### Analysing AwE dataset, Cirulli lab, in contact with Alessandra Berry

In order to analyse the recent test held for the AwE mice,

* Combining the old mice (24, 25 months) with the young mice tests (3 months).
* The main idea here is to focus on the Controlled diet treatment (C), in order to figure out the female and male behaviour.

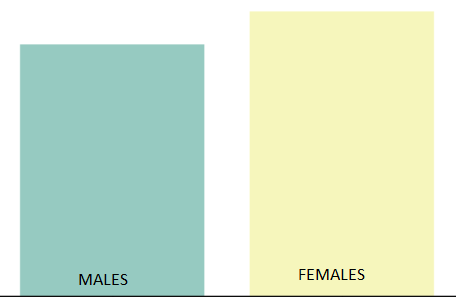
As a baseline Alessandra suggested to find the analysis for the main differences between the physical and mental tests, the second step is to find the power difference between the males and females, as she mentioned Berry et al. (2019), that even for some of the female mice which are suffering from a bad body shape they could pass the big pole test, even better than well-shaped males.

According to Berry et al. (2019) which shows the tests and treatments held for mice in the intellicages as statistical analysis, in my turn now, I need to set the best algorithm for analyzing those measurements.

My goal is to learn the machine to get the best out of those reports.

**Gender as a target**

The first assumption will be is by building a model that test the differences between males and females according to their **mental** and **physical abilities**. We have **128 male** and **145 female**:



I started the test, first by making the mice gender as a target, I did 5 different experiments by using different Features-Engineering algorithms:

**First:** Taking the features (39) feature for the test measurements and one for the treatment (C, Trehalose, Ampel, RA\_D1, RA\_D2, RA\_RA\_D1, RA\_RA\_D2).

#### Features as they are

As feature\_distribution plot function used here to show the significant relationships, as I’ve chosen the highest rank (most important features), where *Random Forest algorithm* used to show the most important features which affect the prediction model, defining the Gender as the class label:

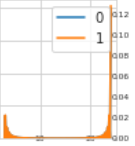
The strongest relationship between:

1. MWM (target time), and the Insulin Sensitivity test. (memory)
2. MWM (LatencyDay2) and the Insulin Sensitivity test.\\ different major (learning),
3. //Elevated plus maze (Sniffing\_D) and elevated plus maze (wall rearing\_F). (exploration)
4. //BASAL grip strength score (s) and the Treatment.
5. POST-BIGp\_TTT and the Bodyweight.
6. Rotarod score 3 (s) and the Glucose tolerance test (AUC).

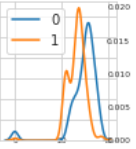
I’d like to show you the relationship between each attribute and its target (male-female):

Treatments (refer above) male:0 Female:1

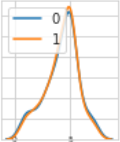
Age



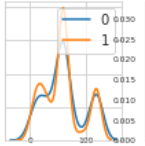
Bodyweight



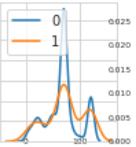
Treatment



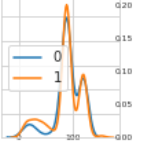
BASAL grip strength score (s)



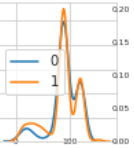
POST grip strength score (s)



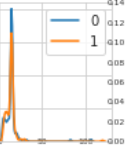
BASAL BIGp\_TTT



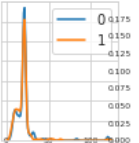
POST BIGp\_TTT



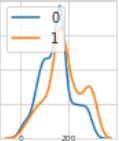
BASAL SMALLp\_TTT



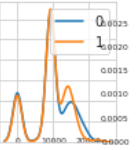
POST SMALLp\_TTT



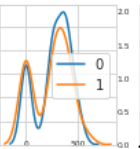
Rotarod score 3 (s)



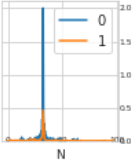
Glucose tolerance test (AUC)



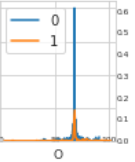
Insulin sensitivity test (AUC)



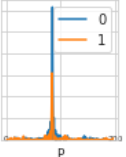
elevated plus maze (open) done



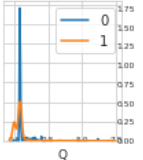
elevated plus maze (closed)



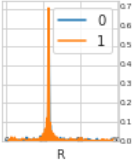
elevated plus maze (Sniffing\_D)



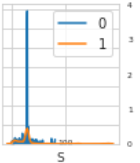
elevated plus maze (Rearing\_D)



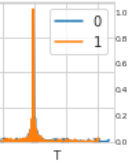
elevated plus maze (Wall rearing\_D)



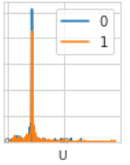
elevated plus maze (grooming\_D)



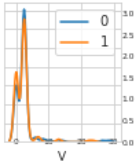
elevated plus maze (SAPD\_D)



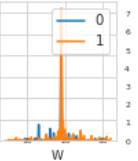
elevated plus maze (head dipping\_D)



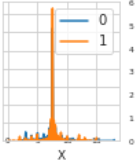
elevated plus maze (immobility\_D)



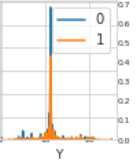
elevated plus maze (centre\_F)



elevated plus maze (open\_F)



elevated plus maze (close\_F)



elevated plus maze (crossing\_F)

elevated plus maze (sniffing\_F)

elevated plus maze (rearing\_F)

elevated plus maze (wall rearing\_F)

elevated plus maze (grooming\_F)

elevated plus maze (SAP\_F)

Please refer to the [running code](https://colab.research.google.com/drive/13rWfJQWBugr1jPPsWzYCt_dwznoJnTe8#scrollTo=1KIfjyer7M0C) section **Plotting data**.

### Wrapper Method

**Second:** Applying (Wrapper) the first feature selection method, which is based on a specific machine learning algorithm that we are trying to fit on our dataset.

It follows a greedy search approach by evaluating all the possible combinations of features against the evaluation criterion.

Applying the wrapper method (Features engineering):

Wrapper method: I used the feature engineering method, to extract the measurement of the test to a lower dimension, so I can extract the best features overall, the figure below shows the only the forward wrapper, as you can see the best independent features overall are:

* **Best features according to the Step Forward Feature Selection, are:** 'C', 'N', 'P', 'Q', 'R', 'S', 'T', 'U', 'V', 'W', 'X', 'AA', 'AD', 'AH', 'AI', 'AJ', 'AK', 'AL', 'AM', 'AN'
* **Best features according to the Step Backwards Feature Selection, are:** ['C', 'E', 'F', 'G', 'H', 'J', 'K', 'L', 'M', 'R', 'T', 'U', 'V', 'W', 'AA', 'AD', 'AH', 'AJ', 'AL', 'AM']
* **Best features according to the Exhaustive Feature Selection, are:** ['E', 'F', 'G', 'H', 'J', 'K', 'L', 'M', 'N', 'P', 'Q', 'R', 'S', 'W', 'X', 'AH', 'AI', 'AL', 'AM', 'AN']

Merging the Forward+Backward+Exhaustive and taking only matching features we have:

'C', 'N', 'P', 'Q', '**R**', 'S', 'T', 'U', 'V', '**W**', 'X', 'AA', 'AD', 'AH', 'AI', 'AJ', 'AK', '**AL**', **'AM**', 'AN'

'C', 'E', 'F', 'G', 'H', 'J', 'K', 'L', 'M', '**R**', 'T', 'U', 'V', '**W**', 'AA', 'AD', 'AH', 'AJ', '**AL**', '**AM**'

'E', 'F', 'G', 'H', 'J', 'K', 'L', 'M', 'N', 'P', 'Q', **'R**', 'S', '**W**', 'X', 'AH', 'AI', **'AL**', '**AM**', 'AN'

**R, W, AL, AM** (I’ve added those features plus the common output featured for each two common) in total are ().

**The most important measurement overall are the 4** Plus the following (common between every two outputs as above)**:**

* Age
* BASAL grip strength
* POST grip strength score
* BASAL BIGp\_TTT
* Rotarod score 3
* Elevated plus maze (sniffing\_F)
* MWM(TARGET ZONE\_D WHOLE%)
* Elevated plus maze (Wall rearing\_D)
* Elevated plus maze (centre\_F)
* LATENCY DAY 1
* LATENCY DAY 2

#### Machine Learning Models

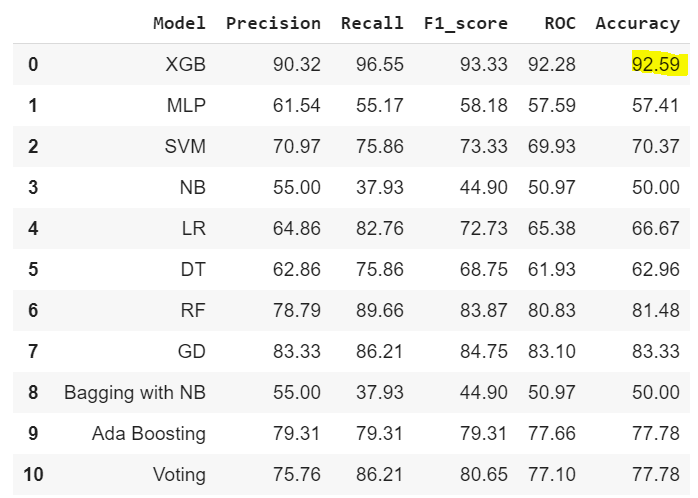
#### Gender as a label

Predicting the mice sex according to its body measurement (physical and mental).

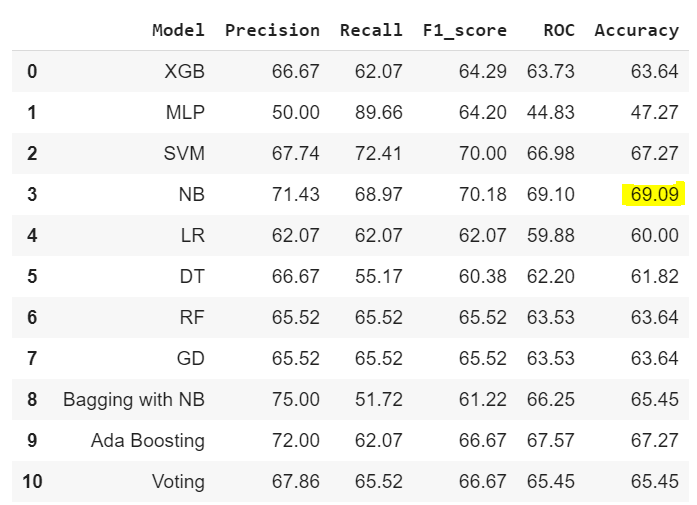
First: taking the values of the original features, as a new sheet called gender.csv, and apply the following models, as you can see below 10 different ML-based models plus the Voting.

Where **XGB** boot recorded the highest score.

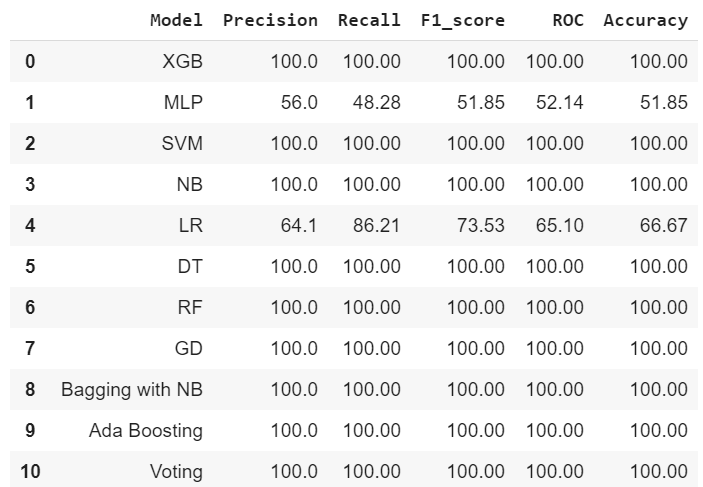
First place goes to the features as they are, the other features selection methods weakness back to a lot of missing data. DoNE 1



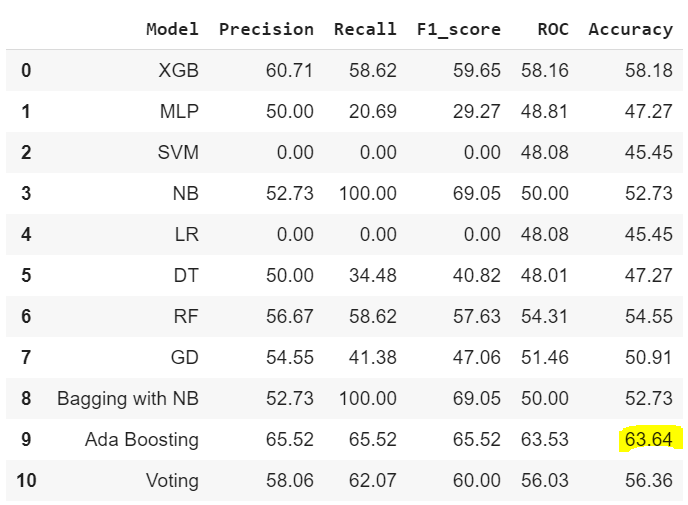
##### Wrapper Features engineering **DONE 6**



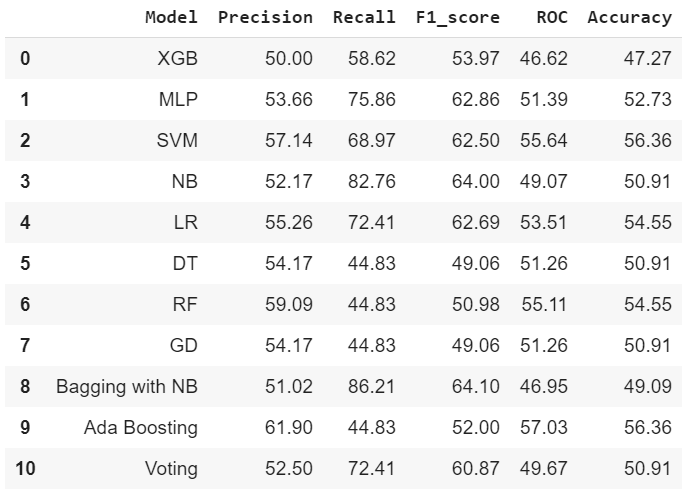
PCA Features Selection



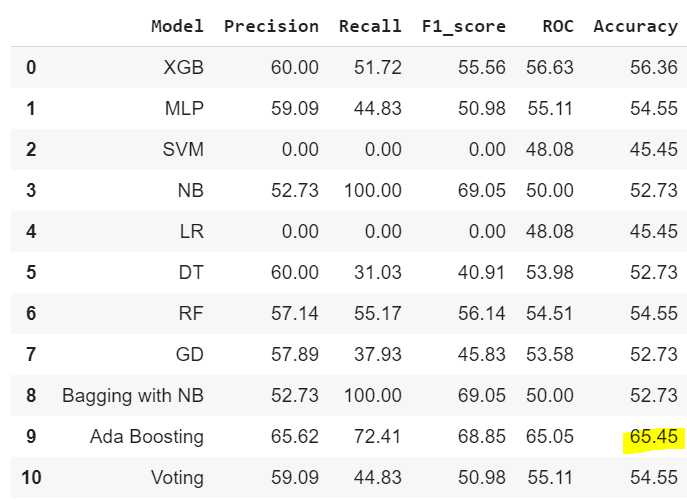
##### TFIDF features selection method DONE 5



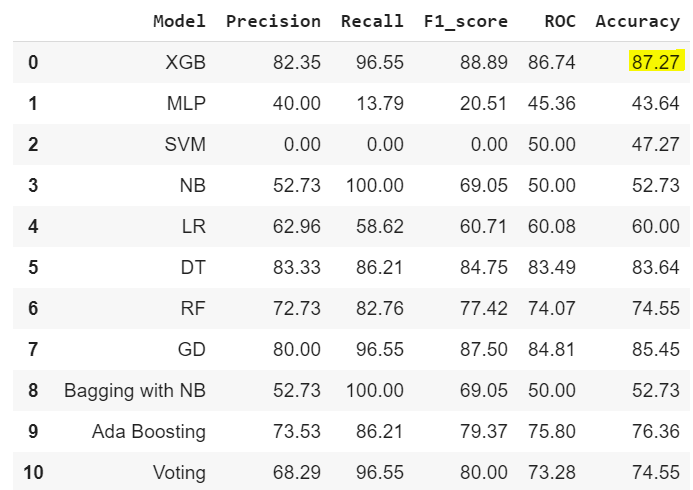
LDA Linear Discriminant Analysis **DONE 4**



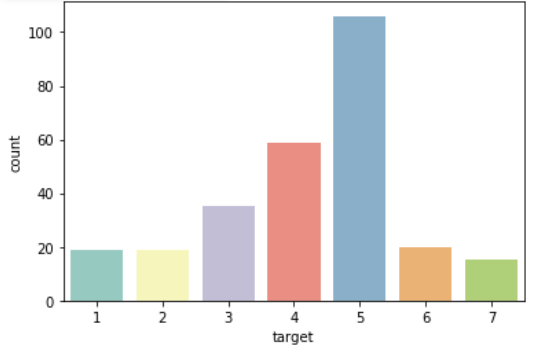
TFIDF+LDA **DONE 3**



All Features Selection methods (second place) **DONE 2**



**Treatment as a target**



1: RA\_D1 2: RA\_D2 3: RA\_RA\_D1 4: RA\_RA\_D2 5: Control

6: Ampel 7: Trehal 8: hfd 9: cd

##### Features as they are

As feature\_distribution plot function used here to show the significant relationships, as I’ve chosen the highest rank (most important features), where Random Forest algorithm used to show the most important features which affect the prediction model, defining the **Treatment** as the class label:

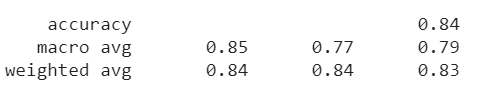
1. The strongest relationship between:
2. MWM (target time), and the BASAL grip strength score (s).
3. Rotarod scores 3 (s) and age. (Inverse relationship)
4. BASAL grip strength score (s) and the Treatment (C).
5. POST-BIGp\_TTT and the Bodyweight.
6. Rotarod score 3 (s) and the Glucose tolerance test (AUC).

#### Treatment as a label

Predicting the mice gender according to its body measurement (physical and mental).

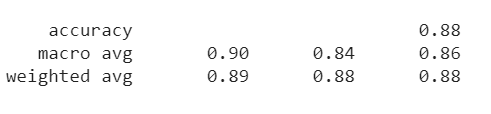
First: taking the values of the original features, as a new sheet called treatment.csv, and apply the following models, as you can see below 10 different ML-based models plus the Voting.

Where **RandomForestClassifier** recorded the highest score.



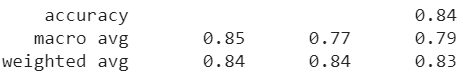
### Wrapper Features engineering

Gradient Boosting with DTC



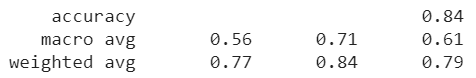
### PCA Features Selection

RandomForestClassifier

****

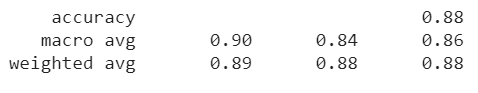
### TFIDF features selection method

Decision tree 2



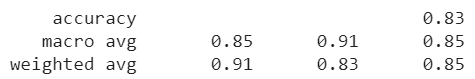
### LDA Linear Discriminant Analysis

Gradient Boosting with DTC



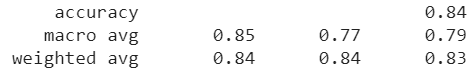
### TFIDF+LDA

GaussianNB



### All Features Selection methods (second place)

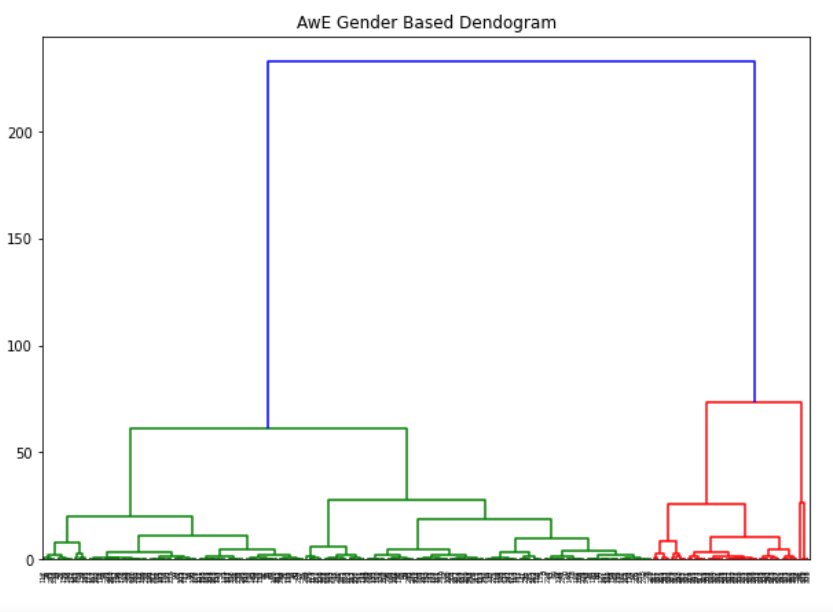
RandomForestClassifier

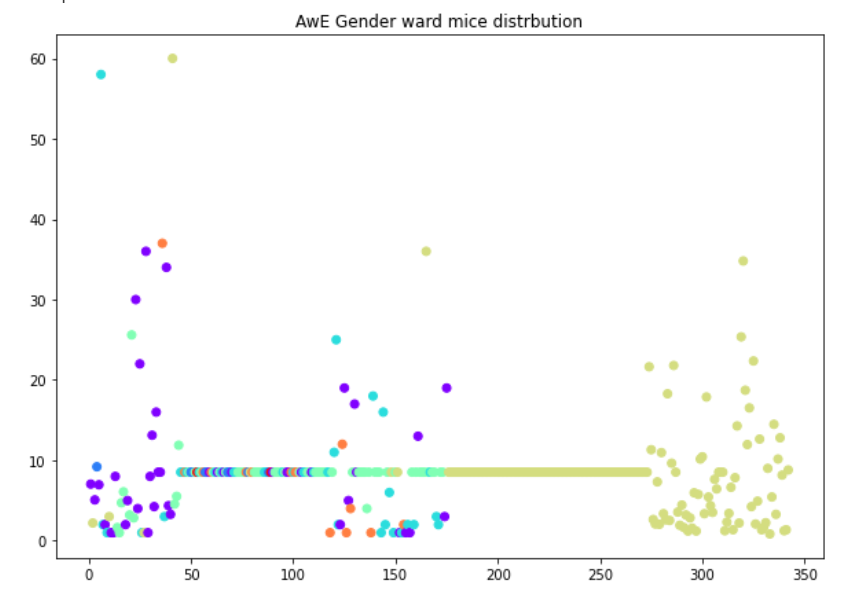


## Hierarchical Clustering Algorithm

Hierarchical clustering algorithms group similar objects into groups called clusters. According to their features, if we use a class label, then it becomes a hierarchical classification.

I created an instance of AgglomerativeClustering using the *euclidean* distance as the measure of the distance between points and ward linkage to calculate the proximity of clusters.

****

****