**The automation of the detection process of the genotype, control and the suitable treatment of the mice using only their protein levels expressions**

*<Please insert your contact details>*

*<Please insert date>*

Table of contents

[**Abstract**](#_w8nkmvj2m4ng) **2**

[**Introduction**](#_en81g6mv4pwj) **2**

[**Analysis**](#_ydhff1dk7mn1) **3**

[**Methodology**](#_n28atgs3y6ix) **10**

[**Results**](#_h080y7p5i6on) **11**

[**Discussion**](#_fn572v1e9rpp) **11**

[**Conclusion**](#_6opbtpidj1xd) **12**

[**References**](#_7ipkeq5cf6v9) **13**

# **Abstract**

Mice are always used in the initial experimental stages to test treatments and their side effects, and also to learn more about some diseases and the best way to treat them. Therefore, in this research, we identify some characteristics of mice such as genotype, behavior, and appropriate treatment with the use of reading 78 different proteins or modifications of the protein levels for all mice participating which are 72 mice that split between two types 38 control mice and 34 trisomic mice in this experiment. This would be done using machine learning classifiers that it first would use a subset of data for training two powerful machine learning which is decision trees and the random forest classifiers those models would be used to identify the genotype, behavior and needed treatment automatically using only the protein levels measures as input which would save a huge time for the scientists who used to make this process manually using analysis and microscopies so, they would avoid all this and all that needed is providing the machine learning system with the protein levels of the mice and it would automatically be decided to which group it belongs.

# **Introduction**

As mentioned above the main goal of the research is to build an accurate and robust machine learning system that is able to detect the behavior, genotype, and suitable treatment. The machine learning system would discriminate between the genotypes as control or trisomic, discriminate between the mice behavior as stimulated to learn and not stimulated and discriminate between the treatment if the mice injected with saline or mice injected with the memantine. So, the researchers combined the three characteristics; genotypes, behaviors, and the treatment types needed together and the label that would be used in the modeling consist of 8 classes:

a- First class, c-CS-s which includes nine mice; that express control, stimulated to learn and injected with saline mice.

b- Second class, c-CS-m which includes ten mice; that express control, stimulated to learn and injected with memantine mice.

c- Third class, c-SC-s which includes ten mice; that express control, not stimulated to learn and injected with saline mice.

d- Fourth class, c-SC-m which includes ten mice; that express control, not stimulated to learn and injected with memantine mice.

e- Fifth class, t-CS-s which includes seven mice; that express trisomy, stimulated to learn and injected with saline mice.

f- Sixth class, t-CS-m which includes nine mice; that express trisomy, stimulated to learn and injected with memantine mice.

g- Seventh class, t-SC-s which includes nine mice; that express trisomy, not stimulated to learn and injected with saline mice.

h- Eighth class, t-SC-m which includes nine mice; that express trisomy, not stimulated to learn and injected with memantine mice.

We would use the expression levels of the protein and the protein modifications to predict to which class of those 8 classes this sample belongs. For more information about the data we would explain some of the features:

A- MouseID: The identifier for each mouse sample in the data

B- DYRK1A\_N: The value of the expression level of the DYRK1A\_N protein.

C- ITSN1\_N: The value of the expression level of the ITSN1\_N protein.

D- BDNF\_N: The value of the expression level of the BDNF\_N protein.

E- NR1\_N: The value of the expression level of the NR1\_N protein.

F- NR2A\_N: The value of the expression level of the NR2A\_N protein.

G- pAKT\_N: The value of the expression level of the pAKT\_N protein.

H- pBRAF\_N: The value of the expression level of the pBRAF\_N protein.

I- pCAMKII\_N: The value of the expression level of the pCAMKII\_N protein.

J- pCREB\_N: The value of the expression level of the pCREB\_N protein.

In the same manner the values of the other 62 expression levels of the protein.

K- Genotype: That expresses the type of gene either control (c) or trisomy (t).

L: Treatment type: That expresses the type of treatment that applied to the mice.

M- Behavior: That shows the different two types of behaviors in the mice which are context-shock (CS) and the shock-context (SC).

N- Class: That the label includes the 8 classes that we have explained earlier (MM,2015).

For building a robust and accurate model there are a set of stages the data should come through to ensure it in their correct and suitable form to be supplied to the machine learning model we would explain some of those step in abbreviation as we would take on it in detail earlier in the methodology. Those steps are:

a- Reading the data in the tabular data structure (dataframe in our case).

b- Ensuring the data free from missing and errors.

c- Imputing the missing values in each feature with the most frequent value in that particular feature.

d- Drop the highly correlated features to avoid the feature redundancy that can lead to overfitting.

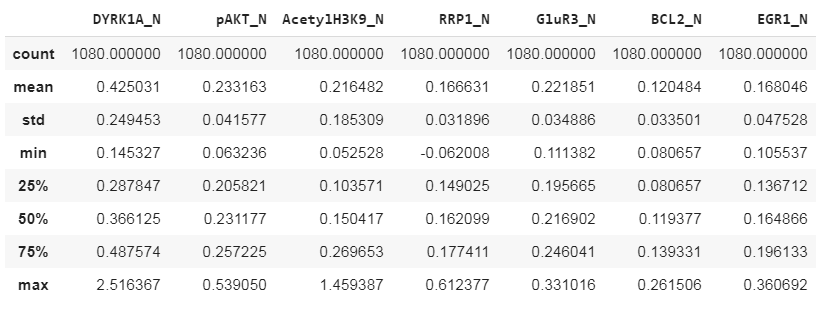
e- Scaling the features using the robust scaler to make them all in the same range of values and also as its name tells it is robust to the outliers.

e- Splitting the data into 20 folds and using them for training and validation of the model using the cross-validation method.

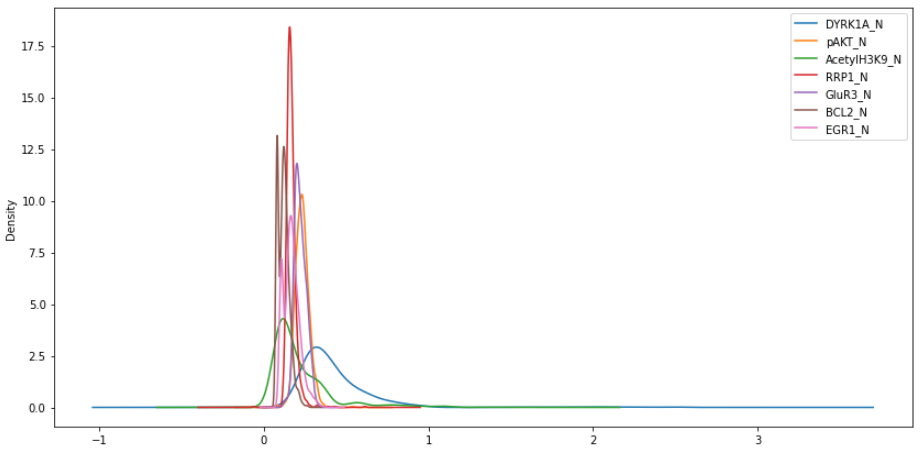
F- Finding the model with the highest average validation score and use it for tuning its hyperparameters.

# **Analysis**

We have constructed both types of analysis; independent features analysis in which we would analyze each feature independently and the pairs of features analysis in which we would study the hypothesis of the correlation between a set of pairs of the data and their relationship with the characteristics of the mice. First, we would start by analyzing each feature in the data independently. As seen in the below figure Fig-1 the mean for the DYRK1A\_N protein expression is higher than the other means of the other protein level expressions with almost the double that means that the DYRK1A\_N protein expression has a higher range of values compared to the other protein expressions as it also has a higher minimum and maximum values that the other protein levels expression. This means that the features despite most of the features have a close range of values but still the range of values is different as can be noticed in the distribution of each feature in figure Fig-2 which means that we need to scale the features as would be shown earlier.



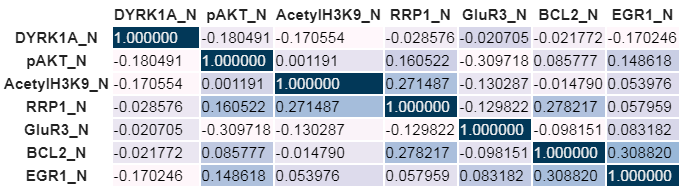
“Fig-1”



“Fig-2”

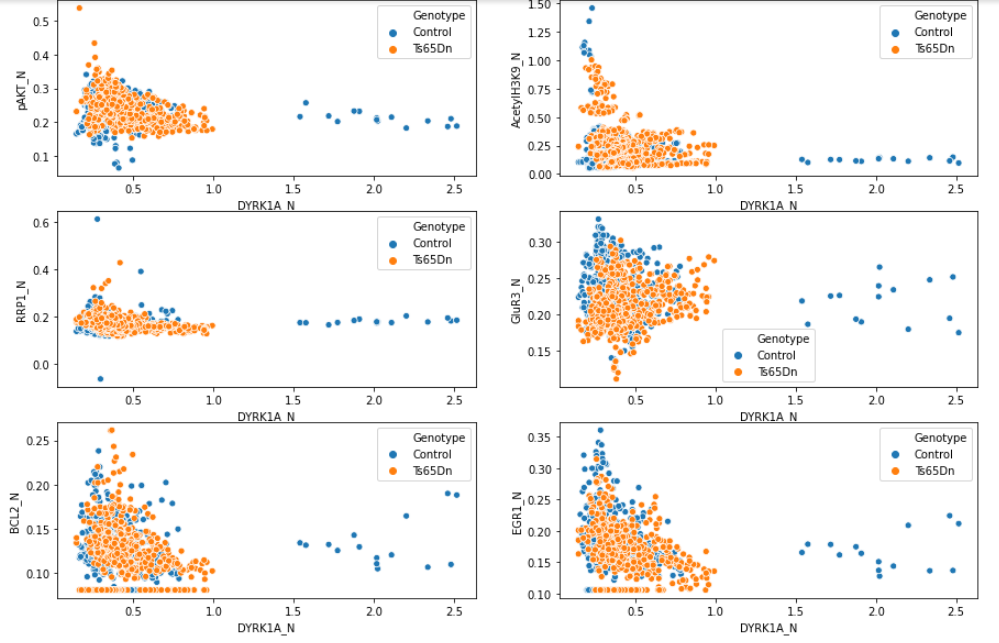
So, the Hypothesis is there is no high positive or negative between the pairs of the remaining features, and from the correlation matrix below we can see that our hypothesis is valid.

As we have mentioned in the introduction we have removed all the highly correlated features by highly correlated we mean the features with a correlation higher than 0.4 or less than -0.4. So, as seen in the below figure Fig\_3 the correlation matrix that shows the correlation between all the remaining pairs of features and as seen the highest positive correlation value is almost 0.3 which between the two features BCL2\_N and EGR1\_N and the highest negative correlation was almost -0.3 which is between GluR3\_N and pAKT\_N and that seems to be intuitive as we have mentioned above that all the highly correlated features have been removed.



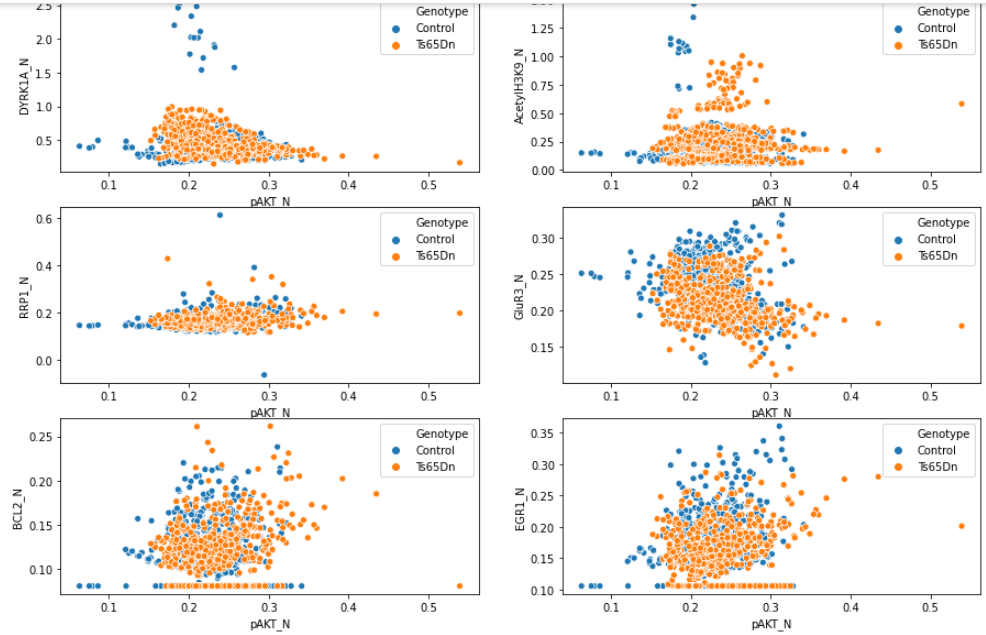
“Fig-3”

Now, we would discuss a set of hypotheses that investigate the relationship between the features; the first hypothesis is that if the value of the expression level protein DYRKIA\_N exceeded the threshold 1.5 then this mouse would be a trisome mouse regardless of the values of other expression levels proteins. As seen in the below figure Fig-4 that regardless of the y-axis value in all subgraphs once the value of the DYRKIA\_N on the x-axis exceeded the threshold 1.5 that the whole samples of mice belong to the trisome genotype which proves the validity of our hypothesis.



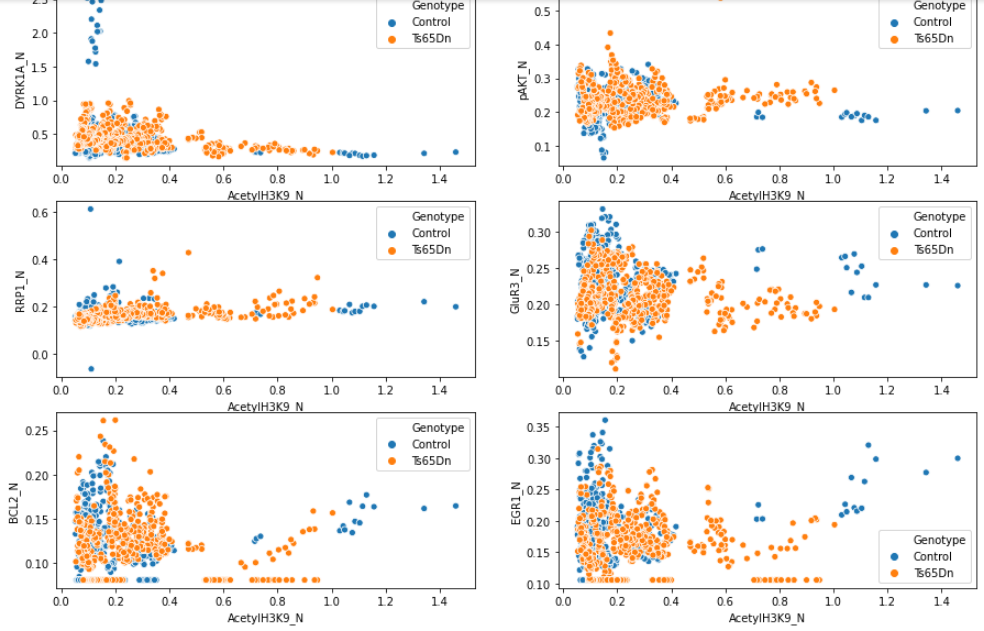
“Fig-4”

The second hypothesis is that contrary to the first hypothesis this hypothesis assumes that if the value of the expression level of the protein pAKT\_N exceeds a specific threshold which is 0.4 all mice would be control mice regardless of the values of the other expression levels proteins values. As seen in the figure below that the hypothesis is a valid hypothesis as for all the subfigures below regardless of the value of the y-axis once the value of pAKT\_N on the x-axis gets through the threshold 0.4 all the mice belong to the control genotype.



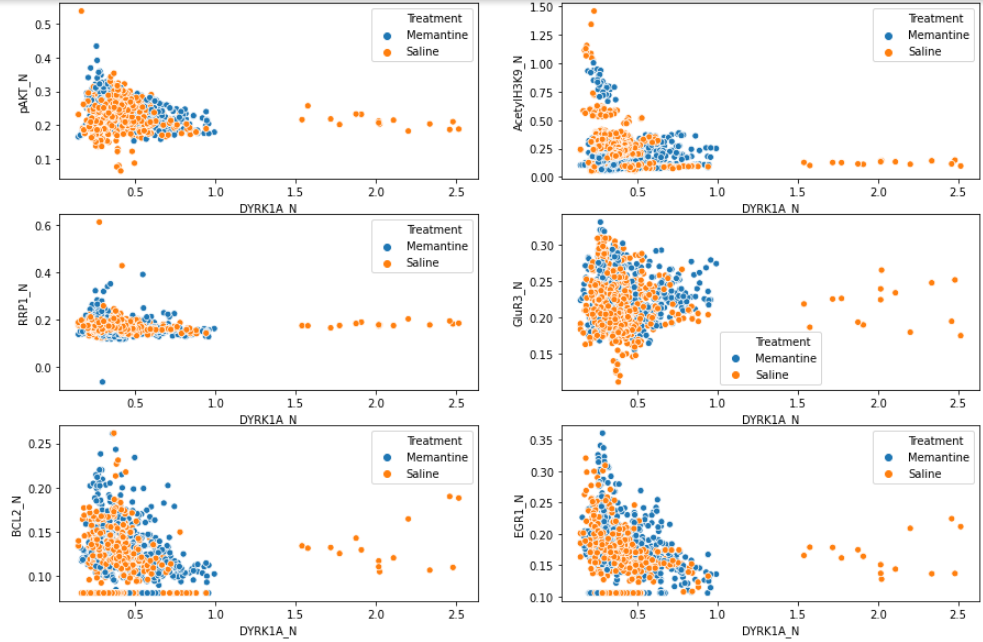
“Fig-5”

The third hypothesis is that if the value of the expression level of the protein AcetylH3K9\_N exceeds a specific threshold which is <=0.4<=1 all mice would be control mice regardless of the values of the other expression levels proteins values and when the value of AcetylH3K9\_N >1 all the mice would be trisomy mice. As seen in the figure below Fig-6 that the hypothesis is a valid hypothesis.



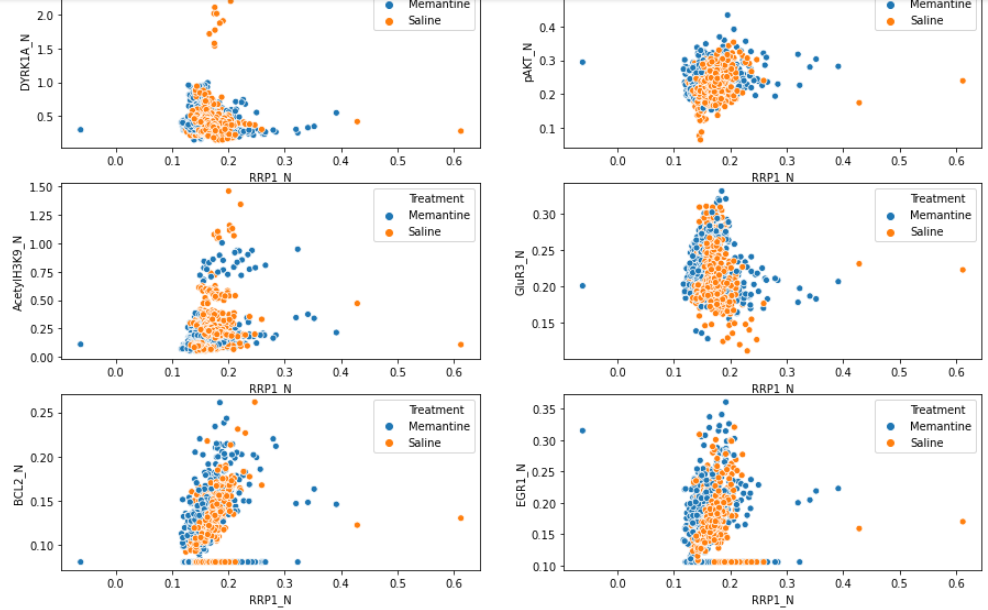
“Fig-6”

The fourth hypothesis is that if the value of the expression level of the protein DYRK1A\_N exceeds a specific threshold which is 1.5 all mice would be treated with saline regardless of the values of the other expression levels proteins values. As seen in the figure below Fig-7 that the hypothesis is a true hypothesis as to when the value of the DYRK1A\_N exceeded the threshold 1.5 all the mice treated with the saline treatment.



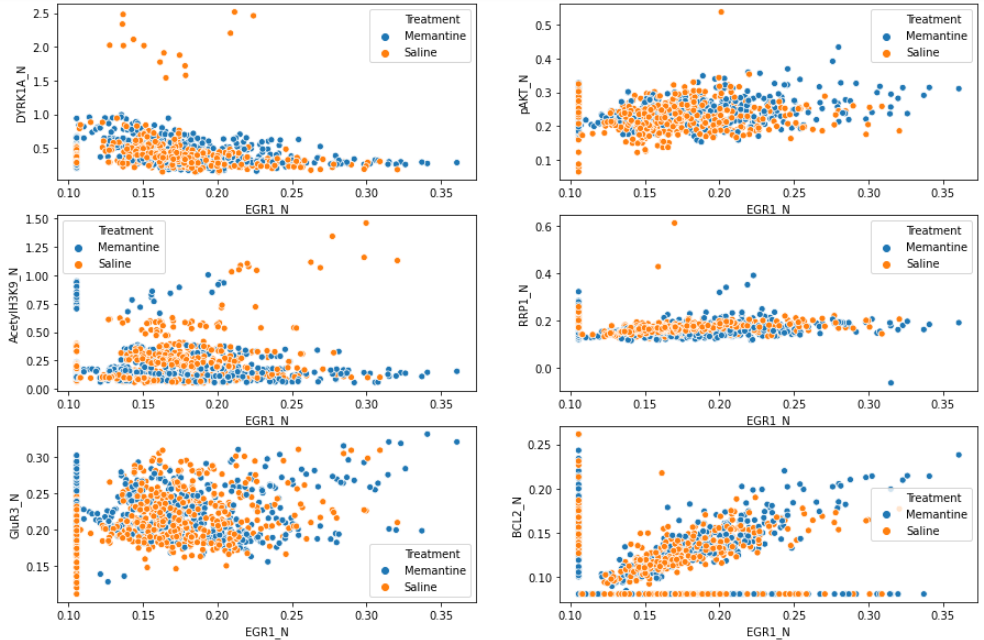
“Fig-7”

The fifth hypothesis is that most of the mice treated with saline tend to have a higher RRP1\_N expression level protein value than the mice that treated memantine regardless of the values of the other features. As seen in the figure Fig-8 that our hypothesis is not valid is that only applied for subfigures that show the relationship between (RRP1\_N and EGR1\_N) but (RRP1\_N and BCL2\_N) not applied to the other subfigures that show that in some figures the mice which treated with memantine have higher RRP1\_N values that mean that our hypothesis is not valid.



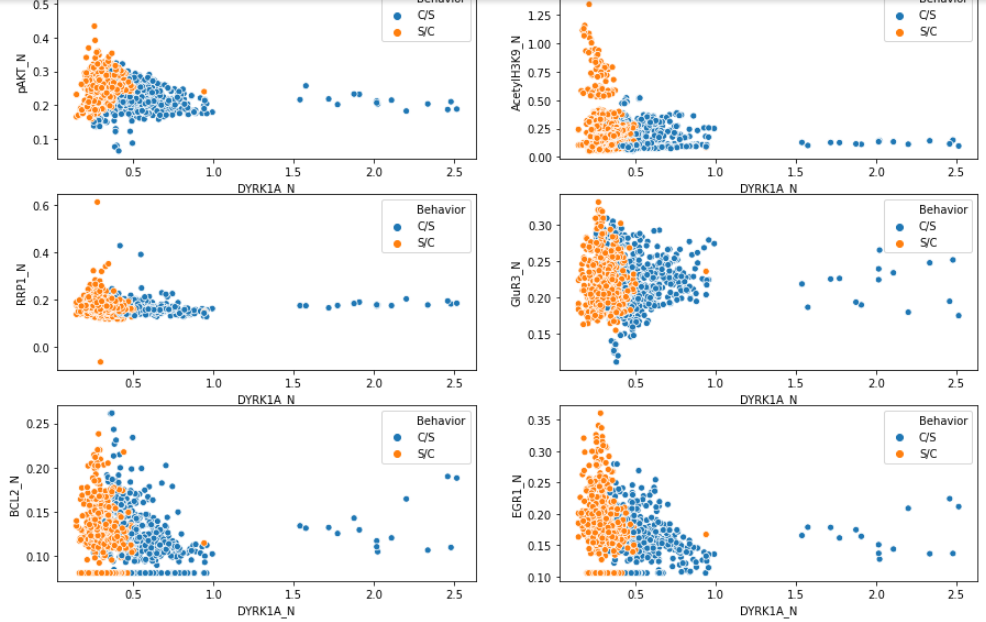
“Fig-8”

The sixth hypothesis is that at a specific threshold for the EGR1\_N expression level protein which is 0.3 when it exceeded it all the mice would be treated using memantine regardless of the other expression level proteins. As seen in the figure below the hypothesis is not valid as there are subfigures that show some of the mice are saline and EGR1\_N expression level protein is more than 0.3.



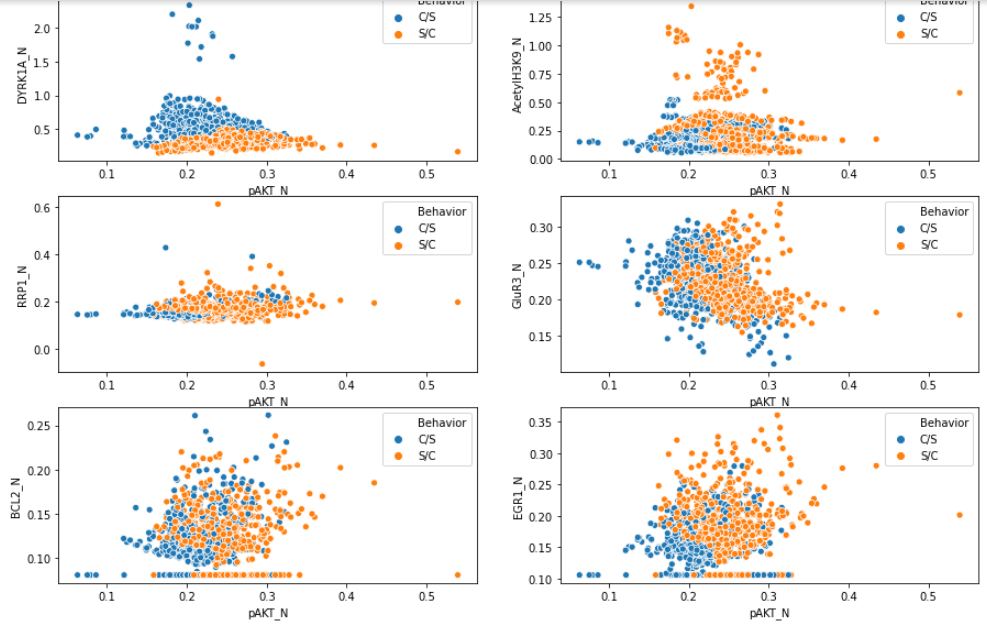
“Fig-9”

The seventh hypothesis is that almost all the mice that behave in a context-shock way have a higher DYRK1A\_N value of the expression level protein than the mice that behave in a shock-context way also all the context-shock mice have more than 0.6 DYRK1A\_N value of the expression level protein while shock-context mice have a DYRK1A\_N value of the expression level protein below 0.6. And, as seen in the below figure Fig-10 that our hypothesis is totally true and valid.



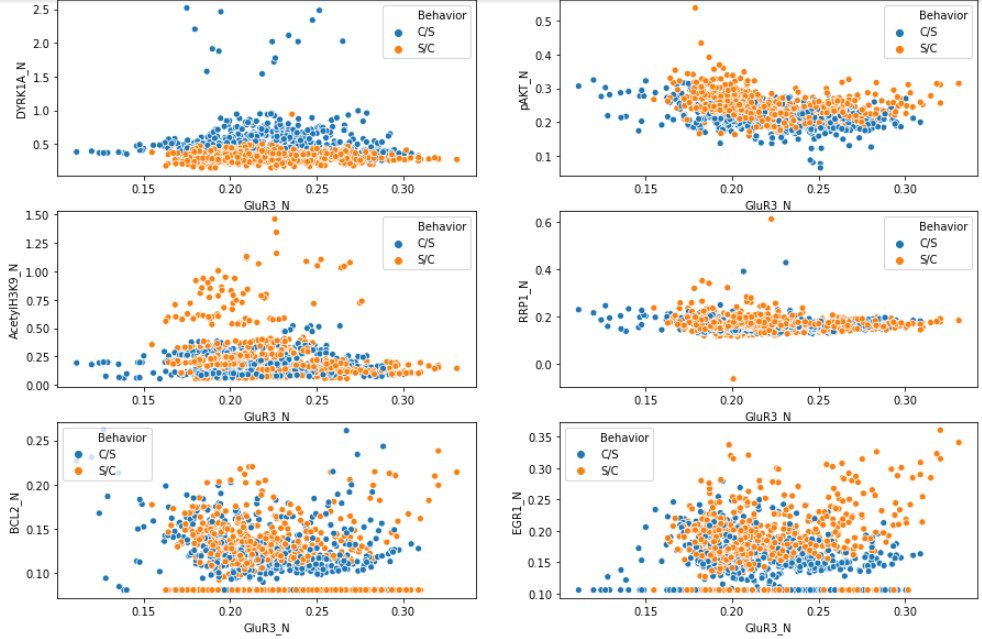
“Fig-10”

The eighth hypothesis is that all the mice that exceeded 0.3 for the value of the pAKT\_N expression level protein are belonging to the shock-context behavior type. As seen in the below figure Fig-11 we can see that the hypothesis is a valid hypothesis.



“Fig-11”

The ninth hypothesis is that all the mice that have the value of the GUR3\_N expression level protein below 0.15 would belong to the context-shock behavior group regardless of the other expression level protein values. From the below figure we can conclude that the hypothesis is a valid hypothesis.



“Fig-12”

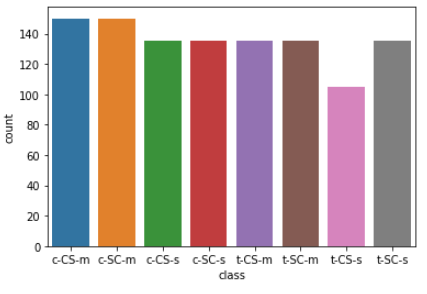
# **Methodology**

As mentioned in the introduction there is a set of strict steps applied to the data to make it in the correct form that is suitable to be fed in the machine learning models. We would explain some of these critical steps:

1- At the first stage we have discovered a set of features that have a set of errors then we have replaced these missings with the mode of that feature with missing by mode we mean the most frequent value. This technique worked well and got rid of all the missings but the new problem resulted in a type of all the features changed to ‘object’ that means they are categorical data types, so we returned them back to the float data type (numerical).

2- The second stage is that as the nature of the data that have 72 different expression levels of protein features, 3 characteristics features, and the class feature which in total are almost 76 different features in total that a high number of features and this data are called a high dimensional data that can lead to an overfitting problem that make the model generalized to the new data. So that we have decided to get rid of a set of these features which we call redundant features and our metric for defining this “redundant features'' is that the features that highly correlated with another feature and removing them won’t affect the variation of the data and in turn won’t cause a high information loss. So, all the features that have a correlation <=0.4 or >=-.04 have been removed. And, the remaining expression levels protein is DYRK1A\_N, pAKT\_N, AcetylH3K9\_N, RRP1\_N, GluR3\_N, BCL2\_N, EGR1\_N.

3- The third stage is that we had ensured that the label or what called class in our case should be well-represented and distributed by that we mean there are no mice types that have the majority of others and that what we could interpret from the below figure Fig-13. As seen in the figure the data is almost balanced, so there is no need to apply any balancing technique. And, then scaled the values of the features using the Robust scaler to make the whole features in the same format and to get rid of the extreme outliers in those features as this scaler as its name tells robust toward the outliers as it uses the interquartile range (IQR) not the min and maxes like min-max scaler or the mean and standard deviation as the standard scaler (Keen, 2017).



“Fig-13”

4- The fourth stage is splitting the data at first to the features which are the remaining expression levels protein and label which is the class and drop all the characteristics as they are combined in the label. Then we have used the cross-validation techniques that split the data into k folds or partitions in our case we have split the data into 20 folds in which 19 fold used in training and 1 used for validation and this process would be iterated until all the folds used in the training and the validation.

5- The fifth stage is the modeling stage in which we have used 2 machine learning models and for each model, we have applied two different heuristic techniques in the training. The two machine learning models are the decision tree which as its name tells construct a tree that would be used to decide the characteristics (type) of the mice using only the protein expression levels that the models trained on, the other model is the randomforest model which is a well-known model as it always gives decent results in many classification problems as it uses the ensembling methods that uses a set of decision trees for deciding the final outcome (mouse type). As mentioned for each model we have used two different heuristic techniques; the first one is one vs all (OVA) technique which builds binary for each class in the label in which this class considers the positive class and the rest of classes considered the negative class which means we would end with a number of classifiers equal to the number of classes in the label which in our case 8 classes that means using this heuristic method would end with building 8 classifier one for each class, the second heuristic method used was the one vs one method (OVO) in which we construct a binary classifier for each unique pair of classes in the label which means we would build (num\_of\_classes \* num\_classes-1) / 2 classifiers that mean in our case we would end with 28 classifiers constructed. In both heuristic methods each classifier for its class outcome specific probability value then the class with the highest probability would belong to it (Brownlee,2020).

# **Results**

As we mentioned above there are two machine learning models used which are the decision tree and the randomforest models and for each model, we have applied two different heuristic techniques OVO and OVA. Also, if first, we have used the default hyperparameters in both models, then after finding the model with the higher accuracy we would be tuning its hyperparameters. The below table Table-1 has each model with a specific heuristic method and its accuracy.

|  |  |
| --- | --- |
| Model | Accuracy |
| OVA Decision Tree | 0.77 |
| OVO Decision Tree | 0.64 |
| OVA Randomforest | 0.88 |
| OVO Randomforest | 0.90 |

“Table-1”

# **Discussion**

As seen in the results the random forest model is better than the decision tree model in the two applications of the models which means that the randomforest model is a totally better model. Also, the random forest model that applied using the one vs one technique was the best was 90% correctly predicted samples of all the predicated samples which is really great accuracy but we still need more improvements in the model so, we have selected a set of so important hyperparameters in the randomforest model to be tuned which are:

* n-estimators: which express the number of trees in the random forest with a 100 as a default value and can accept any non-zero positive integer value.
* criterion: that expresses the quality of split in each tree whose default value is ‘gini’ for Gini impurity and it accepts either gini or entropy for information gain.
* max\_depth: that shows the maximum depth of each tree, its default value is none which means the tree would continue in the expanding process until reaching a minimum number of samples, it could accept any positive integer value.
* min\_samples\_split: The minimum number of the samples needed for the node of the tree to split, its default value is 2 while its accepted float or integer positive values are higher than or equal 2.
* min\_sample\_leaf: show the minimum samples required to be in the leaf node, its default value is 1 and it accepts non-zero float or integer positive numbers.
* max\_features: that expresses the number of features that used in searching for the best split, its default value is auto which means the square root of the total number of the features it can accept non-zero float or integer positive values and also can accept ‘auto’, ‘log’ which get the log of the number of the features and sqrt that get the square root of the number of the features (Probst,2019).

We have tested different combinations of the values for those hyperparameters using a random search method that randomly selects the combination of the hyperparameters and runs the model with it and calculates the accuracy for each training and keeps doing that process for 200 iterations as we have defined earlier. The best combination was as the following:

|  |  |
| --- | --- |
| hyperparameter | best value |
| n-estimators | 170 |
| criterion | entropy |
| max\_depth | 30 |
| min\_samples\_split | 12 |
| min\_sample\_leaf | 3 |
| max\_features | auto |

“Table-2”

That combination resulted in almost 0.97 f1-score which means the hyperparameters tuning massively improved the model performance.

# **Conclusion**

In the end, we can say we were able to successfully build a robust and reliable machine learning system that is capable of predicting the mice’s treatment, behavior, and genotype using only the values of the expression levels of the protein. That considered the initial step for building a similar machine learning system that was able to predict some characteristics in human beings as the results for the system in the mice were really promising results.

# **References**

Ahmed MM, Dhanasekaran AR, Block A, Tong S, Costa ACS, Stasko M, et al. (2015) Protein Dynamics Associated with Failed and Rescued Learning in the Ts65Dn Mouse Model of Down Syndrome. Retrieved from DOI:10.1371/journal.pone.0119491

Ben Keen. (2017). FEATURE SCALING WITH SCIKIT-LEARN. retrieved from <https://benalexkeen.com/feature-scaling-with-scikit-learn/>

Jason Brownlee. (2020). How to Use One-vs-Rest and One-vs-One for Multi-Class Classification. retrieved from <https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/>

Philipp Probst, Marvin Wright, Anne-Laure Boulesteix. (2019). Hyperparameters and Tuning Strategies for Random Forest. retrieved from <https://arxiv.org/abs/1804.03515>