**Data Acquisition & Understanding**

**By Meryl Gabrielle Tubio (100763231)**

**December 1, 2020**

Table of Contents

[1.0 Re-stated Problem Statement and Analytical Statement 3](#_Toc35898687)

[2.0 Identify and Justify Output Variable Class Structure 3](#_Toc35898688)

[3.0 Action Plan for Exploratory Data Analysis (EDA) 4](#_Toc35898689)

[4.0 Analytical Scorecard 9](#_Toc35898690)

# 1.0 Re-stated Problem Statement and Analytical Statement

Princess Margret Hospital (in partnership with the UHN Echocardiography Lab) aims to save lives through early detection of heart murmur. With the heart murmur dataset provided by the Hospital, Researchers aim to detect early signs of heart murmur. In the hopes that this study would lead to better preparation and precaution for identified patients.

This study aims to identify the valve measurements that indicates if the patients has heart defects or not. For the purpose of this study precision and recall would be used to measure results.

The objective of the problem statement is to identify a pre-diagnosis of potential heart issues in potential patients. Improved early detection would increase the awareness and prevention, and ultimately a reduced mortality rate for cardio-pulmonary patients.

# 2.0 Identify and Justify Output Variable Class Structure

Indicated in the Problem Statement, the focus of this study is to identify unique measurements of congenital defect and heart valve defects (Class 1), the data to be used is a Binomial classification structure. The naming convention for this study is as follows:

Class 0 – No Heart Ailment

Class 1 – Heart Ailment (Congenital Defect or Heart Valve Defect)

# 3.0 Action Plan for Exploratory Data Analysis (EDA)

* 1. Action Plan

The following are the detailed descriptions of the Action Plan steps:

**Step 1: Identify the data**

This can be done by executing simple statistics and data information in python.

Identify the dependent and independent variables.

**RESULTS:**

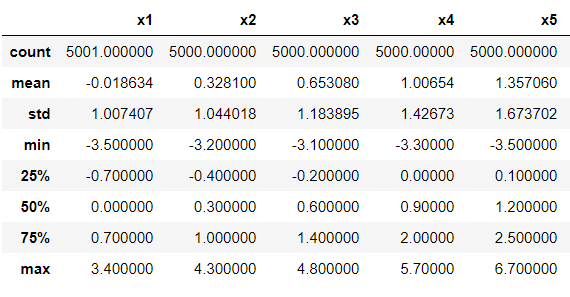


Figure 1

From figure 1, the dataset is imbalanced and has inconsistent fluctuations in standard deviations. Identifying the dependent and independent variables. – X1 to X40 are independent variables, although the column “classes” is the only dependent variable in the dataset.

**Step 2: Run Pandas profiler**

Afterwards run the dataset through pandas profiler (pandas profiler is a separate library that needs to be downloaded for this code to run). Pandas profiler allows you to identify information outside of simple statistics and data type information. This includes: Warnings, Attributes, visualization of items (i.e. column headers, heatmap of correlation) etc. Pandas profiler also helps identify missing values.

**RESULTS:**pandas profiling report, primarily for its quick insight into the dataset. The following are the results that need immediate action:

1. Missing Cells – 450 (0.2%)
2. Duplicate Rows – 9 (0.2%) Duplicate rows are not dropped in this instance because each row signifies a patient in the dataset. Having similar measurements with another patient should remain in the data

**Step 3: Drop/Modify Irrelevant data**

Cleaning the data would be broken down to 3 sub-processes:

Step 3a: First, check for missing values and drop values (if applicable)

Step 3b: Check if dataset is balanced and perform SMOTE (if applicable)

Step 3c: Check for any outliers and remove (if applicable)

**RESULTS:**

Step 3a: Check for missing values – There were 450 missing values or NaNs identified in Python. Although 450 missing values accounts 0.2% of the dataset. The missing values were dropped. Primarily for rows that were completely NaNs. There also one patient that registered only 1 feature out of the 40, particularly the X1 measurement. Therefore, the last patient was dropped from the dataset due to lack of information.

Step 3b: Check if dataset is balanced – The data is imbalanced, therefore there is a need to balance the dataset first. The Researchers did SMOTE to balance the dataset Making the total dataset at 5292 items.

Step 3c: Check for outliers – There were 742 outliers identified through the turkey method. After removing the dataset, there were 4258 patient measurements left.

**Step 4: Visualize Data**

Visualize the data through boxplot and a correlation heatmap. This allows outliers to be highlighted.

**RESULTS:**

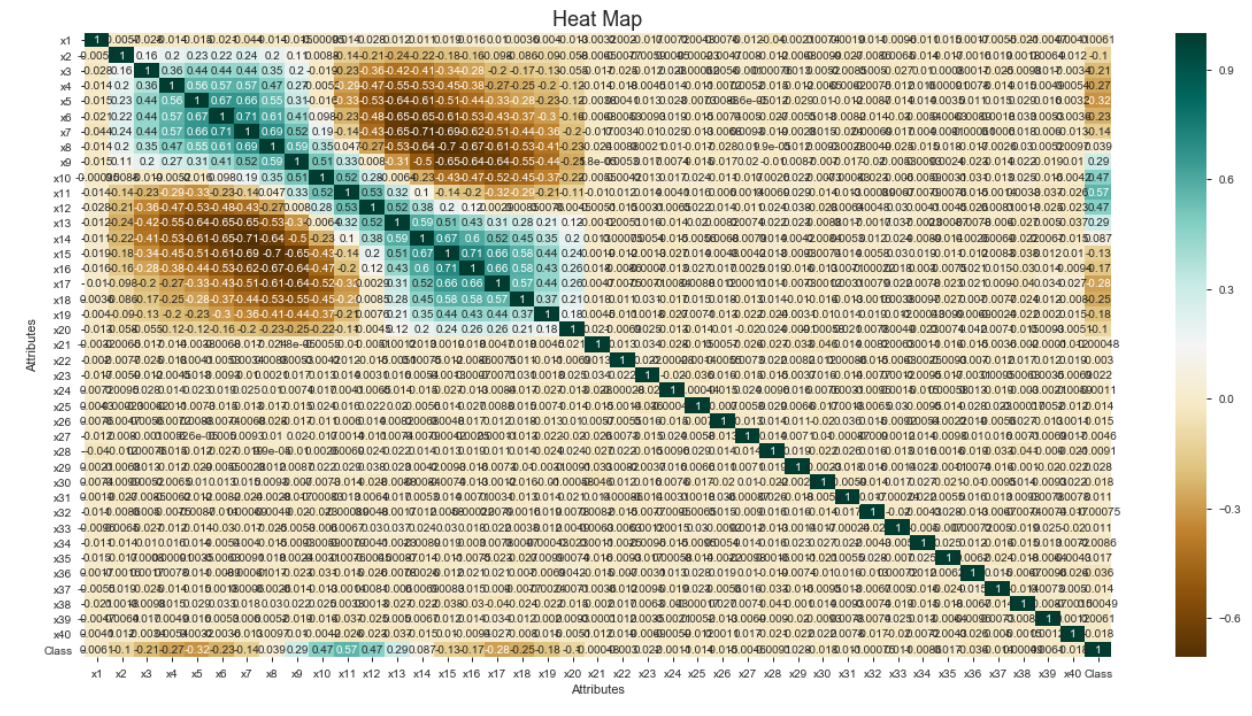


Figure 2

From figure 2, we can identify that there is an extreme range from high to low correlation in the X1 to X20 features. While all other features remain relatively low in terms of correlation.

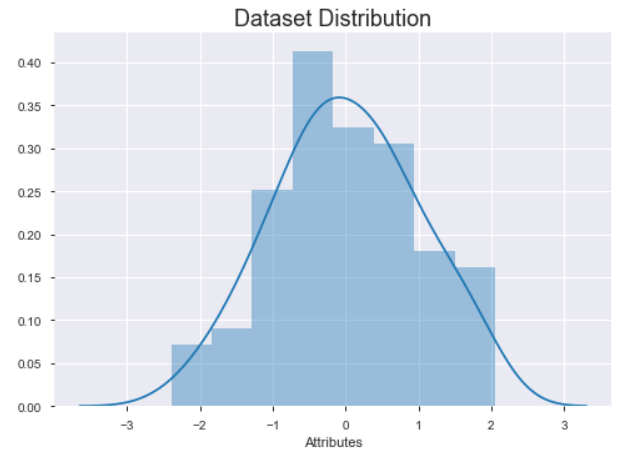


Figure 3

From figure 3, we can identify that the dataset is normally distributed. Although there seems to be a sharp spike in 0.40 and -0.5 features in the dataset.

* 1. Assumptions

There are three assumptions about the dataset:

1. That the data has been vetoed and cleaned
2. That the Murmur data is based on valid individual patient data
3. That the data is normally distributed, unless proven otherwise through the EDA (Exploratory Data Analysis).
   1. Constraints

Basically, a hard constraint has a strong negative impact to the model, a strong negative impact would mean a need to remodel the data if new attributes or information were added. While the soft constraint would have a weaker effect to the model, meaning if there is new information there is no need to remodel the dataset. There are (3) three constraints about the data and within these constraints we can identify (2) two hard constraints and (1) one soft constraint, these are:

1. The dataset cannot be increased. The data was forwarded by the client, if the client forwards additional data this would force the Researchers to perform EDA and data remodeling again. This is categorized as a hard constraint.
2. The dataset cannot have additional features. Again, this data is from the client forcing the researchers to work with the current data. Any additional features not from the client would damage the dataset. In retrospect, if the client also forwarded additional features this would force the Researchers to do EDA and data remodeling again This is also a hard constraint.
3. Limited Knowledge of the dataset. Although this constraint causes a slight inconvenience, if there was more knowledge about the data, depending on the extent of the information, this would still not require the Researchers to recreate EDA and data remodeling. This is a soft constraint.

# 4.0 Analytical Scorecard

As indicated in the problem statement, this study would use the precision and recall metric to evaluate the models. Also, the ROC (Receiver Operator Characteristic) and AUC (Area Under the Curve) learning curves would be utilized, as well.

The precision metric allows one to understand how the model is performing based on the patients correctly diagnosed among all the diagnosed patients diagnosed positive. While recall, also known as sensitivity, focuses on the probability that a patient would have heart murmur out of patients who were misdiagnosed to not have heart murmur. Precision and recall are based on an understanding and measure of [relevance](https://en.wikipedia.org/wiki/Relevance).

ROC and AUC are both good metric for a binary classification. The ROC learning curve highlights the trade-off between recall and specificity (1-False Positive Rate). The AUC learning curve acts as a general measure of predictive accuracy. In other words, the probability of a model to rank a correctly diagnosed patient higher than a patient who does not have heart murmur.

Finally, all model results at 90% and below would be disregarded. This threshold is chosen because of the medical nature of the problem.