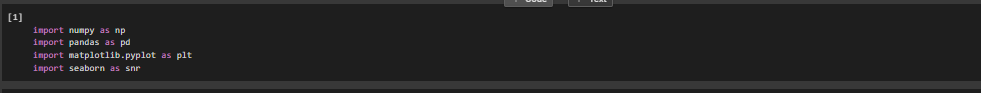
Project on liver cirrhosis prediction using Machine Learning

* **Aim:-**To create a Data science Project, where we will be predicting the the stage of liver Cirrhosis using 18 clinical features. Cirrhosis damages the liver from a variety of causes leading to scarring and liver failure

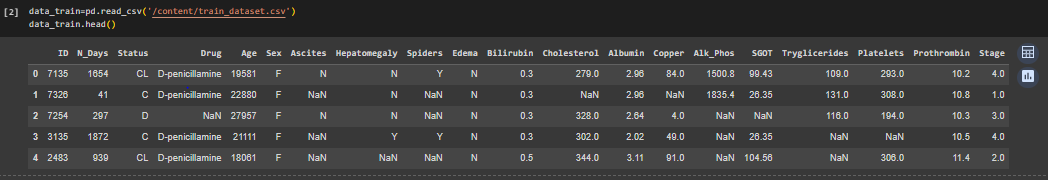
Prediction on liver cirrhosis with help of :-

Train Dataset - It consists of a total of 6801 data points. • Test Dataset - You must predict the stage of cirrhosis of 3201 data points.

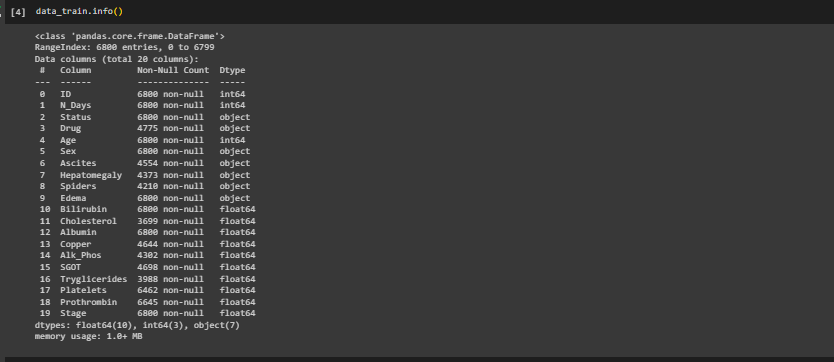
* Steps to be taken in the project is sub-divided into the following sections. These are:
* Importing the libraries such as ‘numpy’, ‘pandas’,‘sklearn. model’ etc.
* Loading Dataset as a CSV file for training & testing the models.
* Splitting the data set into independent & dependent sets.
* Checking if still any null values or any other data types other than float and integers are present into the dataset or not.
* Importing the train\_test\_split model from sklearn.model for splitting data into train & test sets.
* Applying the different kinds of ML Algorithms .which gives Best accuracy of model.
* Also checking with new data set for predicting the values.
* Steps of creating ML model:-
* Importing numpy as np & pandas as pd for loading and reading the data-set & using matplotlib.pyplot and Seaborn for visualization of data.



* Loading the csv-dataset in the variable name ‘data\_train’ Then viewing the data with data\_train.head()



* Checking the data such as number of columns, rows and type of data(float,integer) with help of data\_train.info()

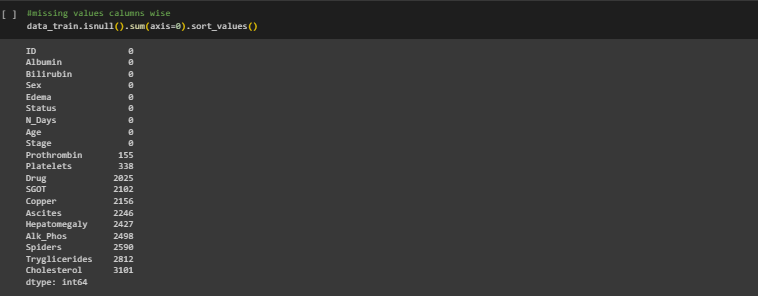


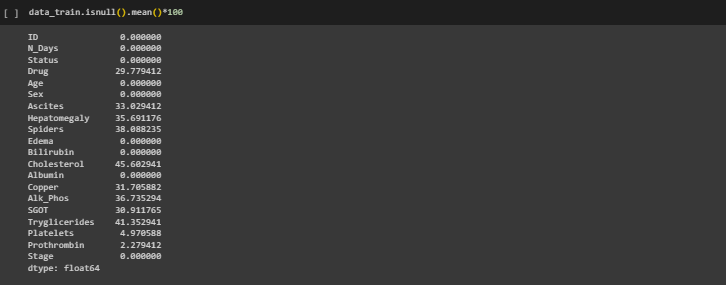
We observe that the above data have integer, object and float.



Train data have 6800 Rows and 20 columns

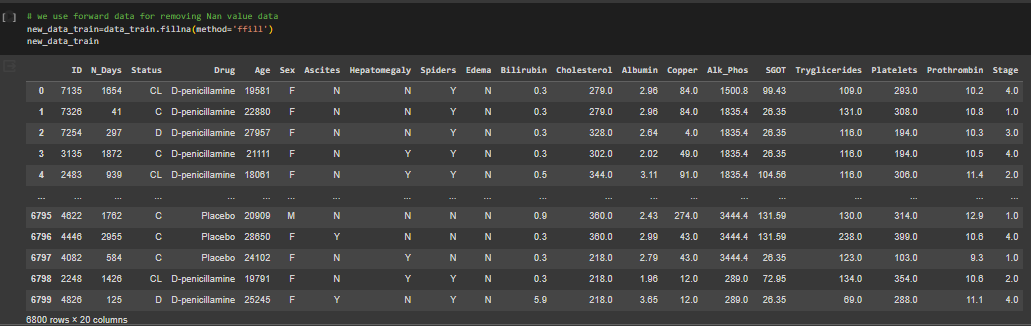
* Now checking data have Nan value or not.

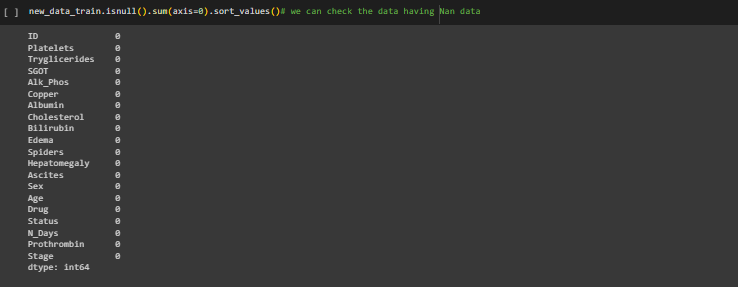




We observe that the above data have Nan value. And we see that maximum 41.5% Nan value in Triglycerides .we cleaning the Nan value before working on it.

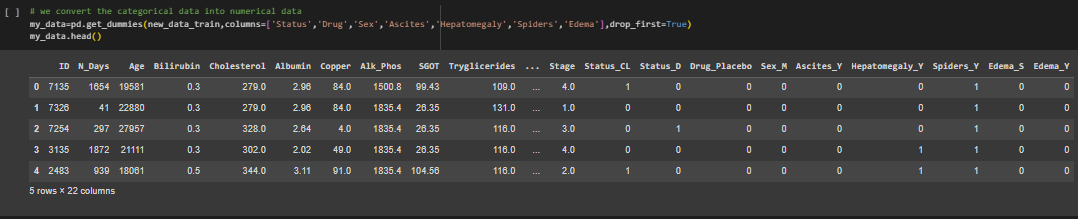
* We remove the Nan value with help of fillna(method=’ffill)

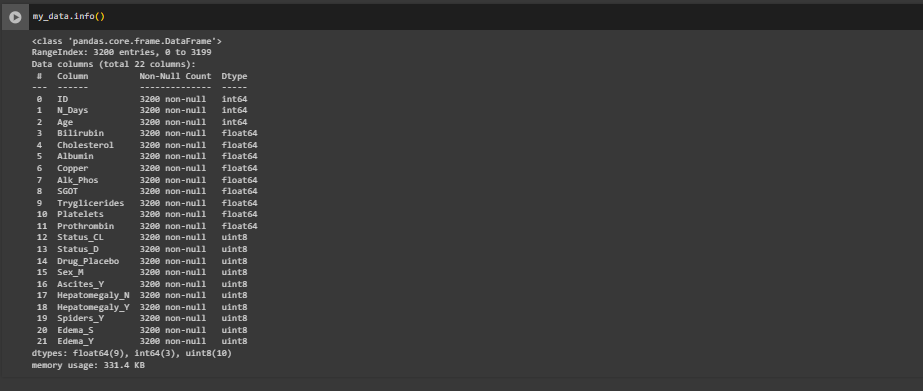




We observe that the above data have fully remove Nan value.

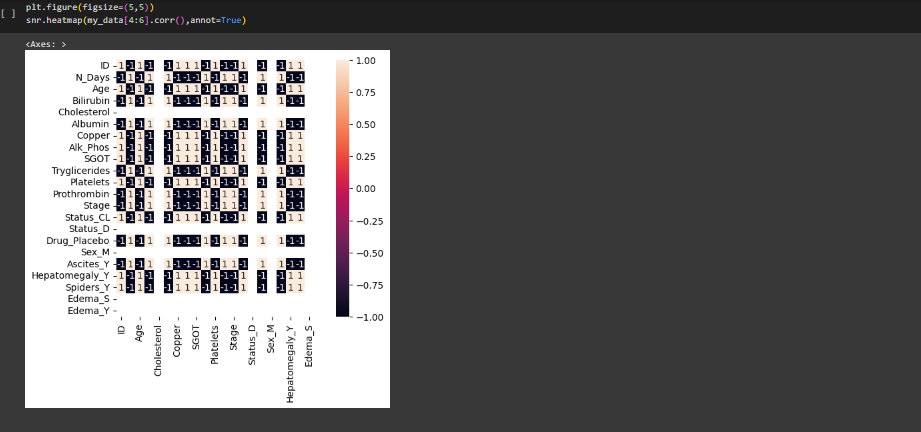
* Now ,Main focus convert the categorical data into Numerical data with help of one hot encoding method.





**Finally we observe the data are fully cleaned.**

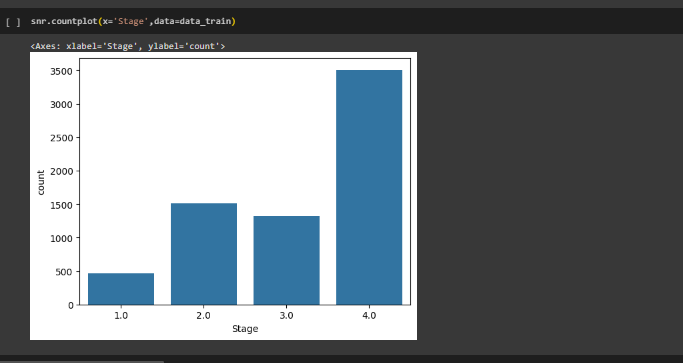
* Now we check the data dependency.



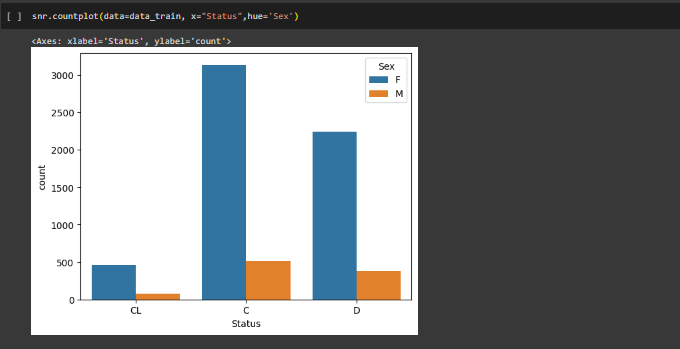
We see that data dependent each other.

* Visualizing the liver cirrhosis with various Stage like Status, Drug, Sex, Hepatomegaly, Cholesterol

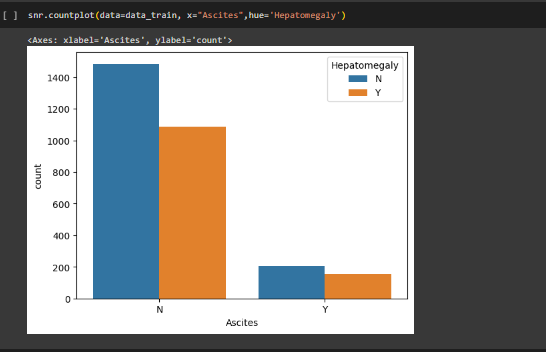
Platelets, Stage etc.



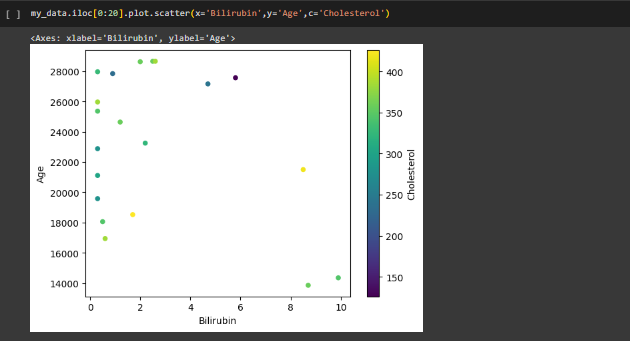
# As per Visualizing the above graph, liver cirrhosis in stage 4 is more than the other stage.



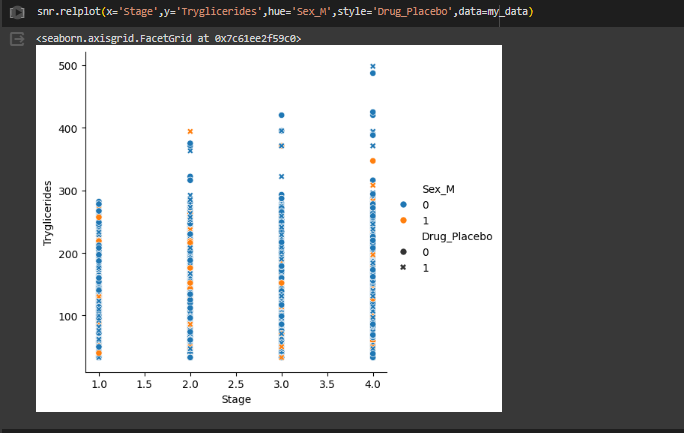
# As per Visualizing the above graph, censored due to liver transplant in female is more.



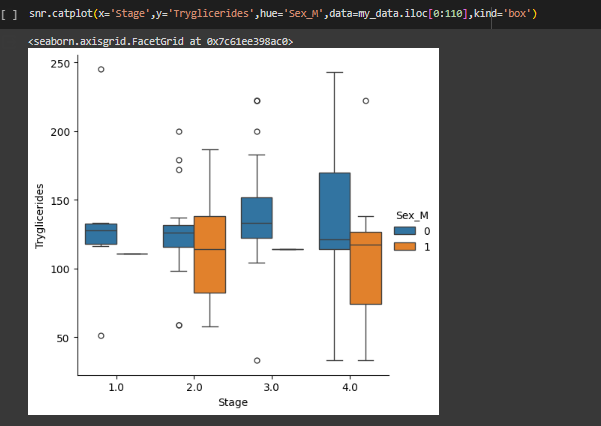
# As per Visualizing the above graph, People have no presence of Hepatomegaly as well as Ascites.

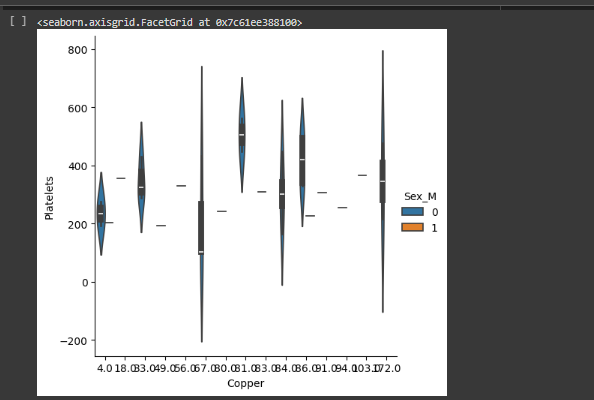


We observed that cholesterol having more with less of Bilirubin

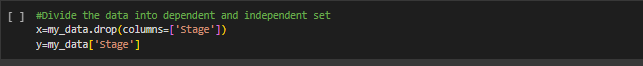


We observe the above graph,stage 4 have more Tryglicerides in female.

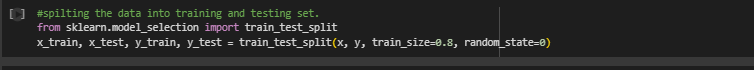




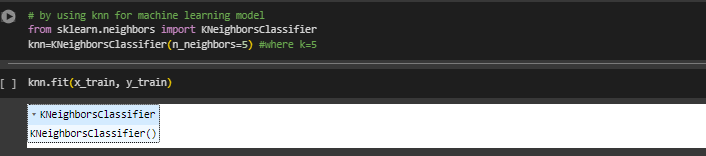
* Splitting the dataset into dependent & independent sets



* Importing train\_test\_split from sklearn.model library for splitting the data into train and test sets. (we consider train dataset).

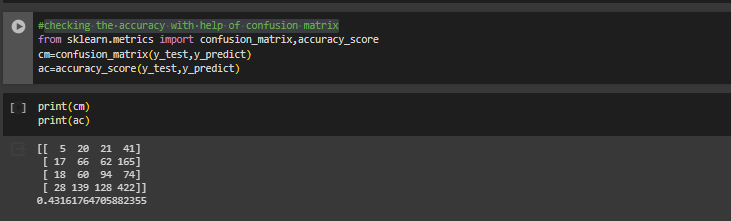


* Importing KNN from sklearn Libaray & then activating the Machine learning Model .Then used knn.fit() to training the model by providing train & test sets as x & y. And then predicted the trained model with help of MLM & the checked score as knn.score(x,y)



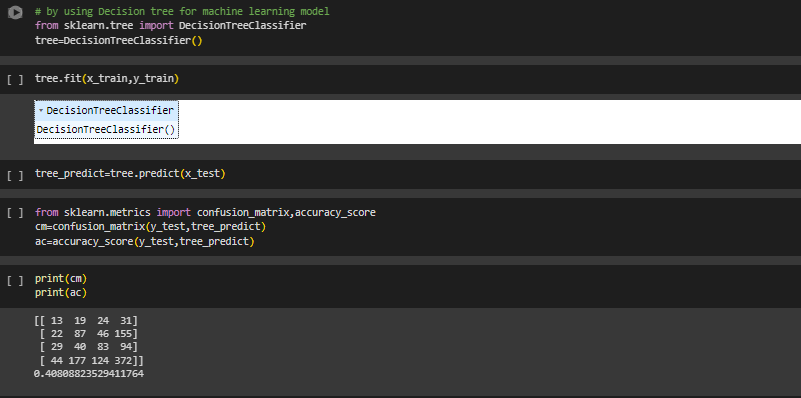


* Checking the accuracy with help of confusion Matrix.



-In the above model we can see that the accuracy obtained is 43% which is not good , we can try using different models to see if we can get better accuracy than this or not.

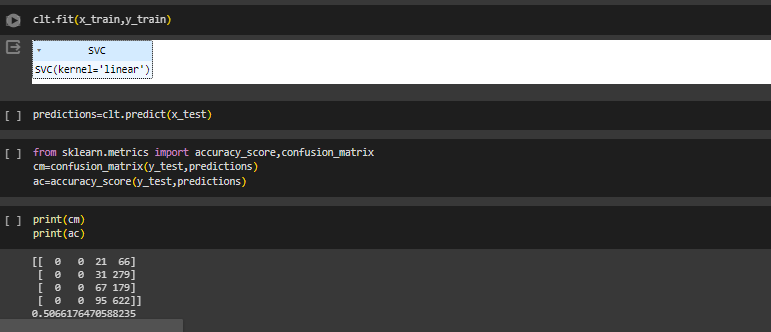
* Now applying new algorithm DecisionTreeclassifier ,then checked score.



In the above model we can see that the accuracy obtained is 40% ,is less than **KNN**.But we can try using **SVM** to see if we can get better accuracy than this or not.

* Applying Support Vector Machines



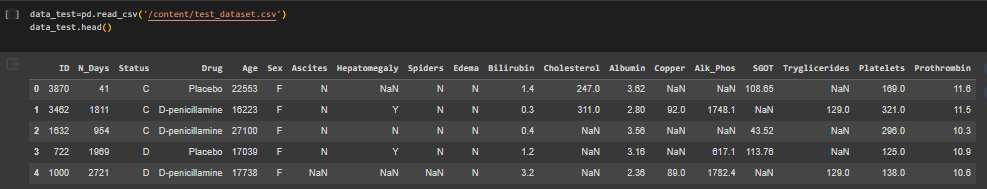


In the above model we can see that the accuracy obtained 50%

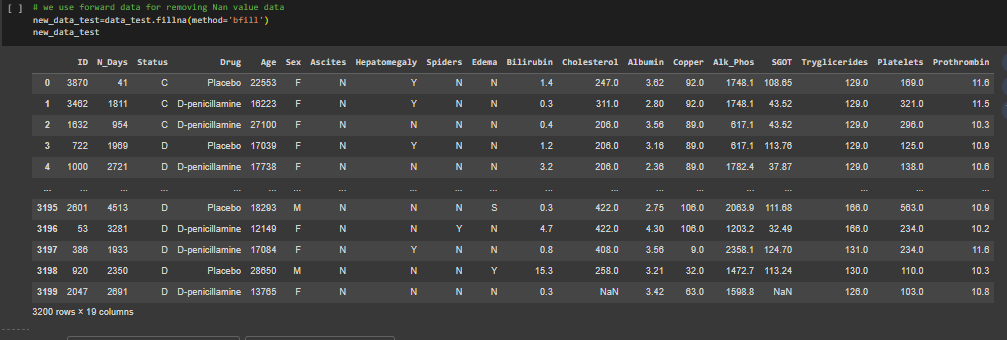
|  |  |
| --- | --- |
| Algorithms | accuracy |
| Support vector machine | 50% |
| KNN | 40% |
| Decision Tree classifier | 43% |

Conclusion-Support vector machine algorithms is better than KNN and Decision Tree.

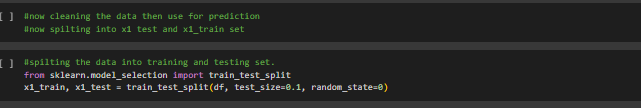
* Now recalling the test data set.
* Loading the csv-dataset in the variable name ‘data\_test’ Then viewing the data with data\_test.head()



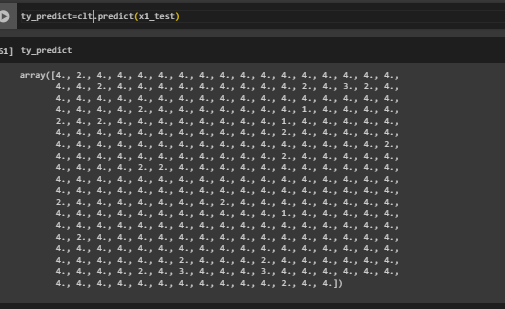
* Cleaning the test dataset.



* Splitting into test & train sets as x1\_test & x1\_train. Then we find the liver cirrhosis prediction using Machine Learning(SVM)



Applying SVM algorithms for predictions.



**Conclusion:- I**n this test data set we analysed the data we found the max. liver disease in stage 4.

Thank you