**PROTEIN EXPRESSION ANALYSIS FOR DOWN SYNDROME**

**1.ABSTRACT:**

Down syndrome (DS) is a chromosomal disorder where organisms have an extra chromosome 21, sometimes known as trisomy 21. Being a syndrome, DS consists of multiple symptoms affecting a large number of systems in the body. It’s effect on learning results in it being an intellectual disability . Memantine is currently proposed as a treatment of the learning deficit symptoms in DS. In this project, we have used several supervised machine learning methods:logistic regression, a random forest classifier,k nearest neighbour and voting classifier, to identify which protein(s) are critical to mice learning ability after being exposed to context fear conditioning (CFC). 77 protein expression levels are analysed from both control and trisomic (Ts65Dn) genotype mice, both with and without treatment from the drug memantine. Results suggest that voting classification approach can identify the most important proteins which may help to identify more effective drugs to help the learning process in people with DS.

**2.GOAL:**

Identifying proteins critical to learning ability and classification of mice based on protein expression values.

**3.DATASET DESCRIPTION:**

* Expression levels of 77 proteins measured in the cerebral cortex of 8 classes of control and Down syndrome mice exposed to context fear conditioning, a task used to assess associative learning.
* The data set consists of the expression levels of 77 proteins/protein modifications that produced detectable signals in the nuclear fraction of the cortex.
* There are 38 control mice and 34 trisomic mice (Down syndrome), for a total of 72 mice.
* In the experiments, 15 measurements were registered of each protein per sample/mouse. Therefore, for control mice, there are 38x15, or 570 measurements, and for trisomic mice, there are 34x15, or 510 measurements.
* The dataset contains a total of 1080 measurements per protein.
* Each measurement can be considered as an independent sample/mouse.
* The eight classes of mice are described based on features such as genotype, behavior and treatment. According to genotype, mice can be control or trisomic.
* According to behavior, 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.

**LIBRARIES USED:**

* numpy
* pandas
* seaborn
* matplotlib
* sklearn
* LogisticRegression
* KNeighborsClassifier
* VotingClassifier
* model\_selection
* confusion\_matrix
* classification\_report
* roc\_curve
* f1\_score
* roc\_auc\_score
* StandardScaler
* train\_test\_split
* GridSearchCV
* SelectKBest,f\_classif
* seaborn
* interp
* cycle
* feature\_selection

**SHAPE OF THE DATASET:**

Total number of rows = 1080

Total number of columns = 82

**COLUMNS OF DATASET:**

'MouseID', 'DYRK1A\_N', 'ITSN1\_N', 'BDNF\_N', 'NR1\_N', 'NR2A\_N', 'pAKT\_N','pBRAF\_N', 'pCAMKII\_N', 'pCREB\_N', 'pELK\_N', 'pERK\_N', 'pJNK\_N','PKCA\_N', 'pMEK\_N', 'pNR1\_N', 'pNR2A\_N', 'pNR2B\_N', 'pPKCAB\_N','pRSK\_N', 'AKT\_N', 'BRAF\_N', 'CAMKII\_N', 'CREB\_N', 'ELK\_N', 'ERK\_N','GSK3B\_N', 'JNK\_N', 'MEK\_N', 'TRKA\_N', 'RSK\_N', 'APP\_N', 'Bcatenin\_N','SOD1\_N', 'MTOR\_N', 'P38\_N', 'pMTOR\_N', 'DSCR1\_N', 'AMPKA\_N', 'NR2B\_N','pNUMB\_N', 'RAPTOR\_N', 'TIAM1\_N', 'pP70S6\_N', 'NUMB\_N', 'P70S6\_N','pGSK3B\_N', 'pPKCG\_N', 'CDK5\_N', 'S6\_N', 'ADARB1\_N', 'AcetylH3K9\_N','RRP1\_N', 'BAX\_N', 'ARC\_N', 'ERBB4\_N', 'nNOS\_N', 'Tau\_N', 'GFAP\_N','GluR3\_N', 'GluR4\_N', 'IL1B\_N', 'P3525\_N', 'pCASP9\_N', 'PSD95\_N','SNCA\_N', 'Ubiquitin\_N', 'pGSK3B\_Tyr216\_N', 'SHH\_N', 'BAD\_N', 'BCL2\_N',

'pS6\_N', 'pCFOS\_N', 'SYP\_N', 'H3AcK18\_N', 'EGR1\_N', 'H3MeK4\_N',

'CaNA\_N', 'Genotype', 'Treatment', 'Behavior', 'class'

### Attribute Information:

[1] Mouse ID

[2:78] Values of expression levels of 77 proteins; the names of proteins are followed by N indicating that they were measured in the nuclear fraction. *For example: DYRK1A\_n*

[79] Genotype: control (c) or trisomy (t)

[80] Treatment type: memantine (m) or saline (s)

[81] Behavior: context-shock (CS) or shock-context (SC)

[82] Class: c-CS-s, c-CS-m, c-SC-s, c-SC-m, t-CS-s, t-CS-m, t-SC-s, t-SC-m

**TARGET COLUMN:**

**c-CS-s:**

control mice, stimulated to learn, injected with saline (9 mice)

**c-CS-m:**

control mice, stimulated to learn, injected with memantine (10 mice)

**c-SC-s:**

control mice, not stimulated to learn, injected with saline (9 mice)

**c-SC-m**:

control mice, not stimulated to learn, injected with memantine (10 mice)

**t-CS-s:**

trisomy mice, stimulated to learn, injected with saline (7 mice)

**t-CS-m:**

trisomy mice, stimulated to learn, injected with memantine (9 mice)

**t-SC-s:**

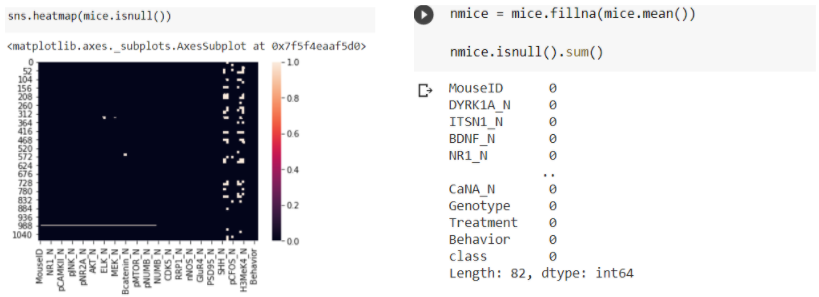
trisomy mice, not stimulated to learn, injected with saline (9 mice)

**t-SC-m:**

trisomy mice, not stimulated to learn, injected with memantine (9 mice)

**4.DATA PREPROCESSING:**

* Checking for null values and replacing it with the mean values of the columns.



* Identification of categorical and numerical attributes

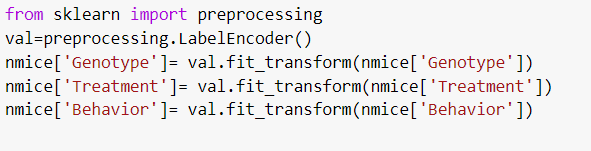
**Categorical attributes:**

|  |
| --- |
| MOUSE ID |
| GENOTYPE |
| TREATMENT |
| BEHAVIOUR |
| CLASS |

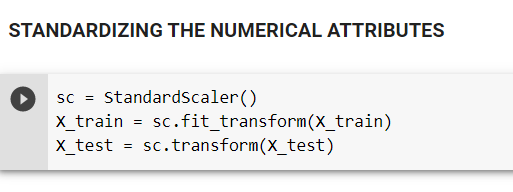
* Finding the values in each of the categorical attributes.

|  |  |
| --- | --- |
| **COLUMN NAME** | **VALUES** |
| GENOTYPE | 'Control' , 'Ts65Dn' |
| TREATMENT | 'Memantine' , 'Saline' |
| BEHAVIOUR | 'C/S' , 'S/C' |
| CLASS | 'c-CS-m', 'c-SC-m', 'c-CS-s', 'c-SC-s' 't-CS-m', 't-SC-m', 't-CS-s', 't-SC-s' |

* Encoding categorical attributes to numerical values using LabelEncoder.



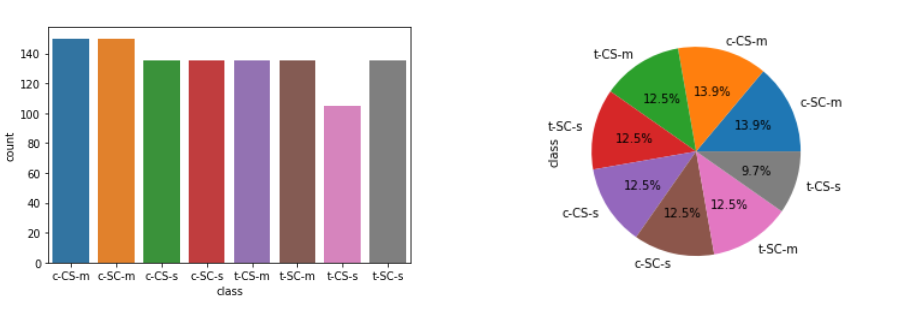
* Standardizing the numerical attributes such that the data is internally consistent using StandardScaler() .
* StandardScaler scales each input separately by subtracting the mean and dividing by the standard deviation to have a mean of 0 and a unit standard deviation.



**5.EXPLORATORY DATA ANALYSIS:**

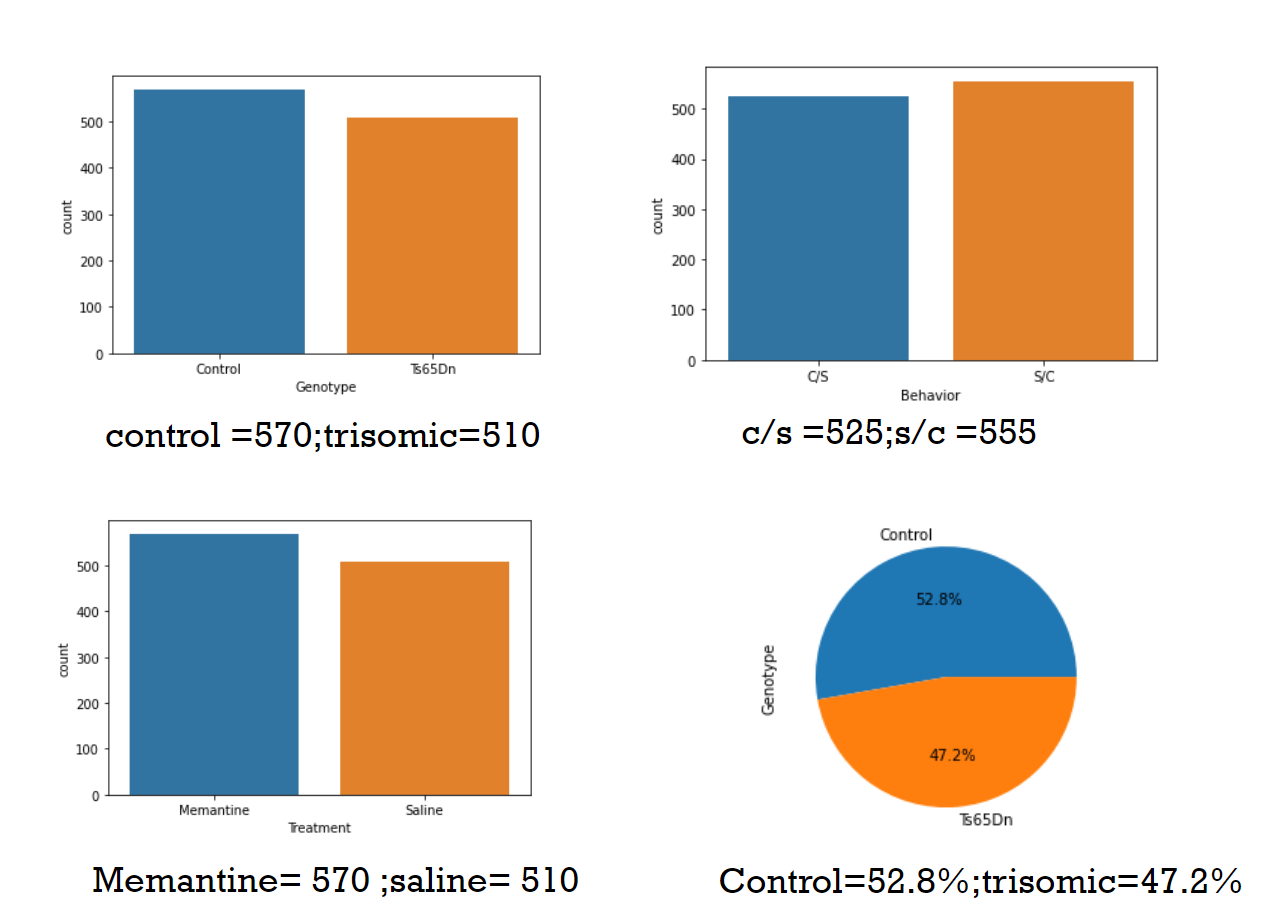
* Value count of each class using countplot and percentage of distribution of values in the class label using pie chart.





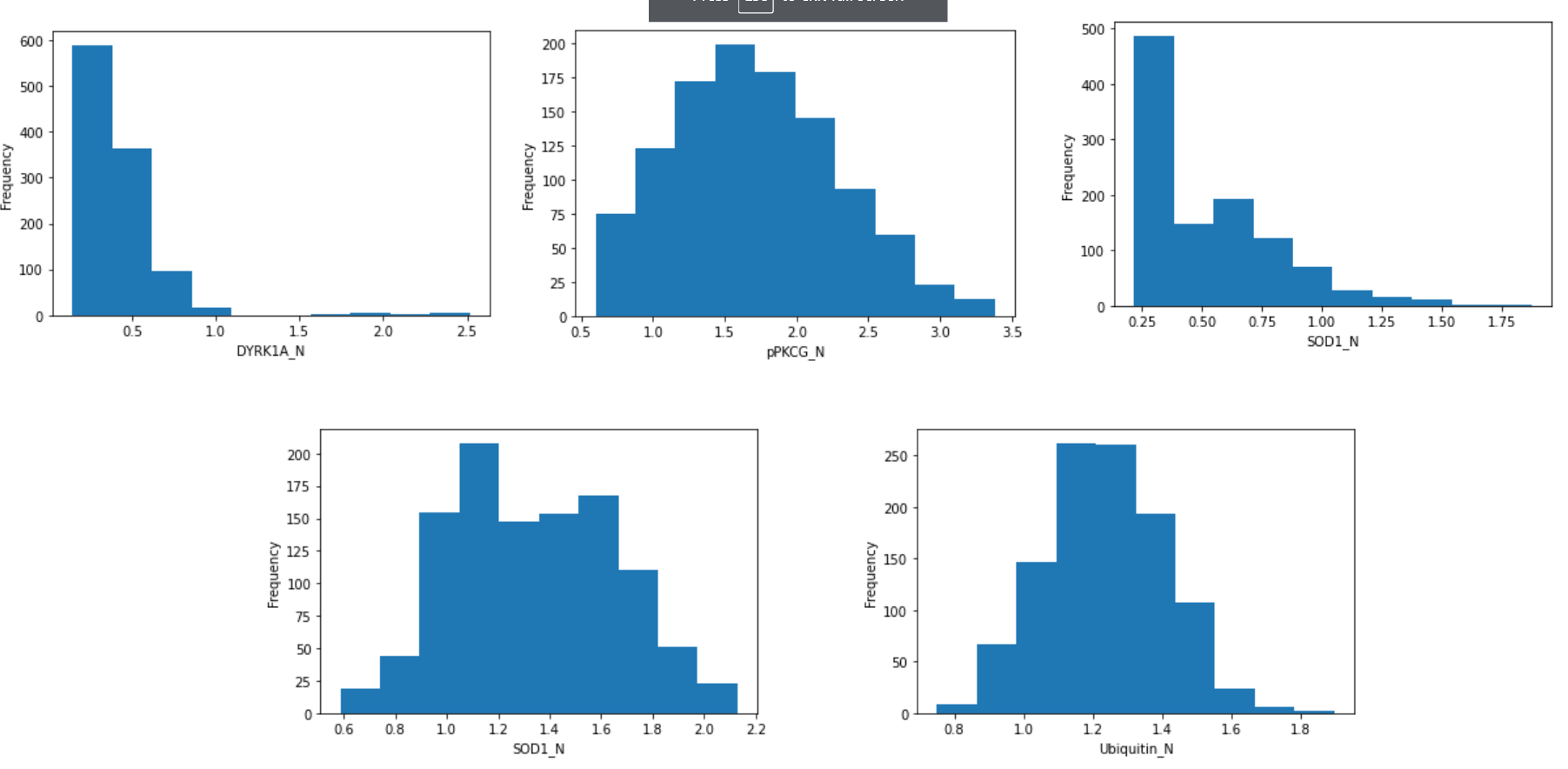
* Value count of each categorical attribute using count plot.



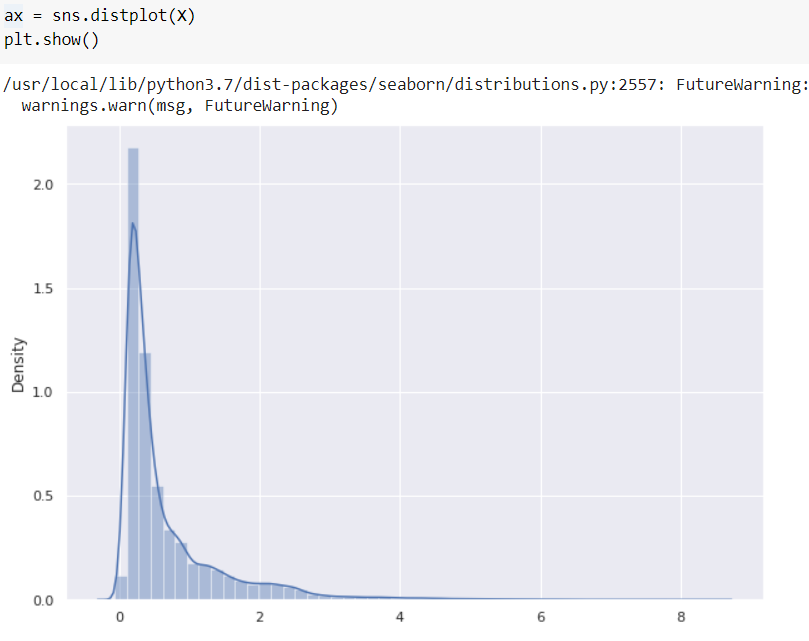


* Histogram for the numerical attributes i.e., the protein values.





* Distplot for the dataset(**distplot** lets you show a histogram with a line on it).

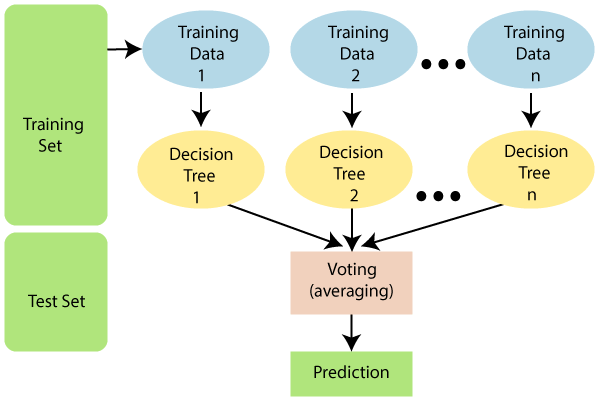
****

we can infer thatthe protein values in the dataset are normally distributed.

**6.MODEL BUILDING PHASES AND MODEL EVALUATION:**

**6.1 RANDOM FOREST CLASSIFIER:**

* Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique.
* It can be used for both Classification and Regression problems in ML.
* Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

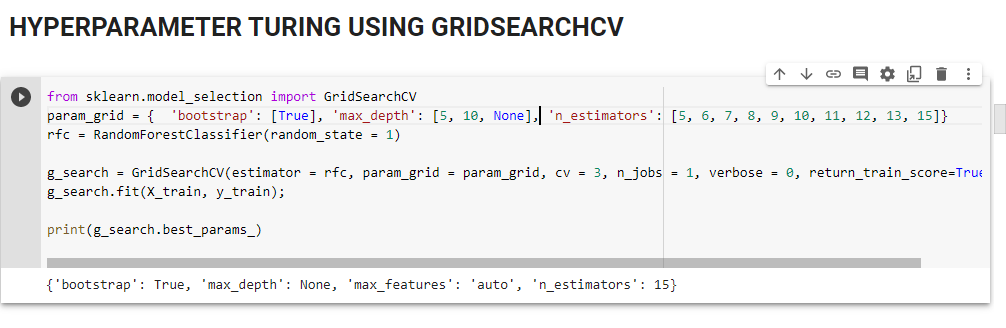


**HYPERPARAMETERS OF RANDOM FOREST CLASSIFIERS:**

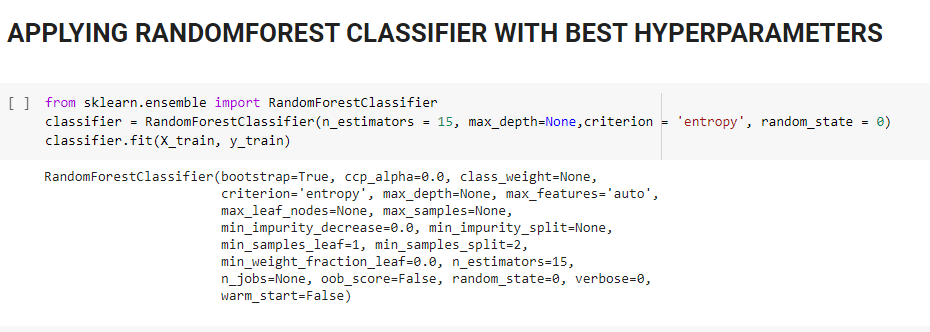
* **n\_estimators** = number of trees in the forest
* **max\_features** = max number of features considered for splitting a node
* **max\_depth** = max number of levels in each decision tree
* **min\_samples\_split** = min number of data points placed in a node before the node is split
* **min\_samples\_leaf** = min number of data points allowed in a leaf node
* **bootstrap** = method for sampling data points (with or without replacement)

**TUNING THE HYPER PARAMETERS USING GRIDSEARCHCV:**

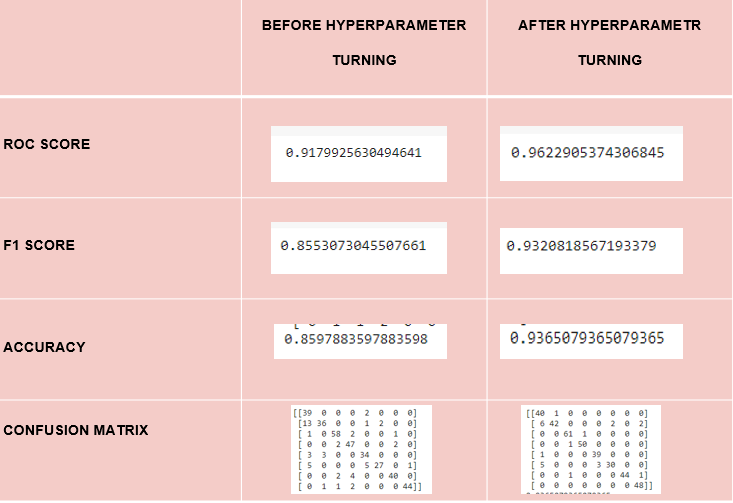
Before applying the RandomForestClassifier,Turning the above parameters to get the optimal values of model.We do this by defining a dictionary in which we mention a practical hyperparameter along with values.GridsearchCV tries all combinations of values passed in the dictionary and evaluates the model for each combination.



**APPLYING THE MODEL WITH THE BEST PARAMETERS RETURNED BY GRIDSEARCH**



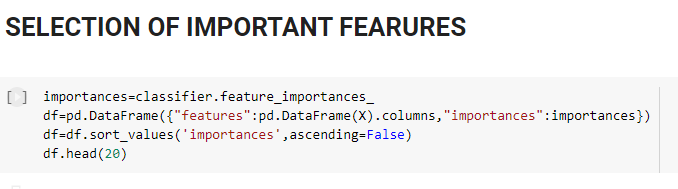
**COMPARISON CHART FOR BEFORE AND AFTER HYPERPARAMETER TUNING:**



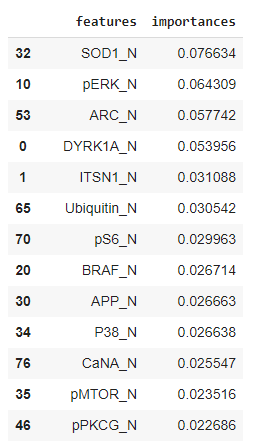
**FEATURE SELECTION :**

In RandomForest there are three inbuilt methods to find the important features.

* Builtin Feature\_importance library
* With permutation method
* Shap values



**THE TOP 20 IMPORTANT FEATURES :**



**6.2 LOGISTIC REGRESSION:**

* Supervised machine learning classification algorithm.
* **Logistic regression** is a statistical model that uses **Logistic** function to model the conditional probability
* Used for predicting the categorical dependent variable using a given set of independent variables
* **Multinomial**

In such a kind of classification, dependent variables can have 3 or more possible *unordered* types or the types having no quantitative significance. For example, these variables may represent “Type A” or “Type B” or “Type C”.

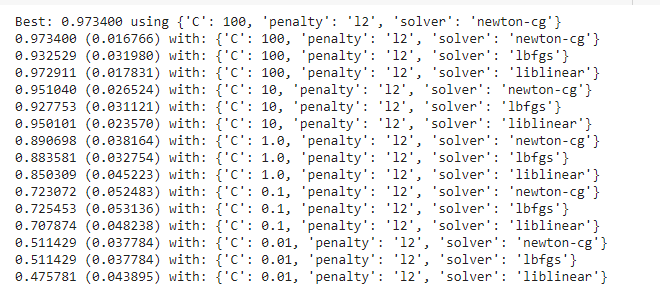
**HYPERPARAMETERS OF LOGISTIC REGRESSION:**

* solver = Algorithm to use in the optimization problem.
* penalty = Used to specify the norm used in the penalization.
* C = Inverse of regularization strength; must be a positive float. Like in support vector machines, smaller values specify stronger regularization.
* max\_iter = Maximum number of iterations taken for the solvers to converge.
* multi\_class = If the option chosen is ‘ovr’, then a binary problem is fit for each label. For ‘multinomial’ the loss minimised is the multinomial loss fit across the entire probability distribution, *even when the data is binary*.

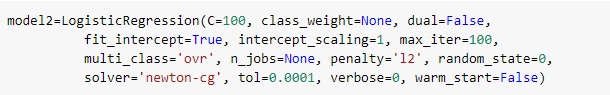
**TUNING THE HYPER PARAMETERS USING GRIDSEARCHCV:**

Before applying the Logistic Regression,Tune the parameters to get the optimal values of model.The parameters of the estimator is applied on the model method and are optimized by cross-validated grid-search over a parameter grid..GridsearchCV tries all combinations of values and evaluates the model for each combination.We select the parameters with highest accuracy.



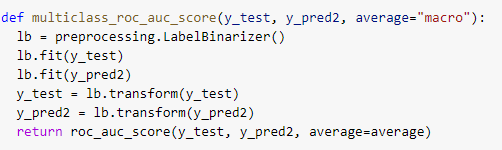


**APPLYING THE MODEL WITH THE BEST PARAMETERS RETURNED BY GRIDSEARCHCV:**

****

**ROC ACCURACY SCORE:**

Compute Area Under the **Receiver Operating Characteristic** Curve (ROC AUC) from prediction scores.



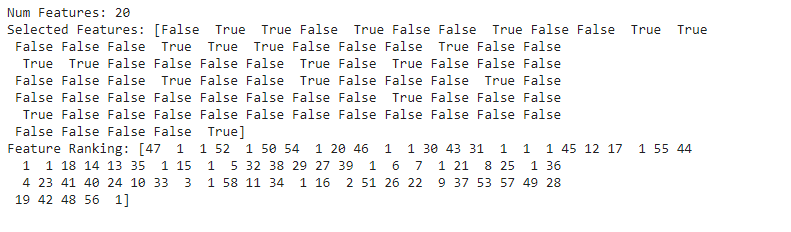
****

**COMPARISON CHART FOR BEFORE AND AFTER HYPERPARAMETER TUNING:**

****

**FEATURE SELECTION (USING RECURSIVE FEATURE ELIMINATION):**

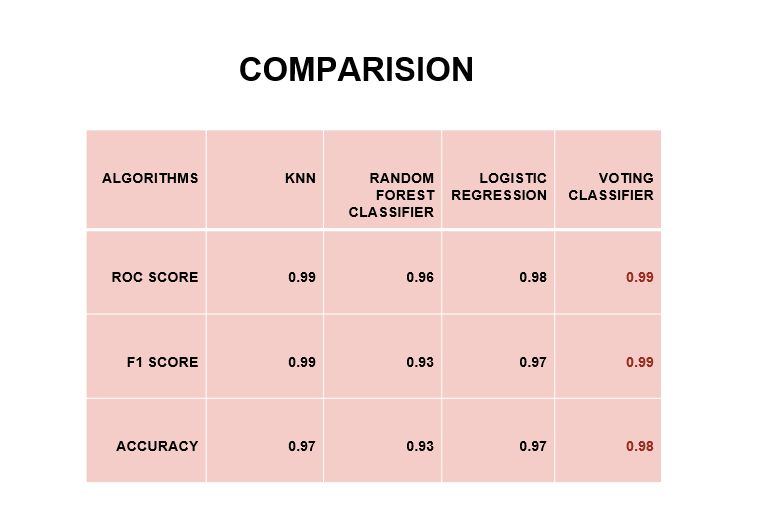
* The goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features.
* First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through any specific attribute or callable.
* Then, the least important features are pruned from the current set of features.
* That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

****

**The features that are selected(best) are:**

ITSN1\_N,'BDNF\_N,'NR2A\_N,pCAMKII\_N,pERK\_N,'pJNK\_N,pNR2A\_N,pNR2B\_N,pPKCAB\_N,CAMKII\_N,ERK\_N,GSK3B\_N,APP\_N,SOD1\_N,pNUMB\_N,pP70S6\_N,Tau\_N,IL1B\_N,CaNA\_N,CDK5\_N

**COMPARISON OF ALGORITHMS:**



**7.CONCLUSION:**

* The top proteins critical to learning ability are as follows:
  + SOD1\_N
  + CaNA\_N
  + Ubiquitin\_N
  + ARC\_N
  + pS6\_N
  + P38\_N
  + S6\_N
  + pPKCAB\_N
  + pGSK3B\_N
  + pERK\_N
* Given the protein expression values of the specified 77 proteins, the voting classifier model predicts the type of mouse whether its genotype is control or trisomic, treatment given is memantine or saline and behavior is context shock or shock context with an accuracy of 99.28%.

**8.REFERENCES:**

<https://www.kaggle.com/ruslankl/mice-protein-expression>

<https://drive.google.com/file/d/1V3FDYNBV9z3-41YcbnYakEGrQSePHI70/view>

<https://www.researchgate.net/publication/329601164_Comparative_Analysis_of_Mice_Protein_Expression_Clustering_and_Classification_Approach>

<https://github.com/nisarachmatika/Machine-Learning-mice-protein-expression->

<https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression>

[https://scikit-learn.org/stable/supervised\_learning.html#supervised-learning](https://scikit-learn.org/stable/supervised_learning.html)