ECG Prediction

# Introduction:

The MIMIC-IV-ECG module contains approximately 800,000 diagnostic electrocardiograms across nearly 160,000 unique patients. These diagnostic ECGs use 12 leads and are 10 seconds in length. They are sampled at 500 Hz. This subset contains all of the ECGs for patients who appear in the MIMIC-IV Clinical Database. When a cardiologist report is available for a given ECG, we provide the needed information to link the waveform to the report. The patients in MIMIC-IV-ECG have been matched against the MIMIC-IV Clinical Database, making it possible to link to information across the MIMIC-IV modules.

# Dataset Overview:

The dataset is related to electrocardiogram (ECG) data, which is used to monitor the electrical activity of the heart over time. Each row in the dataset represents a single ECG recording, and the columns contain various attributes and measurements associated with each recording.

Here's a breakdown of the columns in the dataset:

1. **subject\_id**: A unique identifier for each individual or patient.
2. **gender**: Gender of the individual (e.g., "F" for female).
3. **anchor\_age**: Age of the individual at a specific reference point.
4. **dod**: Date of death of the individual, if applicable.
5. **ecgtime**: Date and time of the ECG recording.
6. **report\_2**, **report\_3**: Additional reports or annotations associated with the ECG recording.
7. **bandwidth**: Bandwidth settings used for filtering the ECG signal.
8. **filtering**: Filtering settings applied to the ECG signal.
9. **rr\_interval**: RR interval measurements, representing the time between successive R waves in the ECG signal.
10. **p\_onset**, **p\_end**: Measurements related to the onset and end of the P wave in the ECG signal.
11. **qrs\_onset**, **qrs\_end**: Measurements related to the onset and end of the QRS complex in the ECG signal.
12. **t\_end**: Measurement related to the end of the T wave in the ECG signal.
13. **p\_axis**, **qrs\_axis**, **t\_axis**: Measurements related to the electrical axis of the P wave, QRS complex, and T wave, respectively.
14. **target\_variable**: The target variable indicating whether the ECG recording is classified as "Normal ECG" or "Abnormal ECG".

Overall, the dataset contains detailed information about ECG recordings, including demographic data, timestamps, signal processing settings, and various measurements derived from the ECG signal. It is structured for analysis and classification of ECG recordings based on their normality or abnormality.

# Exploratory Data Analysis (EDA):

Please refer to files:

1. report.html
2. sweetvizReport.html

# Models & Evaluation Metrics:

1. Base model – Logistic Regression:
   1. Model Accuracy - 0.7115856069592724
2. SVM
   1. Model Accuracy - 0.8124159746935548
3. KNN
   1. Model Accuracy - 0.8409648082245947
4. Random Forest
   1. Model Accuracy - 0.8733096085409253
5. Random Forest – Hyper tuned
   1. Best accuracy found: 0.872175099353461
   2. Accuracy with best parameters: 0.8741004349545275
6. XG Boost
   1. Model Accuracy - 0.8733886911822855
7. XG Boost – Hyper tuned
   1. Best accuracy found: 0.8753386353494716

## Pickle Files:

Finally, we considered to go ahead with XG Boost Model after hyper tuning to create pkl file.