Module-2 EDA PLAN

Name: Kartik Sojitra

Table of Contents

[Problem Statement 3](#_Toc32616368)

[Analytics Rationale Statement 3](#_Toc32616369)

[Output variable class structure with Naming convention 3](#_Toc32616370)

[EDA Action Plan 3](#_Toc32616371)

[Assumptions: 5](#_Toc32616372)

[Hard Constraints: 5](#_Toc32616373)

# 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.

# Output variable class structure with Naming convention

* The output variable will be multi-class, classified as ‘0’, ‘1’, ‘2’
* Where ‘0’ identifies normal patients, ‘1’ will classify kids with Congenital heart defect and ‘2’ will identify Heart valve defect with patients across age group.

This multi-class output variable will be predicted using several predictive variables.

# 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-5 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-4- 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 5 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.

Step 7- Optimization and Preparing for ideal model

* 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:

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.
2. The dataset is well normally distributed.
3. There is no inter correlation among the independent variables.
4. The dataset given are reading of patients across age group.
5. The readings are diagnosed using same technique which may help to nullify any variance in the model.
6. The dataset has patients with all gender, and it is of multi race.

# Hard Constraints:

1. Age group is mentioned which may cause an issue when we wanted to differentiate prediction of Class 1 and Class 2.
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.
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.
4. Model selection may cause some restriction in optimizing the dataset as model like LDA or QDA do not have any hyperparameters to tune.
5. With the available dataset, if none of the model can 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.