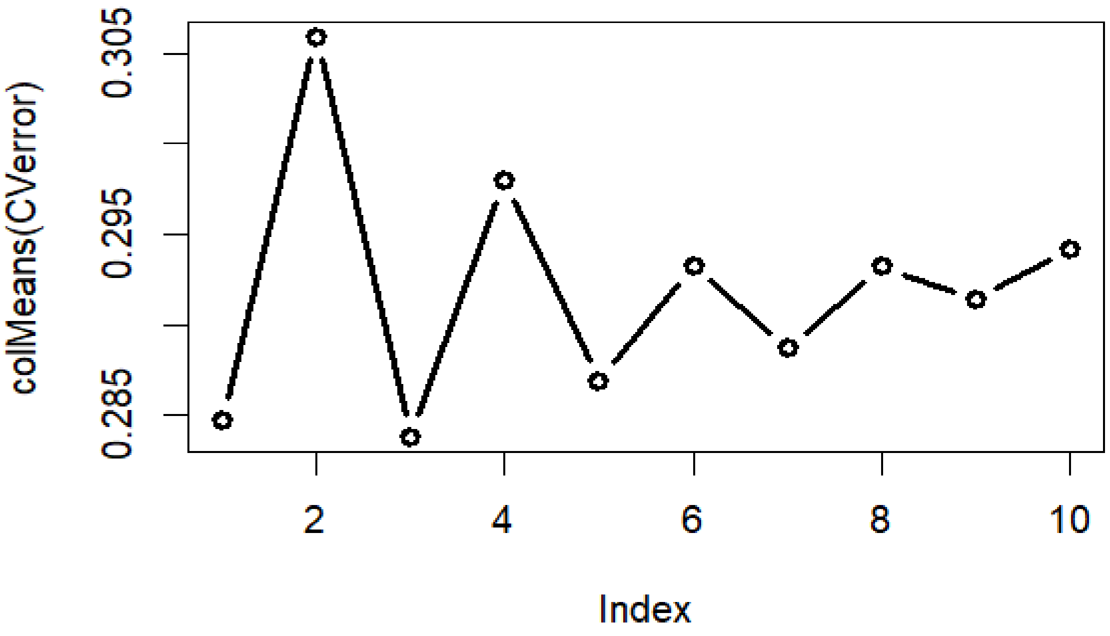
Classification - Logistic

As expected, Full dataset performs better in term of accuracy.
None of the two models performed amazingly: only 63% and 53% accuracy.

It is worth noticing that a perfectly random model (one that assigns randomly a class given the input) would have 50% accuracy, since we only have two classes. Therefore, PCA performed very poorly in this case.



KNN




```
Confusion Matrix and Statistics

          Reference
Prediction     a     c
         a 15246  5803
         c  4877 14010

               Accuracy : 0.7326
                 95% CI : (0.7282, 0.7369)
    No Information Rate : 0.5039
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4649

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7576
            Specificity : 0.7071
```

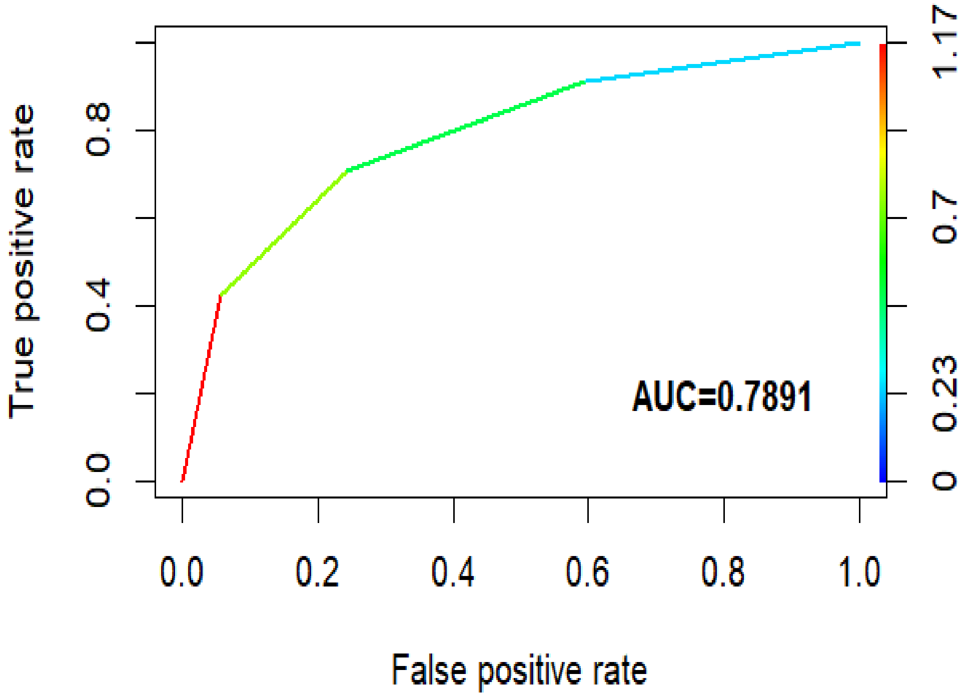
Let’s try with KNN: the search suggest to use k=3.
This time the computational cost was too high for the Full Dataset, therefore I only performed PC.

KNN

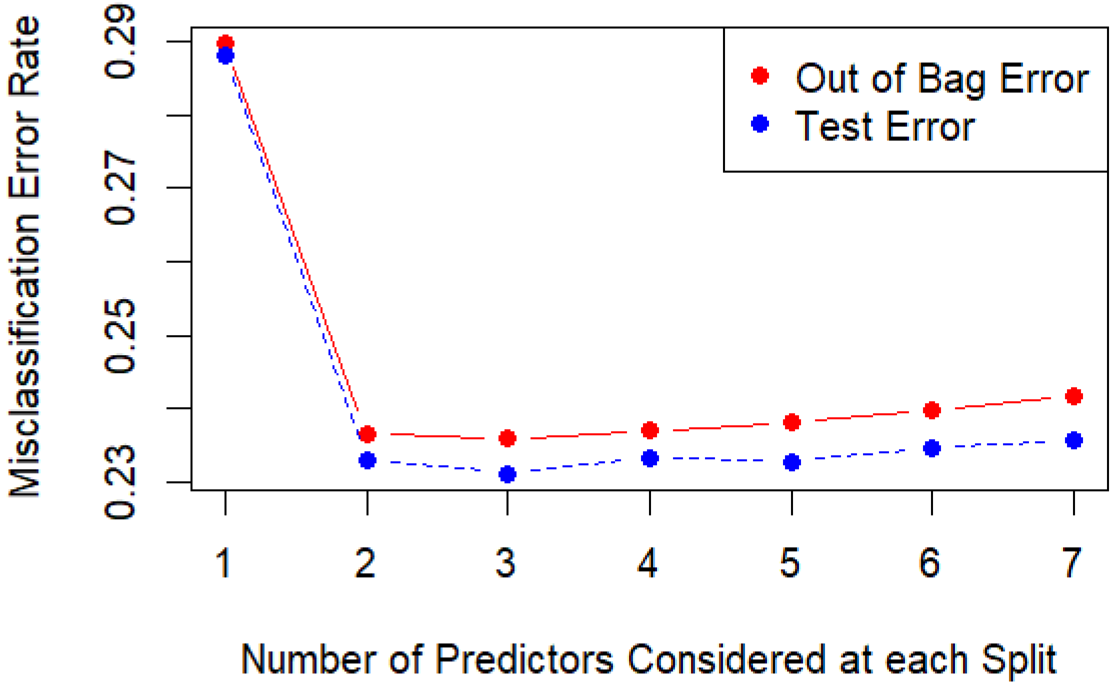
The performance is much better than the logistic regression.

Accuracy, sensitivity and specificity are about 0,7-0,75, which is a 20% improvement.

Even the AUC score from the ROC curve is good.



Random Forest



At last, I tried the classification using a Random Forest: this method is expensive and PCA was used.

The error is stable, we can use Bagging with mtry=3.

Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction     a     c
         a 16103  5399
         c  4020 14414

               Accuracy : 0.7641
                 95% CI : (0.76, 0.7683)
    No Information Rate : 0.5039
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.528

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8002
            Specificity : 0.7275
```

The performance is not bad, but it's similar to KNN, with sensitivity much higher than specificity.

INTRODUCTION

GOAL

EXPLORATORY ANALYSIS

MODEL DEFINITION

RESULTS AND APPLICATIONS

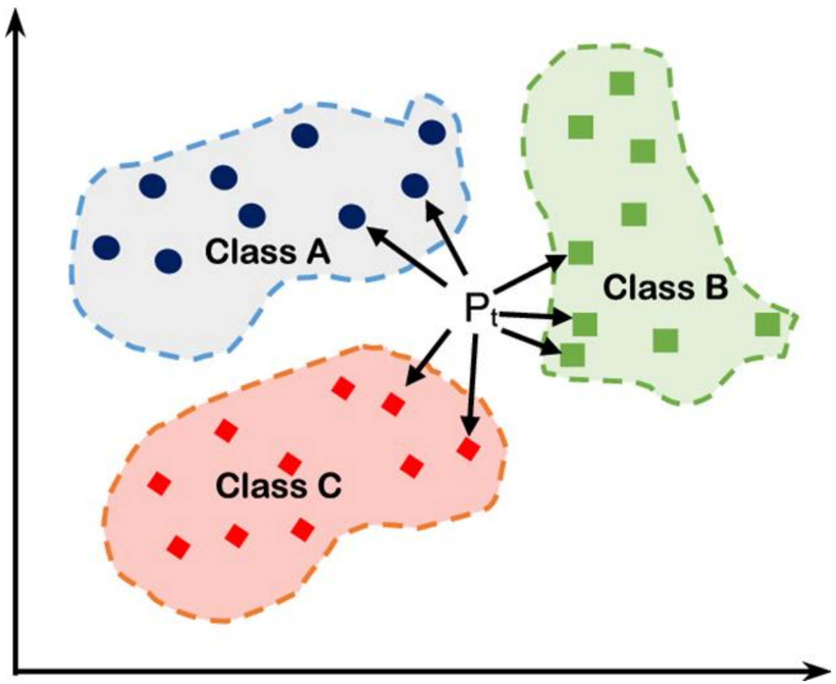
PRELIMINARY RESULTS

It is time to summarize the results:

- when presented with the same images, control and alcoholics brains react in almost opposite ways.
- There's a correlation of electrodes **within** the same area.
- In certain situations, there's a correlation **between** areas

Accuracy wise, KNN and RF are very similar. They have the similar accuracy (which in this case is a useful parameters, since the dataset is perfectly balanced).

Overall, KNN is to be preferred because of the computational time.



IMPROVEMENTS

There are a few things that could be changed in order to improve the results:

To train the model, I used each line individually: the fact that the signal of a single electrode is a time serie wasn’t taken into consideration. If I were to use the functions over time instead of single points, accuracy might improve.

IMPROVEMENTS

Most lines are referred to the same subject:

Having only 16 subjects and thousands of line, many of them refer to the same person: we have around 7500 rows per subject.

The models try to predict the class for each combination of measures, at a fixed time.

The results could be grouped differently, in fact it is more important the accuracy on the individual, not the general one.