Module-2 EDA Plan execution

Name: Kartik Sojitra

Table of Contents

[Problem Statement & Analytics Rationale Statement 2](#_Toc32580262)

[EDA Steps 1 & 2 2](#_Toc32580263)

Step 2-Step7 ……………………………………………………………………………………………………………………………………………3

Assumptions and Constraints…………………………………………………………………………………………………………………..4

EDA execution Step 1………………………………………………………………………………………………………………………….5&6

Step 2……………………………………………………………………………………………………………………………………………….. 7&8

Step 3………………………………………………………………………………………………………………………………………………. 9&10

Step 4……………………………………………………………………………………………………………………………………………………10

Step 5 & 6 …………………………………………………………………………………………………………………………………………… 11

Step 7………………………………………………………………………………………………………………………………………………….. 12

EDA Insights………………………………………………………………………………………………………………………………………… 13

Metric selection and Targets………………………………………………………………………………………………………… 14& 15

Problem Statement

“To identify and determine Children with Congenital heart defects, patients with Heart valve defects across the age group using machine learning algorithm”

* Heart valve defects can be seen at the time of birth with newborn babies and it is seen in adults as well. With this project, the algorithm will find out a **predictive model** to diagnose heart defects in other patients.
* The multi-Class output variable will be evaluated using classification report metrics to find out accurate responding model. The model will look for output variable classified as 0- Normal patients, 1- Congenital heart defect, 2- Heart Valve defect

Analytics Rationale Statement

congenital heart defects have been left undiagnosed which have results existence of heart murmur in 40-45% of Children.

* Heart valve defects in adults are also undiagnosed which results heart murmur in 10% of Adults.
* Early diagnosis of heart defects helps to reduce 90% of heart episodes. So, by analysing the data, the project can identify a pattern using multi-class model, to predict heart valve defects for future patients. A portable machine predicting heart valve defect can be developed applying the established algorithm with better accuracy.

EDA Action Plan

Step-1- Exploring dataset

* The very first step will be to explore the dataset, identify the predictive variables along with the output variable, finding the data types used both for independent and dependent variables.
* Followed by, we can start looking at the statistical values, like ‘Mean’, ‘Median’ which can give a rough idea about the distribution of the variables.

Step-2-Identifiy missing values

* In this step, we will identify any null values present in the dataset, if yes, the number of null values which may cause more variance in the dataset.
* Further it will be removed but also will make sure we are missing any information, will check the learning curve whether we get any difference in the variance.

Step-3 -Dealing outliers

* Using ideal methods like Tukey, the outliers and its #s can be identified.
* Same as null values, we must make sure we are not missing any information or biasing the dataset, before removing the outliers.

Step-4 Splitting the data as Train and test

* Every dataset needs to be trained before validated and hence the dataset will be split using standard split technique of 80% train and 20% Test.
* After this, using standard scaler, the dataset will be scaled.

Step-5- balancing Dataset

* Exploring the dataset also helps to find the size of the dataset and for the outcome variables. If it is not balanced, ideal way is to follow certain feature engineering methods to balance the dataset.

Step 6 Feature selection

* Once the dataset has been balanced, using feature selection techniques, can identify intercorrelation among the independent variables.
* Also, the independent variables which are contributing less to the output can be identified.
* Identifying it, we can later decide whether to keep it or not.

Step 6 Running Visuals

* Running ideals visuals like Histogram, Boxplot, heat maps or q-q plots, we can identify the trend of distribution which can give more clarity about the dataset.
* Also, running learning curve will help to see how the training set is getting validated with the test set with better Variance-Bias balance. (Removed)

Step 7- Optimization and Preparing for ideal model (Removed)

* The next step would be optimizing the dataset using ideal Cross validation method, tune hyperparameters and minimize any error to validate the training set.
* Followed by, we can run an ideal model which can help to predict the output variable better.

Assumptions- Validation:

1. With the given dataset, I am assuming that the dataset will not have any null values or any outliers which can reduce the accuracy and helps to manage time in a better way- **The assumption is not right as we a lot of N/A and outliers**.
2. The dataset is well normally distributed- **After balancing the dataset, it is right.**
3. There is no inter correlation among the independent variables- **No, 7 independent variables identified who are correlated.**
4. The dataset given are reading of patients across age group- **Still assuming the same**.
5. The readings are diagnosed using same technique which may help to nullify any variance in the model- **May not be right with the distribution in the q-q plot.**
6. The dataset has patients with all gender, and it is of multi race- **As we couldn’t identify, still assuming the same.**

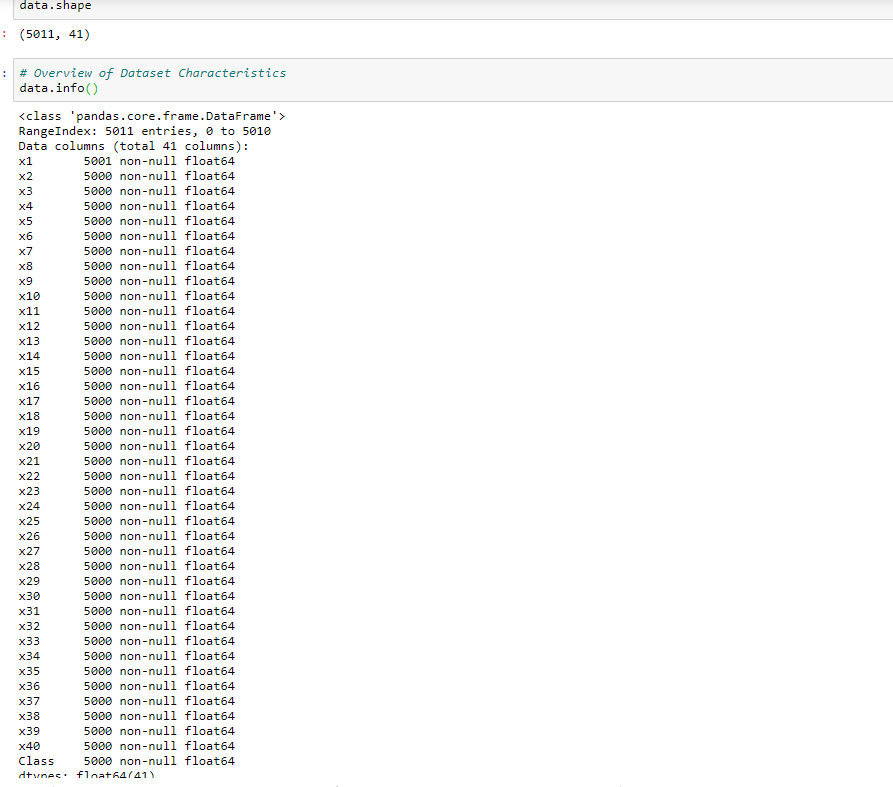
Hard Constraints:

1. Age group is not mentioned which may cause an issue when we wanted to differentiate prediction of Class 1 and Class 2- **Yes, it is an hard constraint.**
2. Though we know the source of the data, we cannot go back to them, to ask for more information if needed and restricted with whatever data available. We are not sure these predictive variables will help to have better accuracy- **Age of the patients, highly needed to differentiate class 1 and 2.**
3. As we developing a predictive model identifying the heart defect, the accuracy must be more than 90% at least. If the patients involved have been used different diagnosing techniques measuring the readings, it may cause imbalanced variance- Bias. This further will restrict to have better accuracy. – **Couldn’t identify n the analysis**
4. Model selection may cause some restriction in optimizing the dataset as model like LDA or QDA do not have any hyperparameters to tune- **SVM may be the right model, then these factors won’t be needed**
5. With the available dataset, if none of the model is able to get accuracy more than 90%, then we will be limited to select one of the better among those least accurate models which in turn may not help developing a better predictive model- **Yet to see this constraint, may be in the coming analysis**

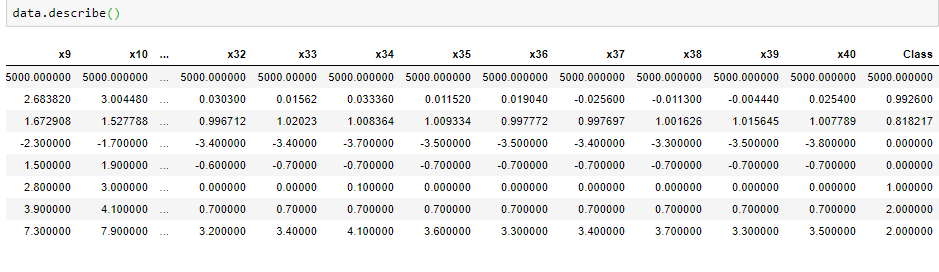
**EDA Execution**

Step 1

In the first very step, the shape of the dataset with # of rows, 5011 and # of columns, 41 can be noticed. Followed by, the datatype of each variable can be identified, as seen everything are float

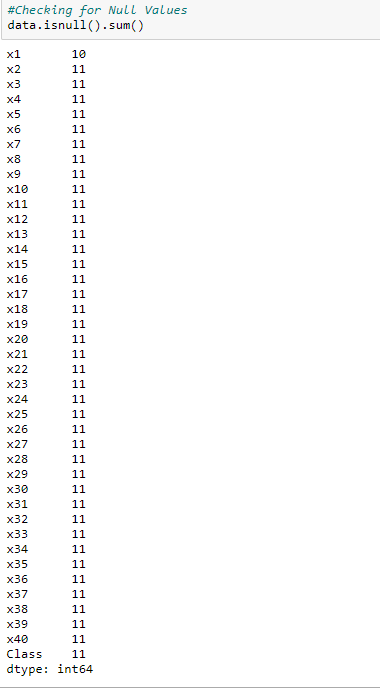


In the next very step, the dataset can be explored to see how the distribution of each column is. This may give a rough idea but not accurate as dependant variable are between 0-2, the mean median may not give an idea about the distribution.

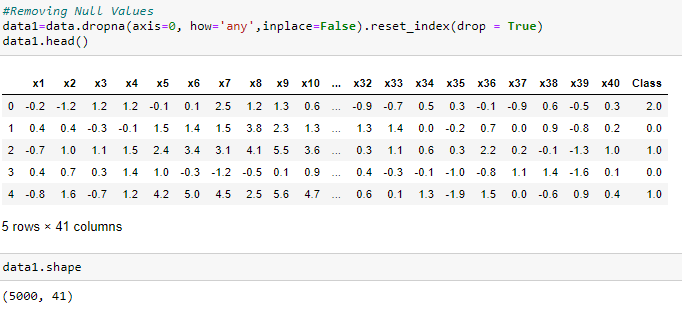


Step-2 – Dealing N/A:

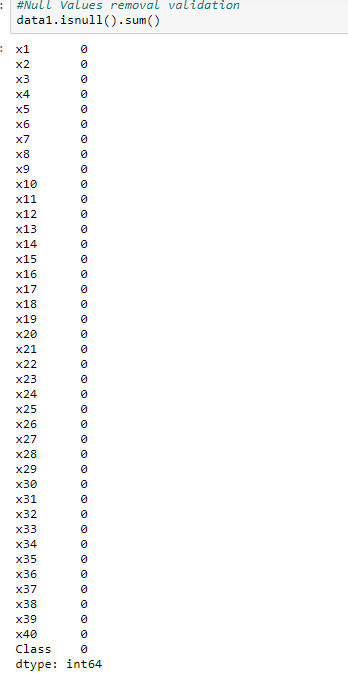
In this step, running appropriate code, the n/A can be identified and removed.



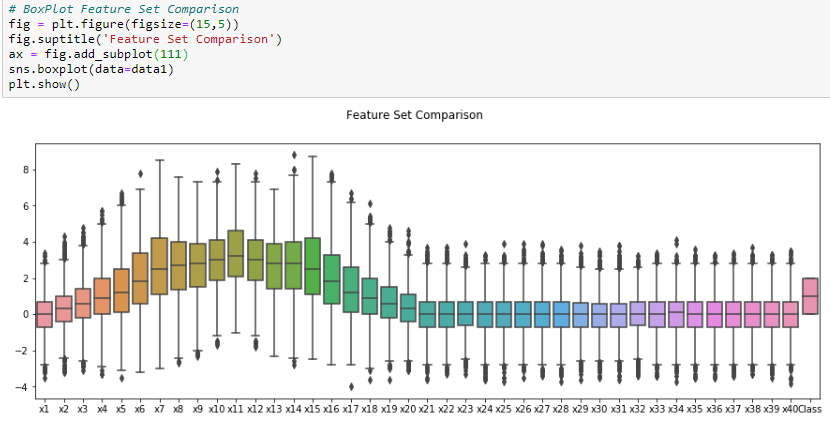
Removing N/As



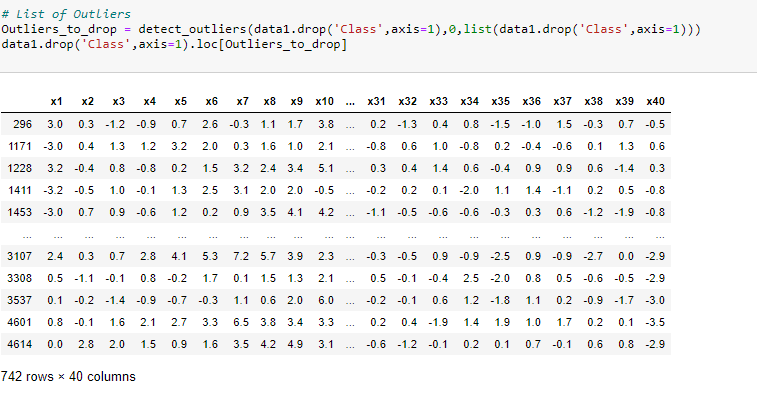
Validating removal of N/As



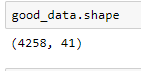
**Step 3 Dealing with outliers**



* A box plot can help to identify the outliers available as above.

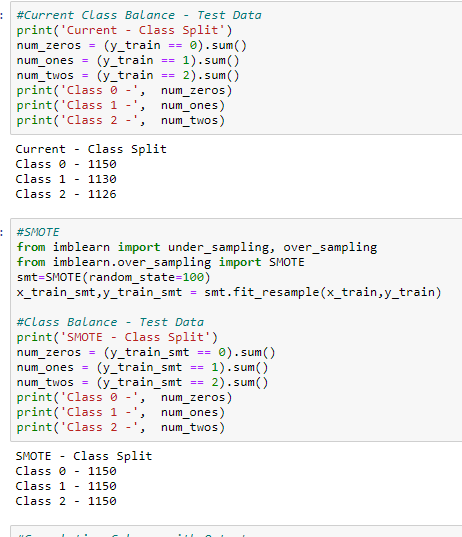


* As noticed above, totally, 742 rows of outliers have been identified, which will be removed, considering it may reduce the accuracy of the dataset.
* Removing the outliers, the dataset will be having 4258 rows as below.



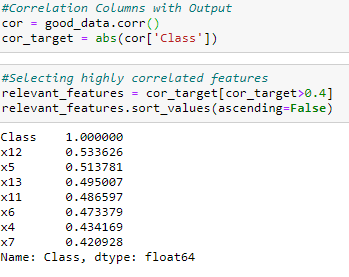
Step 4 & 5-Splitting training set and Balancing the dataset

* Further, splitting the dataset as train and test, in 80/20 ratio, we can check how good the dataset is balanced as below.
* Since there is an imbalance with couple of classes, we run Smote method, to synthetically create data points to balance all 3-output variable.



Step-6 Feature selection

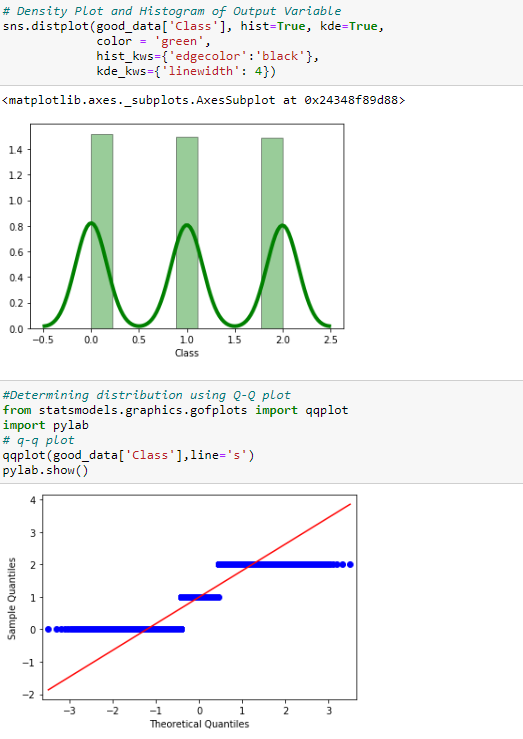
* Next step would be, to check if there variables are inter-correlated. As below. There are 7 variables, out of 40 are inter-correlated getting target 1.



Step 7- Running Visuals

To check, the distribution, a histogram and q-q plot can get an idea.

* The histogram states, the output variables are normally distributed without any skewness.



**Key Insights:**

Data Cleaning:

As noticed, there are a 8 rows of N/A and 742 rows of outliers, which for sure will bring down the accuracy of the dataset. As assumed removing the this will help to improve the accuracy

Correlation:

As discussed in the execution, we can identify a correlation for 7 independent variables in contributing the dependant variable ‘1’, means they are least contributing to the output. Rest 33 variables are independently contributing towards the output. So, the dataset is considered not highly correlated which matches almost the assumption that the dataset is not correlated.

Distribution:

As a lot of outliers can be seen in the boxplot, removing the outliers, the distribution looks normal as discussed in the assumption.

As per the q-q plot, not all the datapoints for all 3 dependent variables are falling on the line. This may be due to different methods, they used to measure the heart valve size for different individuals.

Also, the dataset was balanced using Smote, Synthetic data of human valves may have resulted this.

As described in the constraints, the age of the patients is not provided which if we have, can help to correlate it with getting accurate dependant variable.

We may require some more data, which we cannot get will also be a constraint in dealing with this dataset.

**Analytical score boards and targets**:

* As discussed above, SVM will be the right model in running this classification dataset.
* With the help of confusion matrix and classification report, three metrics can be figured out which can help to get the output correctly.

Precision:

Precision will be one of the important, as how many times the dependent variable, are predicted correctly. As it is medical data, 95% target can be considered as an ideal one for this metric.

Recall:

This metric, defines, how many times every Dependent variable was true positive, when predicted overall. With the measures of the heart valves, the dependent variable can be identified. Again, since we don’t have age, the accuracy in recall for dependent variable one and 2 will be tough. Still targeting 95 % of recall.

F1 Score:

It will be an average of both Precision and Recall, which is again targeted, 95% accuracy.

As the dataset is medical dataset, the accuracy measure must be very high. Using these metrics and the targets, the perfect predictive model can be developed predicting heart murmur.