**Use case study report: Classification of Cardiac Arrhythmia Disease**

**Group No.:** 20

**Student Names:** Aakash Atnoorkar, Athul Sony

**Data Collection and Exploration:**

The project is related to a medical problem – Diagnosis of Cardiac Arrhythmia from 12 lead ECG recordings. Each record contains clinical measurements from ECG signals such as QRS duration, RR, P-R and Q-T intervals and some other information such as sex, age, weight, along with the decision of a cardiologist. There are total 279 attributes (features) per patient in a record. Cardiologist has diagnosed the ECG into either normal or one of the other 15 classes of arrhythmia.

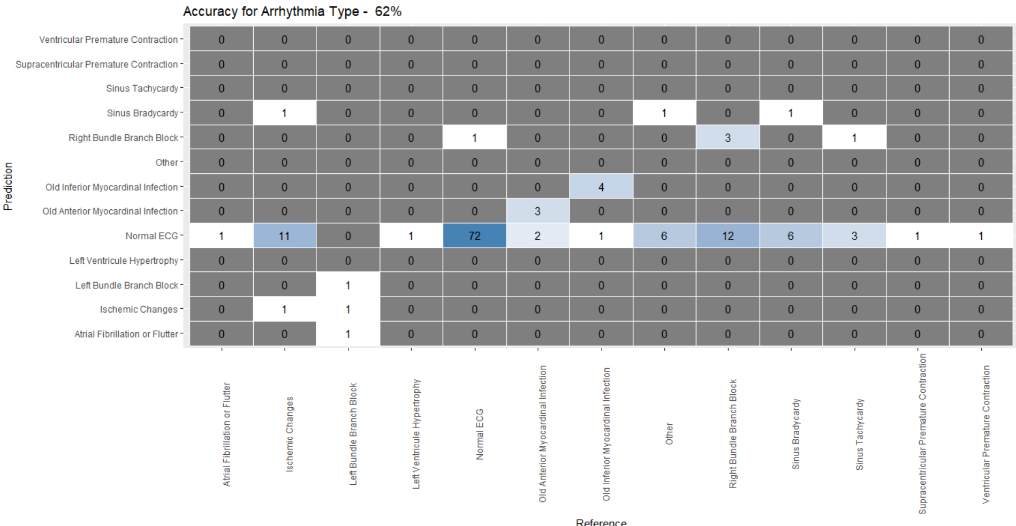
Excluding the categorical variable, we calculated the correlation matrix and we found that the variables are weakly correlated. Based on this, we have decided not to perform PCA on this dataset. The operations performed on the dataset till now include missing-data imputation, identifying and converting columns to factors and scaling the numerical columns.

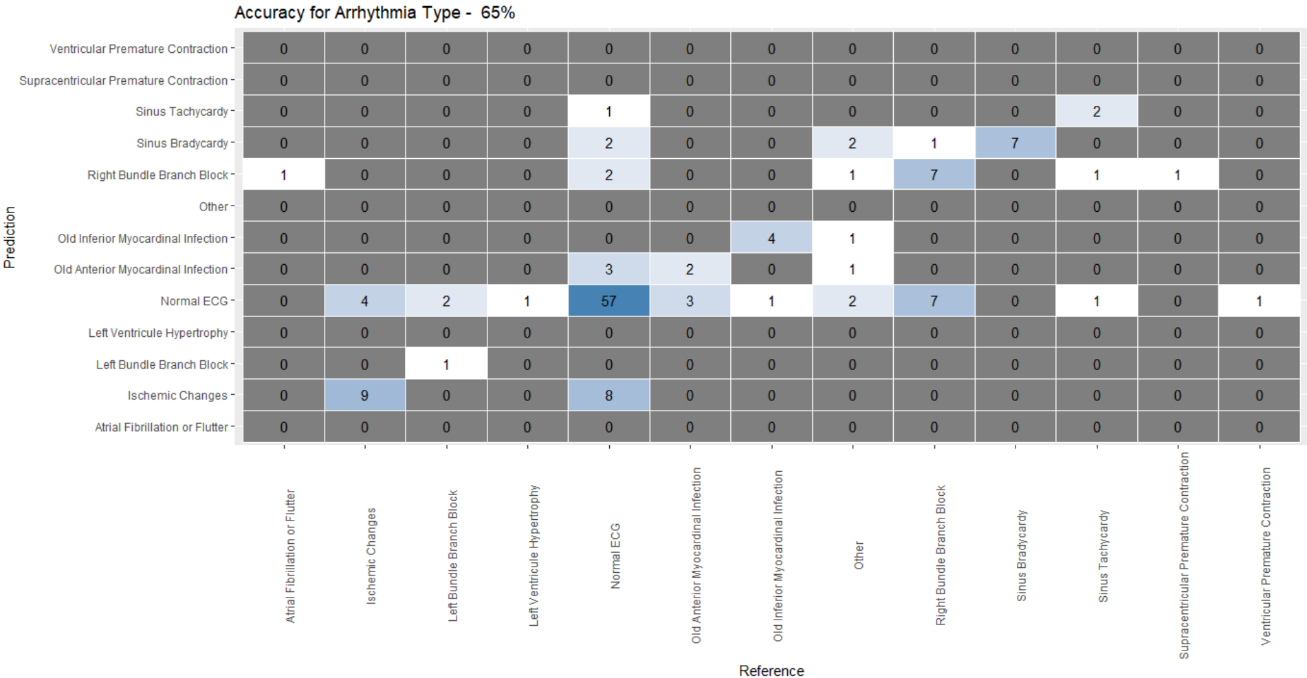
**Algorithm Selection and Implementation:**

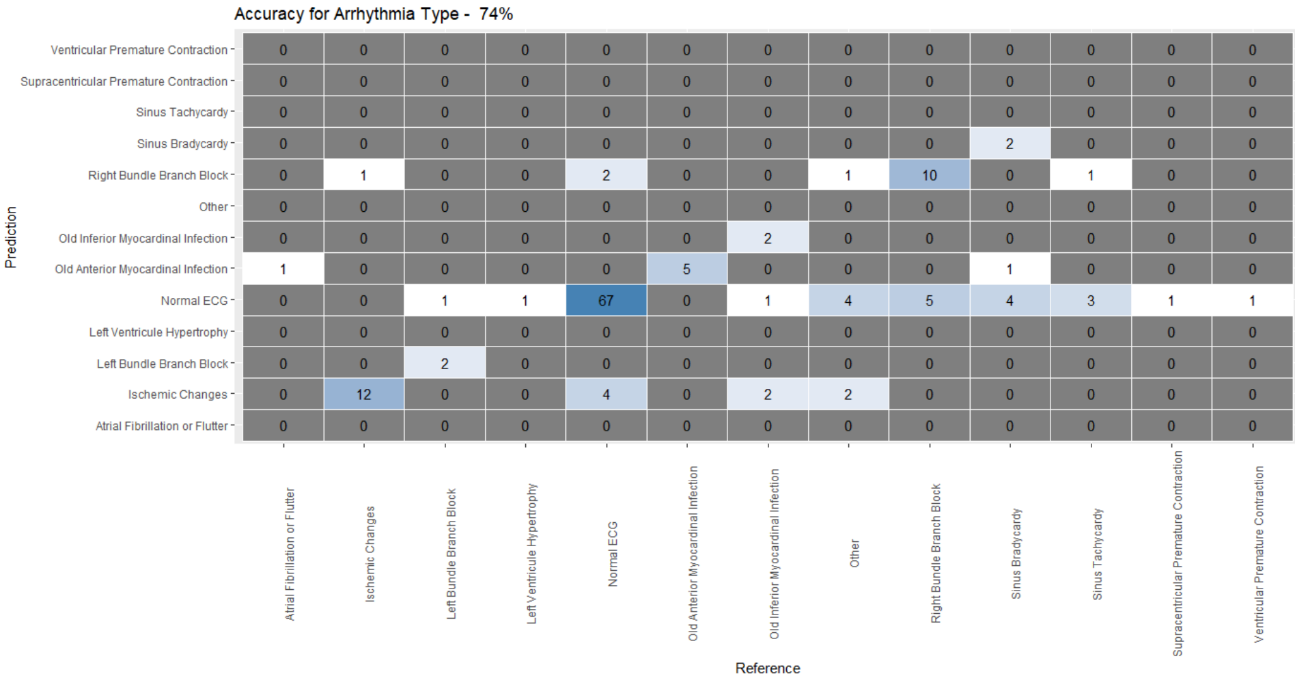
As this is a classification problem, we will be predicting the type of Arrhythmia using following algorithms:

1. K-Nearest Neighbors (KNN) Algorithm

* The scaled data is further prepared for knn function by creating *m* dummy variables for *m* categorical values.
* The dataset is split into training and test sets in the 7:3 ratio
* The accuracy is calculated on test set for k values ranging from 1 to 50. We found that the knn algorithm gives maximum of 62% accuracy for k = 5.
* **Challenge:** As there are multiple number of classes having very high variance in the number of records for each class, balancing and validating its correctness is a tough task



1. Decision Tree Algorithm
   * The data is divided into training and validation set using the split ration of 7:3
   * First the classification tree is fully grown, and we use the *cp* value of this model to prune the tree.
   * This algorithm gives the accuracy of 65% which is better than KNN algorithm.
2. Random Forest Algorithm

* Similar procedure of splitting the data is followed
* The algorithm provides the accuracy of 74%

**Solution Approach – Further Steps:**

1. We will be implementing following algorithms and work on improvement of the algorithms.
   1. Logistic Regression
   2. SVM
   3. Neural Network