Mice Protein Expression Levels

I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honor code by typing "Yes"

COSC 2670 – Practical Data Science
Student ID: s3756188
Student Name: Kanishka Tamang
Assignment 2: Data Modelling
s3756188@student.rmit.edu.au
Wednesday, June 10, 2020

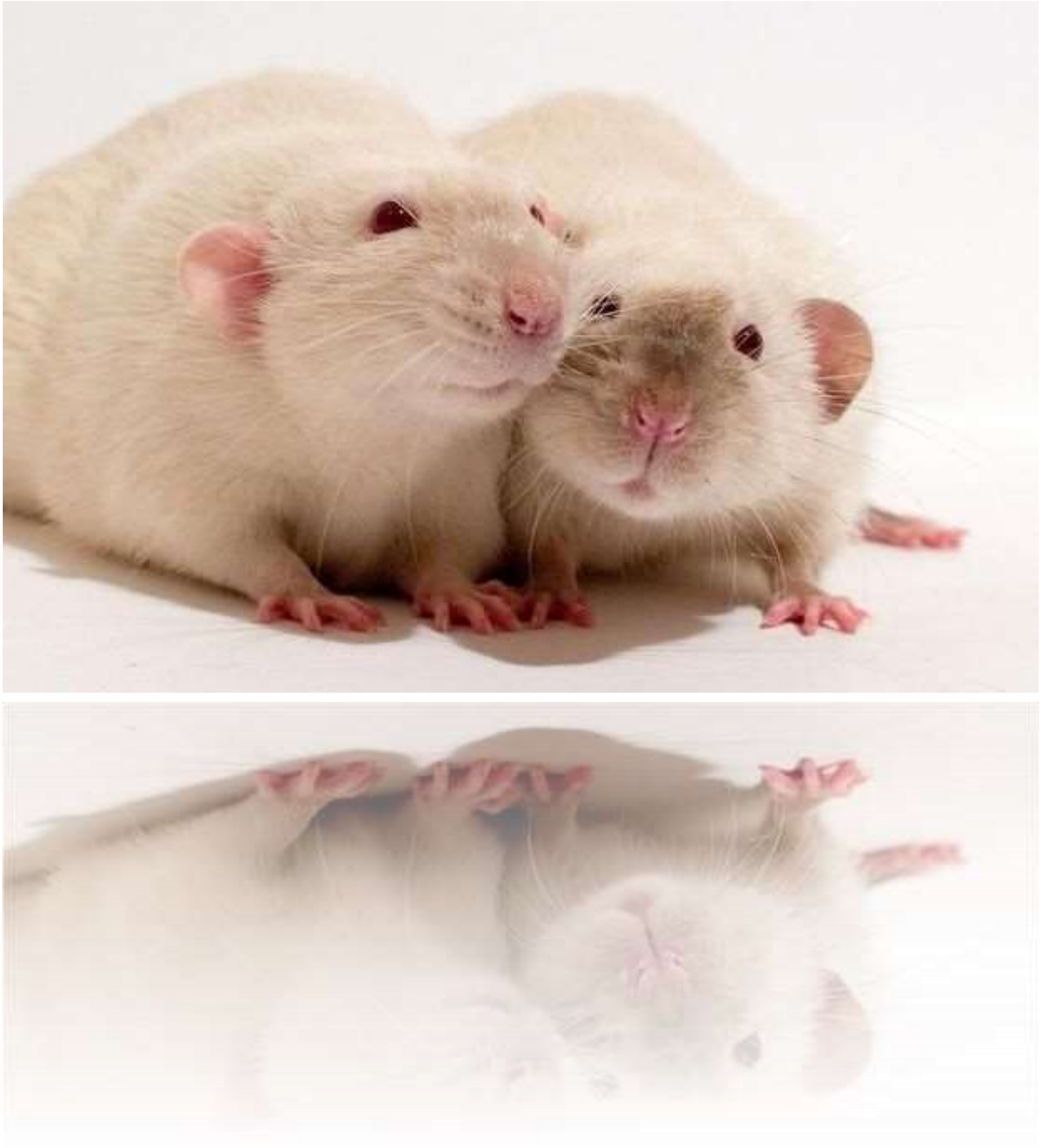


Table of Contents

ABSTRACT

Down Syndrome (DS) is a genetic disorder which is mainly due to extra genetic material from chromosome 21 therefore the people with Down syndrome have 47 chromosomes. The symptoms of Down syndrome include poor judgement, short attention span and cognitive impairment (CPL, 2020). In order to evaluate the effects of the drug to diagnose the DS the protein expression mice dataset was published. The report aims to analyse the effects of the protein which could be an important factor to have an overall effect on the recovering ability to learn amongst the mice with down syndrome. The Data Mining approach adopted for the report would be the Classification Approach.

To elaborate further on the objective of the report, the following Hypothesis will also be investigated:

- **Investigation of Research Question:** "Is there any co-relation among the most important proteins that best determines the class of the mice?".
- **Investigation of Hypothesis:** Seek statistical evidence whether the mean, median and variance is same among the protein expression readings grouped by classes in the mice datasets.

INTRODUCTION

The mice protein expression dataset is sourced from the UCI Machine learning repository. There are 1080 rows and 82 columns in the dataset. It consists of 77 protein expression levels (numerical feature) which produces detectable signals in the nuclear fraction of the cortex. There are 38 control and 34 trisomic mice (with Down Syndrome) amounting to 72 mice in total. In this dataset 15 measurements were recorded for each protein per mouse. The control mice have 38x15 = 570 measurements and the trisomic mice 34x15 = 510 measurements. Therefore, the datasets consist of 1080 measurements in total. There are 4 categorical variables namely Genotype, Treatment, Behaviour and class. The Target feature 'class' is illustrated below:

Classes	Features	Count of Mice_ID per class
c-CS-s	control mice, motivated to learn, injected with saline	135
c-CS-m	control mice, motivated to learn, injected with memantine	150
c-SC-s	control mice, not motivated to learn, injected with saline	135
c-SC-m	control mice, not motivated to learn, injected with memantine	135
t-CS-s	trisomy mice, motivated to learn, injected with saline	105
t-CS-m	trisomy mice, motivated to learn, injected with memantine	135
t-SC-s	trisomy mice, not motivated to learn, injected with saline	135
t-SC-m	trisomy mice, not motivated to learn, injected with memantine	135

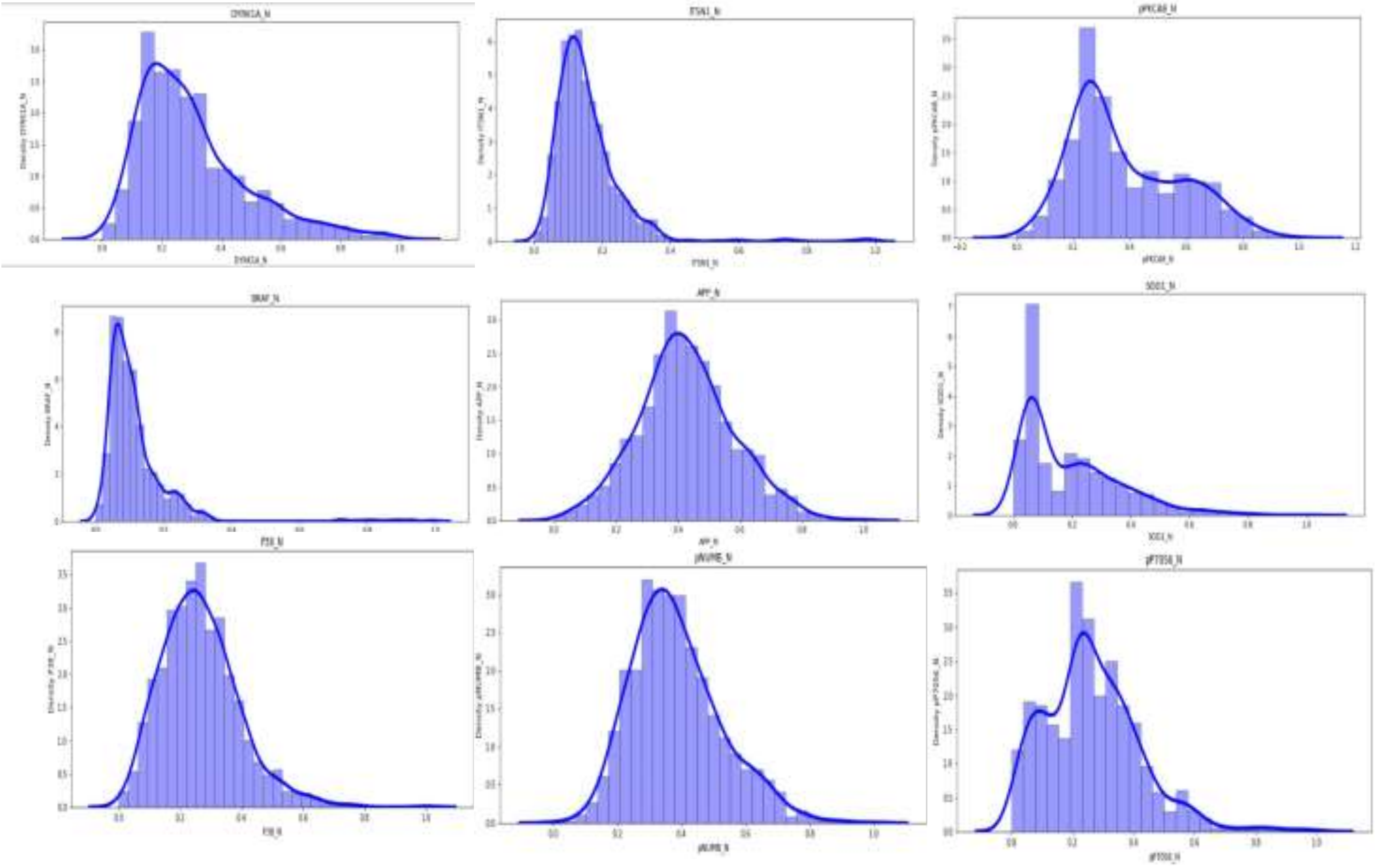
METHODOLOGY

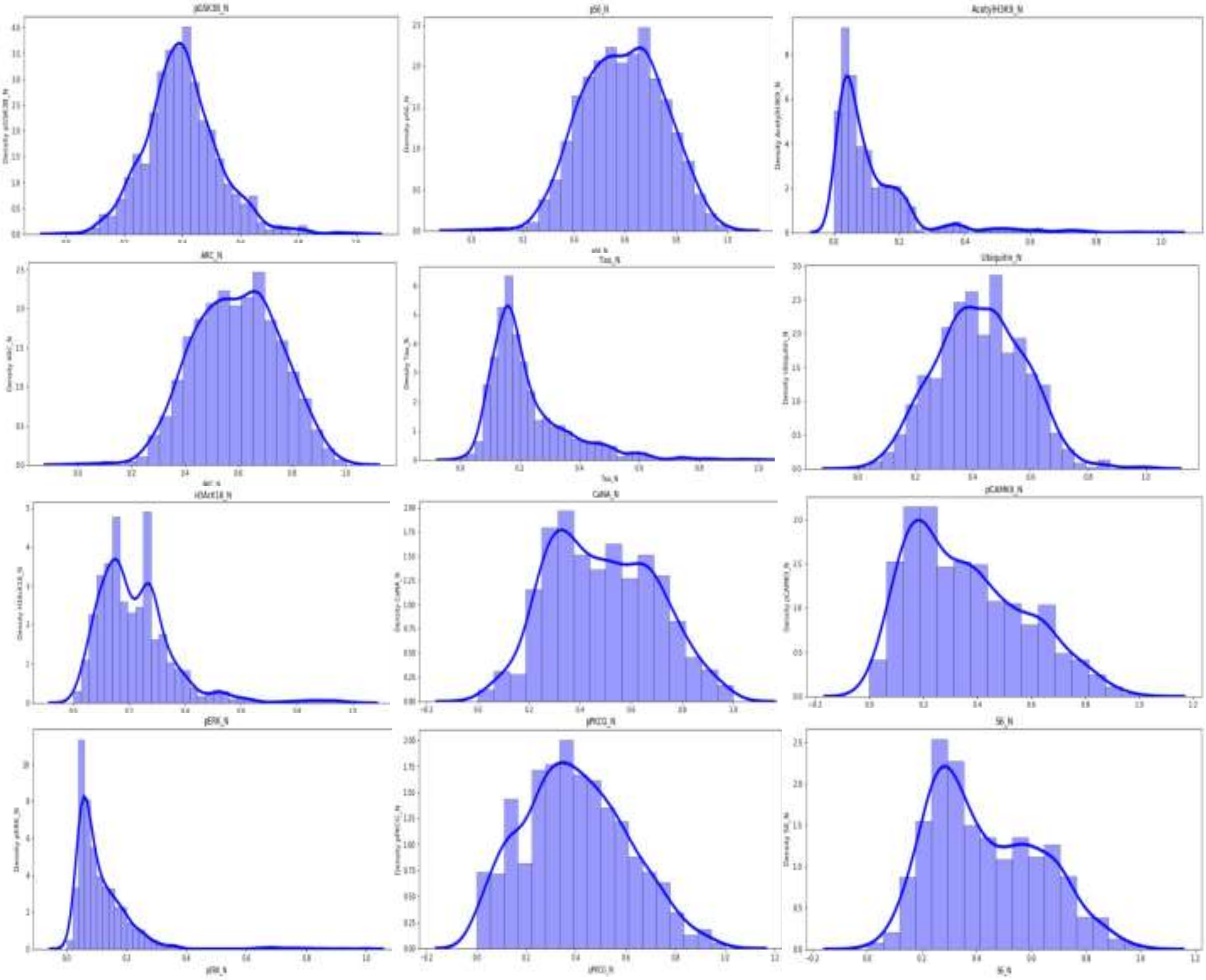
1.1) Retrieving Data: The relevant packages necessary for the analysis is loaded. The raw dataset is imported in to Jupyter Workspace with the read csv function and is stored to a variable name df. The dimensions and the datatypes are re-checked to make sure that the data is correctly imported in to the environment.
1.2) Data Preparation: The objective here is to get rid of all the typos, missing values and detect the outliers.

- Outlier Detection: A list is created and then the descriptive statistics like mean, (2 and 3 Standard Deviation) is calculated based on the classes to detect the outliers. All the outliers are imputed by the median value because about 21% of the dataset consisted of outliers and by removing them would result in a loss of considerable portion of the dataset.
- There were also a lot of missing values in the dataset. Based on each class the protein values were replaced by the median value.
- It is important to note here that the Mouse_ID here is not a redundant feature and have been split in two columns namely ‘Mice_ID’ which is the unique ID for each mouse and ‘Measurement_No’ ranging between 1-15 is the measurement number of each mouse
- The features have been scaled through the technique of standardization with the MinMaxScaler() as it is a pre-requisite measure before fitting any machine learning algorithms. Standardization also ensures that the feature has the property of standard normal distribution with a mean of zero and the standard deviation of one.
- The target column ‘class’ has been encoded to 'c-CS-m': 0,'c-CS-s': 1,'c-SC-m': 2,'c-SC-s': 3,'t-CS-m': 4,'t-CS-s': 5,'t-SC-m': 6,'t-SC-s': 7.

1.3) Data Exploration: In this step, the main focus lies in exploring the data and gain deeper visual insights which would be helpful in the Data Modelling phase. However, due to the limitations on the page count for the report, 21 features which was used for data modelling is chosen to be visualized.

Explore Each Numerical Column: A histogram with a normal overlay curve has been plotted for every column to visualize the distribution of the data for every protein.





It is observed from the above graph that the data is normally distributed for most of the proteins except for ITSN1_N, SOD1_N, BRAF_N, DYRK1A_N, P38_N, Tau_N, AcetylH3K9_N and pERK_N which is right skewed.

1.4) Investigation of Hypothesis: Seek statistical evidence whether the mean, median and variance is same among the protein expression readings grouped by classes in the mice datasets.

In the *Data preparation* phase, the outliers had been imputed by the median of the protein based on each class. However, once the outlier is removed from the dataset and then we plot a boxplot the IQR changes respectively which will keep showing a new outlier. Therefore, in order to test statistically whether the median, mean and variance is same or not among the classes, the following tests were conducted with their respective hypothesis:

Kruskal Test:

H0: The median of all the proteins categorized by classes are equal

H1: The median of all the proteins categorized by classes are not equal

One-way ANOVA test: (F_onewayResult)

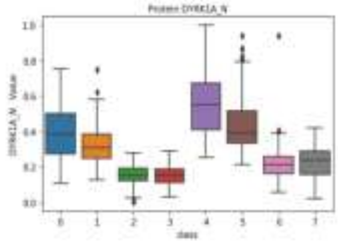
H0: The mean of all the proteins categorized by classes are equal

H1: The mean of all the proteins categorized by classes are not equal

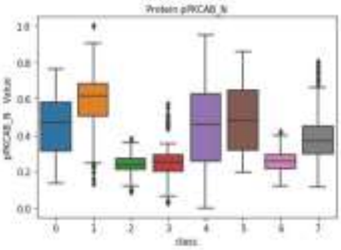
Levene Test:

H0: The proteins categorized by classes have equal variance

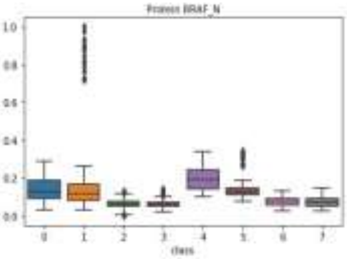
H1: The proteins categorized by classes does not have equal variance



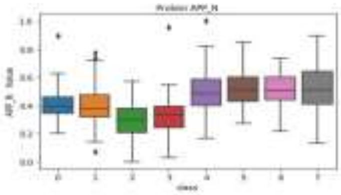
```
Statistical test for : DYRK1A_N
KruskalResult(statistic=685.3688029011459, pvalue=9.837520885304124e-144)
F_onewayResult(statistic=203.69954994636345, pvalue=6.05551966940657e-192)
LeveneResult(statistic=38.52834950853434, pvalue=2.1495769037075046e-48)
```



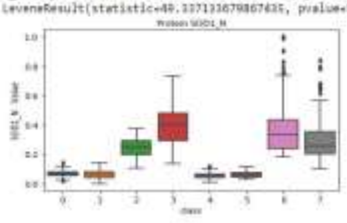
```
Statistical test for : pPKCAB_N
KruskalResult(statistic=419.6086452698364, pvalue=1.4834787328284452e-86)
F_onewayResult(statistic=96.20050445442888, pvalue=6.342083272455478e-109)
LeveneResult(statistic=49.337133679867435, pvalue=5.877347003305401e-61)
```



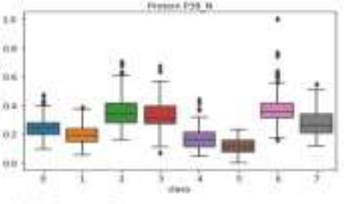
```
Statistical test for : BRAF_N
KruskalResult(statistic=578.7252631031952, pvalue=9.272377708735738e-121)
F_onewayResult(statistic=53.08297148871324, pvalue=3.6410095095623153e-65)
LeveneResult(statistic=22.983922825091582, pvalue=3.6043399049857685e-29)
```



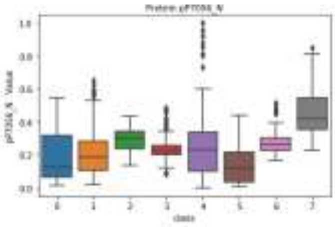
```
Statistical test for : APP_N
KruskalResult(statistic=379.8291227004053, pvalue=5.021503493040242e-78)
F_onewayResult(statistic=75.01228107854279, pvalue=2.0031026558223799e-88)
LeveneResult(statistic=8.608168700962504, pvalue=2.6574860720037425e-10)
```



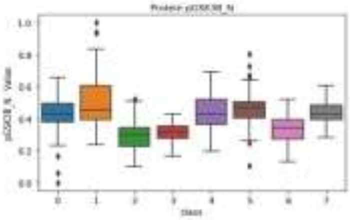
```
Statistical test for : SOD1_N
KruskalResult(statistic=837.4011269221327, pvalue=1.6106529736733600e-176)
F_onewayResult(statistic=290.3165930693143, pvalue=2.3055597842292785e-244)
LeveneResult(statistic=64.47052903434585, pvalue=1.5894260860956353e-77)
```



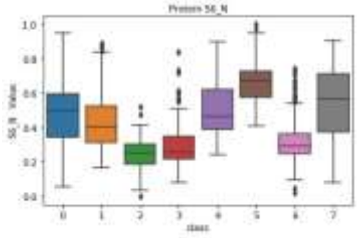
```
Statistical test for : P38_N
KruskalResult(statistic=575.8369951557703, pvalue=3.8811249249200819e-120)
F_onewayResult(statistic=134.76523027537707, pvalue=3.4257870183370984e-142)
LeveneResult(statistic=8.306440633323262, pvalue=5.243400270430512e-10)
```



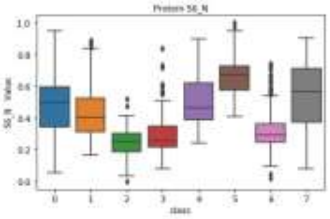
```
Statistical test for : pP70S6_N
KruskalResult(statistic=320.48521842580635, pvalue=2.5390647224229173e-65)
F_onewayResult(statistic=58.37977763604251, pvalue=5.407957606657105e-71)
LeveneResult(statistic=20.667963618561954, pvalue=3.4565571340127126e-26)
```



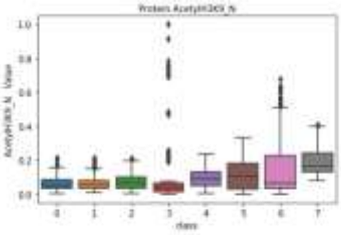
```
Statistical test for : pGSK3B_N
KruskalResult(statistic=444.7034667010371, pvalue=6.093310800543124e-92)
F_onewayResult(statistic=86.61086637452859, pvalue=7.154063353504704e-100)
LeveneResult(statistic=14.833135159845183, pvalue=1.4796298183721987e-18)
```



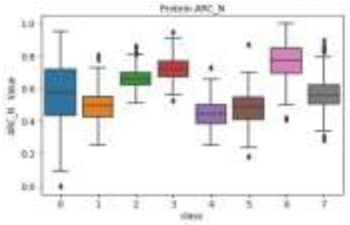
```
Statistical test for : S6_N
KruskalResult(statistic=445.54059536181876, pvalue=4.0274964099536644e-92)
F_onewayResult(statistic=98.0473301477476, pvalue=1.2503547648661810e-110)
LeveneResult(statistic=14.331150557827147, pvalue=6.814817878853922e-18)
```



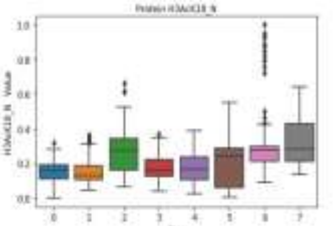
```
Statistical test for : S6_N
KruskalResult(statistic=445.54059536181876, pvalue=4.0274964099536644e-92)
F_onewayResult(statistic=98.0473301477476, pvalue=1.2503547648661810e-110)
LeveneResult(statistic=14.331150557827147, pvalue=6.814817878853922e-18)
```



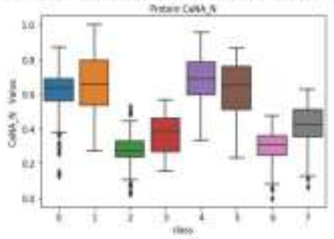
```
Statistical test for : AcetylH3K9_N
KruskalResult(statistic=161.44813340906014, pvalue=7.411626485101566e-30)
F_onewayResult(statistic=22.732667107220437, pvalue=7.566911232133878e-29)
LeveneResult(statistic=17.424853316010232, pvalue=5.768355185927494e-22)
```



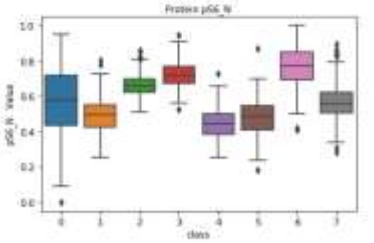
```
Statistical test for : ARC_N
KruskalResult(statistic=554.5108303272655, pvalue=1.510241848117133e-115)
F_onewayResult(statistic=138.89373203702593, pvalue=1.7286138905903922e-145)
LeveneResult(statistic=29.687207155058477, pvalue=1.2472903078444778e-37)
```



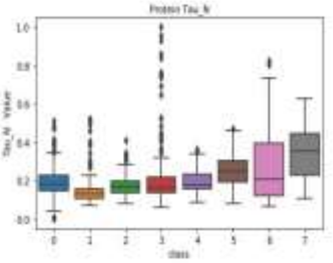
```
Statistical test for : H3AcK18_N
KruskalResult(statistic=265.5924480831325, pvalue=1.324072342415673e-53)
F_onewayResult(statistic=49.076849833340354, pvalue=1.1595700674064142e-60)
LeveneResult(statistic=14.213810943568895, pvalue=9.743439961378521e-18)
```



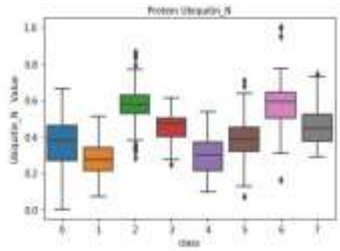
```
Statistical test for : CaNA_N
KruskalResult(statistic=686.2577172474598, pvalue=6.327957117018958e-144)
F_onewayResult(statistic=240.9717677880923, pvalue=5.372624266710476e-215)
LeveneResult(statistic=15.296388480024991, pvalue=3.621152876106118e-19)
```



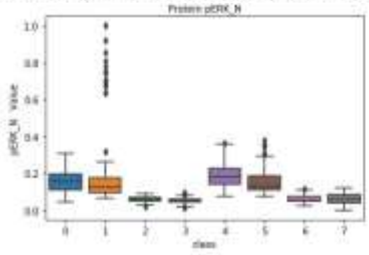
```
Statistical test for : pS6_N
KruskalResult(statistic=554.5108303272655, pvalue=1.510241848117133e-115)
F_onewayResult(statistic=138.89373203702593, pvalue=1.7286138905903922e-145)
LeveneResult(statistic=29.687207155058477, pvalue=1.2472903078444778e-37)
```



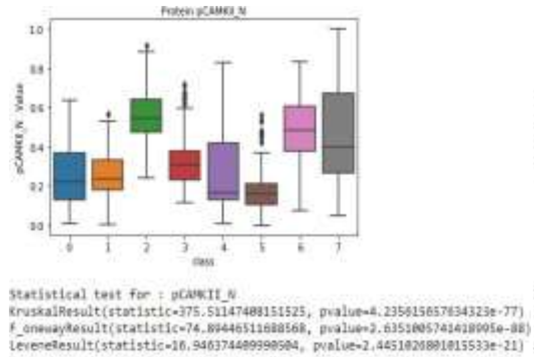
```
Statistical test for : Tau_N
KruskalResult(statistic=240.19005017454447, pvalue=3.3860122517272065e-48)
F_onewayResult(statistic=33.43769050506947, pvalue=2.085743412412002e-42)
LeveneResult(statistic=19.309058760493272, pvalue=2.0021312270595708e-24)
```



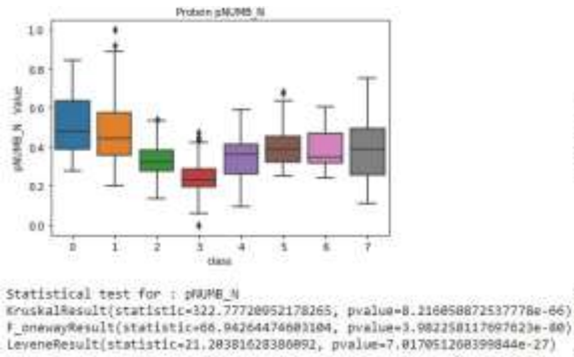
```
Statistical test for : Ubiquitin_N
KruskalResult(statistic=575.8964392975568, pvalue=3.768432545163687e-120)
F_onewayResult(statistic=161.1638003977183, pvalue=1.6289861770779066e-162)
LeveneResult(statistic=3.6934779237237705, pvalue=0.0005019352622096251)
```



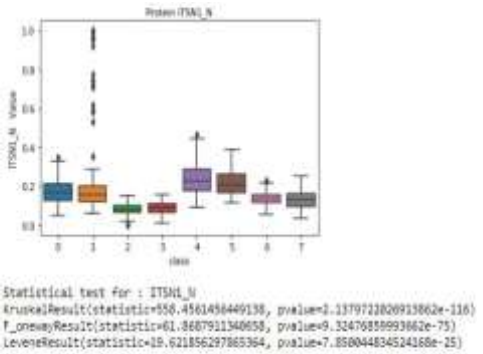
```
Statistical test for : pERK_N
KruskalResult(statistic=756.8158243559748, pvalue=3.8523633677428227e-159)
F_onewayResult(statistic=86.04090475484287, pvalue=2.5708531225069517e-99)
LeveneResult(statistic=24.642534230252196, pvalue=2.7527189560095423e-31)
```



```
Statistical test for : pCAMKII_N
KruskalResult(statistic=375.5114748815152, pvalue=4.23561565783432e-77)
F_onewayResult(statistic=74.8944651168856, pvalue=2.635100574141899e-88)
LeveneResult(statistic=16.946374409990504, pvalue=2.445102680101553e-21)
```



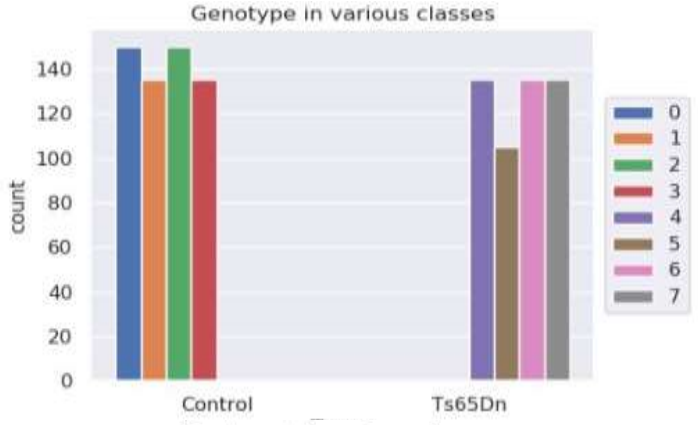
```
Statistical test for : pNUMB_N
KruskalResult(statistic=322.7772095217826, pvalue=8.216058872537778e-66)
F_onewayResult(statistic=66.94264474660310, pvalue=3.982258117697623e-80)
LeveneResult(statistic=21.20381628386092, pvalue=7.017051260399844e-27)
```



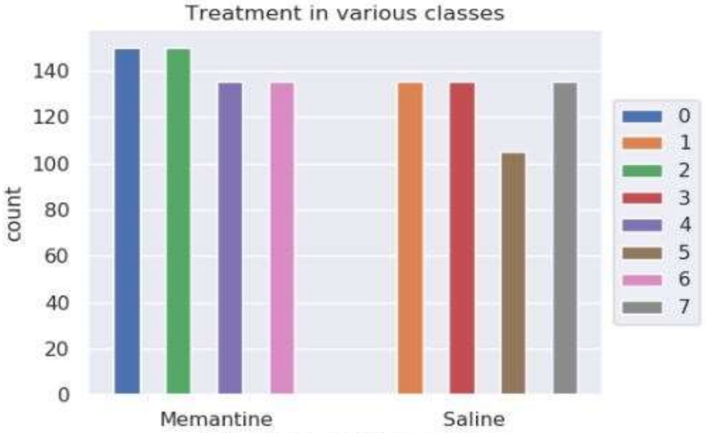
```
Statistical test for : ITSN1_N
KruskalResult(statistic=558.4561456449138, pvalue=2.137972202691386e-116)
F_onewayResult(statistic=61.8687911340658, pvalue=9.324768559993862e-75)
LeveneResult(statistic=19.62185629786536, pvalue=7.850044834524168e-25)
```

In all the test performed, it is noticed that the p-value is lesser than the significance level of 5 %. Therefore, it is statistically significant and we reject the null hypothesis. Therefore, the median, mean and the variance are not same among all the groups for each of the protein.

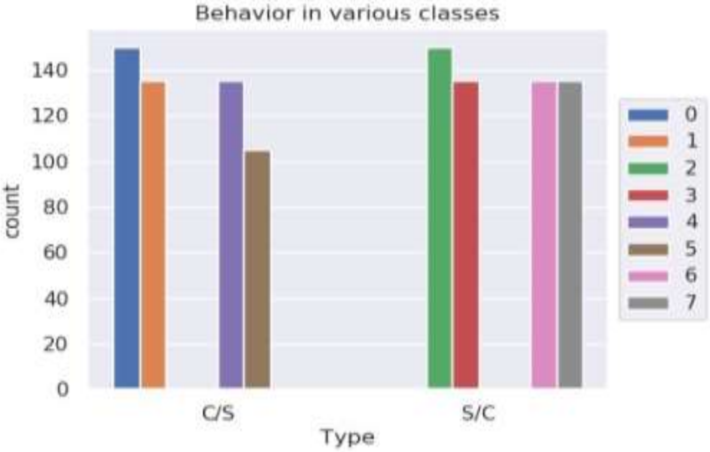
1.5) Exploring relationship between categorical columns in pairs with the Target Variable



Genotype in various classes: It can be observed from the bar graph that Trisomic (with down syndrome) as ‘Ts65Dn’ type of mouse are lesser in number across all the classes in comparison to type control. Additionally, the class 0 which is the 't-CS-s' in Ts65Dn has the lowest number of observations.



Treatment in various classes: Memantine is used to treat the mice with Down syndrome. It is observed from the graph that more of the mice are injected with memantine among which the class 0 which is c-CS-m' is injected the most.

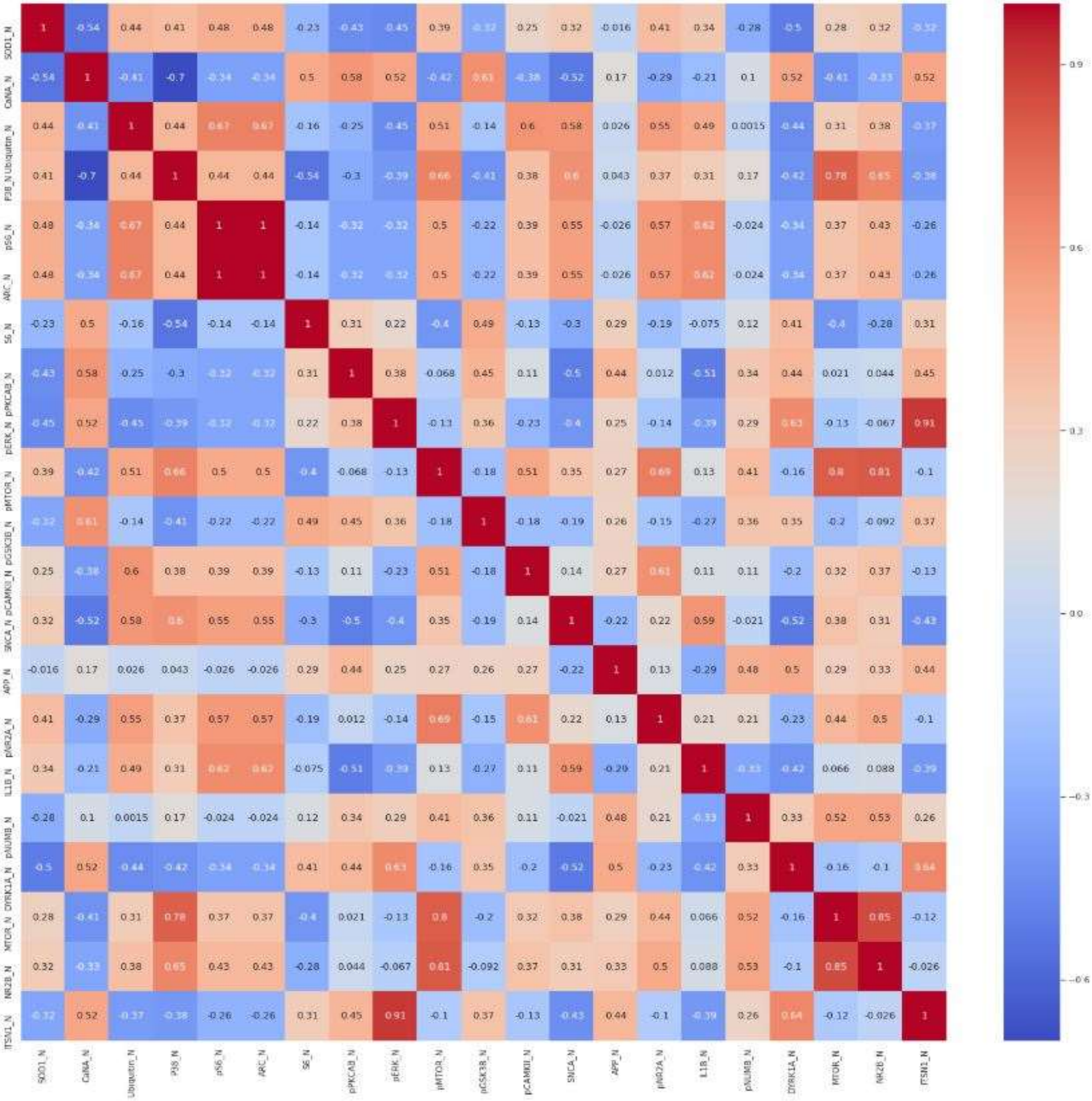


Behavior in various classes: There are two types of groups namely, the (context-shock) where the mice was given a chance to explore for 3 minutes and was given a shock. On the contrary, the other group namely (shock-context) were immediately given a shock and later was allowed to explore for 3 minutes. In terms of classes, it is observed in the graph that there are more mice belonging in the shock context group.

1.6) Investigation of Research Question: "Is there any co-relation among the most important proteins that best determines the class of the mice?".

Relationship between Attributes: To explore the relationship between attributes the Correlation matrix is plotted which gives us a good visualization of the relation of the 21 features with their correlation. Correlation coefficient ranges from-1 to 1. -1 indicates a strong negative correlation and +1 indicates a strong positive correlation and coefficients close to 0 indicate no linear correlation. It can be observed from the graph below that most of the proteins are moderately correlated.

Scatter Plot Matrix of the 21 features selected for Data Modelling



2) Data Modelling: Important insights have been drawn during the process of Data Exploration which would now help to model the data. Firstly, the Hill Climbing Method was chosen to derive the best features from the Dataset. However, Hill climbing cannot reach the optimal/best state (global maximum) if it enters any of the following regions which are as follows, Local maximum, Plateau and ridge.
Therefore, feature selection in this report, is chosen to be based on Random Forest Classifier, an ensemble model made of many decision trees which also reduces the variance of a single decision tree resulting in a more accurate selection of features. After the feature selection, the hyper-parameters of K-NN, Decision Tree and Random Forest Classifier have been tuned. Also, a t-test has been conducted to compare if the accuracy scores of the models fitted are statistically significant or not. Then the Models have been evaluated based on the confusion matrix and a cross validation approach.

The Data has to be split into training and test set before selecting the features which would be used to build the model. The model learns in the train set in order to be generalized in the test set. Generally, feature selection is performed to reduce the dimensionality of the dataset and select the best descriptive features that explain the target feature. However, there were two important concerns: with less training data, the parameter estimates will have a greater variance and with less testing data, the performance statistics will have greater variance. Therefore, it is very important to note this while dividing the data to make sure that neither variance is too high. In this report, prior to the feature selection an experiment has been conducted to see the different accuracy scores of K-NN Model and Decision tree model ***with all the features*** (proteins) in the dataset. It is to be noted here that these are the cross-validated scores.

K-NN Model	**Score Metric: Accuracy**	**Decision Tree Model**	**Score Metric: Accuracy**
80% Train and 20% Test Split	0.67	80% Train and 20% Test Split	0.67
70% Train and 30% Test Split	0.922	70% Train and 30% Test Split	0.731
60% Train and 40% Test Split	0.968	60% Train and 40% Test Split	0.733
However, with the Random forest with only 21 features in 70% Train and 30% Test Split the score is 0.935.			

It is noticed here that as the training set decreases the cross- validated mean scores gradually increases in the test set. Similar trends were observed with the Decision Trees and K-NN models. In this report, the split: Training 70% and Testing 30% is used for the feature selection

- **2.1) Feature selection**: Firstly, the Random Forest Feature Importance is found out for all the proteins to find the minimum threshold. The minimum threshold is set to 0.01687021. The top 21 features chosen based on the minimum threshold is as follows: DYRK1A_N, pPKCAB_N, BRAF_N, APP_N, SOD1_N, P38_N, pP70S6_N, pGSK3B_N, pPKCG_N, S6_N, AcetylH3K9_N, ARC_N, Tau_N, Ubiquitin_N, pS6_N, H3AcK18_N, CaNA_N, pERK_N, pNUMB_N, ITSN1_N and pCAMKII_N. Therefore, by performing feature selection the dimensions of the dataset have been reduced from 77 to 21 features. The cross-validated mean score after feature selection is 0.935 which is just 0.033 less while all the features had been selected. Also, in order to have a control on the randomness, a random state of 1 is defined in all the models.

- **2.2) Hyper-parameter Tuning:** It is very important to tune the Hyper-Parameter of the model as it solves the problem of efficiently detecting the set of optimal hyperparameter for a learning algorithm. In this step, a cross validation strategy via a "grid search" is adopted to detect the most optimal hyperparameter values for the machine learning Algorithms. The scoring metric used in this step is accuracy, which searches through all the combinations of hyper parameters in the train set and calculates the accuracy of each model.
A stratified 5- fold cross validation is used where the train set is randomly split in 5 partitions of equal size. While stratifying the data it helps to preserve the ratio of each sample for each class. Then, one partition of the 5 equal parts is used to test the model and the remaining 4 partitions is used to train the data. This process of cross-validating is repeated 5 times using each of the partition as a validation set. Each of the combinations will give a score and then an average of the scores is used to depict the general performance of the models with their respective hyper-parameters.

The most important point is that by using a cross validation approach the chances of having a "lucky-split" is decreased which in turn helps the model to generalize it well on the unseen data. The table below will summarize the input parameters passed in the model and the optimum parameter detected with the help of a grid search.

Classifier	Input Parameters	Optimal Parameters Detected
K-Nearest Neighbor (K-NN)	Number of Neighbors: 1 to 7 P (Distance Used): 1: Manhattan Distance 2: Euclidean Distance	Number of Neighbors = 1 P: 1
Decision Tree (D.T)	Criterion: Gini, Entropy Max Depth:3,4,5,6,7,10,12 Minimum Samples Split:2,5,15,20,25	Criterion: Gini Max Depth:10 Minimum Samples Split:2
Random Forest (R.F)	Number of estimators:50,100,200 Max Depth: 3, 4, 5, 8, 10	Number of estimators:200 Max Depth:10

- **2.3) Model Evaluation:** The mice dataset has 8 classes and is considered to be a multinomial classification problem. Therefore, for each of the model used the evaluation is based on the hold-out test data. The average class accuracy is computed because the dataset is considered to be a multinomial classification problem. In the data exploration, we found out that some of the classes were not equally represented thus micro averaging is chosen to be the input parameter to take into account the class imbalance in the dataset. In this step, the micro average F1 Score is considered to evaluate the models because the F1 score is the result of the arithmetic mean of recall and precision.

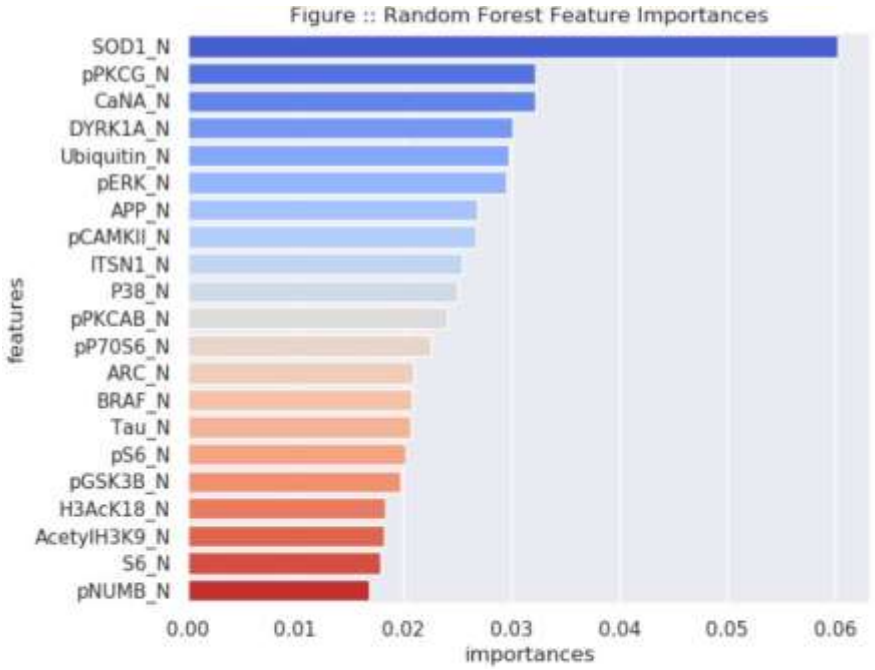
  In the process of evaluating the model on the basis of test set, a powerful approach namely K-Fold Cross validation is adopted to avoid the chances of having a lucky split.

RESULTS

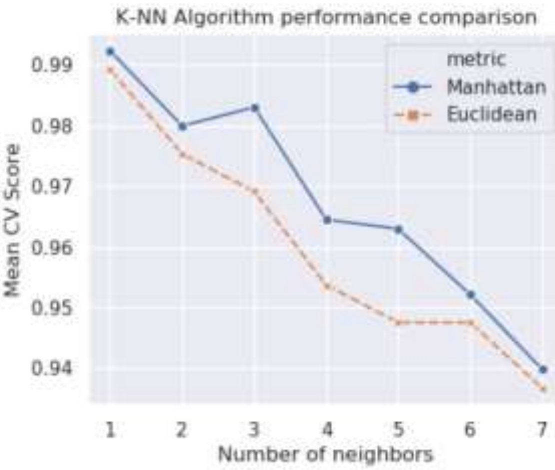
In this section, the results of the following will be discussed elaborately:

A. Feature Selection
B. Hyper-parameter Tuning and
C. Model Evaluation**.**

A. Feature Selection: Based on the Random Forest Classifier 21 features out of 77 were chosen to build the models with K-NN and Decision Tree. The graph below depicts the Random Forest Importance for the 21 proteins in descending order. To make a strong visualization the proteins in the blue color has a higher importance value in comparison to the proteins highlighted in red color.



B. Hyper-parameter Tuning: The table below depicts the result of the Hyper-Parameter Tuning process and the corresponding scores of the various models at a glance.



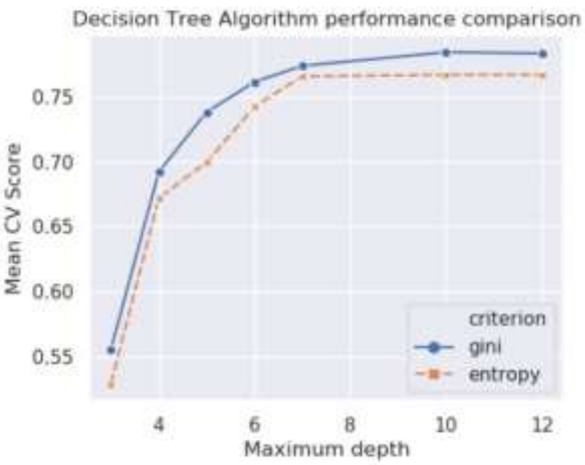
<u>Best Parameters and the corresponding cross validated scores</u>

Number of Neighbors = 1

P: 1(It is seen that Manhattan Distance is a better metric)

It can be noticed that as the number of neighbors increases the mean cross validated scores keeps decreasing.

The highest mean cross validated score is 0.992.



<u>Best Parameters and the corresponding cross validated scores</u>

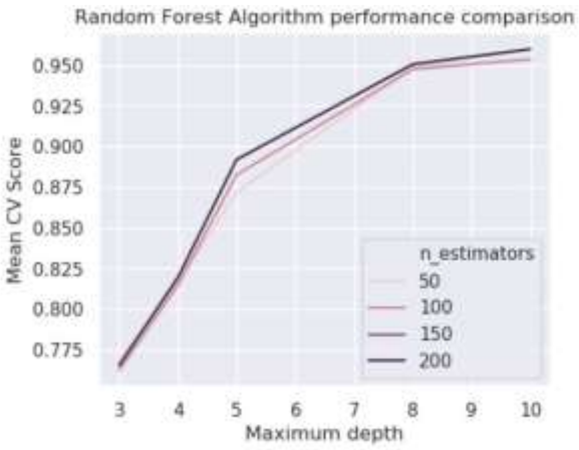
Criterion: Gini

Max Depth:10

Minimum Samples Split:2

It can be noticed that as the maximum depth increases the mean cross validated scores keeps increasing.

The highest mean cross validated score is 0.81.



<u>Best Parameters and the corresponding cross validated scores</u>

Number of estimators:200

Max Depth:10

It can be noticed that as the maximum depth increases the mean cross validated scores keep increasing for all the number of estimators passed as input feature.

The highest mean cross validated score is 0.96.

C. Model Evaluation: The model is evaluated on the basis of

I. Classification Report

II. Cross- Validation

III. Paired T-Test

<u>Classification Report</u>

Evaluation Metrics (Micro Average)	**K-NN**	**D.T**	**R. F**
precision	**1**	**0.86**	**0.97**
recall	**1**	**0.86**	**0.97**
f1-score	**1**	**0.86**	**0.97**

It is seen in the Classification Report that micro average of precision, recall and F1 score is taken into account as it solves the problem of class imbalance in the dataset. In short, Precision is basically the measure of exactness and recall is a measure of completeness. In this step, the micro average F1 Score is considered to evaluate the models because the F1 score is the result of the arithmetic mean of recall and precision. The K-NN classifier has an overall F1 score of 1. It may be the case of overfitting of the model. However, various ratios of splits of the dataset were tried out to avoid the overfitting of the model. The random forest also has a high F1- score of 0.97 followed by the Decision Trees with the F1- score of 0.86.

II. Cross- Validation:

It is important to note here that the cross-validated scores here is based on the 70% training and 30% test split and the scoring metric is ‘balanced accuracy’ which helps us to handle if the class is imbalanced. On the basis of the cross-validated table, K-NN has again out-performed all the models with a score of 0.992 followed by Random Forest with 0.909 and Decision trees with 0.992.

Cross- Validation Table

Models	**Cross- Validated Scores**
K-NN	**0.992**
Decision Trees	**0.766**
Random Forest	**0.909**

III. Model Evaluation: Paired T-Test with a significance level at 0.05.

H0: The accuracy achieved with K-NN is equal to the accuracy achieved with Decision Tree and Random Forest

H1: The accuracy achieved with K-NN is not equal to the accuracy achieved with Decision Tree and Random Forest

K-NN vs Decision Tree: P-value = 0.001

K-NN vs Random Forest: P-value = 0.017

In both the scenario the p-value is lesser than the significance level, meaning that the differences between classifiers performances are statistically significant. Thus, we reject the Null Hypothesis and accept that K-NN is the best classifier among the three.

DISCUSSION

In the above analysis, the mice dataset has 1080 observations and 82 variables. In the Data Exploration step, two main hypotheses were set. Firstly, “Is there any co-relation among the most important proteins that best determines the class of the mice?” It was evident that most of the proteins were moderately correlated. It is also interesting to note that ARC_N and ps6_N were highly co-related. Secondly, seek statistical evidence whether the mean, median and variance is same among the protein expression readings grouped by classes in the mice datasets. It was statistically proved that the median, mean and variance are not same among all the groups for each of the protein.

In the dataset, the target feature is the class of the mouse. There is a total of 77 proteins which was selected as the feature of the model. At first, before selecting the features to build the model, the cross-validated scores were calculated based on the different ratios of train and test split. It was done so that we do not under fit or over fit the model. It was then decided to move forward with the train- test split of 70:30. The random forest Importance was then calculated to set a minimum threshold for all the proteins to be considered in the model. The minimum threshold was set to 0.01687021. The 21 features among the 77 others were based on the random forest classifier. The hyper parameters for the respective models were tuned and the best scores for each of the model was calculated. At last, the models were evaluated based on three approaches namely, classification report, cross-validation scores and a Paired T-Test. All the approaches confirmed that K-NN would be the best classifiers to predict the class of the mouse based on the protein reading.

CONCLUSION

In this report a strong focus has been given to Data Exploration in order to have a strong understanding of the dataset before modelling it. Meaningful graphs have been plotted which has helped to answer the research question and the hypothesis with a statistical evidence. However, while modelling the data it was really hard to dig into depths due to the lack of expertise in the domain knowledge. Also, after reading a lot of research reports it was evident that clustering could also help to predict the class of the data. In the future, more powerful algorithms and clustering approaches could be adopted to model the data.

REFERENCES

Aksakalli, V., 2020. *Feature Ranking.* [Online]
Available at: https://www.featureranking.com/[Accessed 8 June 2020].

CPL, 2020. *CPL.* [Online] Available at: https://www.cpl.org.au/resources/understanding-disability/what-is-down-syndrome?gclid=Cj0KCQjww_f2BRC-ARIsAP3zarHe5OD06KeF_k4lvCKDSWWtVmsuGOAige0jmXUaSPo808gACNepAVsaAnwdEALw_wcB[Accessed 8 June 2020].

Rachmadita Andeswari, A. M. Z. S., 2020. *International Journal of Integrated Engineering.* [Online]
Available at: https://publisher.uthm.edu.my/ojs/index.php/ijie/article/view/2779
[Accessed 8 June 2020].