JUNE 10, 2020

Student ID: s3827495
Student Name: Kaushik Sunil Anagarkar

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 honour code by typing "Yes": *Yes*.

ASSIGNMENT 2 PRACTICAL DATA SCIENCE WITH PYTHON

DATA MODELLING AND PRESENTATION

KAUSHIK SUNIL ANAGARKAR
MASTER OF DATA SCIENCE,RMIT UNIVERSITY
S3827495@STUDENT.RMIT.EDU.AU
+61 424973698

Contents

Abstract:

In this assignment, we are analysing the dataset based on Expression levels of 77 proteins measured in the cerebral cortex of 8 classes of control and Down syndrome mice exposed to context fear conditioning, a task used to assess associative learning. Based on this dataset, we carried out the steps of the data science process, including the cleaning, exploring, modelling of data using classification models. The aim of the project is to identify subsets of proteins that are discriminant between the classes. In this report we discuss how to predict subsets of proteins using the data modelling techniques which are belonging to specific class and recommending the best model to be used for identifying subsets of proteins that are discriminant between the classes. Based on the protein levels we can predict the mouse is down syndrome.

Introduction:

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. Based on these genotype, behavior and treatment there are eight different classes produced. In this report we discuss the subset of proteins that can identify the class of mice.

Methodology:

Data Collection

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. Therefore, for control mice, there are 38x15, or 570 measurements, and for trisomic mice, there are 34x15, or 510 measurements. The dataset contains a total of 1080 measurements per protein. Each measurement can be considered as an independent sample/mouse.

Data Preparation

Data cleaning is one of the important and fundamental tasks to remove the errors and accuracies in the datasets which could have led to inaccuracy in data modelling.

1.1 Data Retrieving - Imported the Mice dataset using pandas using pd.read_csv() command by providing the accurate file path of the csv file to load in the notebooks .

1.2 Check data types- Printing Mice dataframe and verifying if all the data and the data types of the loaded data are equivalent to the data in the source file(csv file) using dtypes() functions in pandas.

1.3 Typos and Extra-whitespaces - Generate the frequency of the categorical columns in mice dataframe using value_counts() and unique() function ,if there are any data entry error replace the value with replace() function and if any extra whitespaces present remove it using the strip() function.

1.4 Missing Values**-** Checked the missing values present in the dataset using isna() function and also the sum of missing values in all the columns and rows of dataset using isna().sum() function for columns and isna().sum(axis =1) for rows function .Replace the null values in the mice protein dataframe with 0 because the protein levels after performing experiment gives some measurements

replacing null values with 0 means there was no protein for null values ,rather than giving mean/median for no protein values.

1.5 Upper Case – Casted all of text data to Upper case by using lambda function first converting the data type to string and later apply upper () function.

Data Exploration

The aim of data exploration is to get a deep understanding of the data.

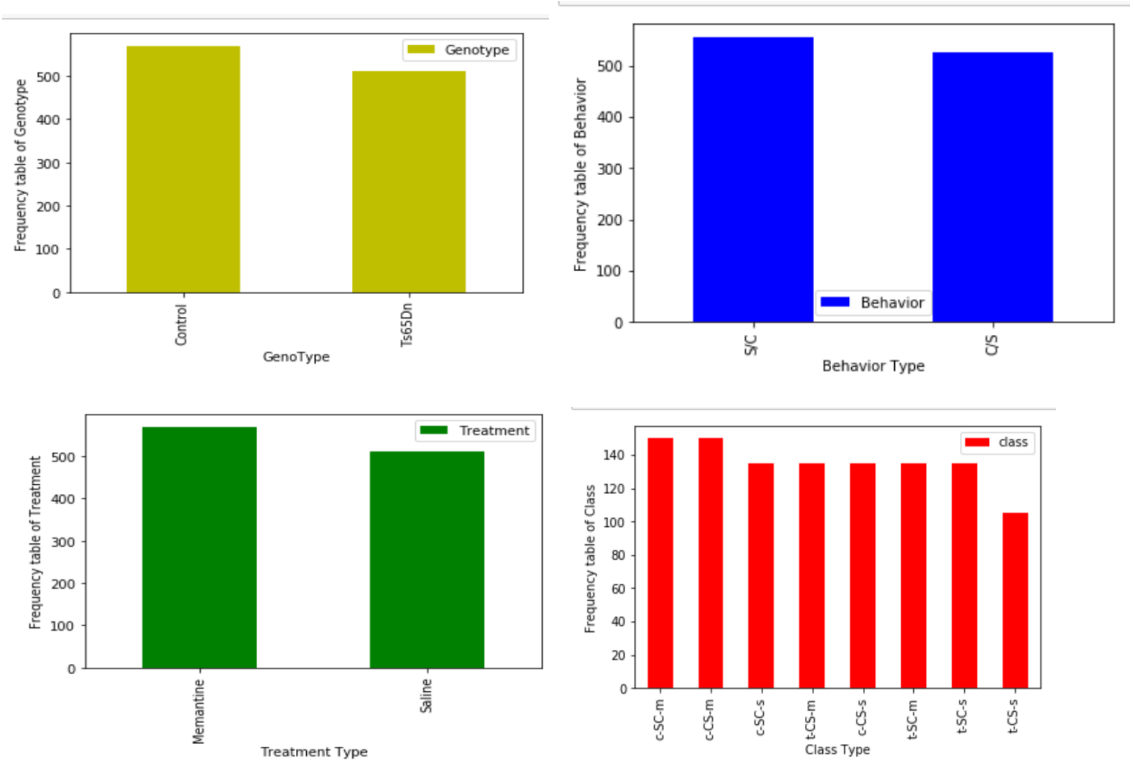
Exploration of the single columns of the Mice dataset:

Genotype Column: After plotting the graph we can state that the count of the mice of control genotype are more when compared to the mice of trisomic genotype.

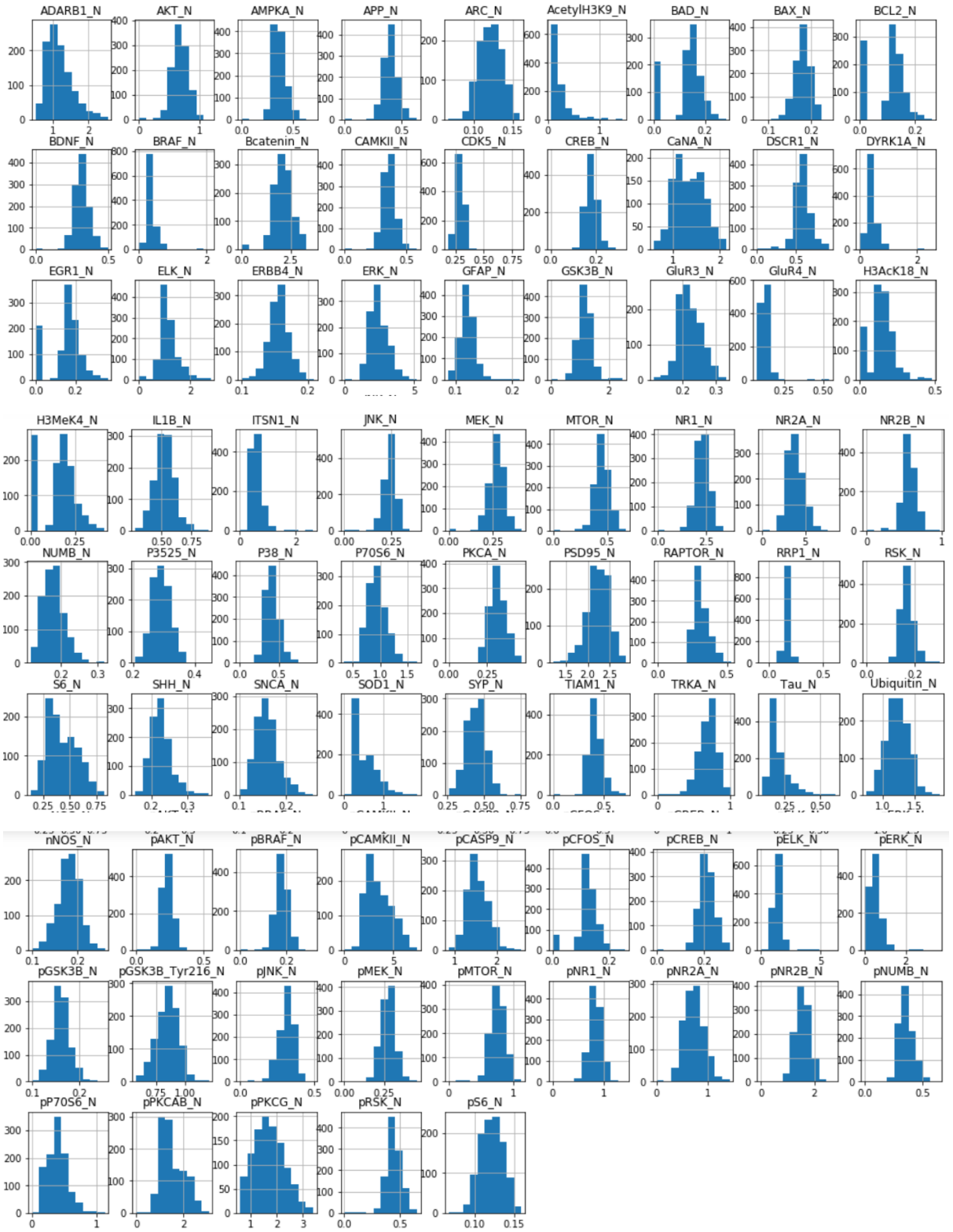
Behavior Column : After plotting the graph we can state that according to behaviour count of mice which were simulated to learn (Context-Shock) were less than those compared to mice's which are not simulated to learn(Shock-Context).

Treatment Column: After plotting the graph we can state that according to treatment count of mice which were injected with mematine drug are more than those injected with saline drug

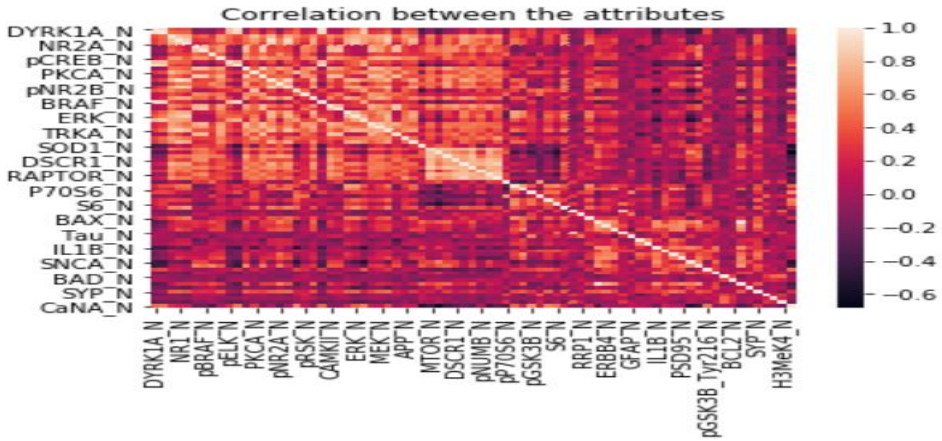
Class column: After plotting the graph we can state that c-CS-s(control mice, stimulated to learn, injected with saline) and c-CS-m (control mice, stimulated to learn, injected with memantine) classes has more count of proteins than compared to other 6.



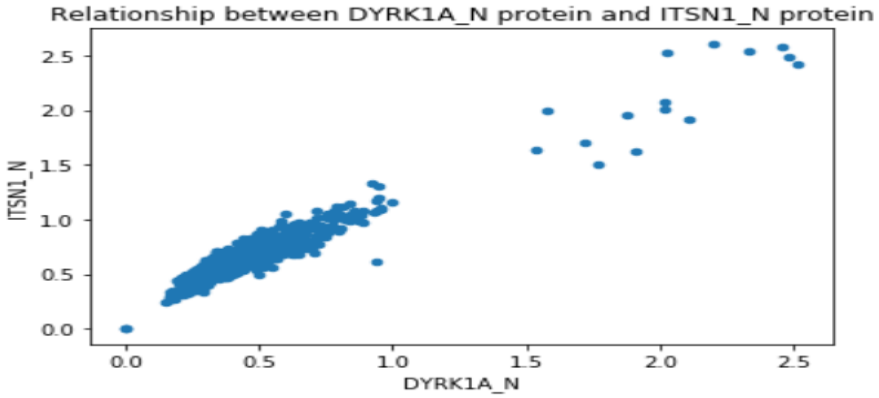
For the exploration of all the 77 expression levels of proteins we visualize them using histogram : The histogram provides the distribution of the all 77 protein expression level and average expression levels of proteins and the range where most of its datapoints lie .



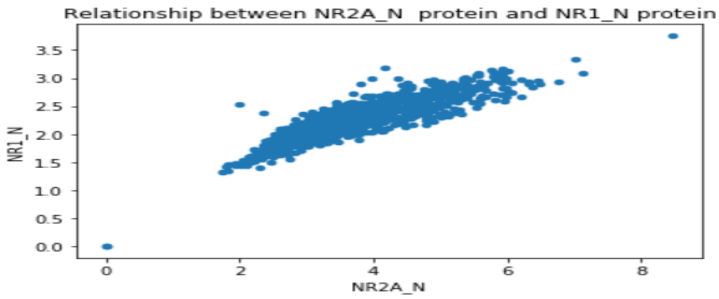
Explore the relationship between attributes of the Mice dataset **:** In order to get the correlation between the 77 protein features we create a correlation matrix and explore it using heatmap() which is imported from seaborn library. But due to large volume of features ,heatmap does not give clear values .



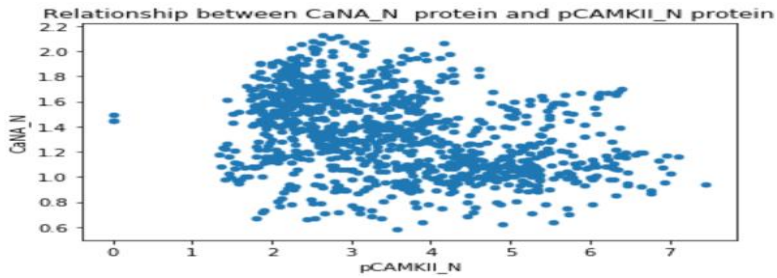
Relationship between DYRK1A_N protein and ITSN1_N protein: From the correlation matrix we can see that expression levels of DYRK1A_N protein and ITSN1_N protein had the positive correlation coefficient of 0.95.After plotting the graph we can state that both the protein levels have the linear relationship between them.



Relationship between NR2A_N protein and NR1_N protein: From the correlation matrix we can see that expression levels of NR2A_N protein and NR1_N protein had the positive correlation oefficient of 0.87. After plotting the graph we can state that both the protein levels have the linear relationship.



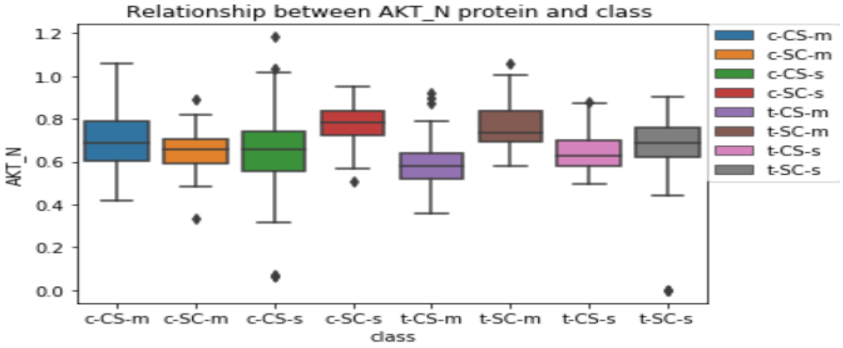
Relationship between NR2A_N protein and NR1_N protein: From the correlation matrix we can see that expression levels of CaNA_N protein and pCAMKII_N protein has the negative correlation coefficient of -0.37.After plotting the graph we can say there is negative relation between them .As the plot is not linear that suggests relationship between two variables is not that strong.



Relationship between Protein expression levels of AKT_N protein and class :

Hypothesis : We can assume that the control mice's which are context-shock(stimulated to learn) injected to saline drug produce more amount of AKT_N protein than mice which are shock-context(not stimulated to learn) injected to saline drug.

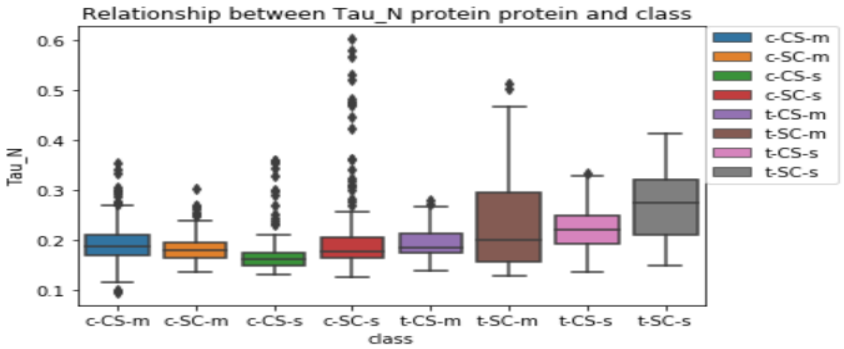
According to graph we can say that more more than 50% of protein level measurements of AKT_N protein belonging to the control mice which was not simulated to learn(Shock-context) injected with saline drug class has relatively high protein measurements than other classes



Relationship between the protein expression levels of Tau_N protein and class:

Hypothesis: We can assume that the trisomy mice's which are shock-context(not stimulated to learn) injected with saline drug produce more amount of Tau_N protein than control mice which are context-shock(stimulated to learn) injected with saline drug

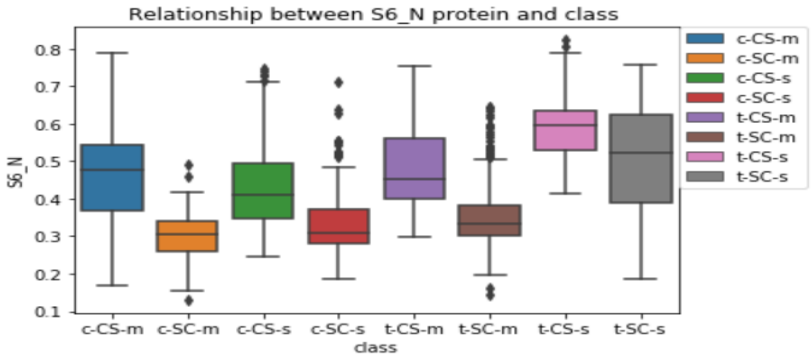
According to graph we can state that 75% of protein level measurements of Tau_N protein belonging to the trisomy mice which was not simulated to learn(Shock-context) injected with saline drug class has relatively high protein measurements than other classes .



Relationship between protein expression levels S6_N protein and class:

Hypothesis: We can assume that trisomy mice's which are context-shock(stimulated to learn) injected to mematine drug produce more amount of S6_N protein than trisomy mice which are context-shock(stimulated to learn) injected to saline drug

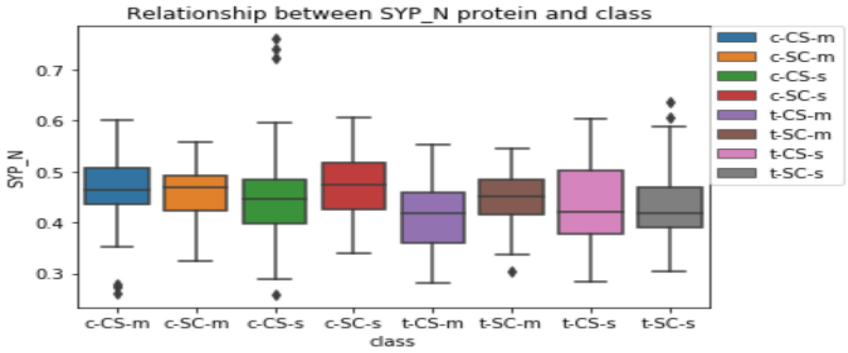
According to graph we can state more than 50% of protein level measurements of S6_N protein belonging to the trisomy mice which was simulated to learn(context-Shock) injected with saline drug class has relatively high protein measurements than other classes



Relationship between protein expression levels S6_N protein and class:

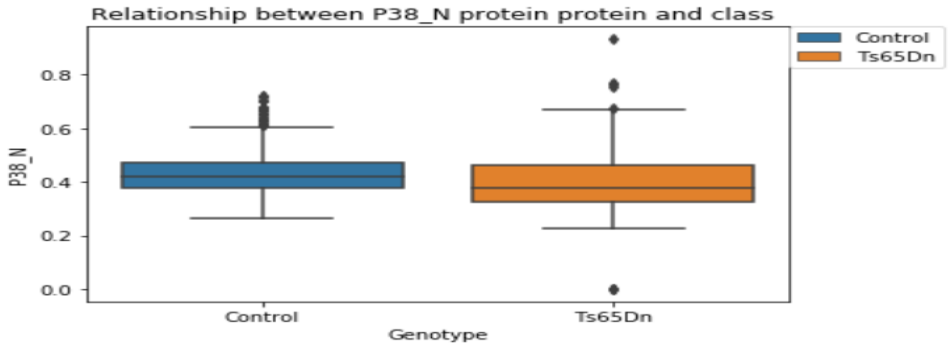
Hypothesis: We can assume that the control mice's which are context-shock(stimulated to learn) injected to mematine drug produce more amount of SYP_N protein than control mice which are context-shock(stimulated to learn) injected to saline drug.

According to graph we can state that more than 50% of protein level measurements of SYP_N protein belonging to the control mice which was not simulated to learn(Shock-context) injected with saline drug class has relatively high protein measurements than other classes .



Relationship between protein expression levels P38_N protein and Genotype:

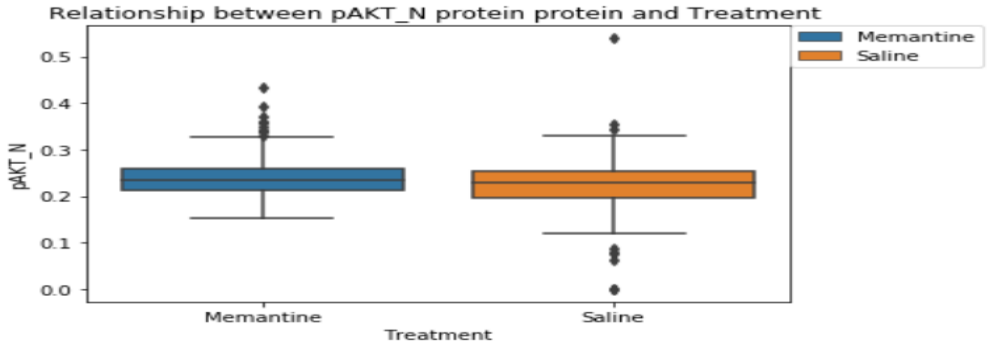
Hypothesis: We can assume that the mice's which are of control genotype tend to produce more amount of P38_N protein than trisomy mice. According to graph we can say that that the more than 50% of the control mice produces 0.5 expression level of P38_N protein and more than 50% of trisomy mice produces 0.4 expression level of P38_N protein. Control mice induces more amount of P38_N protein than trisomy mice.



Relationship between protein expression levels pAKT_N protein and Treatment:

Hypothesis: We can assume that the mice's which are injected with mematine drug tend to produce more amount of pAKT_N protein than the mice's injected with saline drug .

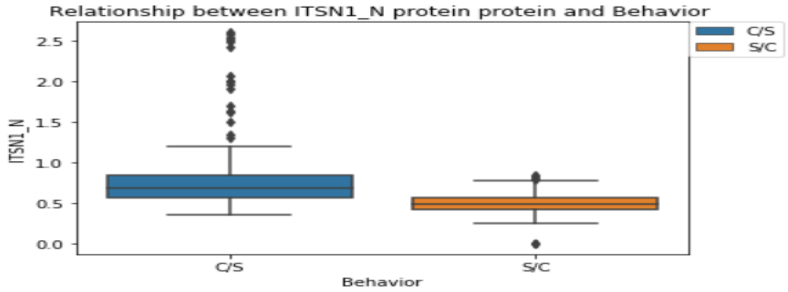
According to graph we can say that he more than 50% of the mice injected with memantine drug induce more than 0.25 expression level of pAKT_N protein and more than 50% of trisomy mice produces around 0.25 expression level of P38_N protein.Mices injected with Mematine drug produces more pAKT_N protein than mices injected with saline drug



Relationship between protein expression levels ITSN1_N protein and Behavior:

Hypothesis: We can assume that the mices which are context shock (stimulated to learn) tend to produce more amount of ITSN1_N protein than the mice's which are shock context(not stimulated to learn).

According the graph we can say that the more than 50% of the mice which are context shock induce more than 0.75 expression level of ITSN1_N protein and more than 50% of mice produces which are shock context induces around 0.25 expression level of ITSN1_N protein. Mice's which are context-shock produce more amount of ITSN1_N than mice's which are shock-context



As we have large volume of features (77 protein features),it is very hard to extract relationship between the attributes as most of the data are numerical data.

Data Modelling

Data Modelling is the process of extracting best predictors and better model selection as per requirements, training the model, validating the model using several techniques and applying the trained model to the unused .

Feature engineering and Model Selection :In order to get the best features among the 77 protein features , we use simple hill climbing technique to extract the best features for the model , firstly we create a new empty list to add the features which will be extracted from the technique then evaluate the initial first feature of protein list , and make initial model score as current score , loop until there are no new features present which can be applied to current state. Select the other features randomly which can be applied to current state and produce new model score .If the currrent score is less than goal score then remove the feature from the new list, if it is greater keep in the list .After performing simple hill climbing technique we extracted 12 features from the 77 protein feature dataframe which provided the good model score(accuracy).Create a new dataframe protein_feature which consists of the 12 features extracted .(Bcatenin_N', 'NUMB_N', 'JNK_N', 'GFAP_N', 'pERK_N', 'ARC_N', 'P38_N','BDNF_N', 'pPKCG_N', 'CaNA_N', 'SOD1_N', 'AcetylH3K9_N').

Training the Model : From sklearn import the test_train_split package which splits the features data containing the 12 protein features extracted from the simple hill climbing technique and target data containing the class data into Training and Testing sets with test size as 20% .Import the Decision tree classifier and fit the training data of features and target variable to the classifier. Labels of target variable are the eight classes (c-CS-s, c-CS-m, c-SC-s, c-SC-m, t-CS-s, t-CS-m, t-SC-s, t-SC-m) . In order to increase the predictive power of model, decrease the classification error rate and avoid Overfitting we give right set of tree constraints to Decision tree train the model. The next step is import the K-Nearest Neighbour classifier and fit the training data of features and target variable to the classifier by providing the p-value and weight parameter train the data. The parameters in the KNN model are used to increase the accuracy of the model and decrease the classification error rate .

Model Validation and Selection : We can use k-folds cross validation by taking the value of k between 3-10 and splitting the data into test and train for k-times and checking the accuracy score for all the data values to check how well the model is predicting on test data .In this assignment we are taking k value to be 5 and splitting the feature and target data into train and test for 5 -times and checking the accuracy of the data .When we had validated the Decision tree model ,for k=5 we had got the maximum accuracy score of 68.9% and the maximum accuracy score of K -Nearest Neighbour model was 97.2%.

Applying the trained model to unseen future data: Input variable is test data of 12 features of proteins and Output variable is Predicted data of 'Class' of type object. Apply the test data to the model to predict the unused data . By Implementing confusion matrix we can compare the data of the actual value and predicted value and validate the actual data to be predicted true .By Implementing classification report of the actual data and predicted value we can calculate the precision of the model. Precision of Decision tree model on an average was found to be 72% and the accuracy of the model was found to be 66.7% . Precision of K-Nearest Neighbour model on an average was found to be 95% and the accuracy of the model was found to be 94.4% .

We can recommend KNearest Neighbor Classifiers over the Decision tree Classifier as the accuracy score and the classification report of the KNN model is higher than that of Decision tree model.

Results :

The DYRK1A_N protein and ITSN1_N protein features have a positive linear relationship between them with correlation coefficient of 0.95 .

Results from the Simple hill climbing technique:

'Bcatenin_N', 'NUMB_N', 'JNK_N', 'GFAP_N', 'pERK_N', 'ARC_N', 'P38_N','BDNF_N', 'pPKCG_N', 'CaNA_N', 'SOD1_N', 'AcetylH3K9_N' are the extracted features from the simple hill climbing technique .

Feature Engineering	**Count of features before Simple hill Climbing technique**	**Count of Features after Simple hill Climbing technique**
Simple hill climbing technique	77	12

Results from the Classifiers:

The precision and accuracy score of the Decision tree model and KNN model in predicting the subset of proteins that are discriminant between the classes are as follows .

Models	**Precision**	**Accuracy score**
Decision Tree Model	0.72	0.667
K-Nearest Neighbour Model	0.95	0.94

Discussion:

Data Modelling technique helps us in predicting the subset of proteins that are discriminant between the classes. During the data collection we had around 77 protein features, using the simple hill climbing technique we got the best predictor features which will increase the accuracy of the model using that 12 features were extracted from 77 features. Next step is splitting the 12 protein features and class data into train and test data. Then, training the model using the Decision tree classifier with the right parameters and fitting the Decision tree model. We train one more model with K-Nearest Neighbour classifiers with weight parameter and p-value and fitting the model .Using the k-fold validation ,split the train and test data into k -times and check the score of the Decision tree model and K-Nearest Neighbour model. Applying the trained model to unused test data to predict the classes .Both the models predicted well with Decision tree predicting the subsets of proteins with 66.67% accuracy and KNN model predicting the subset of proteins with 94% accuracy. Based on the protein levels we can predict the if the mices are down syndrome.

Conclusion:

The model predicts 94% accurately the subset of proteins that are discriminant between the classes using the K-Nearest Neighbor model. Since most of the data in the mice data was numerical ,we can state that K-Nearest Neighbour model works better than the Decision tree as the

accuracy score for KNN model is better than Decision tree model and we can choose KNN over Decision tree .

References :

pandas.DataFrame.hist — pandas 1.0.4 documentation ***pandas.DataFrame.hist — pandas 1.0.4 documentation* (2020). Available at: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.hist.html (Accessed: 10 June 2020).**

seaborn.boxplot — seaborn 0.10.1 documentation ***seaborn.boxplot — seaborn 0.10.1 documentation* (2020). Available at: https://seaborn.pydata.org/generated/seaborn.boxplot.html (Accessed: 10 June 2020).**

seaborn.heatmap — seaborn 0.10.1 documentation *seaborn.heatmap — seaborn 0.10.1 documentation* (2020). Available at: https://seaborn.pydata.org/generated/seaborn.heatmap.html (Accessed: 10 June 2020).

pandas.DataFrame.corr — pandas 1.0.4 documentation

pandas.DataFrame.corr — pandas 1.0.4 documentation (2020). Available at: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html (Accessed: 10 June 2020).

sklearn.tree.DecisionTreeClassifier — scikit-learn 0.23.1 documentation

sklearn.tree.DecisionTreeClassifier — scikit-learn 0.23.1 documentation (2020). Available at: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html (Accessed: 10 June 2020).

sklearn.neighbors.KNeighborsClassifier — scikit-learn 0.23.1 documentation

sklearn.neighbors.KNeighborsClassifier — scikit-learn 0.23.1 documentation (2020). Available at: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html (Accessed: 10 June 2020).