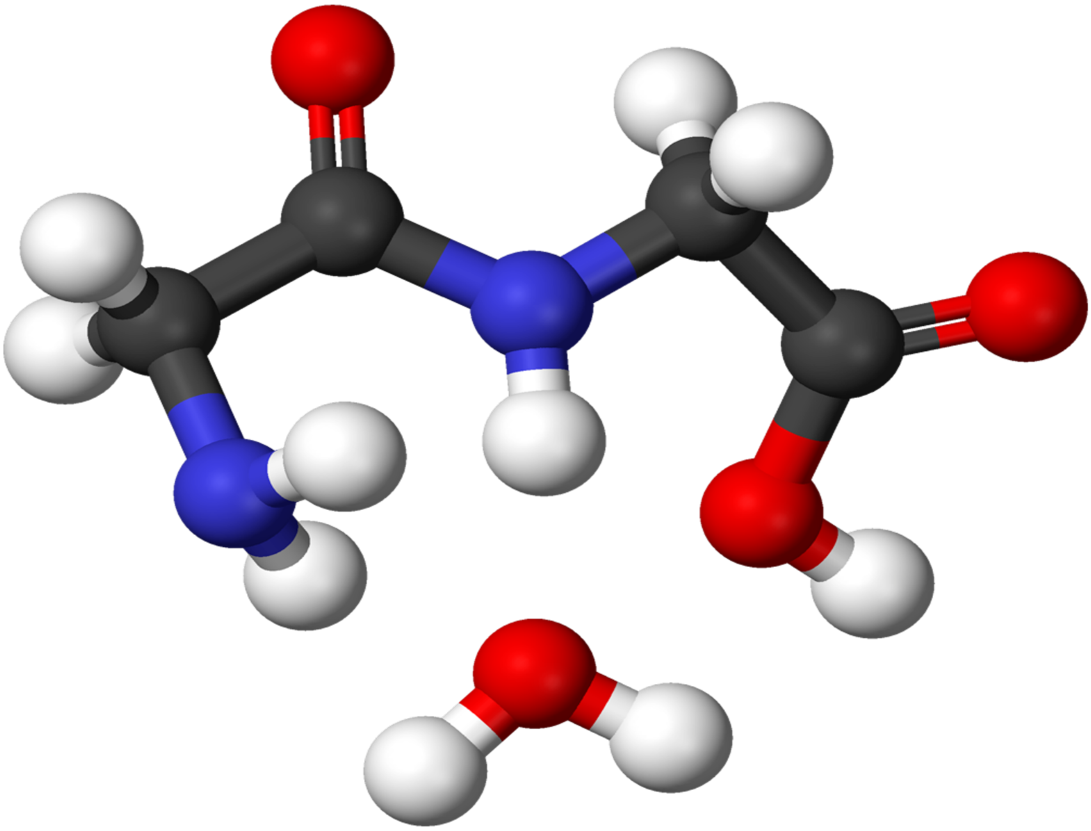


Comparative Classification Analysis to Identify the Subset of Mice based on the Protein Data:

KNN and Decision Tree

Table of Contents

I. Abstract: 2

II. Introduction: 2

III. Methodology: 2

3.1 Dataset: 2

3.2 Data Pre-processing: 2

3.3DataExploration: 3

3.3.1 Exploration of each attribute: 3

3.3.2 Exploration of pair of attributes: 5

3.4 Data Modeling: 7

3.4.1 K – Nearest Neighbour 7

3.4.2 Decision Tree 7

IV. Results: 8

4.1 KNN Results: 8

4.2 Decision Tree Results: 9

V. Conclusion: 10

VI. References 10

I. 1. Abstract:

Mice protein expression data was created to study the effect of learning between normal and trisomic mice or mice with Down Syndrome (DS). This research will help us to analyse and classify the protein samples based on the influence of proteins. Two classification models were used, and they classified the protein samples into 8 different class of mice with high accuracy.

II. 2. Introduction:

Down syndrome is a chromosomal abnormality, which occurs when there is an extra copy of chromosome 21. Down syndrome is a lifelong condition that is associated with cognitive disabilities and physical abnormalities.

For the preclinical evaluation of effectiveness of the drug, [1] had created a protein expression data of 38 control mice and 34 trisomic mice i.e. with Down Syndrome. This dataset is categorised based on their Genotype (control (c) or trisomy (t)) , Treatment Type (memantine (m) or saline (s)), Behavior (context-shock (CS) or shock-context (SC)).

This research is analyse the protein samples to classify the mice into one of the 8 types using two data mining classification techniques: K- Nearest Neighbour and Decision Trees. A comparative analysis has been performed between these two techniques.

III. 3. Methodology:

3.1 Dataset:

The dataset has been taken from UCI Machine Learning Repository. The data set consists of the expression levels of 77 proteins/protein modifications that produced detectable signals in the nuclear fraction of cortex. Based on the features such as genotype, behaviour and treatment mice are classified into 8 types. According to genotype, mice can be control or trisomic. According to behaviour, 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. The dataset contains a total of 1080 observations with 77 numeric attributes and 4 categorical attributes.

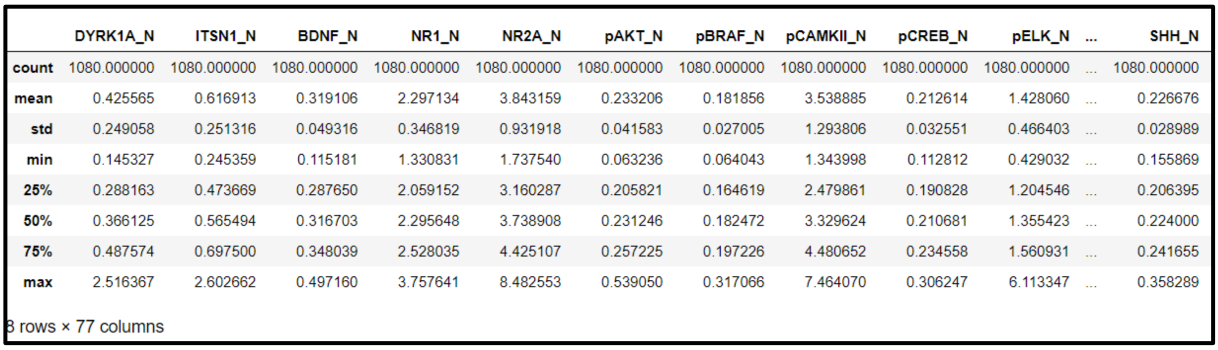
3.2 Data Pre-processing:

Some of the data attributes have missing values. The missing values are replaced with the mean value of the that expression level of the same class i.e. if the a protein expression has missing value then that value is replaced with the mean value of the proteins belonging to the same class.

3.3DataExploration:

3.3.1 Exploration of each attribute:

Statistical description of the numerical attributes can be found in the Fig.1. It contains the count of number of observations, mean, standard deviation, minimum and max of that attribute. Also, it contains 0.25 percent, 0.5 percent and 0.75 percent quartile of that attribute.



	DYRK1A_N	ITSN1_N	BDNF_N	NR1_N	NR2A_N	pAKT_N	pBRAF_N	pCAMKII_N	pCREB_N	pELK_N	...	SHH_N
count	1080.000000	1080.000000	1080.000000	1080.000000	1080.000000	1080.000000	1080.000000	1080.000000	1080.000000	1080.000000	...	1080.000000
mean	0.425565	0.616913	0.319106	2.297134	3.843159	0.233206	0.181856	3.538885	0.212614	1.428060	...	0.226676
std	0.249058	0.251316	0.049316	0.346819	0.931918	0.041583	0.027005	1.293806	0.032551	0.466403	...	0.028989
min	0.145327	0.245359	0.115181	1.330831	1.737540	0.063236	0.064043	1.343998	0.112812	0.429032	...	0.155869
25%	0.288163	0.473669	0.287650	2.059152	3.160287	0.205821	0.164619	2.479861	0.190828	1.204546	...	0.206395
50%	0.366125	0.565494	0.316703	2.295648	3.738908	0.231246	0.182472	3.329624	0.210681	1.355423	...	0.224000
75%	0.487574	0.697500	0.348039	2.528035	4.425107	0.257225	0.197226	4.480652	0.234558	1.560931	...	0.241655
max	2.516367	2.602662	0.497160	3.757641	8.482553	0.539050	0.317066	7.464070	0.306247	6.113347	...	0.358289

8 rows × 77 columns

Fig.1

Each categorical attribute can be explored by pie charts. Pie chart provides us the details of the percentage of the existence of each label. Fig.2 represents the pie chart representation of attribute 'class'.

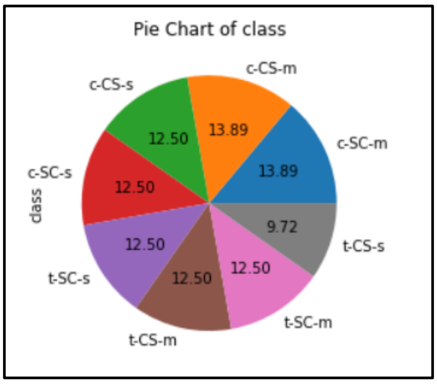


Fig.2.

Each numerical attribute can be explored by either histogram, line graphs, density graphs or boxplots. Histograms provides us the information of the distribution of data. Fig.3 represents the histogram of 'Protien – pCREB_N'. From it we can observe that the data follows normal distribution. Maximum number of mice has the protein values between 0.18 and 0.24.

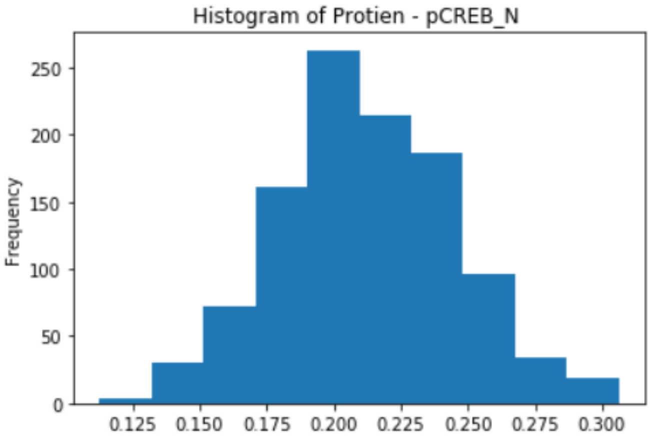


Fig.3.

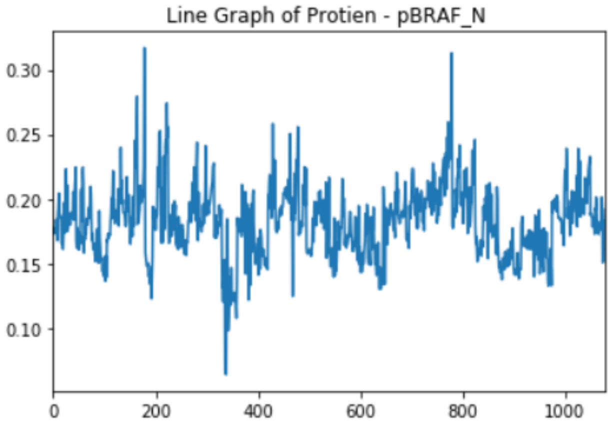


Fig.4.

Fig.4 represents the Line graph of the protein – 'pBRAF_N'. From the graph we can observe that the data oscillates mostly between 0.15 and 0.23

Boxplot drafts the key figures in the distribution helps us to spot outliers. Fig.5. provides us the information of protein – 'NR1_N'. From the boxplot we can observe that the attribute has 3 outliers.

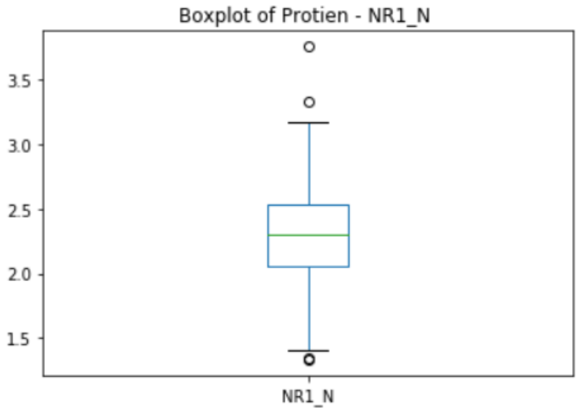


Fig.5.

When we observe the line graph in Fig.6. (represents the line, boxplot ad histogram of protein – ‘PCAMKII_N’) , data is highly distributed and makes us feel the existence of outlier, whereas when we explore the protein the boxplot we can see that there are no outliers and from the histogram we can see that the protein follows right skewed distribution.

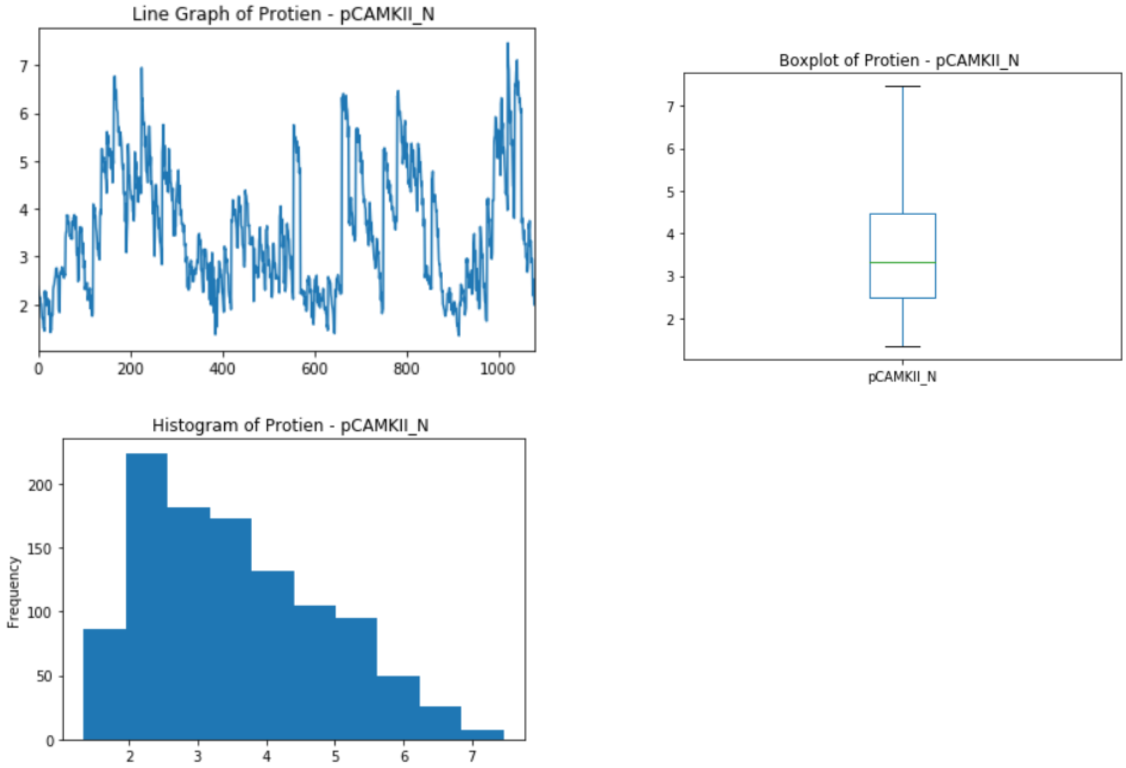


Fig.6.

3.3.2 Exploration of pair of attributes:

Relationship between attributes can be explored with Boxplots by groups, scatter plots and bar graphs.

Considering Box plot of protein ‘DYRK1A_N’ grouped by class can be view in Fig.7. The plot provides us the information of protein with respect to each class. In the figure we can observe that class ‘c=CS-m’ has the no outliers whereas the class ‘c-CS-s’ has many outliers for the same protein data.

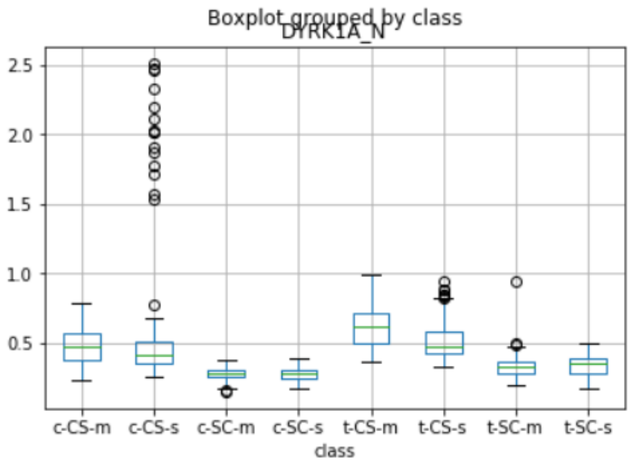


Fig.7.

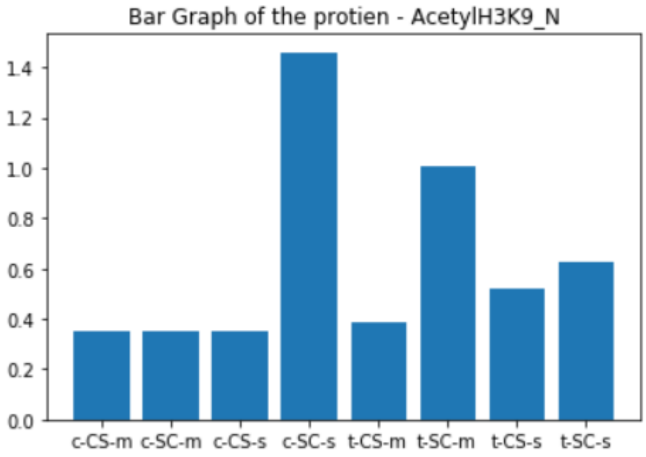


Fig.8.

Bar graph in Fig.8. provides us the details of the count of protein data with respect to the class. From the graph we can clearly state that the maximum number of the protein belongs to the class ‘c-SC-s’, followed by ‘t-SC-m’ and the classes ‘c-CS-m’, ‘c-SC-m’ and ‘c-CS-s’ project that the same number of proteins belongs to each one of class respectively.

Scatter plot between proteins 'pELK_N' and 'ITSN1_N' can be observed in the Fig.9. Most of the protein data are distributed closed to each other but a few of them are distributed very far, representing the outliers.

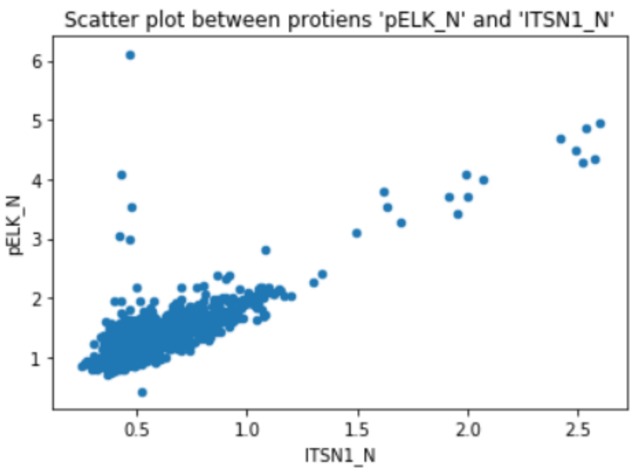


Fig.9.

3.4 Data Modelling:

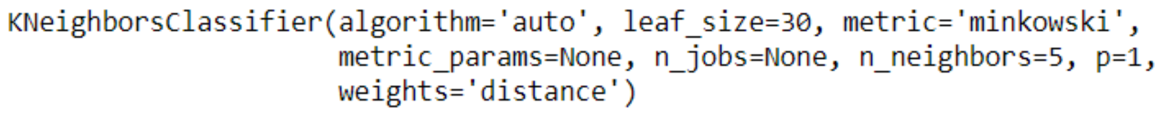
This paper is about comparative classification analysis on data mining classification models: K-Nearest Neighbouring and Decision Tree classification models.

To feed the data to the models we need initially set the data for it. All the numerical attributes are taken as feature variable and 'class' attribute is taken as the target variable. Then the data is split 80% of data as training and 20% as testing data with random_state(seed) as 7 i.e. each time data is takes same number of observations into account.

3.4.1 K – Nearest Neighbour

K- Nearest Neighbours classifies the data based on the 'k' data points similar to the given testing data. It has various other parameters into account such as distance and algorithm to calculate the distance.

For a testing data tuple, KNN takes the 'n_neighbours' data points similar to it. And these data points are weighted based on the parameter 'weight' which could be 'uniform', where are the similar data points are weighted equally and for 'distance' data points are weighted by a distance mentioned by another parameter 'p' i.e. Manhattan (p=1) or Euclidean (p =2). For our data model we have choose '5' neighbours which are weighted by a distance calculated by following Manhattan [Fig.10].



```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=None, n_neighbors=5, p=1,
                     weights='distance')
```

Fig.10.

3.4.2 Decision Tree

Decision Tree is a flow-chart like structure, where features are represented as internal nodes, decision rule is represented as branch and respective outcome is represented by the leaf nodes. The topmost node is called the root node and leaf nodes are labels of target variables. The partition is made based on the attribute value by calculating information gain (GINI index or Entropy). Decision tree visualization helps the human level thinking to make decisions easily.

Fig.11. represents the Decision tree representation of mice protein expression dataset.

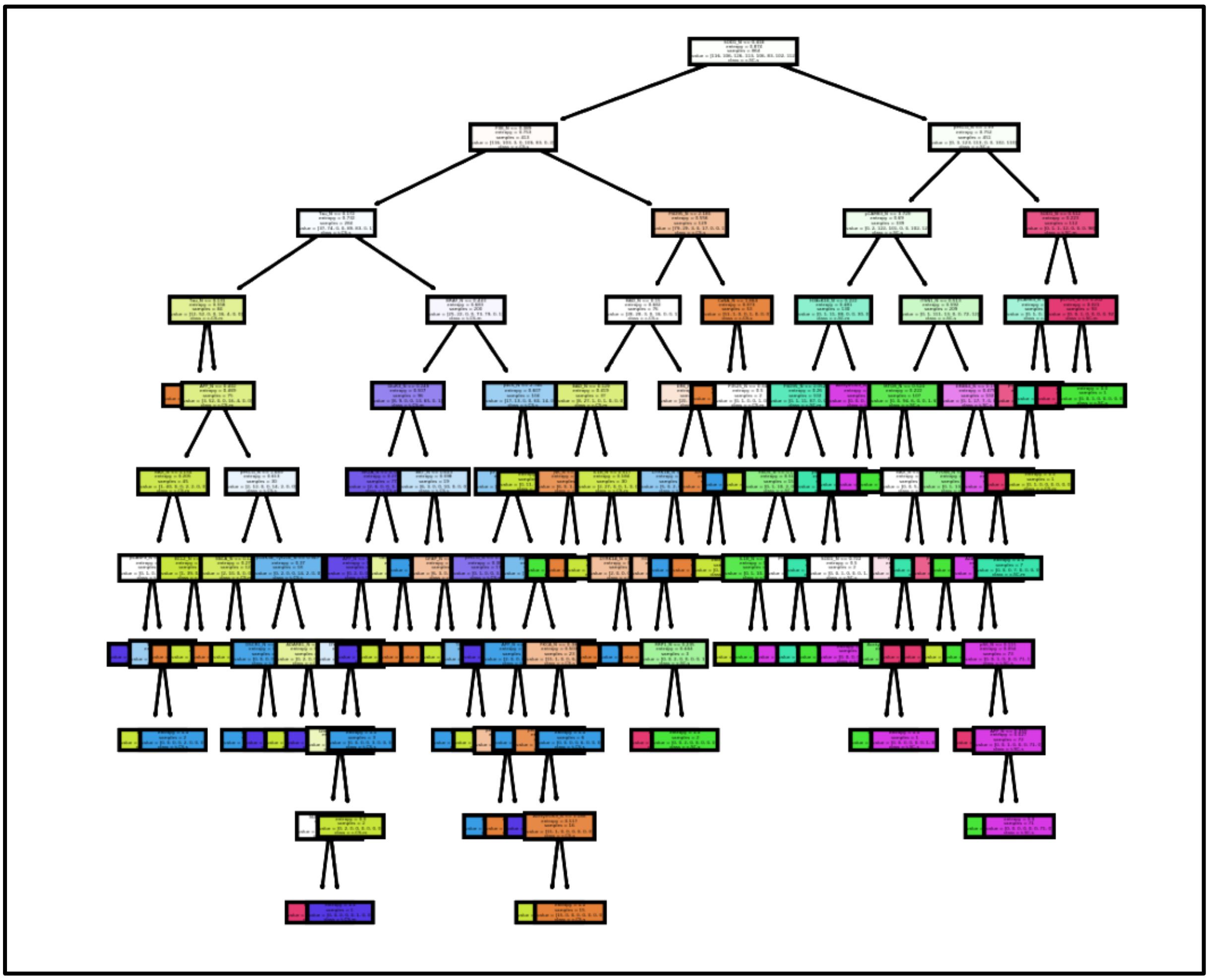


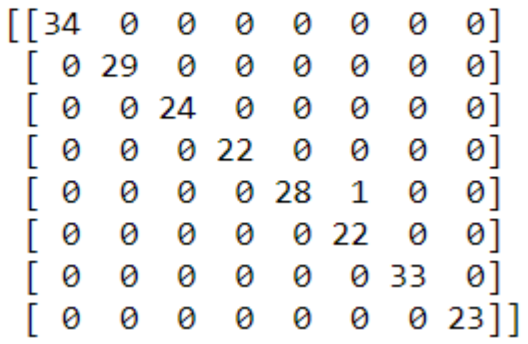
Fig.11.

IV. 4. Results:

After evaluating and predicting the test data on the respective models, confusion matrix and classification reports were generated for each one of them.

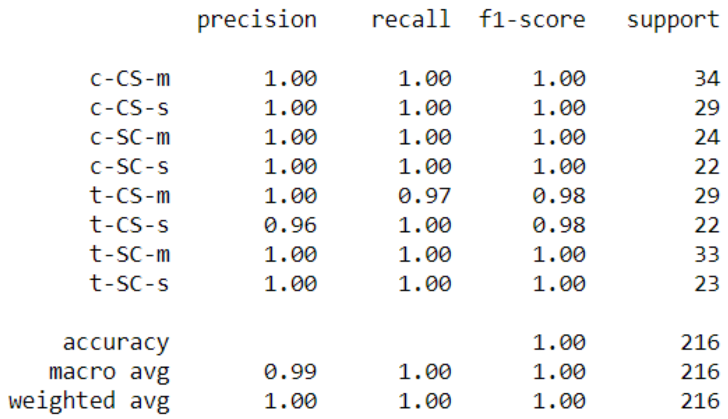
4.1 KNN Results:

Confusion matrix, has not recorded any discrepancies between actual and predicted values and is as follows:



```
[[34  0  0  0  0  0  0  0]
 [ 0 29  0  0  0  0  0  0]
 [ 0  0 24  0  0  0  0  0]
 [ 0  0  0 22  0  0  0  0]
 [ 0  0  0  0 28  1  0  0]
 [ 0  0  0  0  0 22  0  0]
 [ 0  0  0  0  0  0 33  0]
 [ 0  0  0  0  0  0  0 23]]
```

And Classification report is as follows:

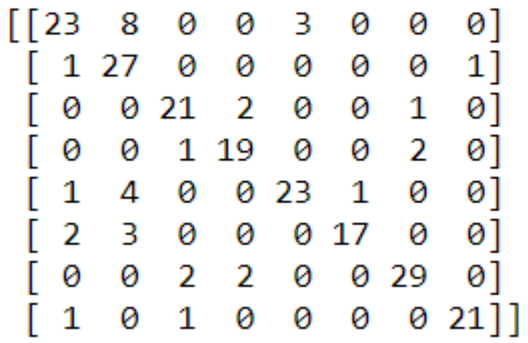


```
              precision    recall  f1-score   support

      c-CS-m       1.00      1.00      1.00        34
      c-CS-s       1.00      1.00      1.00        29
      c-SC-m       1.00      1.00      1.00        24
      c-SC-s       1.00      1.00      1.00        22
      t-CS-m       1.00      0.97      0.98        29
      t-CS-s       0.96      1.00      0.98        22
      t-SC-m       1.00      1.00      1.00        33
      t-SC-s       1.00      1.00      1.00        23

    accuracy                           1.00       216
   macro avg       0.99      1.00      1.00       216
weighted avg       1.00      1.00      1.00       216
```

KNN records the accuracy score as 99.53%.

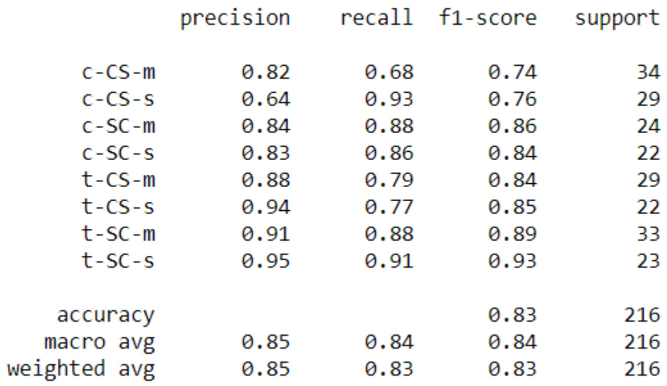
4.2 Decision Tree Results:

Confusion matrix of decision tree has recorded small discrepancies between actual and predicted is as follows:



```
[[23  8  0  0  3  0  0  0]
 [ 1 27  0  0  0  0  0  1]
 [ 0  0 21  2  0  0  1  0]
 [ 0  0  1 19  0  0  2  0]
 [ 1  4  0  0 23  1  0  0]
 [ 2  3  0  0  0 17  0  0]
 [ 0  0  2  2  0  0 29  0]
 [ 1  0  1  0  0  0  0 21]]
```

Classification report is as follows:



```
              precision    recall  f1-score   support

      c-CS-m       0.82      0.68      0.74        34
      c-CS-s       0.64      0.93      0.76        29
      c-SC-m       0.84      0.88      0.86        24
      c-SC-s       0.83      0.86      0.84        22
      t-CS-m       0.88      0.79      0.84        29
      t-CS-s       0.94      0.77      0.85        22
      t-SC-m       0.91      0.88      0.89        33
      t-SC-s       0.95      0.91      0.93        23

    accuracy                           0.83       216
   macro avg       0.85      0.84      0.84       216
weighted avg       0.85      0.83      0.83       216
```

Decision Tree records the score as 83.33%

V. 5. Conclusion:

As the KNN records the highest score with 99.5% when compared to 83.33% for decision tree. Also, precision, recall and F1 are 1.0 each for the KNN but decision tree has 0.82 , 0.68 and 0.74 respectively.

On comparing both the classification models we get better results for K-Nearest Neighbour and hence it is the better classification model out of two.

VI. 6. References

1. https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression#
2. Higuera, C., Gardiner, K. J., & Cios, K. J. (2015). Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. PloS one, 10(6).
3. Saringat, M. Z., Mustapha, A., & Andeswari, R. (2018). Comparative Analysis of Mice Protein Expression: Clustering and Classification Approach. International Journal of IntegratedEngineering,10(6).