Student ID: s3781718

Student Name: Yashwin Sudarshan

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": *Yes*

Classification of Mice Class according to Protein Expressions

Practical Data Science with Python

Author Information: Yashwin Sudarshan (s3781718)

Affiliations: RMIT University

Contact Details: s3781718@student.rmit.edu.au

Date of Report: 10th June 2020

Table of Contents

Abstract

Mice with Down syndrome, namely, whose genotype is trisomic, exhibit different levels of protein expressions in the cortex nuclear fraction, as compared to mice which are not trisomic. The protein expression levels of these two genotypes groups of mice are further affected according to certain behaviour stimulation learning methods, and the type of drug injected into these mice to recover learning ability in trisomic mice. Thus classes of mice are formed according to genotype, behaviour, and the drug or chemical given to the mice. My investigation into the data that contained various protein levels of the different classes of mice indicated certain subsets of protein groups which were able to powerfully classify mice between the classes. The results of this investigation project can be applied to learning which proteins, or groups of proteins, and the levels of these proteins, can be utilised to predict whether a mouse will have a trisomic genotype, which will allow for greater understanding between the effect of different protein levels on the genotype of mice.

Introduction

Being able to classify the genotype of a mouse, whether or not it has Down syndrome (trisomic), has high importance in understanding the factors that dictate trisomy in mice, and also in other organisms, potentially, human beings. Data on protein expression levels in trisomic and normal mice exist online. The utilised dataset for this project, includes 77 mouse proteins and protein modifications, and 1080 protein expression level measurements, which correspond to 15 protein measurements per protein per mouse for 72 mice, 38 of which are control mice, and 34 of which are trisomic mice. The dataset also contains the genotype, drug treatment, and behaviour, and consequently, the class of each mouse included in the experiment. Classification models that can successfully predict trisomy in mice, built from different groups of proteins and their respective expression levels, would be highly valued when attempting to extrapolate the findings to other organisms other than mice, using models similar to the aforementioned hypothetical classification models. As such, the research goal for this project is to classify the classes of mice, which contain their genotype (c - Control, t - Trisomic), behaviour (CS - Context Shock, SC - Shock Context), and recovery injection drug (s - Saline, m - Memantine), according to subsets of different proteins and their respective expression levels in the cortex nuclear fraction.

Methodology

The first step after defining the research goal is to perform the data retrieval process. The dataset is downloaded from online, and loaded into a Pandas data frame. Next, this retrieved data is then compared to the data source file, to check whether the data is equivalent, in structure and data types.

The next step is to conduct the data preparation process. This involves performing data cleansing and transformation, in order to present a suitable format of the dataset to the visual and modelling tools to be utilised in the later data exploration and data modelling stages, so that no unexpected usage tool errors occur. In the dataset, the number of missing values in each feature column was checked. After checking, it was discovered that a number of features only had a small amount of missing values, but there were some protein columns (BAD_N, BCL2_N, H3AcK18_N, EGR1_N, and H3MeK4_N) with hundreds of missing values. At this point in the project, it was decided not to deal with these null values yet, as it was not desired to drop large amounts of rows, and to fill large amounts of null values with a column mean or median, as it would largely negatively affect data exploration and statistical analysis of the data columns. Then, the numerical protein feature columns of the dataset were checked for impossible values. It was found that only one protein column had a negative impossible value, namely, the RRP1_N protein. This row was thus dropped to remove the impossible value. Then the categorical columns were checked for typos and extra whitespace, of which none could be found.

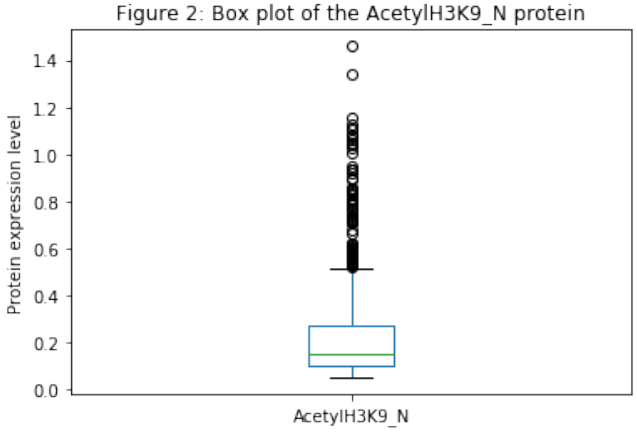
The next step after this is to perform the data exploration stage. 11 individual data columns were desired to be explored employing statistical visualisations, and 10 pairs of data columns were also wanted to be investigated to expose potential relationships between data attributes, reflecting several hypotheses made about certain pairs of data attributes. After the individual data columns were explored, and before pairs of columns were investigated, the individual data visualisations exposed many outliers in those respective columns. Since there were many outliers, it was desired to remove the outliers that were further from the other outliers, as it was assumed that the many smaller outliers that were grouped together were not data entry errors, and were the result of normal non-erroneous data measurement and entry procedures. For more information on the data attributes investigated in the data exploration stage, refer to the Results and Discussion sections of this report.

The next part of the research project is to start the data modelling stage. Here, two classification models are built, and compared to see which model has greater classification predictive potential to classify mice classes according to protein levels. In order to build these models, the categorical columns except the target class column is removed from the dataset (as these columns directly predict the class column and are thus have predictive redundancy), and then the dataset is split into training and testing data subsets, according to the standard 80/20 proportion, i.e. 80 percent of the data is used to train both classifiers, and 20 percent of the data is used to test the classifiers. The classification models chosen for this project are the k-Nearest Neighbours classifier and the Decision Tree classifier. After each train and test process, for each classifier, a classification report is generated in order to evaluate the performance of each classifier for comparison purposes. The first train and test iteration utilised all the features of the dataset to train and test each classifier, and was performed for each classifier using their respective default parameters. The next train and test iteration again used all the dataset features (by all it is meant that all the numerical protein attributes, and the target class attribute), but the parameters of the k-Nearest Neighbours and Decision Tree classifier were altered in an attempt to obtain a better performance. Then, in order to obtain the best possible classifier performance, a Hill Climbing heuristic data selection algorithm was employed to iteratively select numerical protein features for each classifier. After this, the end performance of each classifier was compared.

Results and Discussion

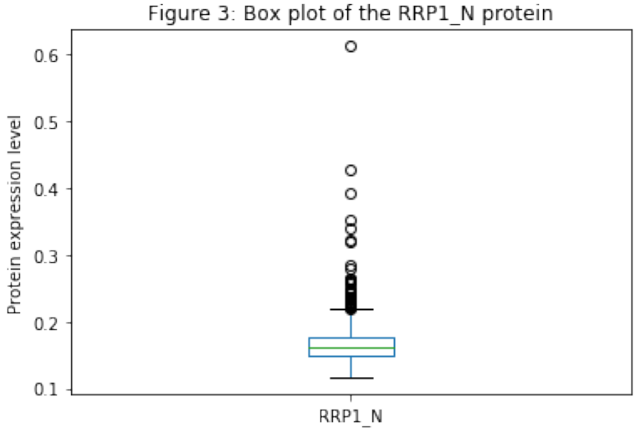
Figures 2, 3, and 6 below are 3 of the 11 individual protein graph visualisations performed during data exploration. These figures are box plots since they effectively visually show important values of the numerical distribution of the proteins, such as the median, the whiskers, and the interquartile range.

Figure 2: Box plot of the AcetylH3K9_N protein



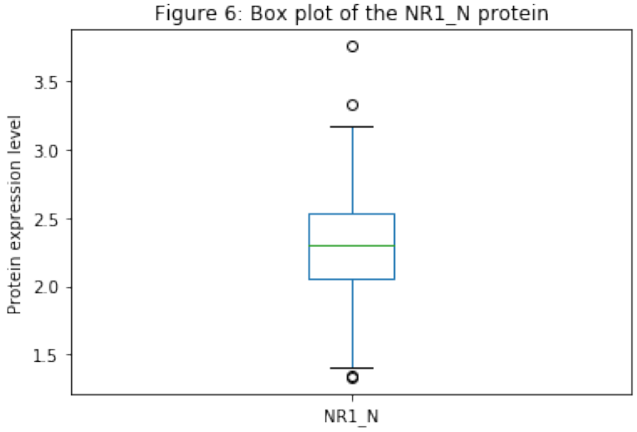
According to Figure 2 on the left, it can be seen that the distribution of the AcetylH3K9_N protein is positively skewed, with many outliers past the last whisker. There are also two extreme outliers that were later removed from the column itself.

Figure 3: Box plot of the RRP1_N protein



According to Figure 3 on the left, it can be seen that the distribution of the RRP1_N protein is very small, with very little spread, and one extreme outlier.

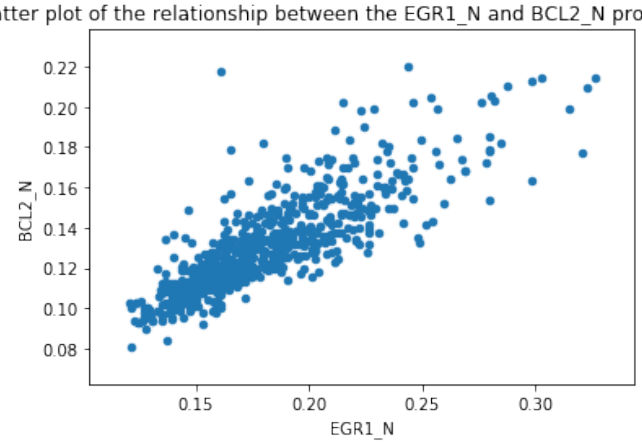
Figure 6: Box plot of the NR1_N protein



According to Figure 6 on the left, the distribution of the NR1_N protein is symmetrical with a median of around 2.3. The distribution also spreads across a large range of values than the proteins investigated prior.

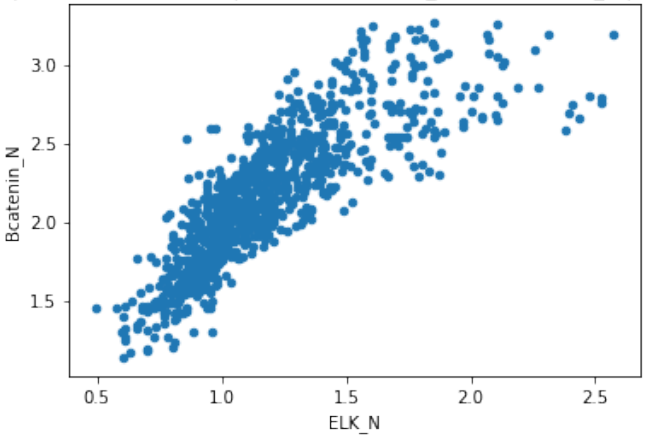
Figures 13, 18, 19, and 20 are 4 of the 10 pair plots investigated, with individual hypotheses for each visualisation.

Figure 13: Scatter plot of the relationship between the EGR1_N and BCL2_N protein levels in the mice



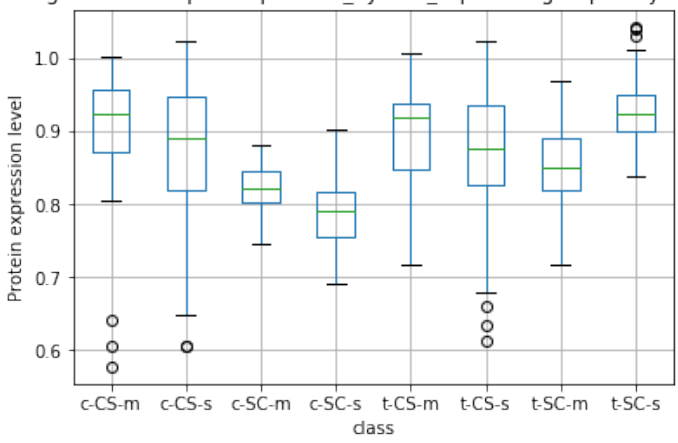
According to Figure 13 on the left, it can be seen that there seems to be a linear correlative relationship between the BCL2_N and EGR1_N proteins. The hypothesis investigated here is whether there is any relationship between the proteins, and the visualisation seems to support this.

Figure 18: Scatter plot of the relationship between the ELK_N and Bcatenin_N protein levels in the mice



According to Figure 18 on the left, there seems to be a seemingly interesting linear-like or logarithmic-like relationship between the Bcatenin_N and ELK_N proteins. The hypothesis investigated here is if there is a linear relationship between the two proteins.

Figure 19: Box plot of pGSK3B_Tyr216_N protein grouped by class

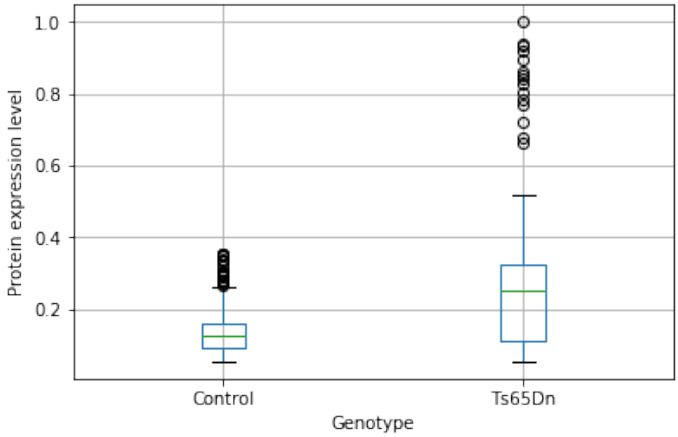


According to Figure 19, it can be clearly seen that the distributions of the pGSK3B_Tyr216_N protein are different according to the class of the mice. It can also be seen that the mice with CS behaviour have similar distributions, and the mice with SC behaviour have more varied distributions. The hypothesis being investigated here was whether the pGSK3B_Tyr216_N protein had a great impact on the class of the mice, and it seems that this is supported heavily by the boxplot visualisation.

According to Figure 20 below, the hypothesis that the AcetylH3K9_N protein has a relationship with the genotype of the mice is supported by this boxplot. The median for the trisomic boxplot is larger as well as the spread, as opposed to the control mice, which suggests that higher levels of this protein are correlated with a higher

amount of mice having Down syndrome.

Figure 20: Box plot of AcetylH3K9_N protein grouped by Genotype



The following discussion concerns the model building and validation and feature selection process. Starting with the k-Nearest Neighbours classification model, in the first training and testing model iteration, all the protein features were used in the model training and testing, with default parameter values set for the classifier. This resulted in a weighted average f1-score of 0.95 for the model. In the second training and testing iteration, the parameters for the k-Nearest Neighbours classifier were tweaked; the number of neighbours was reduced from the default number of 5 (as a lower number gives allows more distinct classification boundaries), the weights parameter was made to be distance (as this will allow closer points to have more of an influence), and the p value was lowered from 2 to 1, just to experiment with the performance results. This tweaked model now gave a weighted average f1-score of 0.98. In the third training and testing iteration, this tweaked model was used in the feature selection Hill Climbing algorithm, in an attempt to improve the performance by selecting a more predictive subset of the proteins (currently, all the proteins are being utilised). After the Hill Climbing was performed, the model was able to achieve a score of 0.995 with 67 proteins : Bcatenin_N, Bcatenin_N, Bcatenin_N, NUMB_N, JNK_N, EGR1_N, GluR3_N, GluR4_N, P3525_N, RRP1_N, DSCR1_N, GFAP_N, pERK_N, ARC_N, P38_N, BDNF_N, pCFOS_N, MEK_N, pPKCG_N, pNR2A_N, pGSK3B_N, AKT_N, CaNA_N, pS6_N, PSD95_N, H3MeK4_N, pCASP9_N, pMTOR_N, NR2B_N, nNOS_N, MTOR_N, SHH_N, RAPTOR_N, CAMKII_N, BAD_N, Tau_N, CDK5_N, pCREB_N, DYRK1A_N, BCL2_N, S6_N, H3AcK18_N, BAX_N, pMEK_N, CREB_N, SOD1_N, APP_N, TIAM1_N, ERBB4_N, ELK_N, pGSK3B_Tyr216_N, pNR1_N, pP70S6_N, IL1B_N, pJNK_N, Ubiquitin_N, pRSK_N, BRAF_N, AcetylH3K9_N, GSK3B_N, pBRAF_N, SNCA_N, pAKT_N, pELK_N, SYP_N, PKCA_N, and AMPKA_N.

The same iteration process was applied to the Decision Tree classification model. In the first iteration, the model obtained a weighted average f1-score of 0.81. In the second iteration, the max_features parameter was altered to be 31, as this was 40 percent of all protein features, and it is a good rule to consider 30-40 percent of the dataset features for optimal splitting. The weighted average f1-score obtained from this iteration was 0.84, an improvement from the previous iteration. In the final iteration, the Hill Climbing algorithm was used on the Decision Tree classifier to choose the most optimally predictive subset of protein features. After running the algorithm, the performance score obtained was 0.892, with 21 protein features: Bcatenin_N, NUMB_N, JNK_N, EGR1_N, GluR3_N, GluR4_N, P3525_N, RRP1_N, pERK_N, P38_N, pCFOS_N, pPKCG_N, pS6_N, PSD95_N, H3MeK4_N, pMTOR_N, nNOS_N, P70S6_N, pCAMKII_N, pMEK_N, NR2A_N, and AMPKA_N. So, the final prediction score for the k-Nearest Neighbours model was 0.995 with 67 proteins, and the score for the Decision Tree classier was 0.892 with 21 proteins. If one wishes for the most apparent accuracy performance in classifying mice classes, then the k-Nearest Neighbours model should be chosen. If one wishes for a high performance, but with less protein features to handle, then the Decision Tree classifier should be chosen, as although its score was about 0.1 lower than the k-Nearest Neighbours model, it used less than a third of the amount of protein features than the k-Nearest Neighbours model used. My recommendation would be that the Decision Tree model should be used to classify mice classes. This model uses far fewer proteins to classify mice classes to a good enough extent, and shows that these fewer proteins are better discriminant proteins between the mice classes.

Conclusion

Mice classes were able to be classified to a 0.892 success rate using a subset with a total of 21 out of the 77 proteins in the original dataset. These proteins are: Bcatenin_N, NUMB_N, JNK_N, EGR1_N, GluR3_N, GluR4_N, P3525_N, RRP1_N, pERK_N, P38_N, pCFOS_N, pPKCG_N, pS6_N, PSD95_N, H3MeK4_N, pMTOR_N, nNOS_N, P70S6_N, pCAMKII_N, pMEK_N, NR2A_N, and AMPKA_N. The proteins in this subset together have a powerful predictive potential, and can be used to better understand which proteins or groups of similar proteins and their expression levels are directly correlated with Down syndrome in mice, and potentially, the findings of this report can be helpful in classifying Down syndrome in other organisms, such as humans.

References

‘Activity 1: Decision Trees’, course notes, COSC2670, RMIT University, viewed 9 June 2020, <https://rmit.instructure.com/courses/67430/files/12058206?module_item_id=2369057>

‘Activity 1: k Nearest Neighbour Classifier’, course notes, COSC2670, RMIT University, viewed 9 June 2020, <https://rmit.instructure.com/courses/67430/files/11552799?module_item_id=2314376>

Dr. Ren, Y 2020, ‘Practical Data Science: Classification’, PowerPoint slides, COSC2670, RMIT University, viewed 28 April 2020, <https://rmit.instructure.com/courses/67430/files/11273015?module_item_id=2333118>

Higuera, C & J. Cios, K & J. Gardiner, K 2015, *Mice Protein Expression Data Set*, data file, UCI Machine Learning Repository, viewed 3 June 2020, <https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression#>