Student ID: S3823274

Student Name: Khushbu Manojkumar Patel

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*.

PRACTICAL DATA SCIENCE WITH PYTHON

(COSC2670)

PROJECT REPORT BY

KHUSHBU MANOJKUMAR PATEL

MASTER OF DATA SCIENCE

RMIT UNIVERSITY

MELBOURNE, AUSTRALIA

S3823274

S382327@STUDENT.RMIT.EDU.AU

JUNE - 2020

Contents

Abstract

Down syndrome is an adverse condition prevailing worldwide. As stated by World health Organization, approximately 1 in 1000 babies tend to suffer from this condition. According to the researchers, error in cell division commonly referred as 'nondisjunction' is the main cause for this condition. This nondisjunction results in an additional partial or whole copy of chromosome 21. It was observed that the chances of conceiving a child with down syndrome increases from 1 among 350 to 1 in 30 from age 35 to 45. Since postponing parenting is becoming more and more common among couples, this condition is expected to increase in coming years. In this report I use the *Mice Protein Expression Data set* provided by UCI. It consists of 77 protein expression levels for control and trisomic mice (suffering from Down syndrome). This report aims to analyse the effect of drug memantine on the trisomic mice and identify a set of proteins which can be used to differentiate between classes of control and trisomic mice.

Introduction

Down syndrome is one of the most common chromosomal condition. It identified among 1 in every 700 babies in the United States as stated by the Centres for Disease Control and Prevention. A normal human being consists of a usual 46 chromosomes in each cell. However, research observed that 47 chromosomes were present in each cell of an individual suffering from Down syndrome. It was identified that an extra partial or entire copy of chromosome 21 results in the characteristics associated with this condition. Characteristics of a human suffering from down syndrome varies from person to person however, a short neck, a flattened bridge of nose, upward slanting eyes, etc are some of the physical traits associated with Down syndrome. Mainly there are three types of Down syndromes amongst which *Trisomy 21* accounts for almost 95% of the cases. There is no standard treatment to cure Down syndrome and each treatment is based on the individual's physical and intellectual needs. In this report I analysis the *Mice Protein Expression Data set* provided by UCI. This data set consists of 77 protein's expression level measured in nuclear fraction. The expression levels were recorded on 72 mice out of which 34 were trisomic mice (Down syndrome) and 38 were control mice. The expression levels were recorded 15 times for a single mouse there by making a total of 1080 independent observations i.e. (15 X 38) + (15 X 34) = 1080. The 72 mice are further divided into eight sub classes based on their genotype, behaviour, and treatment. According to the genotype a mouse can be control (not suffering from Down syndrome) and trisomic (suffering from Down syndrome). Behaviour indicates whether the mouse has been simulated to learn (context- shock) or not simulated to learn (shock-context). Treatment determines if the mouse is injected with drug Memantine or Saline. Memantine is a medication used to treat moderate to severe Alzheimer's patients. Whereas saline which is mixture of Sodium chloride in water is used to dilute the effect of other medications injected. The main aim of the analysis is to identify a subset of proteins that can be used to differentiate between eight classes of mice. As classes to be predicted are categorical, a supervised machine learning classification algorithm is used to predict the target classes.

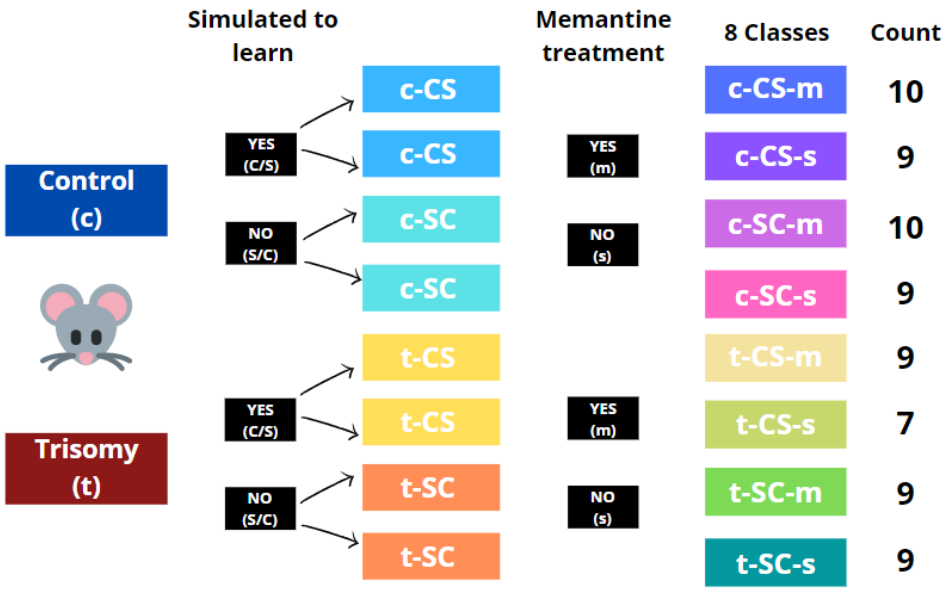
Methodology

Know your data

Before analysing the effects of different proteins on different classes of mice it is important to know what attributes are present in the data set and what they mean. The data set contains 82 attributes in total, comprising of MouseNo, 77 protein expressions measured in nuclear fraction, Genotype, Treatment, Behaviour and Class. The following image shows the of sample of 3 observations.

	Mouse_ID	TestNo	DYRK1A_N	ITSN1_N	...	pCFOS_N	SYP_N	H3AcK18_N	EGR1_N	H3MeK4_N	CaNA_N	Genotype	Treatment	Behavior	class
0	309	1	0.503644	0.747193	...	0.108336	0.427099	0.114783	0.131790	0.128186	1.675652	Control	Memantine	C/S	c-CS-m
1	309	2	0.514617	0.689064	...	0.104315	0.441581	0.111974	0.135103	0.131119	1.743610	Control	Memantine	C/S	c-CS-m
2	309	3	0.509183	0.730247	...	0.106219	0.435777	0.111883	0.133362	0.127431	1.926427	Control	Memantine	C/S	c-CS-m

The below image shows the count and classification of different mice based on their genotype, treatment, and behaviour.



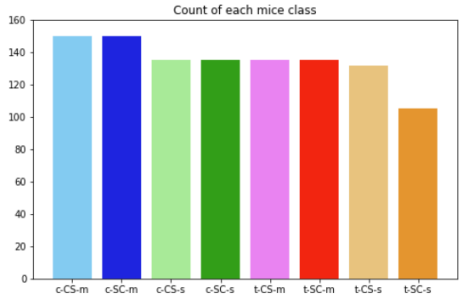
Data Preparation

To perform the analysis and modelling part to identify the set of proteins which can discriminate between different classes of mice it is important to load the dataset properly and prepare it for further analysis. The column MouseNo in original data set has been divided into Mouse_ID and TestNo to perform analysis on individual mouse. The missing values present in the data set are visualized using heatmap in the following image. We can observe that the last few protein columns contain high amount of missing values whereas a thin horizontal line indicates that few rows have missing values for majority number of proteins. All these columns and three rows were dropped as replacing such large number of NAs with any statistical methods could hamper the results. However, the 2 dots shown in the figure were replaced with mean values of that protein in that mice class.

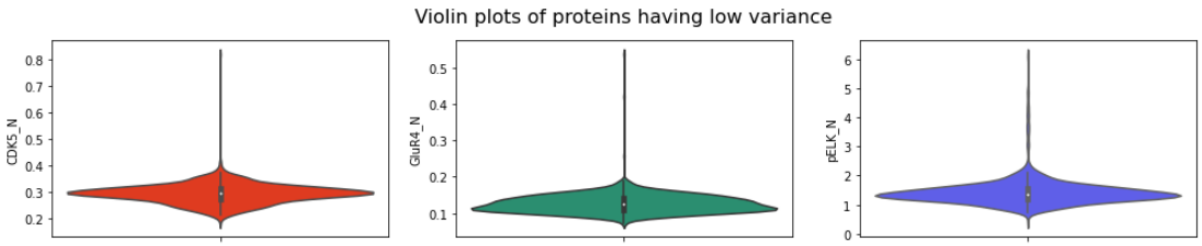
Data Exploration

To identify the subset of proteins that determine the class of mice, it is important to analyse the descriptive statistics of each protein expression levels. The descriptive statistics gives us a broad summary view of entire data set and different visualizations help us identify important patterns concerning the data. After analysing all the protein columns using summary statistics and visualizations it turned out that some features might be more useful than others whereas indicating some features that might be of no use.

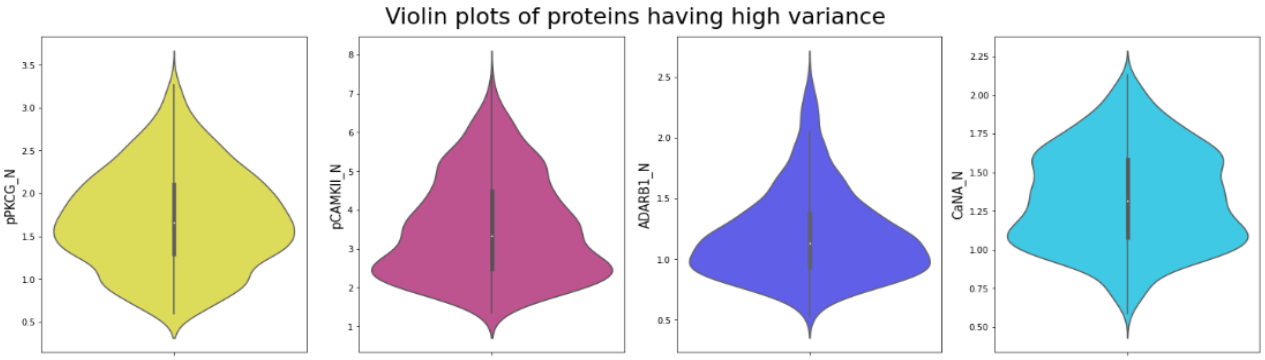
The below bar graph shows the count of each mice subclass. It is clear from the bar chart that highest No. of samples were collected for two mice classes i.e. c-CS-s and c-SC-m. On contrary, 75 samples were collected of mice belonging to t-CS-s class there by indicating lowest sample size among the other mice classes.



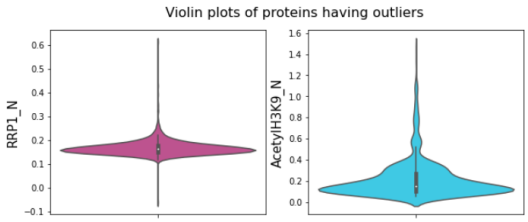
The columns CDK_N, GLuR4_N, and pELK_N were the proteins that showed least variance when compared to other protein expression levels. The protein expression level of all the three columns are visualized using violin plots which serve as a combination of box plot and histogram. As we can see from figure CDK5_N ranges from 0.2 to 0.4 for all the classes of mice. GluR4_N ranges merely between slightly lesser than 0.1 to slight greater 0.2. Apart from that pELK_N ranges from 1 to 2 excluding the outliers. Low variance proteins indicate that they might not be much useful in determining the class of mice as they vary very less will not be able to distinguish between different classes.



In contrast to the above proteins, the expression levels of pCAMKII_N protein (magenta color) showed the highest variance compared to all other proteins among different subclasses of mice. It is visible in the below violin plot that the protein levels vary from 0.5 to as far as 8. In addition to that, the expression level of pPKCG_N, ADARB1_N, and CaNA_N also showed significant amount of variance as compared to other protein expression levels. Proteins having high variance can be useful in distinguishing mice belonging to different classes.

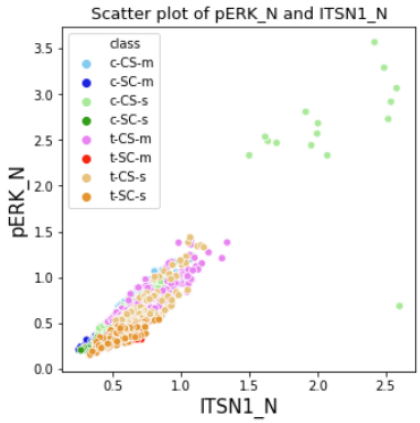


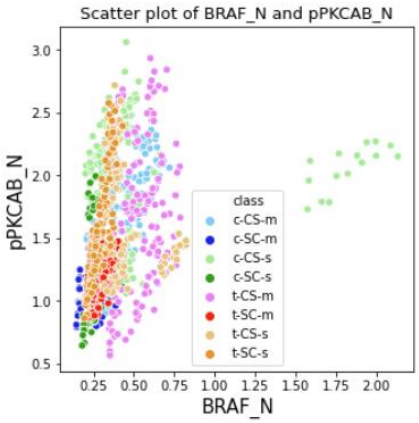
Out of all proteins only RRP1_N protein expression levels showcased a negative minimum value. On further analysis it appeared that only single row contained a negative value there by indicating it might be a typing error. However, as there was no information regarding the absence of negative values it was left unchanged. Apart from that it also shows very less variance. The distribution and spread are shown in the below violin plot. AcetylH3K9_N showcases significant no of outliers.



ITSN1_N – Perk_N

"*Protein levels of ITSN1_N increases along with an increase in protein levels of Perk_N and vice versa.*". The protein levels for both the proteins ITSN1_N and Perk_N range in 0.2 to 1.5 nuclear fraction. The scatter plot visualization shows that both the proteins share a positive linear relationship with each other i.e. if one protein level increases so does the other. However, few green dots indicate the outliers. All the outliers are of single mouse of class c-CS-s (control mice, simulated to learn, saline treatment). Thus, we can that this plot supports the above-mentioned hypothesis.



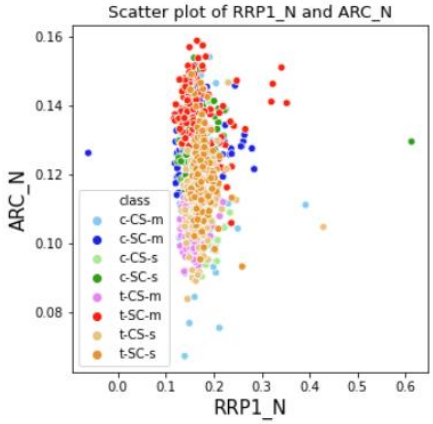


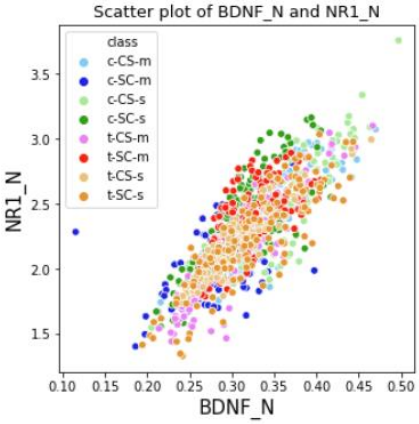
BRAF_N – pPKCAB_N

"Protein levels of BRAF_N tend to remain constant with increase in pPKCAB_N levels." It is observed that protein levels of BRAF_N remains almost constant with increase in protein levels of pPKCAB_N. The protein levels of BRAF_N ranges in between 0.2 to 0.80 whereas that of pPKCAB_N ranges from 0.6 to 3.2. Few green dots on the right (~15 samples of same mouse) of class c-CS-s represents high level of BRAF_N i.e. almost double than that of other mice there by indicating outliers. Also, the above plot indicates that the scatter groups are slightly inclined towards the left. Thus, we cannot ensure that this hypothesis holds for the given data set.

RRP1_N – ARC_N

"Protein ARC_N and RRP1_N are not much dependent on each other". We can say that this hypothesis holds on the given dataset by looking at the below scatter plot on the left which shows that the expression levels of RRP1_N almost remains constant across small changes in ARC_N levels.



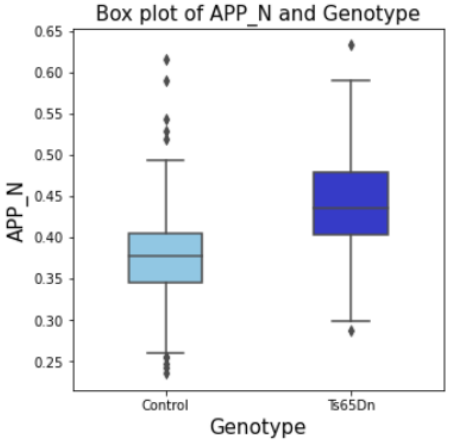


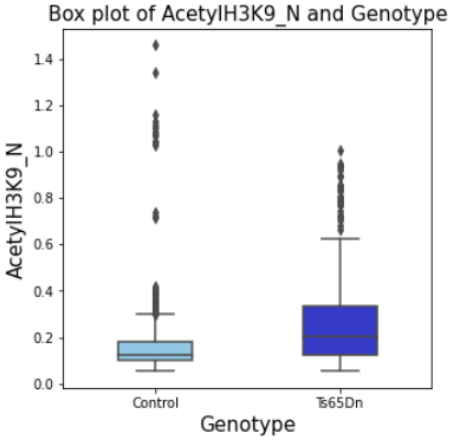
BDNF_N - NR1_N

"Small increase in BDNF_N protein level showcases higher increase in NR1_N protein expression levels". Protein expression levels of BDNF_N approximately ranges between 0.2 to 0.47 whereas the expression levels of NR1_N lies approximately between 1.4 to 3.4 in nuclear fraction. It is evident from the above plot on the right that both these proteins share a positive linear relationship. Also, there are hardly any outliers present for this relationship.

APP_N – genotype

"APP_N shows higher expression levels in trisomic mice than that of in control mice." This hypothesis is justified by visualising the APP_N protein expression levels using box plot grouped by Genotype. It is clear from the plot that the inter quartile range of the trisomic mice is higher than that of control mice. Hence, APP_N might seem useful in differentiating mice based on the genotypes.



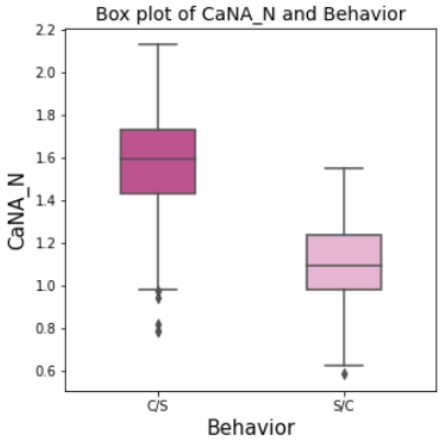


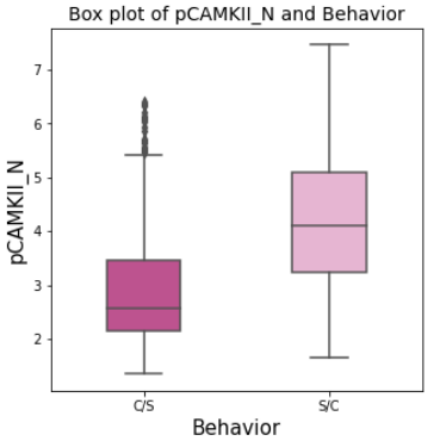
AcetylH3K9_N and genotype

The protein expression levels of AcetlyH3KP_N are smaller in magnitude with low variance for control mice. On the other hand, the trisomy mice have higher range of this protein levels. With this we can state the hypothesis that *"Trisomy mice tend to have higher levels of AcetlyH3K9_N than that of control mice"*. However, the no of outliers present in control mice for this protein expression are more as compared to that of trisomy mice hence there might be situations where this hypothesis might hold. It is visualized in the above box plot on right which is grouped by Genotype.

CaNA_N - Behaviour

"*CaNA_N plays an important role in identifying mice which are simulated to learn from mice which are not simulated to learn*". Plotting the visualisation of expression levels of CaNA_N protein grouped by Behaviour of mice, highlighted the fact that the mice which are not simulated to learn i.e. shock-context have lower values of CaNA_N in contrast to higher expression levels of mice who are simulated to learn i.e. context-shock. Thus the below box plot provides justification of the hypothesis.



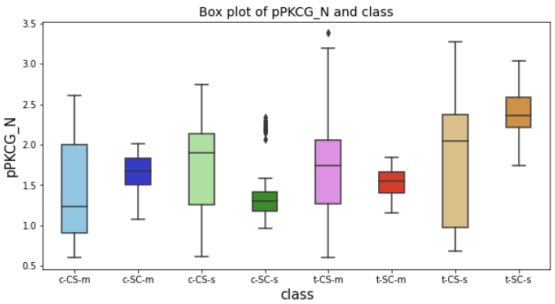


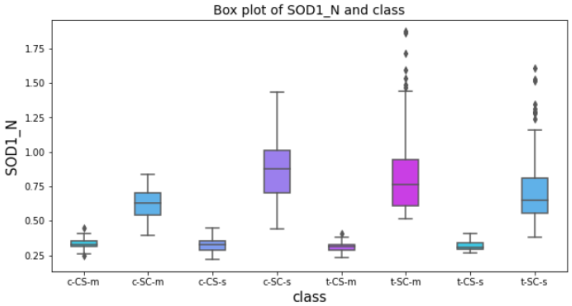
pCAMKII_N – Behavior

"Pcamkii_N can be used to distinguish between mice that are simulated to learn from mice which are not simulated to learn". Above box plot on the right shows that the inter quartile range of pCAMKII ranges from approximately 2.2 to 3.5 for mice which are simulated to learn whereas it approximately ranges from 3.5 to 5.1 for mice which are not simulated to learn. However, as the range for both types of mice behaviour lies close to each other there might be cases when pCAMKII might not be able to distinguish between both type of behaviours and hence we cannot ensure that the hypothesis holds on the given data set.

pPKCG_N – class

"pPKCG_N is useful in discriminating mice into one of the eight subclasses." The below image on the left represents the visualization of protein expression levels of pPKCG_N using a boxplot grouped by eight subclasses of mice. The image shows that the hypothesis holds. Also, it is evident from the box plot that the interquartile range of mice showing C/S behaviour is much higher as compared to that of mice showing S/C behaviour. Also, the trisomic mice which are not simulated to learn and given saline treatment tend to have higher levels of pPKCG_N protein whereas the mice control mice which are simulated to learn and given saline treatment tend to have lower levels of this protein. Apart from that the inter quartile range of mice belonging to the subclass t-CS-s is highest among all other subclasses. Thus, we can say that pPKCG_N might be useful in distinguishing mice based on their class.





SOD1_N – class

"SOD1_N plays a significant role in discriminating mice into different subclasses". The above boxplot on the right shows SOD1_N protein expression levels grouped by class. It is evident from the plot that SOD1_N levels

show significant difference in the expression levels among different subclasses of mice there by implying that this hypothesis holds.

Data Modelling

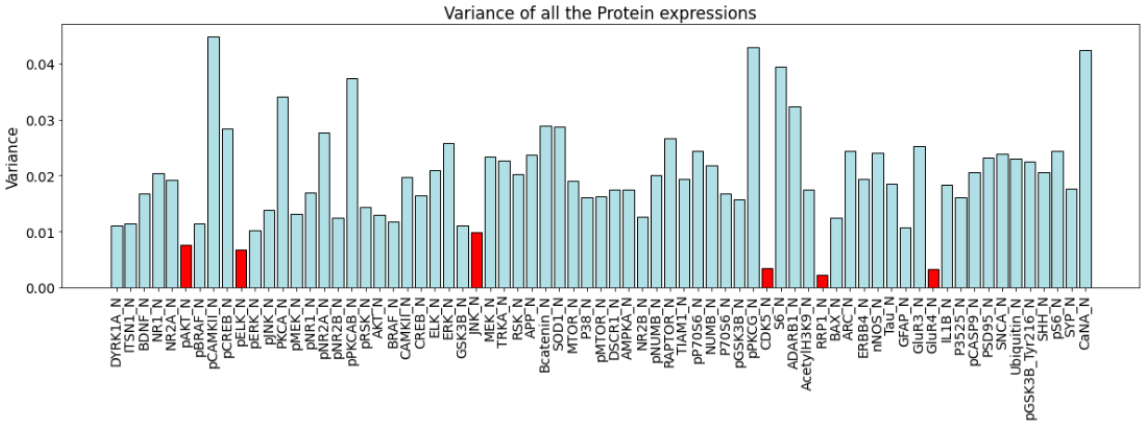
Feature selection

After performing the exploratory analysis, we can see that certain features showcased high level of importance whereas certain features were deemed to show less importance. Moving further we need to create a classifier to predict the eight sub classes of mice depending on the subset of proteins. However, as we know there are 77 protein columns so training a model on all the protein columns is a very expensive task to perform and also the features which are of less importance will degrade the accuracy of the classification model. Thus, we will apply various feature selection techniques to discard the features having less importance. The strategies used for feature selection are as follows:

1. Filter strategy
    - Missing values ratio: This strategy removes features containing missing values above the threshold values. For this task, the threshold is kept at 75 count
    - Low variance filter: This strategy removes features having variance below the mentioned threshold
2. Wrapper strategy
    - This strategy calculates feature importance for each feature and thereby removing low scoring features

Filter strategy

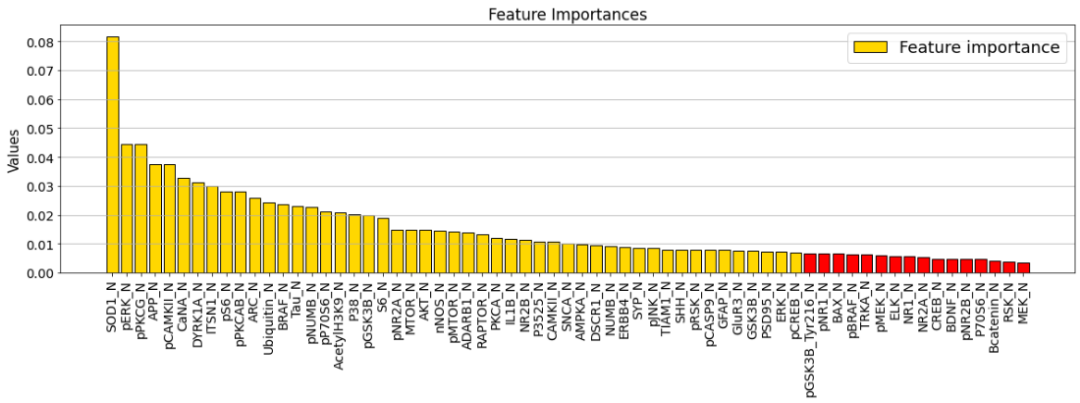
Before applying low variance filter strategy, the protein levels are scaled using the Standard scaler provided by the sklearn library as the protein expression of different proteins are of different range. Once, the proteins were standardised, all the proteins having variance less than the threshold variance which was set at 0.01 were removed. The below image shows the variance of all 77 proteins. Red coloured bars indicate the features that were filtered using this strategy.



Wrapper strategy

The main reason of using this strategy was to remove low scoring features corresponding to the target feature which was class of mice in our case. For this strategy, a tree-based model was used to calculate the feature importance scores of each protein w.r.t to the target variable. The below bar chart shows the feature importance sorted in descending order. The red colour bars indicate the features which were dropped because of very low importance. However, from the plot it is evident that SOD1_N has the highest feature importance score. This

was also justifiable from the above figure in data exploration part showing the box plot of SOD1_N grouped by class.



Selecting the model

Once the features have been selected for the data modelling task the next step is to choose the model to predict the target variable. Here as mentioned previously *class* is the target variable as we aim to determine the set of proteins to distinguish between eight subclasses of mice. As the classes are categorical and the data set is labelled, we will use supervised machine learning classification algorithms. There are various classification models out of which for this task, K-nearest Neighbours and Decision tree provided by the sklearn library are used.

Training the model

Once the models are selected to perform the classification task the next step is to train the model. First step of training the model comprises of splitting the given data set into train and test part. Here the test size is taken as 20% of that of the total data set. The below image shows the training data size and test data size.

```
Training data set size: (861, 49)
Testing data set size: (216, 49)
```

First, we consider Decision tree classifier. The model is trained with different settings by tuning various parameters. The process of experimenting with different parameters to find the parameters which gives the highest accuracy is commonly known as hyper tuning of parameters. Apart from that k-folds cross validation strategy is used to evaluate the models. 10 split k-folds is performed on each model in which 9 folds data is used for training whereas the model is scored on remaining 1-fold data. After that, the mean accuracy score is calculated by averaging the accuracy of all the 10 splits. The below table shows mean accuracy scores of Decision tree classifier with hyper tuned parameters.

	Model	Parameters	Test_mean_accuracy
0	DecisionTreeClassifier	(random_state=42)	0.845496
1	DecisionTreeClassifier	(criterion='gini', max_depth=5, random_state=42)	0.754878
2	DecisionTreeClassifier	(criterion='gini', max_depth=10, random_state=42)	0.840871
3	DecisionTreeClassifier	(criterion='gini', max_depth=15, random_state=42)	0.845496
4	DecisionTreeClassifier	(criterion='gini', max_depth=15, max_features=...	0.797862
5	DecisionTreeClassifier	(criterion='gini', max_depth=15, max_features=...	0.851310
6	DecisionTreeClassifier	(criterion='gini', max_depth=15, max_features=...	0.847821
7	DecisionTreeClassifier	(criterion='entropy', random_state=42)	0.837450
8	DecisionTreeClassifier	(criterion='entropy', max_depth=10, random_sta...	0.835124
9	DecisionTreeClassifier	(criterion='entropy', max_depth=20, random_sta...	0.837450
10	DecisionTreeClassifier	(criterion='entropy', max_depth=15, max_featur...	0.801323
11	DecisionTreeClassifier	(criterion='entropy', max_depth=15, max_featur...	0.852513

From above image we can see that the decision tree classifier predicts the mice classes with highest accuracy of 0.85. We initially start with by training the model with only default parameters with criterion as 'gini'. However, the accuracy of model comes to 0.84. So next the 'max_depth' parameter is tuned. Values 5, 10, 15 are used for tuning this parameter. A depth of 5 shows very low accuracy whereas the accuracy remains almost

contant for values 10 and 15. Hence we set the max_depth value as 10. Further, the max_features parameter is tuned with values 10, 20, and 40. By setting max_features as 10 the accuracy again decreases to 0.79. However, value 20 again increases the accuracy to 0.85 and remains almost same by setting the values as 40. Next, the model is trained with criterion as 'entropy' and the accuracy comes to 0.83. For the criterion as 'entropy' the parameters are tuned in the above fashion, and, the highest accuracy after that comes to 0.85.

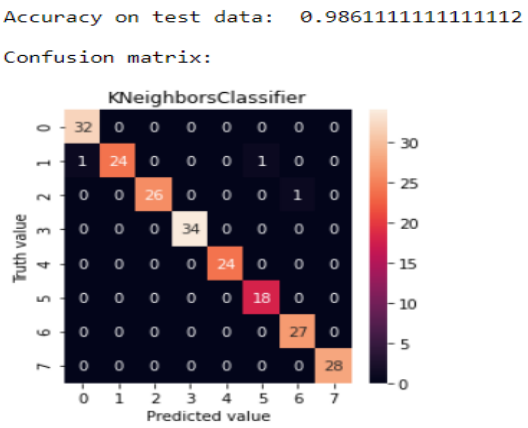
Next, we will consider K-nearest neighbours' model. The model is trained with different settings by tuning various parameters. The process of experimenting with different parameters to find the parameters which gives the highest accuracy is commonly known as hyper tuning of parameters. Apart from that k-folds cross validation strategy is used to evaluate the models. 10 split k-folds is performed on each model in which 9 folds data is used for training whereas the model is scored on remaining 1-fold data. After that, the mean accuracy score is calculated by averaging the accuracy of all the 10 splits for different tuned classifiers. The below table shows k-folds mean test accuracy for differently tuned k-nearest neigbours classifiers.

12	KNeighborsClassifier	()	0.945416
13	KNeighborsClassifier	(n_neighbors = 3)	0.972133
14	KNeighborsClassifier	(n_neighbors = 3, p = 1)	0.987223
15	KNeighborsClassifier	(n_neighbors = 3, p = 2)	0.972133
16	KNeighborsClassifier	(n_neighbors = 3, p = 3)	0.955921
17	KNeighborsClassifier	(n_neighbors = 6)	0.928014
18	KNeighborsClassifier	(n_neighbors = 3, p = 1, weights='distance')	0.987223

The highest accuracy achieved using K-nearest neighbours is 0.987 which is much higher as compared to that of decision tree classifier. Even with the default k neighbours classifier an accuracy score of 0.94 is achieved. Next, we start by setting the parameter n_neighbors. Different values like 3, 6, 10 are used for this parameter. By setting value as 3 it gives higher accuracy as compared to that of the accuracy obtained with the default k-nearest classifier. However, the accuracy decreases with increase in number of n_neighbors. Next the parameter p is tuned for the classifier having n_neighbors as 3. However, it is evident that the classifier with p=1 and n_neighbors as 3 has the highest value as compared to all other classifiers. Apart from that the classifier with n_neighbors = 3, p=1 and weights='distance' has same accuracy as that of classifier with n_neighbors = 3 and p=1.

Results

As the accuracy of K-nearest neighbour classifier with n_neighbors=3 and metric p=1 showed the highest accuracy amongst all we will use that model to predict the mice classes for X_test data set. This is called model scoring. The below image shows the accuracy and the confusion matrix of the classifier.



Discussion

The given dataset consisted expression levels of 77 proteins for eight subclasses of mice based on their Behaviour, Genotype, and treatment. The aim was to identify a subset of proteins that can

distinguish between different classes of mice. Firstly, the exploratory data analysis was conducted on the given data set to visualize the data set which helps to understand the dataset from different perspectives. Some interesting patterns and insights were gained from the plots. It appeared that SOD1_N was one of the most important protein that helped in differentiating mice into different classes. The analysis also highlighted various short comings as few mice showcased outliers in the same class for e.g. Mouse with id 3484 of class 'c-CS-s' appeared as an outlier in various protein expression readings. The analysis helped in identifying linear relationships between different set of proteins as well. Apart from that the no of proteins to be considered for the data modelling part were reduced to 35 using various strategies like filter strategy and wrapper strategy. This helped increase the performance of the machine learning models. Once, the feature selection was done, two classification models namely Decision tree and K-nearest neighbours were used to train the models on the training data set by tuning various parameters. After training the different models they were evaluated using k-folds cross validation scores by taking the mean accuracy of the k-folds split. K-nearest neighbour classifier with parameters n_neighbours = 3 and p=1 outperformed other classifier with mean accuracy score of 0.987 on training data and accuracy score of 0.986 on the test data. With this accuracy we can say that model was able to discriminate between the mice classes for most of the part. However, 100% accuracy could not be achieved, and it is justifiable from the analysis performed in the above steps as it indicated of few erroneous readings (reading with negative values), NAs, and outliers.

Conclusion

Mice protein expression data set provided by UCI was used to perform to exploratory data analysis with the aim of identifying a subset of proteins that can be used to differentiate different classes of mice derived from their Behaviour, Genotype, and Treatment. The data was prepared to perform the analysis. As the number of features to be considered for data modelling part were large, various techniques like filter strategy and wrapper strategy were used to reduce redundant features, features with low variance and features of less importance. A total of 35 features were used to create a model to predict the target classes. K-nearest neighbours' classifier was used after comparing it with Decision tree classifier, and it was successfully able to differentiate the target classes with an accuracy of 0.986 on the test data. To conclude, we can say that it was possible to identify different subclasses of mice using the k-nearest neighbours machine learning classification algorithm. Future work included to dive deep into this subset of proteins to identify if the drug memantine was effective on the trisomic mice.

References

[1]C. Higuera, K. Gardiner and K. Cios, "UCI Machine Learning Repository: Mice Protein Expression Data Set", *Archive.ics.uci.edu*, 2020. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression. [Accessed: 01- Jun- 2020].

[2]J. Thomas, "Down Syndrome: Facts, Statistics, and You", *Healthline*, 2020. [Online]. Available: https://www.healthline.com/health/down-syndrome/down-syndrome-facts#2. [Accessed: 01- Jun- 2020].

[3]"What is Down Syndrome? | National Down Syndrome Society", *NDSS*, 2020. [Online]. Available: https://www.ndss.org/about-down-syndrome/down-syndrome/. [Accessed: 01- Jun- 2020].

[4]"MyApps Portal", *Rmit.instructure.com*, 2020. [Online]. Available: https://rmit.instructure.com/courses/67430. [Accessed: 01- Jun- 2020].