Student ID: s3758710

Student Name: Samalka Wedaratne Nanayakkara Talpe Merenchige.

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 BASED ON PROTEINS PRESENT IN THEM

By: Samalka Wedaratne

Date: 10-06-2020

CONTENTS

ABSTRACT

The aim of this project was to examine the relationship between different variables in the Mice Protein data and to build two classifier models to classify the class of the mice based on the proteins presented in their bodies. Exploration aided with visualization techniques were carried out for examining the variables. Two classifiers named Decision Tree and K-Nearest Neighbour were modelled and tuned for testing the results. The results showed that many variables didn’t or had only a weak relationship with each other. The KNN classifier yielded the highest prediction power in terms of accuracy from this study. It is recommended to explore more variables and to use different feature selection techniques and classifiers as well as to tune the Decision tree classifier for more accuracy.

INTRODUCTION

Mice are often used for experimental purposes for scientific research. There are practical reasons associated with this such as the size, the weight and their harmlessness compared to many other animals such as tigers or elephants in which case it would be physically impossible to experiment with. Apart from these practical reasons, we share a subsequent number of our human genes with mice. Therefore if we want to understand the human brain function, we might as well start with mice. Similarly in this project, results obtained from an experiment done on mice has been used. According to the original data, there are eight classes of mice that have been described based on features such as genotype, behaviour pattern and the treatment given to them. This report will discuss about classifying the mice into these eight classes, depending on the proteins presented in them.

Classes:

- c-CS-s: control mice, stimulated to learn, injected with saline
- c-CS-m: control mice, stimulated to learn, injected with memantine
- c-SC-s: control mice, not stimulated to learn, injected with saline
- c-SC-m: control mice, not stimulated to learn, injected with memantine
- t-CS-s: trisomy mice, stimulated to learn, injected with saline
- t-CS-m: trisomy mice, stimulated to learn, injected with memantine
- t-SC-s: trisomy mice, not stimulated to learn, injected with saline
- t-SC-m: trisomy mice, not stimulated to learn, injected with memantine

METHODOLOGY

The main data resource for this project comes from the UCI Machine Learning Repository under the name “Mice Protein Expression Data Set”. The original dataset was obtained to identify subsets of proteins that are discriminant between the classes which was created by Katheleen J. Gardiner. However for this particular project I will be using this dataset for a classification problem of classifying the mice into their respective classes based on proteins present in them. The tasks were carried out using Python.

The methodology is divided into sub tasks as shown below.

1. **Data Pre-processing** – The necessary packages were imported. The original dataset was imported as csv file to a pandas data frame. The shape of data and the data types were then checked. Missing values were then handled by imputing them with the mean respective class for the numerical columns (proteins). Outliers are not dealt with as the original dataset had no mention on the appropriate values for the proteins. However value counts for the categorical features were obtained to check for any irregularities. The MouseID column was not in a tidy format as it essentially had 2 variables included in it, so this was separated to two columns. Summary statistics of the descriptive features were then obtained. More processing will be done when dealing with the modelling section.

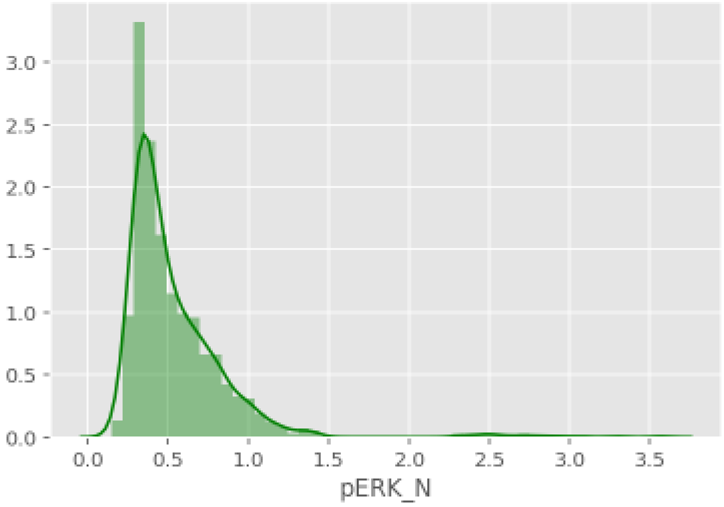
2. **Data Exploration** – First, 10 columns were explored individually. Visualization techniques such as bar plots, boxplots and histograms with kernel density maps were used for better understanding of the features that are dealt with. Next, 10 pairs of columns were explored to understand the relationship between two such columns. Hypothesis were assumed when carrying out these explorations. Visualization techniques such as Boxplots, pie charts, and scatter plots were used to aid this.

3. **Data Modelling** – The mouseNo, Mouse Version columns were dropped as they represent an ID which is not essential for modelling. The columns Genotype, Behavior and Treatment were also dropped as these 3 essentially makes the target variable “class”. The target variable “class” was then encoded. In order to understand the individual contribution of the descriptive features, feature selection method Random Forest Identifier (RFI) was used. The best 20 features were visualized. Then, the dataset was separated to train and test data with 30% being on the test data. The model evaluation strategy was then defined. Repeated Cross validation was used. The two chosen classifiers that were used in this project are, KNN and Decision Tree. Pipelines were used to optimize the modelling process. Decision Tree parameters, minimum sample split, maximum depth were tuned in the Hyper parameter tuning process. The KNN was tuned by changing the number of neighbours and the distance metric. A pipeline with gridsearch was used to efficiently do this. I kept the combination that gave the highest accuracy in both cases from what I tried. Lastly the performances of the two classifiers were compared by introducing the test data and the classification report and the Confusion matrix were used for this purpose. A paired t-test is also carried out to see whether the results are statistically significant.

RESULTS

DATA EXPLORATION RESULTS- SINGLE COLUMN EXPLORATIONS.

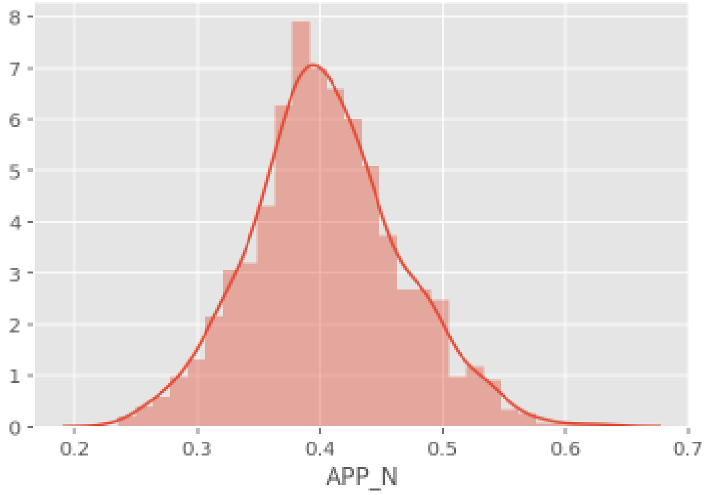
Column Name:pERK_N



The histogram was plotted the protein pERK_N. The plot clearly shows that it is a positively skewed distribution.

Figure 1: Histogram of pERK_N Protein

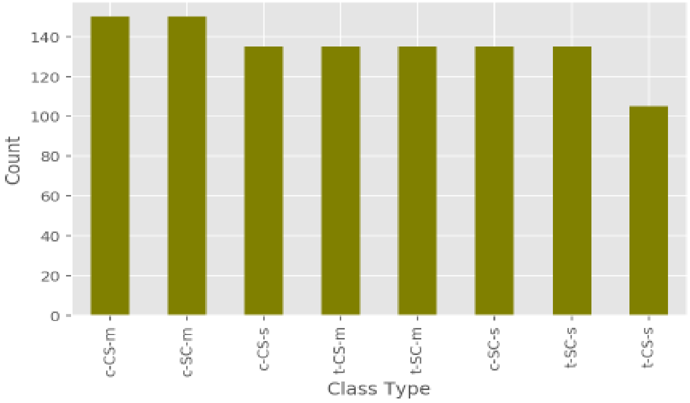
Column Name: APP_N



This protein has a symmetric distribution which can be considered adequate especially for model building purposes.

Figure 2: Histogram of APP_N Protein

Column Name: Class

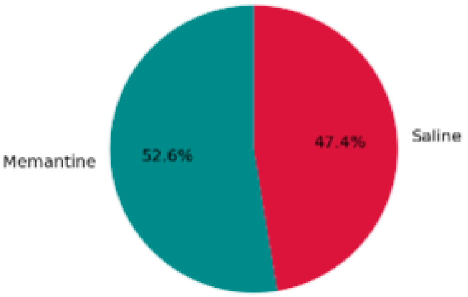
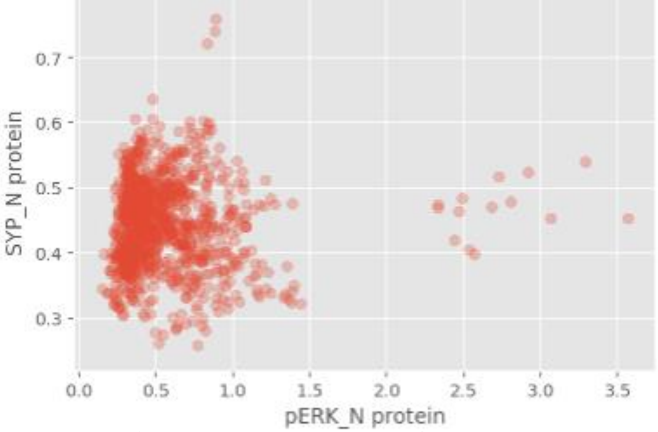
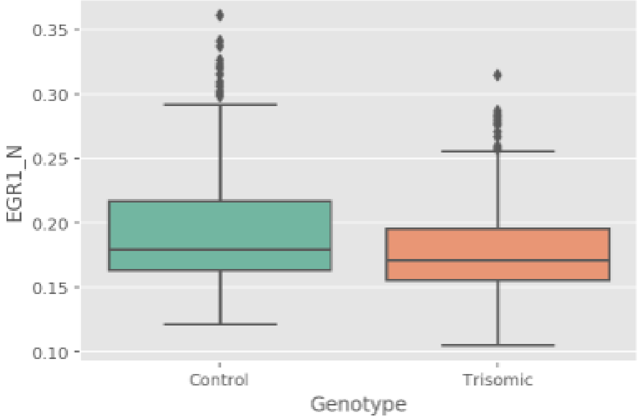


This will be the target variable in modelling, the class of the mouse. The type t-CS-s are relatively low compared to others.

Figure 3: Bar plot of Class Type

DATA EXPLORATION RESULTS- TWO COLUMN EXPLORATIONS.

Several Hypothesis were used in exploring. Only a few of them will be included in this report.

Hypothesis	**Visualization**	**Outcome**
Columns: Treatment, Genotype H0 = The proportion of Memantine treatment mice are larger than Saline treated mice in the Control group	 Figure 4: Memantine vs Saline in Control group	A higher proportion of mice are treated with memantine in the control group. Therefore there is no enough evidence to reject the null hypothesis.
Columns: pERK_N and SYP_N H0 = There is a strong correlation between protein pERK_N and SYP_N	 Figure 5: pERK_N vs SYP_N	A strong correlation cannot be seen between the two proteins, hence reject the null hypothesis.
Columns: ERG_1 and Genotype. H0= EGR1_N is more present in Control Mice than Trisomic Mice group	 Figure 6: ERG_1 in Control and Trisomic mice	From the two boxplots the control group seems to have more EGR1_N protein, therefore there is no enough evidence to reject the Null Hypothesis.

DATA MODELLING RESULTS

Feature Selection Result – using Random Forest Feature Importance (RFI)

RFI was used for feature selection, the following plot was obtained to just to get an idea on which features seems the most representative of the class.

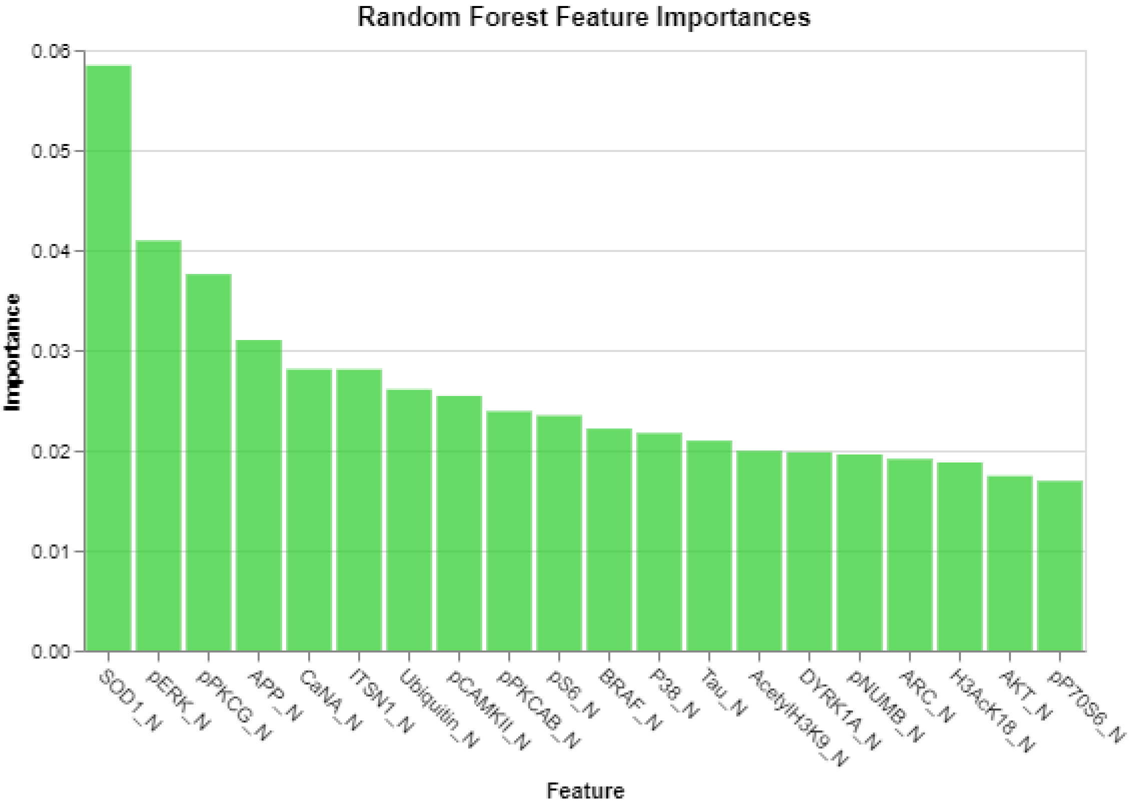


Figure 7: RFI Features

Decision Tree Best Parameter combination and Accuracy for Training data.

- Decision Tree Split Criterion: Gini
- Decision Tree Maximum Depth: 12
- Decision Tree Minimum Samples Split: 2
- Number of features: 20
- Accuracy: 83%

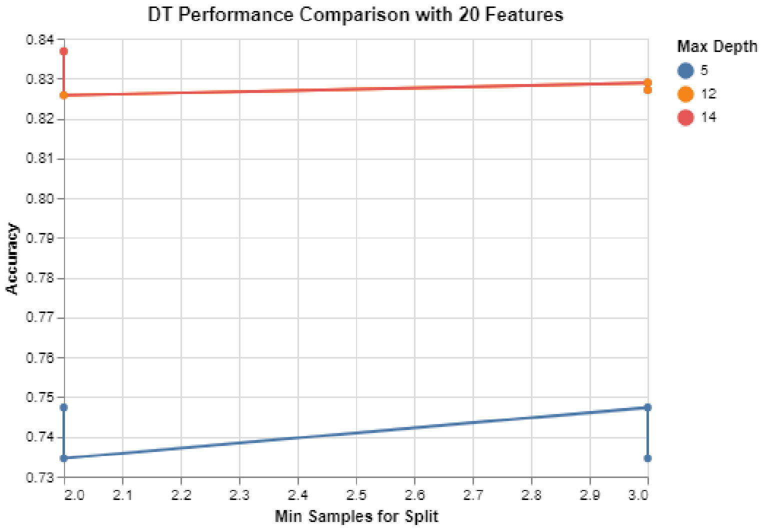


Figure 8: DT Performance plot in Tuning

K-Nearest Neighbour Best Parameter combination and Accuracy for training data.

- Neighbours: 1
- Distance: 1 (Manhattan)
- Number of features: 20
- Accuracy: 99%

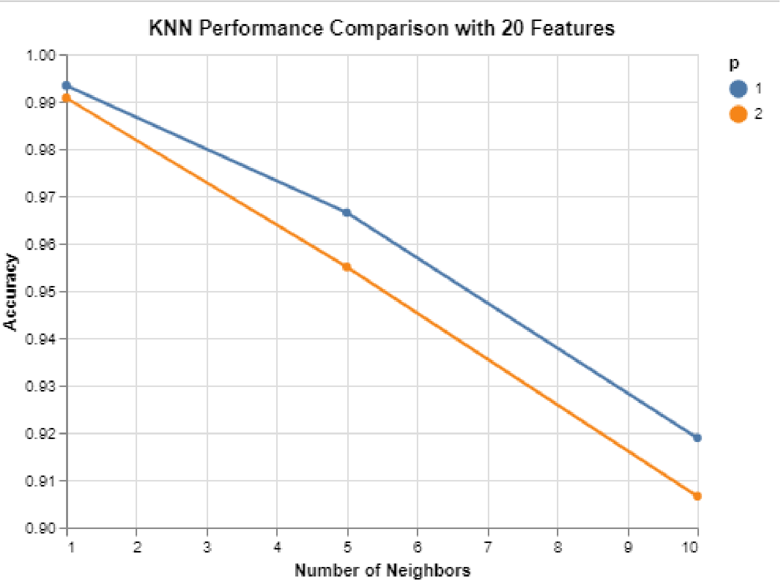
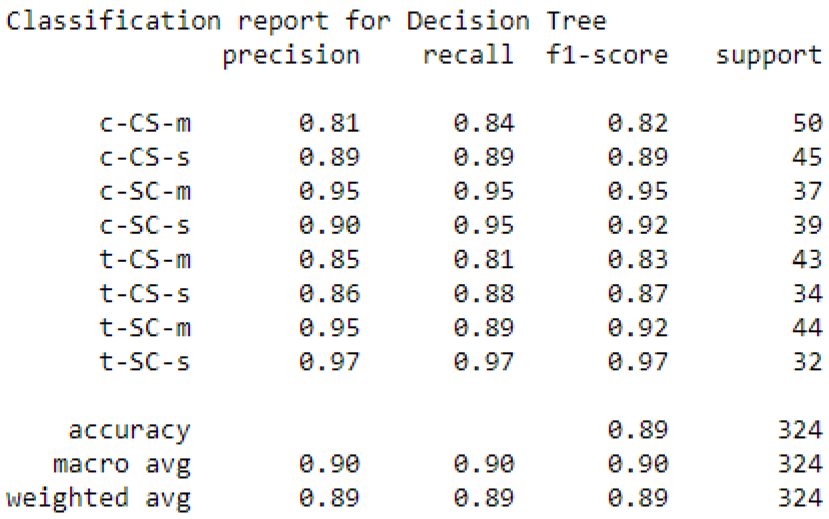
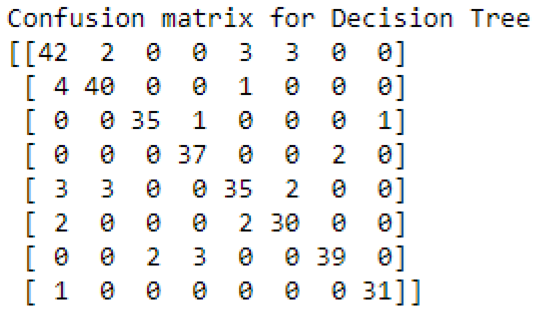


Figure 9: KNN Performance plot in Tuning

Performance of Model – Decision Tree

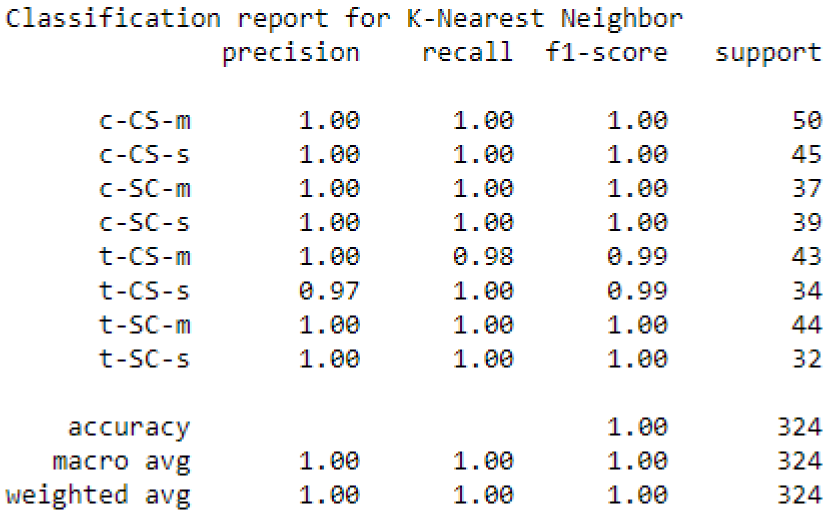


```
Classification report for Decision Tree
              precision    recall  f1-score   support

      c-CS-m       0.81      0.84      0.82        50
      c-CS-s       0.89      0.89      0.89        45
      c-SC-m       0.95      0.95      0.95        37
      c-SC-s       0.90      0.95      0.92        39
      t-CS-m       0.85      0.81      0.83        43
      t-CS-s       0.86      0.88      0.87        34
      t-SC-m       0.95      0.89      0.92        44
      t-SC-s       0.97      0.97      0.97        32

    accuracy                           0.89       324
   macro avg       0.90      0.90      0.90       324
weighted avg       0.89      0.89      0.89       324
```

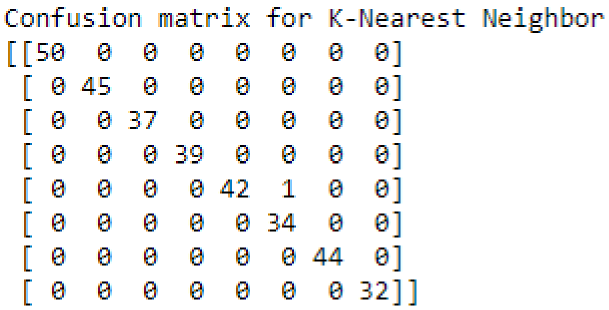


```
Confusion matrix for Decision Tree
[[42  2  0  0  3  3  0  0]
 [ 4 40  0  0  1  0  0  0]
 [ 0  0 35  1  0  0  0  1]
 [ 0  0  0 37  0  0  2  0]
 [ 3  3  0  0 35  2  0  0]
 [ 2  0  0  0  2 30  0  0]
 [ 0  0  2  3  0  0 39  0]
 [ 1  0  0  0  0  0  0 31]]
```

Performance – K-Nearest Neighbour.

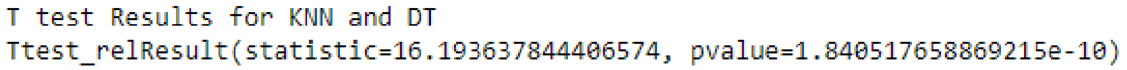


```
Classification report for K-Nearest Neighbor
              precision    recall  f1-score   support

      c-CS-m       1.00      1.00      1.00        50
      c-CS-s       1.00      1.00      1.00        45
      c-SC-m       1.00      1.00      1.00        37
      c-SC-s       1.00      1.00      1.00        39
      t-CS-m       1.00      0.98      0.99        43
      t-CS-s       0.97      1.00      0.99        34
      t-SC-m       1.00      1.00      1.00        44
      t-SC-s       1.00      1.00      1.00        32

    accuracy                           1.00       324
   macro avg       1.00      1.00      1.00       324
weighted avg       1.00      1.00      1.00       324
```



```
Confusion matrix for K-Nearest Neighbor
[[50  0  0  0  0  0  0  0]
 [ 0 45  0  0  0  0  0  0]
 [ 0  0 37  0  0  0  0  0]
 [ 0  0  0 39  0  0  0  0]
 [ 0  0  0  0 42  1  0  0]
 [ 0  0  0  0  0 34  0  0]
 [ 0  0  0  0  0  0 44  0]
 [ 0  0  0  0  0  0  0 32]]
```

The paired t-test results of DT and KNN



```
T test Results for KNN and DT
Ttest_relResult(statistic=16.193637844406574, pvalue=1.840517658869215e-10)
```

DISCUSSION

The main objective of this project was to classify eight types or classes of mice based on the proteins present in them. Apart from that a minor goal of exploring and finding relationships or answers to Hypothesis were also taken in to account. Many proteins showed very low to no relationship or correlation with another protein. For the main goal of the project, which was to classify the mice to 8 classes, was successfully completed by introducing two well-known classifiers in Classification Modelling namely Decision Tree and K-Nearest Neighbour. The classifiers were tuned as much as possible with the limited time frame and with the limited computation power to derive the optimum results. The KNN model won was deemed optimal in terms of accuracy as well as by looking at the confusion matrix itself. The Decision tree had a few mismatches in the matrix while KNN had only 1 mismatch. The paired t-test showed that there is a statistical significant difference in using KNN over DT with an extremely smaller p value. (less than 0.05 for 95% Confidence Interval means it's statistically significant) But considering the test sample size, it should be noted that these results may slightly change with introduction of new samples which is encouraged. Whether the KNN was over fitting the model is a slight concern and therefore testing with newer data would help us understand this much more.

CONCLUSION

As I said in the very beginning of the report, mice are extremely useful for experimentation purposes. For this project, a dataset based on characteristics primarily on mice with down syndrome and mice without down syndrome were used. The task was to classify the mice into 8 classes derived from normal or down syndrome which was completed with acceptable results. Though the project mainly focused on a classification problem, the same data can be applied for clustering purposes as well. Also, other tasks such as finding out the proteins presented mainly in down syndrome mice (Vaguely explored here) would lead to more scientific discovery of this genetic disorder which is even presented in humans. Human brain and cognition may not be exactly similar to that of a mice but most genomes are presented in both parties, so using data science to explore such important medical issues would be highly appreciated by the society.

REFERENCES

Katheleen J. Gardiner (n.d). Mice Protein Expression Dataset. Retrieved June 10, 2020, from https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression

V. Aksakalli,Z. Yenice, Y. Kai Wong, I. Ture, M. Malekipirbazari (n.d.). Feature Selection and Ranking in Machine Learning. Retrieved June 10, 2020, from http://www.featureranking.com

Kate Kreshnar (n.d.), Why do we experiment on mice?. Retrieved June 10, 2020, from https://science.howstuffworks.com/innovation/scientific-experiments/experiment-on-mice.htm