CardioViglant: Cardiovascular Decompensation Forecasting

PROJECT PHASE #1

Aishwarya Chand, Prajakta Jhade, Shreya Thakur

PROJECT OVERVIEW

*One person dies every **33 seconds** in the United States from cardiovascular disease. About 695,000 people in the United States died from heart disease (according to the data collected in 2021)—that's 1 in every 5 deaths. Globally, the scale of mortality due to this disease is even more surprising with a record of 20.5 million.*

Cardiovascular diseases are one of the major healthcare concerns and leading causes of mortality globally. Early intervention plays a crucial role here and enables healthcare providers to tailor treatment plans and adjust medications accordingly. Despite advancement in cardiovascular care, predicting and preventing cardiovascular decompensation remains a significant challenge.

Our Web Application : CardioVigilant aims to transform cardiovascular healthcare, equipping healthcare providers with an unparalleled tool for precision forecasting and redefine standards in patient care, ultimately saving lives and enhancing the quality of cardiovascular health globally.

Background and significance of the problem:

Cardiovascular disease is a life-threatening condition that affects the ability of the heart to maintain an adequate blood flow, and it is caused by changes in its structure or function. It is the leading cause of hospitalizations in developed countries putting a high burden on the healthcare system, families, and caregivers.
The annual cost in healthcare for cardiovascular diseases is $30,000 in the USA with an estimate of over 8 million people suffering from heart issues by 2030.
With such statistics, forecasting the cardiovascular decompensation can play an important role in the healthcare system.

Potential of the project and its contribution:

Early diagnosis plays an important role to implement effective disease management strategies and prove timely treatment. Cardio Vigilant will help in early deduction of cardiovascular by using classification models which categorizes and labels the data among 2 or more categories and predicts the correct label of the given input data. Classification problem is best suited to specify if the patient is vulnerable to heart disease or not.

The primary goal of Cardiovascular decomposition forecasting is to explore the available data of the patient to identify patterns or signals that may indicate the risk factor associated with the cardiovascular events. This includes monitoring factors like chest pain type, blood pressure, cholesterol level and maximum heart rate by leveraging models like Logistic Regression, Support Vector Machines, LightGBM and Artificial Neural Networking.

Further , heart disease can have a widespread effect on different organs of our body and can be associated with various cardiovascular diseases and conditions such as Hypertension, Obesity, Diabetes, Arrhythmias(Irregular heart rhythms), Chronic kidney and many more.
Therefore, the long-term goal of the web application is to access the data associated with cardiovascular decompensation forecasting and provide the likelihood of developing other such heart related conditions.

Also, one pivotal enhancement in the pipeline involves implementation of a real-time based alert system that could detect the heart issues and notify the users with proactive steps in managing their heart health and thereby improving their overall user experience.

DATA SOURCE AND ATTRIBUTE DETAILS

The dataset comprises a range of attributes that are clinically relevant in the diagnosis and analysis of heart failure detection. The dataset provides inclusive information about various attributes related to patients' health and potential indicators of heart disease.

The dataset used in this analysis was retrieved from Kaggle. Kaggle is a web platform that hosts the world's largest Data Science community.

As understanding the attributes is crucial for analysis and constructing accurate predictive models, we tried to gain an insight on the meaning of the columns of the dataframe . Each attribute provides important insights into different aspects of heart-related complications.

Dataset Details:

- Age: of the patient in years.
- Sex: of the patient, categorized as Male (M) or Female (F).
- ChestPainType: Describes the chest pain categorized as Typical Angina (TA), Atypical Angina (ATA), Non-Anginal Pain (NAP), or Asymptomatic (ASY). Different types of chest pain may indicate different heart conditions.
- RestingBP: Represents the resting blood pressure of the patient measured in mmHg. Blood pressure is a common risk factor for heart disease.

- Cholesterol: Serum cholesterol level measured in mm/dl. High cholesterol levels are linked with an increased risk of developing heart disease.
- FastingBS: Fasting Blood Sugar Indicates whether the patient has fasting blood sugar or not. Levels > 120 mg/dl (1) or not (0).
- RestingECG: resting electrocardiogram as Normal, showing ST-T wave abnormalities (ST), or indicating left ventricular hypertrophy (LVH).
- MaxHR: Represents the maximum heart rate achieved. This attribute's numeric value lies between 60 and 202.
- ExerciseAngina: Indicates whether the patient experiences exercise-induced angina [Y: Yes, N: No].
- Oldpeak: Represents the ST depression measured during exercise. ST depression can indicate myocardial ischemia, which is a lack of blood flow to the heart muscle.
- ST_Slope: Describes the slope of the peak exercise ST segment as Upsloping, Flat, or Downsloping. [Up: upsloping, Flat: flat, Down: downsloping]
- Heart Disease: Output of predicting heart disease based on the input features mentioned above. [1: heart disease, 0: Normal]

```
print("The size of the data is: " ,df2.shape)
```

```
The size of the data is :  (2003, 12)
```

Experimental Evaluation

DATA CLEANING

Analyzing Null Values in the Dataframe

```
df2.isnull().sum()
```

```
Age              0
Sex              4
ChestPainType    0
RestingBP        8
Cholesterol      2
FastingBS        0
RestingECG       0
MaxHR            3
ExerciseAngina   0
Oldpeak          0
ST_Slope         0
HeartDisease    18
dtype: int64
```

- On Analysis we found , that the data has some null values in columns : Sex , RestingBP , Cholesterol , MaxHR , Heart Disease. To work on that data, we perform further analysis.
- Further, on analyzing we found that some of the categorical columns that are supposed to be binary have more that category.

```
# Check no of unique values in each column
print(df2.nunique())
```

```
Age               67
Sex               11
ChestPainType      5
RestingBP         68
Cholesterol      229
FastingBS         10
RestingECG        15
MaxHR            120
ExerciseAngina     8
Oldpeak           54
ST_Slope           8
HeartDisease       4
dtype: int64
```

Analyzing Heart Disease Column : Dropping Null Values – using dropna()

- Records with no Target variable will not help us train our model . Further it will distort calculations, resulting in inaccurate conclusions.
- Therefore, all the null values are removed from the column - Heart Disease.

```
df['HeartDisease'].isna().sum()
```

```
0
```

Analyzing the Column for any unnecessary values.

```
df['HeartDisease'].unique()
```

```
array(['0', '1', '.', '...'], dtype=object)
```

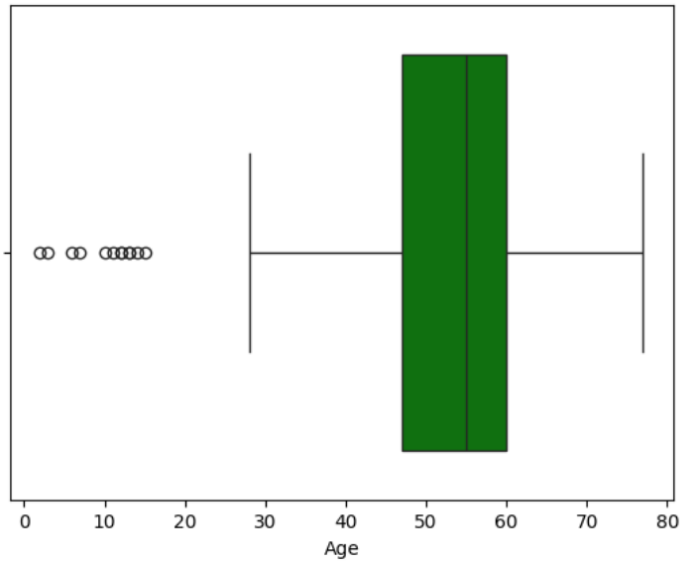
- After removing the unnecessary values in the Column: Heart Disease

```
df['HeartDisease'].unique()
```

```
array(['0', '1'], dtype=object)
```

Analyzing & Removing Outliers

- Based on the graph below, we can see that the Age of the Patients is generally centered around 25+ to 75+. However, there are some records with Age lesser than 20. These records are outliers because they deviate significantly from the rest of the data.



Analyzing & Renaming Column: Sex

```
df['Sex'].unique()
```

```
array(['M', 'F', 'Mr', 'Mrs', 'Fem', 'female', 'feml', nan], dtype=object)
```

According, to the observation, we find that the Column consisting gender of the patients have some discrepancy.

Null Values in the Column: Sex

After analysis, it was found that some of the records also had null values for the gender. In this column we cannot impute the null value with any mean or mode value. Hence these records have to eliminated.

```
 # Drop rows where "Sex" is null
df = df.dropna(subset=["Sex"])
```

Using Groupby and Rename Cleaning the Column: ChestPainType

Checking the value counts of each Chest Pain Type.

```
count_by_value = df.groupby(df['ChestPainType'])
count_by_value.size()
```

```
ChestPainType
ASY      585
ATA      355
NAP      493
NAPPP      1
TA       545
dtype: int64
```

After Renaming

```
#Checking the unique values in Chest Pain Type after Renaming Column
df['ChestPainType'].unique()
```

```
array(['ATA', 'NAP', 'ASY', 'TA'], dtype=object)
```

Analyzing for duplicate records

- After analyzing we found some duplicate records in the dataframe that have to be removed.

```
# Check for duplicate rows
df.duplicated().sum()
```

```
0
```

Renaming Values in ExerciseAngina Column

- The column: Exercise Angina is supposed to have only two categories namely yes or no however, the column has some discrepancy in the naming.

```
df['ExerciseAngina'].unique()
```

```
array(['N', 'NO', 'Yes', 'Never', 'Y', 'no', 'ye'], dtype=object)
```

After renaming them to the correct

```
df['ExerciseAngina'].unique()
```

```
array(['N', 'Y'], dtype=object)
```

Imputing Null Values with mean Values of Resting BP

```
# impute 0 values with mean Resting BP
mean_resting_bp = df['RestingBP'].mean()
df['RestingBP'] = df['RestingBP'].fillna(mean_resting_bp)
```

```
df['RestingBP'].isnull().sum()
```

```
0
```

Renaming Resting ECG

```
df['RestingECG'].unique()
```

```
array(['Normal', 'normal', 'ST', 'normallll', 'aroundnormal', 'LVH',
       'normalll', 'Never', 'lvh'], dtype=object)
```

```
df['RestingECG'].unique()
```

```
array(['Normal', 'ST', 'LVH'], dtype=object)
```

Checking the remaining null Values in MaxHR Column

```
df['MaxHR'].isna().sum()
```

```
3
```

Label Encoding Dataset

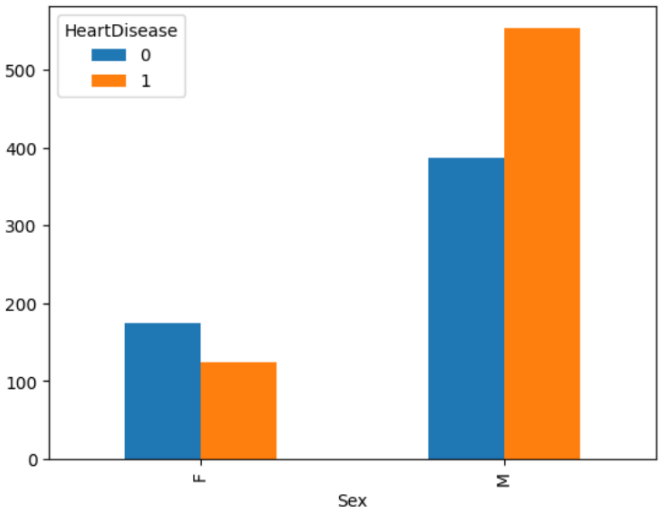
```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df1 = df.copy(deep = True)

df1['Sex'] = le.fit_transform(df1['Sex'])
df1['ChestPainType'] = le.fit_transform(df1['ChestPainType'])
df1['RestingECG'] = le.fit_transform(df1['RestingECG'])
df1['ExerciseAngina'] = le.fit_transform(df1['ExerciseAngina'])
df1['ST_Slope'] = le.fit_transform(df1['ST_Slope'])
```

- Creating a deep copy of the original dataset and label encoding the text data of the categorical features. Modifications in the original dataset will not be highlighted in this deep copy. Hence, we use this deep copy of dataset that has all the features converted into numerical values for visualization & modeling purposes.

EXPLORATORY DATA ANALYSIS

What is the effect of Gender on Heart Disease?



- Male population has more heart disease patients than no heart disease patients. In the case of Female population, heart disease patients are less than no heart disease patients.

What is the most common type of chest pain that affects the disease ?

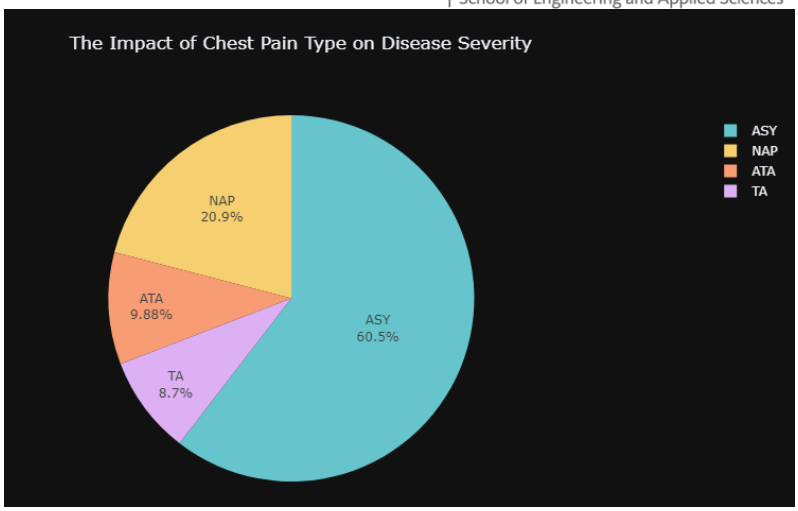
```
df['ChestPainType'].value_counts()
```

```
ASY    527
NAP    292
ATA    230
TA     190
Name: ChestPainType, dtype: int64
```

```
df['new_Col'] = pd.to_numeric(df['HeartDisease'])
```

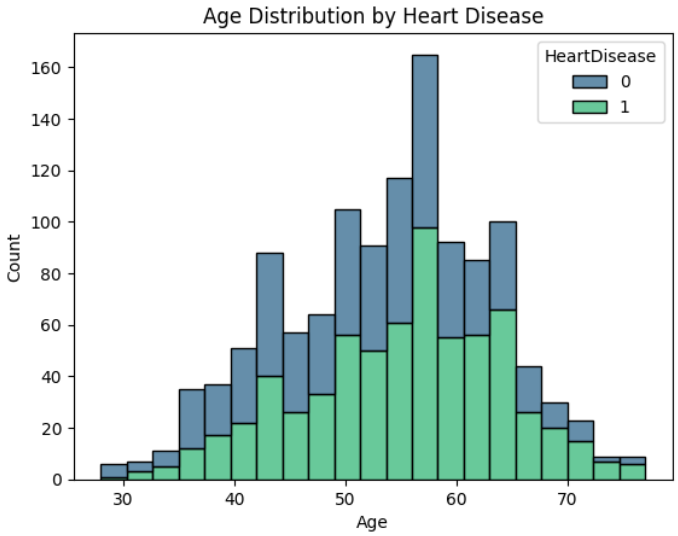
```
df[df['new_Col']==1]['ChestPainType'].value_counts()
```

```
ASY    410
NAP    142
ATA     67
TA      59
Name: ChestPainType, dtype: int64
```



- There are four types of chest pain that one can suffer from - Asymptomatic (ASY), Atypical angina(ATA), Non-Anginal pain(NAP) and Typical Angina(TA).
- According to the plot above, the order of the type of chest pain on the disease is:
  - ASY
  - NAP
  - ATA
  - TA
- Therefore, the people who suffer from ASY-type chest pain are more prone to have heart diseases like heart attack, blockage of blood flow or possible heart damage as compared to others. While with TA-type chest pain shows least sign of heart diseases.

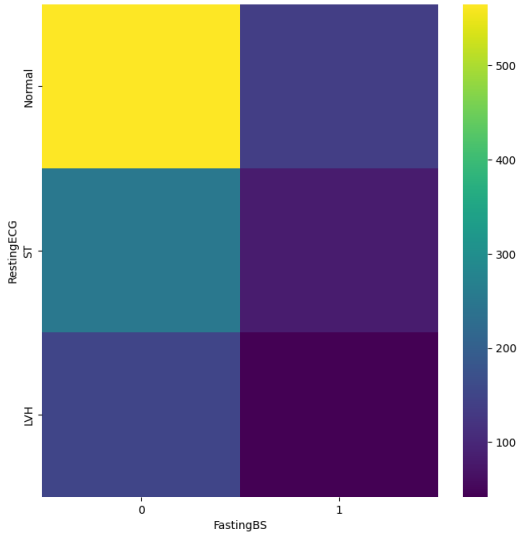
Relation of Age on Heart Disease



- According to the graph above, heart disease is very common in the seniors who are in the age group of 50 and above and common among adults who belong to the age group of 41 to 50.

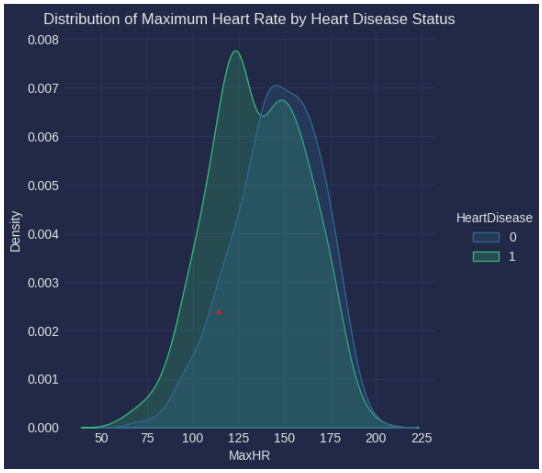
- However, it's rare among the age group of 19 to 40 and very rare among the age group of 0 to 18.

Resting ECG VS FastingBS



- There are three types of fasting ECG:
  1 . Left Ventricular hypertrophy
  2 . ST
  3 . Normal
- According to the plot above, people with resting ECG of type - LVH are highly likely to have a positive fasting blood sugar while people with Normal ECG type are least prone to fasting blood sugar.

What is the distribution of Maximum Heart Rate by Heart Disease Status.



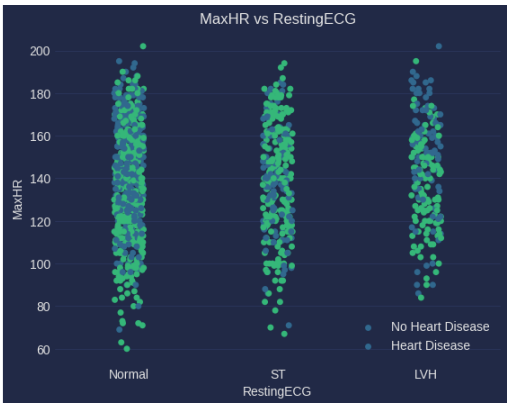
- According to the plot, higher maximum Heart Rates are more susceptible to heart disease.

What is the Relation btw the RestingECG and gender(M/F) ?

RestingECG	LVH	Normal	ST
Sex			
F	0.296296	0.363636	0.600000
M	0.636986	0.546125	0.654762

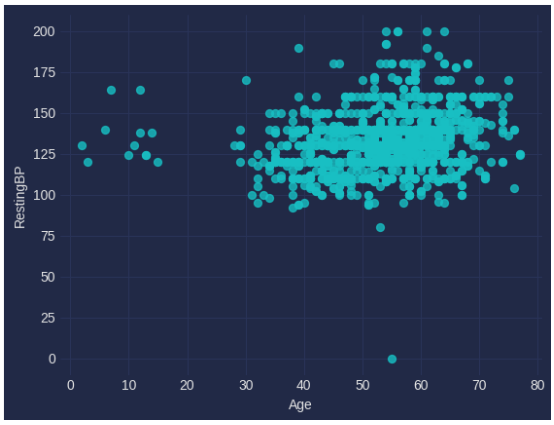
- As per the above table, RestingECG of male is higher than females which implies that heart disease is more common in males as compared to females.

MaxHR VS Resting ECG



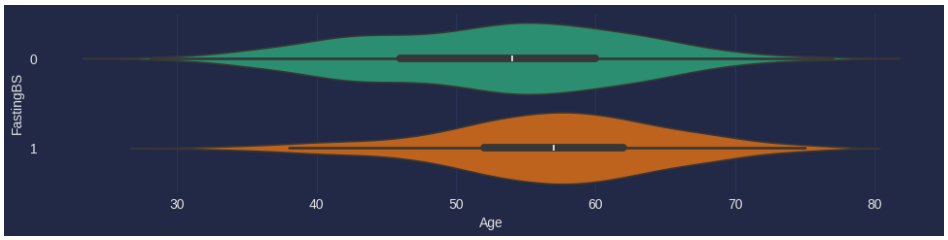
- For maximum heart rate values, heart disease is detected below 140 points and **Normal** RestingECG. ST and LVH throughout the maximum heart rate values display heart disease cases.

AGE VS RestingBP



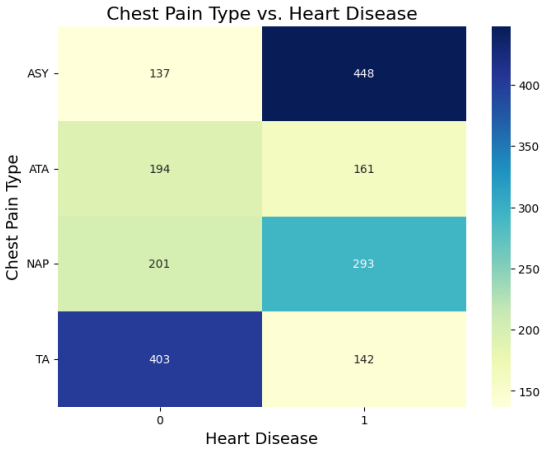
- According to the graph generated above, we can conclude that higher resting BP is more common among the adults and seniors in the age group of 30 to 80 while higher resting BP is not that common among young people with the age group of 0 to 30.

What is the relation between fasting Blood sugar with the age pf the patient ?



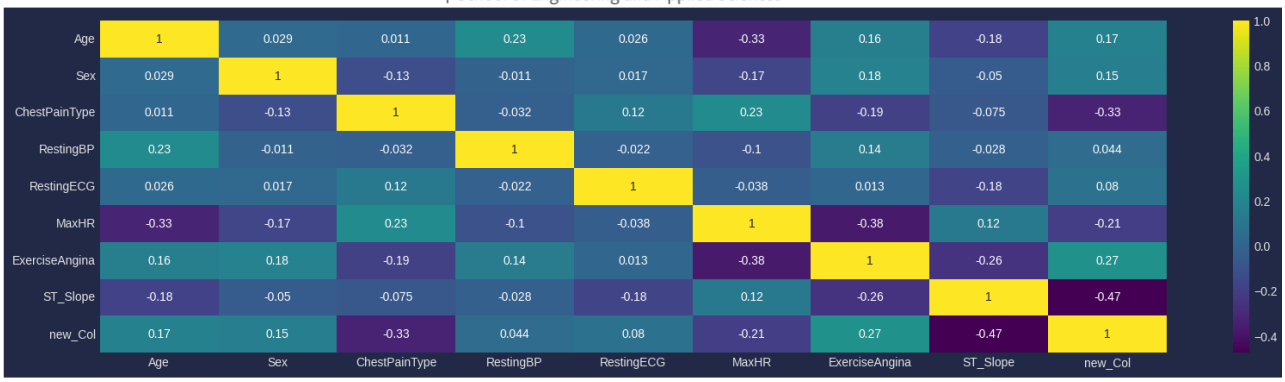
- According to the graph above, FastingBS > 120 mg/dl fasting blood sugar is associated with the median age group of 50 to 60. This implies that high blood sugar level is more susceptible to adults with age groups of 50 to 60.

CHEST PAIN TYPE VS HEART DISEASE



CORRELATION MATRIX

The graph depicts the correlation of all the feature with the Target Variable : Heart Disease.



Future Work

After cleaning the data and conducting exploratory analysis, we will proceed to implement machine learning models such as:

- Logistic regression - To estimate the probability if the user is having heart health issues or not with a Boolean outcome of true or false.
- Random forest classification - Multiple decision trees are created using random subsets of data and the prediction is made by calculating the prediction for each tree and choosing the most popular result.
- K-nearest - Neighbors to predict cardiovascular failure. Subsequently, we will develop a web application where users can access information by inputting the relevant attributes.