**BT5450**

**PROJECT**

**MICE PROTEIN EXPRESSION**

**ATHUL KRISHNA KS**

**CE18B111**

**ABSTRACT**

This project contains the data analysis of expression levels of proteins measured in the cerebral cortex. The dataset is about the expression levels of different proteins that will produce significant signals during the nuclear fraction of cortex. 1080 samples are measured. Mice can be classified based on factors like genotype , treatment and their behavior. The features of normal and trisomic mice or mice with down syndrome can be learned from the analysis of dataset. Special drugs are given to the mouse for observing the capacity of learning in trisomic category. We have to classify the protein samples into eight classes. The aim of this project to identify the trisomy protein that influence the success and failure of mice learning. Down syndrome is a situation in which chromosomes are disordered by the presence of an extra chromosome.

**INTRODUCTION**

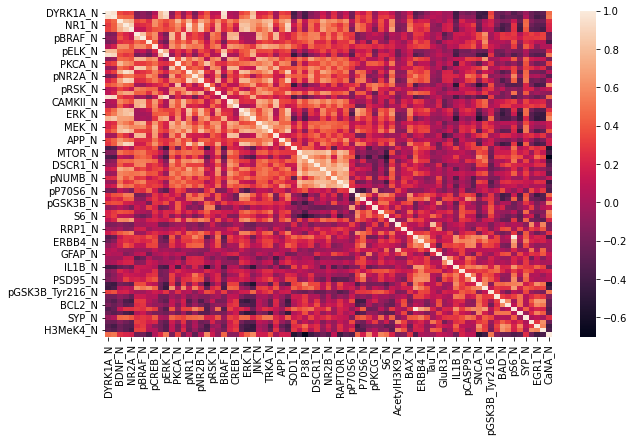
This project explores the dataset of mice protein expression.

The success rate of each protein has to be learned using their expression levels.The results will be very useful in researches related to nulear fraction of cortex or manufacturing of drugs.the signals produced by the protein cells during nuclear fraction can be detected using equipments.The dataset contain expression levels of 77 proteins.

The aim is to identify the trisomy protein classes that is more importantto the success and the failure of mice learning. Classification model will predict the relevant classes of mice analysing their expression levels. We can then identify the proteins which are significant in the predictions. wSuch proteins will be having a huge impact on learning in trisomic mice.

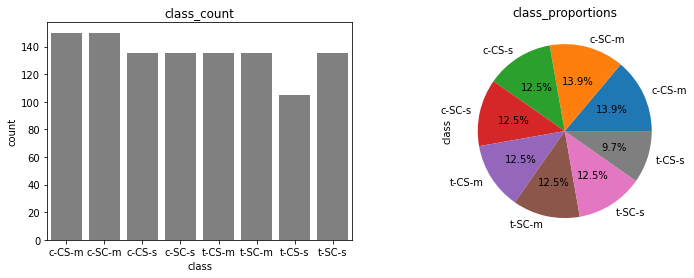
**DESCRIPTIVE STATISTICS AND VISUALIZATION**

The relation between the classes in dataset can be examined by plotting a correlation heatmap.



The correlation among classes can be seen from the above heatmap

Generate a bar chart to give the proportion of different classes in the dataset.



Expression value of proteins can be visualized using violin plots.

Violin plots are useful to see full distribution of the data with their probability density.

The mean value of each protein is first calculated and the data frame

is sorted based on their mean magnitudes.

|  |  |
| --- | --- |
|  |  |

Fig:Violin plots to visualize the data

The violin plots shows the importance of data standardization based on the expression values of each protein.

Also, the plots show most of the data have large ranges of outliers. For learning about the outliers , we can do inspection of protein by using Box plots

The two proteins with highest range of outliers can be examined:

DYRK1A\_N and AcetylH3K9\_N.

|  |  |
| --- | --- |
|  |  |

These pairwise explorations show most of the big range of data magnitudes caused by outliers that usually dominated by one of the class. For example, big range of data magnitudes in protein

DYRK1A\_N is caused by outliers from c-CS-s class.

Now we will go deeper into each class that contains those outliers by examining the mouse numbers in those class . The result is below:

|  |  |
| --- | --- |
|  |  |

Most of the relevant outliers in each class comes from specific mouse numbers.

Analysing the type of mice the obtained mouse numbers to get more details of the outliers that will affect the data magnitudes. The results are shown below using the graphs:

|  |  |
| --- | --- |
|  |  |

**PROBLEM FORMULATION AND PREDICTIVE MODELS**

Classification model can be used to formulate the mice proteins.

Data for x axis and y axis has to be generated from the dataset. The input data can be scaled using standard scalar library. The model should need training data and testing dataset to check its accuracy after the formulation. I have used the support vector classifier for classification. The early generated testing and training data is passed into the classifier.

Cross\_val\_score is used to check the accuracy of the model on different folds.Below is the accuracy on 5 selected folds:

[0.54054054 0.85585586 0.78181818 0.91818182 0.78181818]

The accuracy on training dataset is 1

Final score of the classifier is 0.9879518072289156

**Model 1: Random Forest Classification**

This model helps to analyse the relation between two or more

Independent variables.It will find out the important factors affecting the dependent variable and also the relation between dependent

and independent variables.

Random Forest Classifiaction generate a tree like structure and each branch will return the most important classes in the data.

Threshold of 0.014 is used to find out the important proteins.Below is the list of important proteins in the data with an importance value above 0.014

**No Important Proteins Importance Value**

32 SOD1\_N 0.058882

46 pPKCG\_N 0.039323

76 CaNA\_N 0.036684

30 APP\_N 0.033064

65 Ubiquitin\_N 0.032740

7 pCAMKII\_N 0.027612

53 ARC\_N 0.026766

10 pERK\_N 0.025673

0 DYRK1A\_N 0.025500

1 ITSN1\_N 0.025262

17 pPKCAB\_N 0.025006

56 Tau\_N 0.022867

20 BRAF\_N 0.022289

34 P38\_N 0.021871

42 pP70S6\_N 0.020636

45 pGSK3B\_N 0.019774

70 pS6\_N 0.019454

19 AKT\_N 0.018948

39 pNUMB\_N 0.018189

50 AcetylH3K9\_N 0.017311

48 S6\_N 0.017119

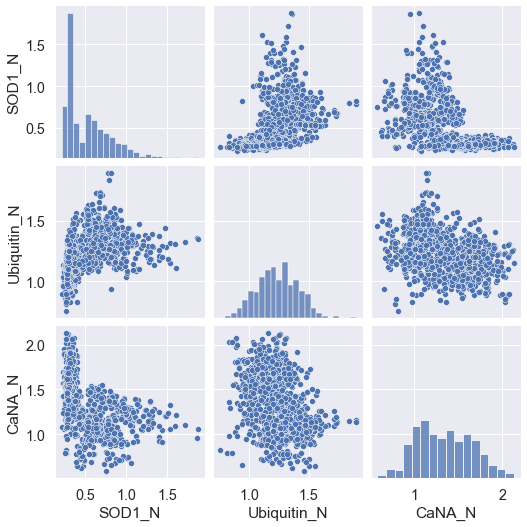
**The accuracy of the model is 0.9907407407407407**

**Model 2: K-Nearest Neighbours**

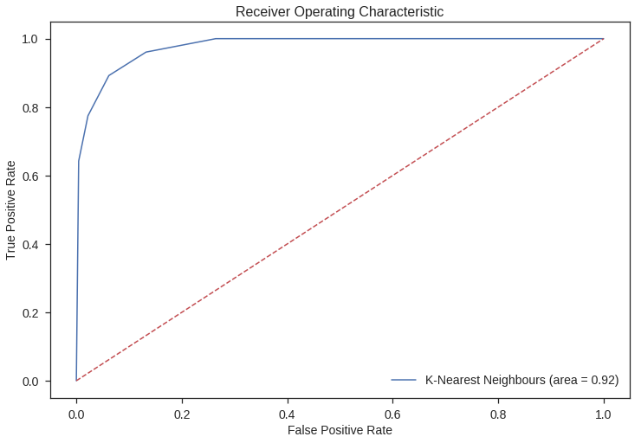
This model identifies the data points that are separated into several classes to predict the classification of a new sample point.

The value of k has to be initialized.Calculate the distance between the query point and current point for each sample in the training data.Sort the ordered collection of distances and indices in ascending order.Select first k entries and their labels from the collection.Then return the mode of k labels.

Correlation chart between top three proteins:

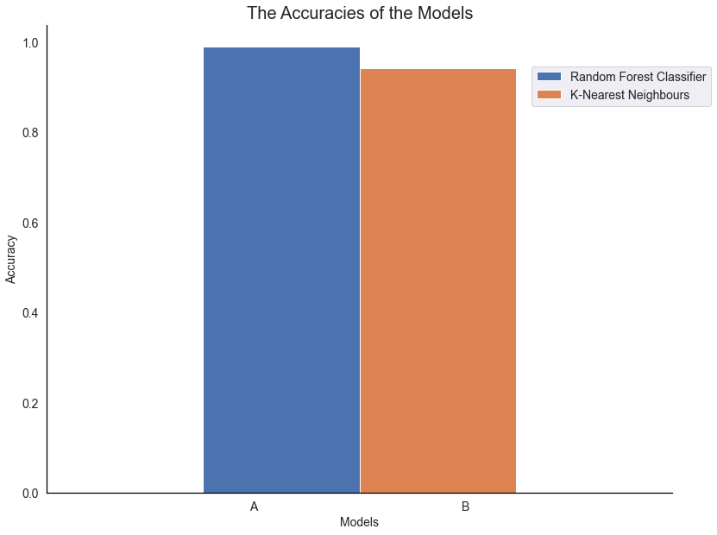
****

Receiver operating characteristic curve of the KNN model



**RESULTS**

Classification using two predictive models has been performed.The below graph depicts the accuracy of the two models.

****

The results obtained after performing classification :

* Top three protein behaviour in mouse are SOD1\_N, pPKCG\_N, and ITSN1\_N
* The Relative Importance of features in Random Forests drop off a lot more quickly than in K Nearest Neighbours

**CONCLUSION**

After performing the analysis using the predictive models,we undersand the important proteins that affect the behaviour of mouse. The top 3 proteins for Treatment and Behaviour were SOD1\_N, pPKCG\_N and pCAMKII\_N. The protein expression levels of these proteins are significantly affected by treatment and behaviour.

Two classification models were used.better results are given by the random forest model when compared to the k nearest neighbour model.The main reason behind this is that the random forest model is using multiple decision trees to avoid overfitting of data.

The correlation between top three proteins was also studied and the corelation graph shows how much they are correlated to each other.

Both models identified the proteins that are important contributers to the learning of mice

**REFERENCES**

* https://seaborn.pydata.org/generated/seaborn.boxplot.html
* https://www.simplilearn.com/tutorials/machine-learning-tutorial/classification-in-machine-learning
* https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest
* https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761