Data set name: Mice Protein

So this is an experimental dataset which has various combinations of proteins on mouse. This is a test dataset. It is all about biology where there are all types of test run is performed on a mouse

Coming to the data set:

* It is a **multivariate data analysis**
* It has 1080 instances and 82 attributes
* The characteristics which is represented in the data set is real and it is performed in the year 2015 August 04

|  |  |
| --- | --- |
| Attributes | Count |
| Number of instances | 1080 |
| Number of Features | 82 |
| Number of categorical columns | 4 |
| Number of Numerical columns | 78 |
| Number of class | 8 |

What are the 82 attributes telling us?

* 1 Mouse ID which can be removed
* 2 to 78 Values of expression levels of 77 proteins; the names of proteins are followed by â€œ\_nâ€ indicating that they were measured in the nuclear fraction. For example: DYRK1A\_n\
* 79th attribute is Genotype: control (c) or trisomy (t)
* 80th attribute is Treatment type: memantine (m) or saline (s)
* 81st attribute is Behavior: context-shock (CS) or shock-context (SC)
* 82nd attribute is 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

so class is my target variable which tells the **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

Steps

1. let's eyeball the datasets

Insights: According to the jupyter notebook we don't have any missing values as the value is "?" it is considered as a category

1. Importing data set by imparting missing value as “?” to get null values

Insights:

* As all the missing values are numeric I can use mean/median/mode. mostly it looks data which is familiar as a chance I am imparting mean to cover the inequality in the dataset
* I am just trying as i have 45 columns as missing values i cannot impart or fix the missing value one by one. So, let me do it at once

1. Missing values are fixed
2. Dropping the variables which are not important i.e., Mouseid, genotype, treatment and behavior

Reason: It seems that all the genotype, treatment and behavior effects the class variable and the combination of these 3 is the class variable so we can drop all these columns

1. Converting the variable names to the range 1-8 for variables in class
2. Performing OneHotEncoding for variable class
3. Now performing MinMaxScaling for only X variables and reshaping Y variables
4. Performing logistic regression using the test size of 20,25, 18,15 where the accuracy is 88.4%, 87.4%, 89.2%, 89.59% so to get the accuracy to 90% I have increased the test variables more than train
5. Performing Neural networks

Insights:

* Hidden layers’ sizes let it be (100,200), (200,100), (30,60) with the number of iterations being constant to 1000 the accuracy is 100%

1. So I performed SVM on the data set

Insights:

Performed polynomial and sigmoid functions for test size of 20. I got the accuracy of 11.11% and 26.85% which I couldn’t understand why?