MICE PROTEIN

Data Science Project

MICE PROTEIN EXPRESSION DATA SET

GOAL

Our goal is to predict the class of the mice out of 8 characters in the class column using the protein combination measurements dataset.

CHOSEN DATA SET

- The data set consists of the expression levels of 77 proteins/protein modifications that produced detectable signals in the nuclear fraction of cortex.
- There are 38 control mice and 34 trisomic mice (Down syndrome), for a total of 72 mice. In the experiments, 15 measurements were registered of each protein per sample/mouse.
- The dataset contains a total of 1080 measurements per protein. Each measurement can be considered as an independent sample/mouse.
- The eight classes of mice are described based on features such as genotype, behavior and treatment.
- According to genotype, mice can be control or trisomic.
- According to behavior, some mice have been stimulated to learn (context-shock) and others have not (shock-context) and in order to assess the effect of the drug memantine in recovering the ability to learn in trisomic mice, some mice have been injected with the drug and others have not

DATA PREPARATION

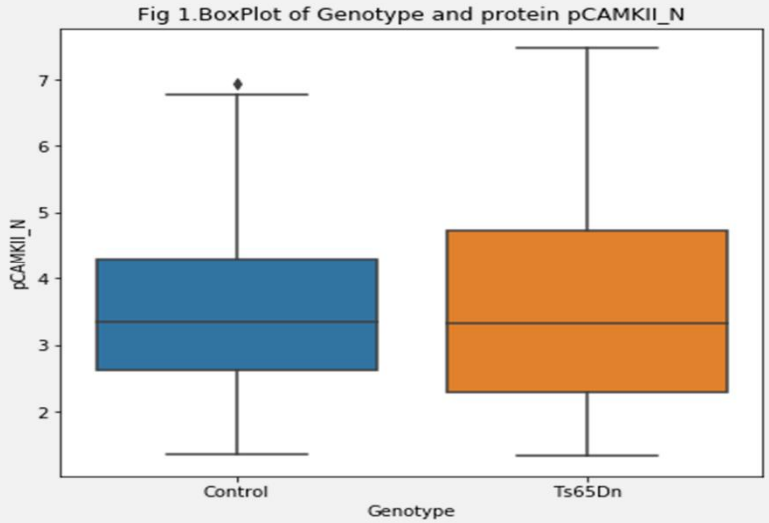
- **Data_Cortex_Nuclear.csv** file is imported from the UCL machine learning repository for the mice protein expression using the function **read_csv()**. It is separated by "," as it is comma separated value file.
- It has total of 82 Columns and 1080 Rows in the entire dataset.
- Later the data types of all attributes are obtained using the function dtypes().
- Removing the ID like Columns which is "MouseID" in the data set as it is not useful for our analysis.
- Now checking the number of missing values in each column by using the function isna().sum().
- So, for all the missing values we are replacing it with the `**mean**` of that corresponding column to the data frame for the further analysis.
- All types of Sanity checks, typos errors, whitespaces checks have been done.
- Once all the checks have been done after missing values have been replaced now the data is ready for further process.

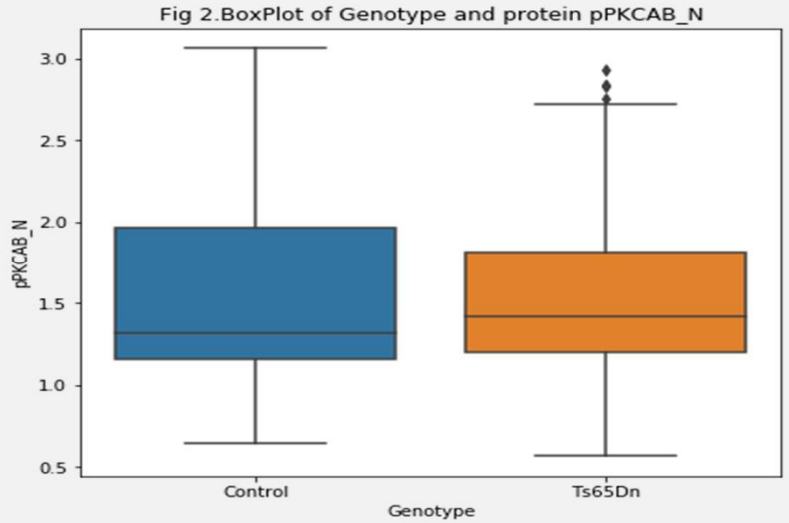
- The columns ‘Genotype', 'Treatment', 'Behaviour’ been removed from the dataset as they are redundant
- Target column has been encoded manually
- c-CS-m': 0, 'c-SC-m': 1,
  c-CS-s': 2, 'c-SC-s': 3,
  t-CS-m': 4, 't-SC-m': 5,
  t-CS-s': 6, 't-SC-s': 7

HYPOTHESIS QUESTIONS

Hypothesis 1:

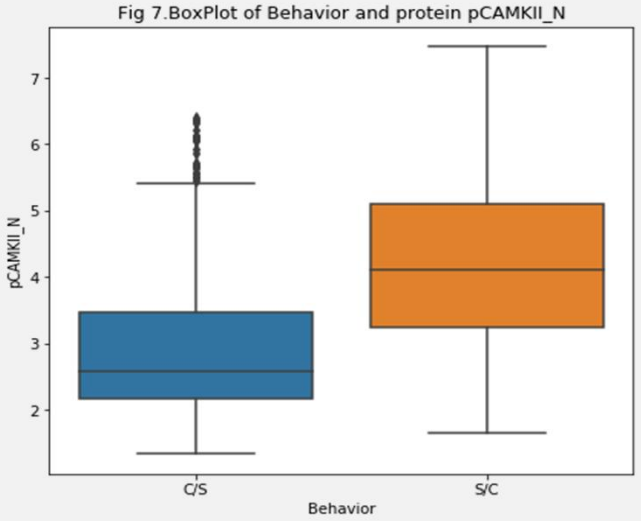
- GENOTYPE of the mice depends on Proteins. As the protein combination changes Genotype also changes.

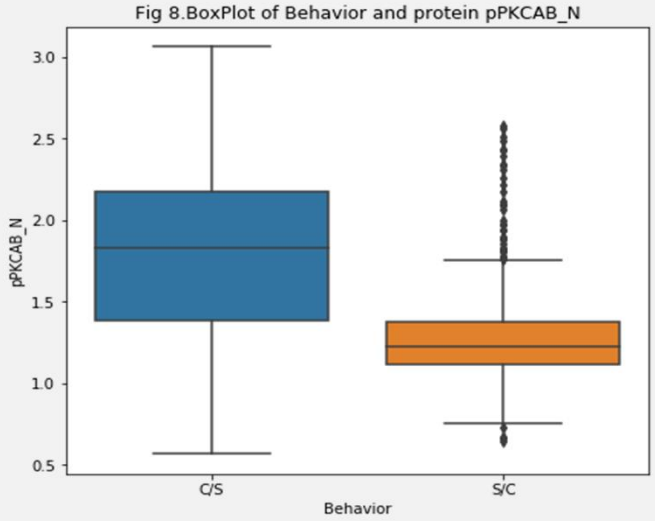




From the above two bar graphs we can say that Protein level impacts the Genotype hence, stated hypothesis is accepted as there is enough evidence.

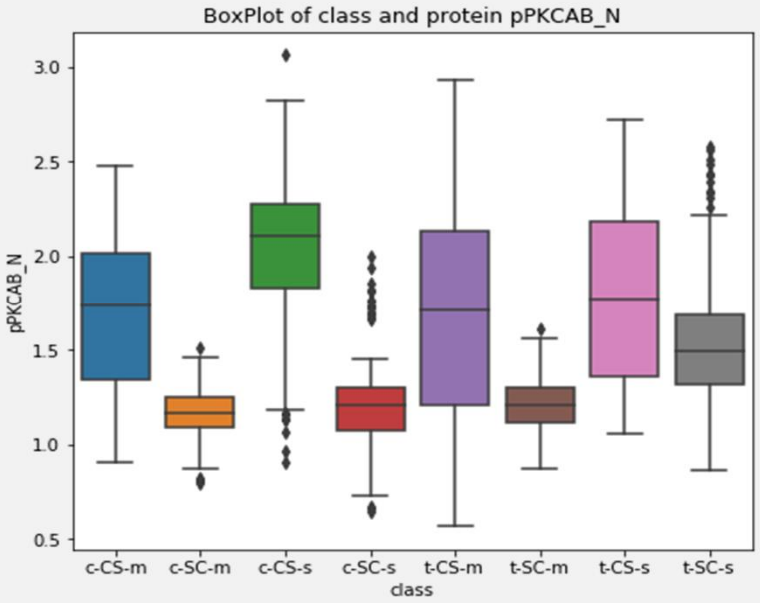
- Hypothesis 2: Behaviour of the mice depends on proteins.

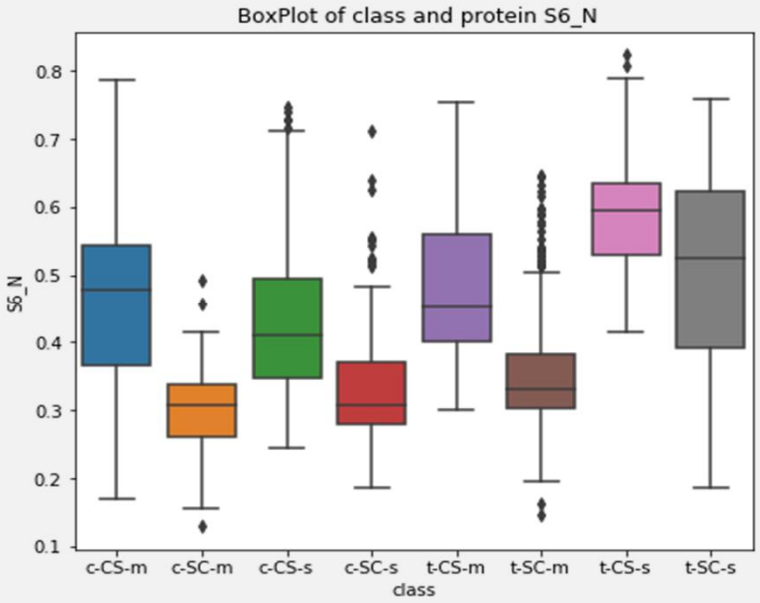




From the above two bar graphs we can say that Protein level impacts the Behaviour of mice. Hence, stated hypothesis is accepted as there is enough evidence.

- **Hypothesis 3**: Class of the mice does not depend on the proteins.





- From the above class level bar graph on two different protein shows that it does depend on the protein and hence we reject the Hypothesis.

MODELLING STEPS

- For this Classification problem we used two Classifiers:
- KNN and Decision Tree
- Once the pre-processing of data is done now all categorical variables are encoded for fitting to the model.
- After that data is scaled to maintain the uniformity among all columns.
- The data is splitted in to test and train data.
- We used 40% ratio of test data and 60% ratio for the train data.
- Feature selection of the data is done to get the maximum accuracy with minimum number of features.
- Train data of the target and the dataset is fitted to the model with default parameters.
- Then the model is tuned for the optimum parameters to get the highest accuracy.

RESULTS

KNN Model's Classification report

```
print(classification_report(y_test,y_pre))
```

	precision	recall	f1-score	support
c-CS-m	1.00	1.00	1.00	58
c-CS-s	1.00	0.98	0.99	53
c-SC-m	1.00	0.98	0.99	62
c-SC-s	0.98	1.00	0.99	58
t-CS-m	0.98	1.00	0.99	59
t-CS-s	1.00	1.00	1.00	43
t-SC-m	1.00	1.00	1.00	46
t-SC-s	1.00	1.00	1.00	53
accuracy			1.00	432
macro avg	1.00	1.00	1.00	432
weighted avg	1.00	1.00	1.00	432

- Decision Tree's Classification report

	precision	recall	f1-score	support
c-CS-m	0.78	0.86	0.82	58
c-CS-s	0.73	0.68	0.71	53
c-SC-m	0.85	0.89	0.87	62
c-SC-s	0.84	0.90	0.87	58
t-CS-m	0.84	0.83	0.84	59
t-CS-s	0.82	0.77	0.80	43
t-SC-m	0.89	0.87	0.88	46
t-SC-s	0.90	0.83	0.86	53
accuracy			0.83	432
macro avg	0.83	0.83	0.83	432
weighted avg	0.83	0.83	0.83	432

Model	Accuracy
Default KNN	94%
Tuned KNN	99%
Feature Selected Default KNN	94%
Feature Selected Tuned KNN	99%

Model	Accuracy
Default Decision Tree	78%
Tuned Decision Tree	81%
Feature Selected Default Decision Tree	80%
Feature Selected Tuned Decision Tree	83%

CONCLUSION AND RECOMMENDATIONS

- Taking into consideration the performance of both classifiers it can be concluded that KNN is giving the highest accuracy with all 33 features i.e. 99%. It can be noted that Decision Tree with 27 features giving efficiency of 83%. Out of 77 features KNN model is giving highest accuracy considering only 33 features is the best model we can use to classify the class of the 8 different mice.

THANK YOU