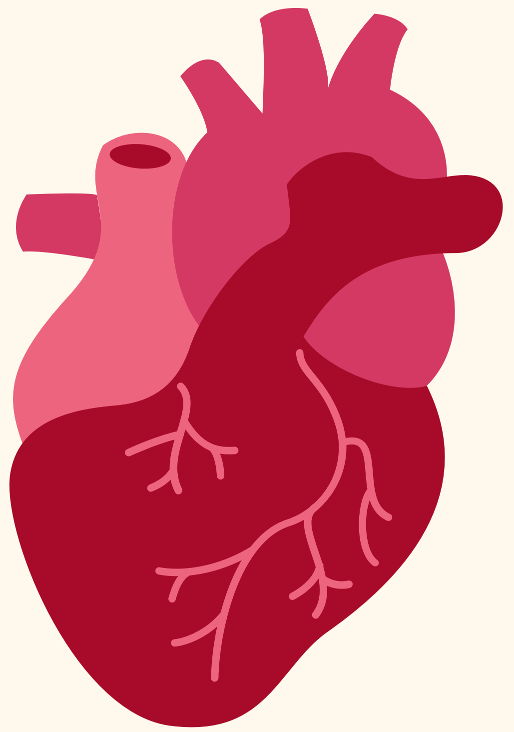
SC1015 MINI PROJECT



~ HEART DISEASE ANALYSIS & PREDICTION ~

KOH WEE XUAN - U2320197F
WOO WENG TAI - U2322615J
TEO LIANG WEI, RYAN - U2321344G

TABLE OF CONTENTS

01
Problem Statement

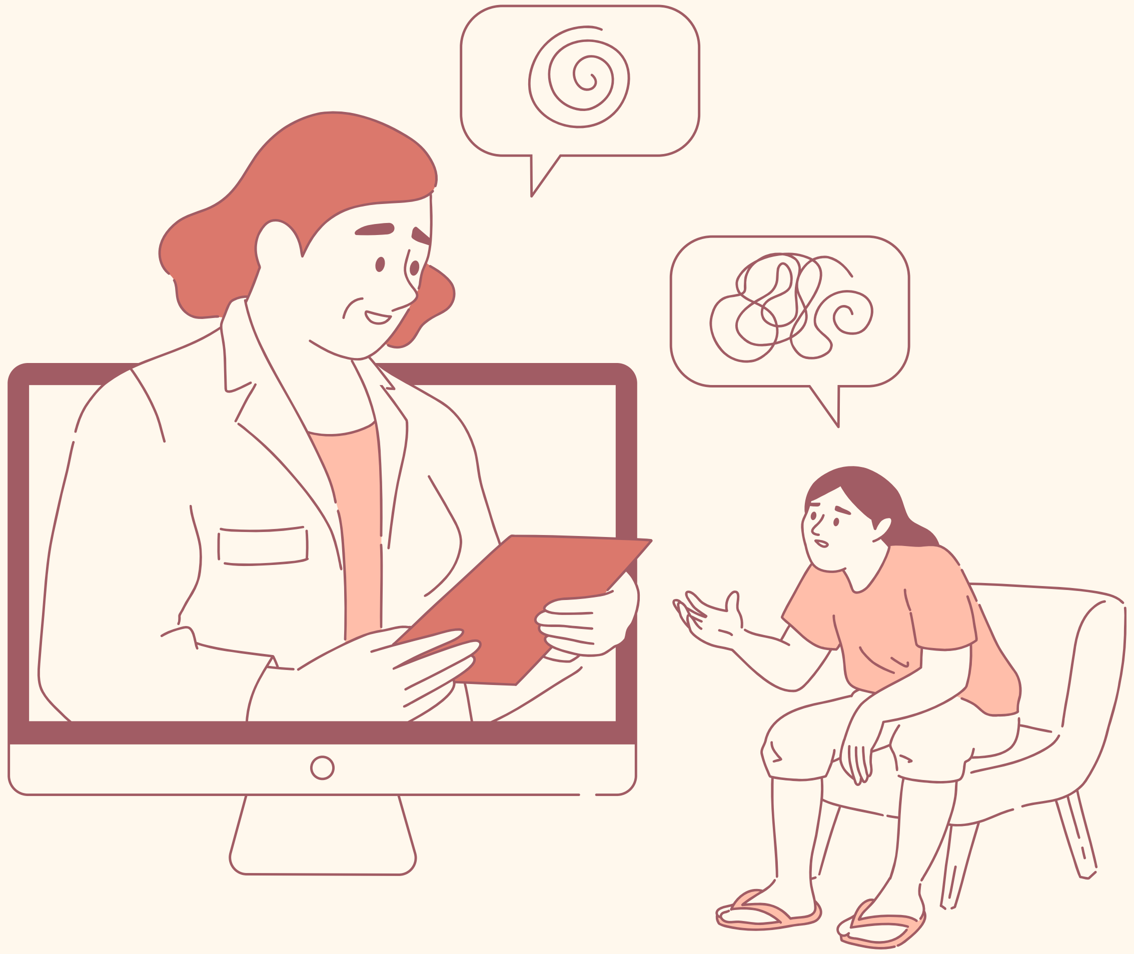
02
Data Preparation

03
Exploratory Analysis

04
Machine Learning Techniques

05
Data Driven Insights

06
Conclusion + Future Possibilities



PROBLEM STATEMENT

DID YOU KNOW?

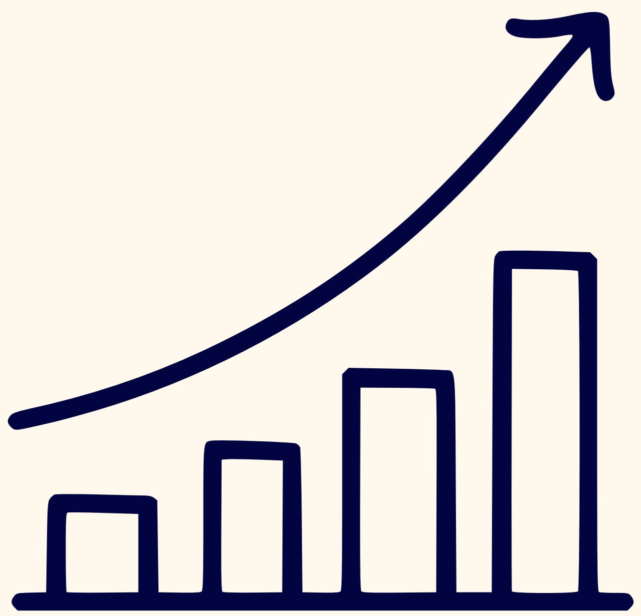
17.9 MILLIONS

Lives are lost to cardiovascular per year!

CHALLENGES

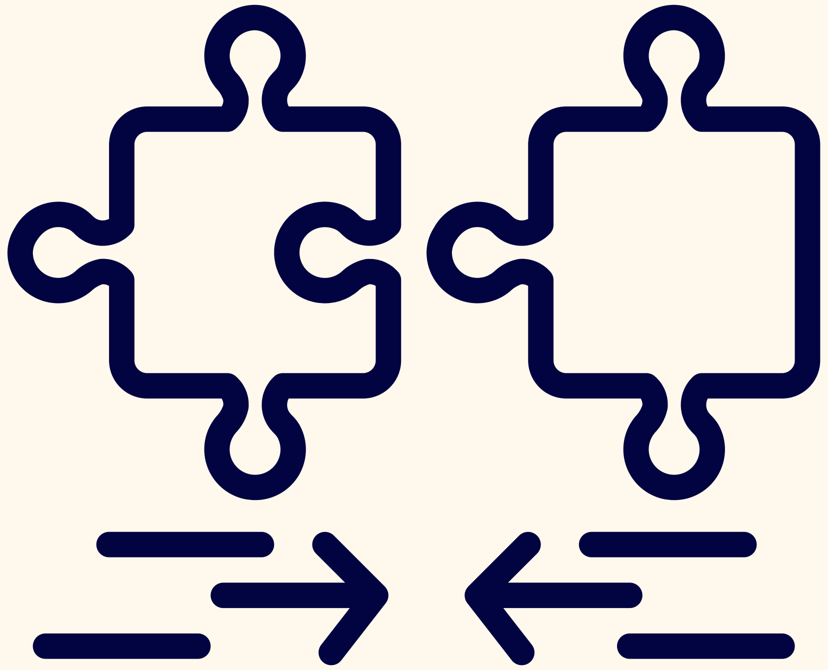
Heart disease is a leading cause of mortality worldwide, it necessitates the development of effective prediction models to support timely interventions and reduce the associated health burden.

OUR GOALS



TRENDS

PREDICTIONS



IMPLICATIONS

SAMPLE COLLECTION

- Framingham Heart Study dataset
- 4238 samples
- 16 different variables
- Identifies the risk of coronary heart disease in the next ten years.
- Goal - To identify trends and to train a prediction model for early prevention

male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0

SAMPLE COLLECTION

- Cardiomegaly Disease Prediction Dataset
- 4438 train images
- 1114 test images
- Goal - to train a Convolutional Neural Network for early detection

Cardiomegaly Disease Prediction Using CNN

Cardiomegaly Disease

kaggle.com

DATA PREPARATION

CLEANING PROCESS

Exploring the variables, there are a few issues:

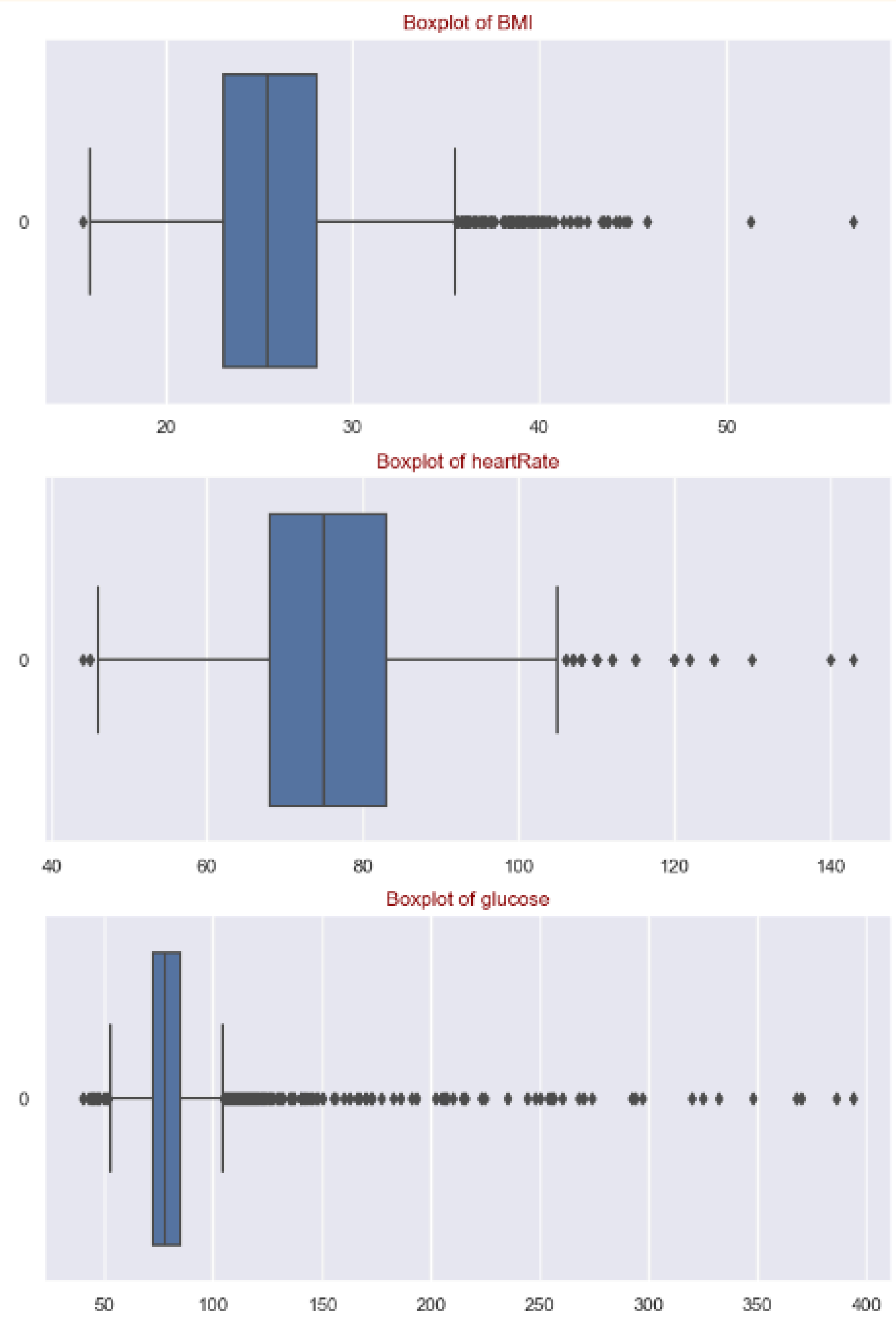
- Certain rows have missing or null values
- The numerical variables have quite many outliers seen from the box graphs plotted

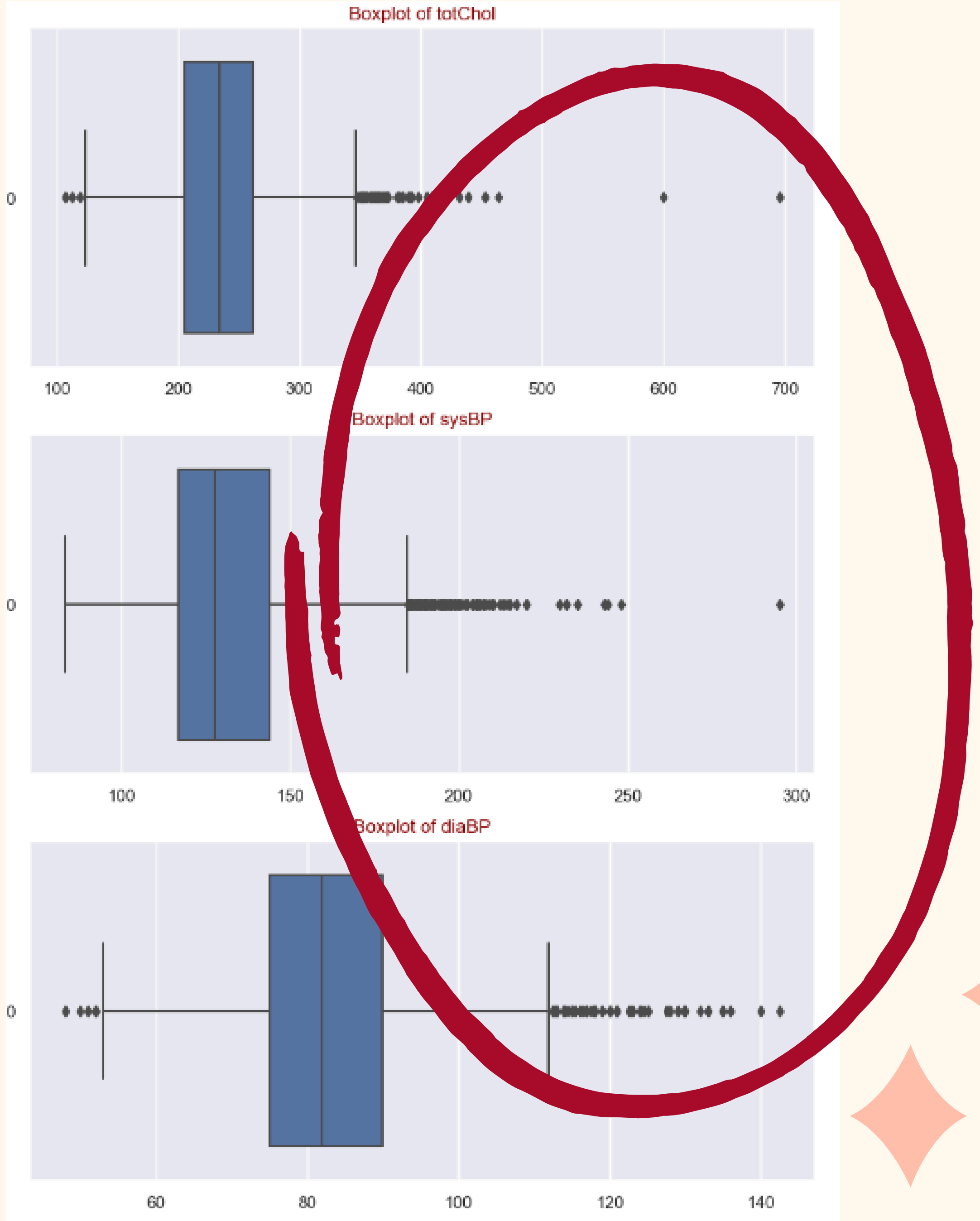
Missing Value Cleaning

Check to see if there is any null values in the data set

```
# Percentage of null values in each column
(heartData.isnull().sum()/heartData.shape[0])*100
```

```
male               0.000000
age                0.000000
education          2.477584
currentSmoker      0.000000
cigsPerDay         0.684285
BPMeds             1.250590
prevalentStroke    0.000000
prevalentHyp       0.000000
diabetes           0.000000
totChol            1.179802
sysBP              0.000000
diaBP              0.000000
BMI                0.448325
heartRate          0.023596
glucose            9.155262
TenYearCHD         0.000000
dtype: float64
```





MISSING/NULL VALUES

- Filling up these missing slots with values
- Either with a 0 if the variable type is binary categorical
- Or with a median value of that variable if its a numerical type
- Some numerical variables have a separate binary categorical value that must be true in order for it

```python
#Median to fill 0 values if diabetes is 1
diabetesIs1_data = heartData[heartData['diabetes']==1]
median_glucose_diabetes_1 = diabetesIs1_data['glucose'].median()
#Median to fill 0 values if currentSmoker is 0
currentSmokerIs0_data = heartData[heartData['currentSmoker']==0]
median_heartRate_currentSmoker_0 = currentSmokerIs0_data['heartRate'].median()

heartData.fillna({'education': 0,
                  'cigsPerDay': heartData['cigsPerDay'].where(heartData['currentSmoker'] == 1).median(),
                  'BPMeds': 0,
                  'totChol': heartData['totChol'].median(),
                  'BMI': heartData['BMI'].median(),
                  'heartRate': heartData['heartRate'].where(heartData['currentSmoker'] == 1).median(),
                  'glucose': heartData['glucose'].where(heartData['diabetes'] == 0).median()},
                  inplace=True)
# Additional step for filling missing values with median values for heartRate and glucose when 'currentS
heartData.fillna({'heartRate': median_heartRate_currentSmoker_0,
                  'glucose': median_glucose_diabetes_1},
                  inplace=True)
```

OUTLIERS TREATMENT

- Removing outliers beyond 1.5 times quantile gap

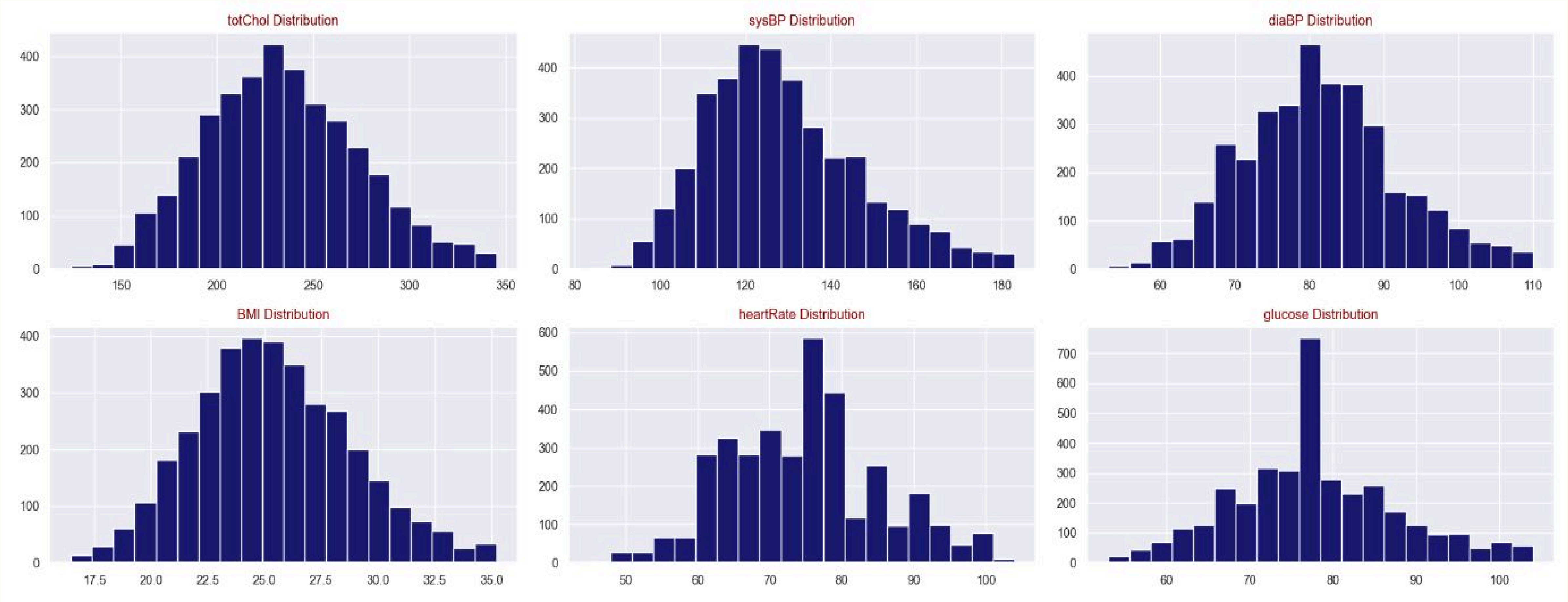
```
There were 4238 rows before outlier treatment.
There are 3620 rows after outlier treatment.
After outlier treatment number of rows lost are 618.
```

EXPLORATORY ANALYSIS

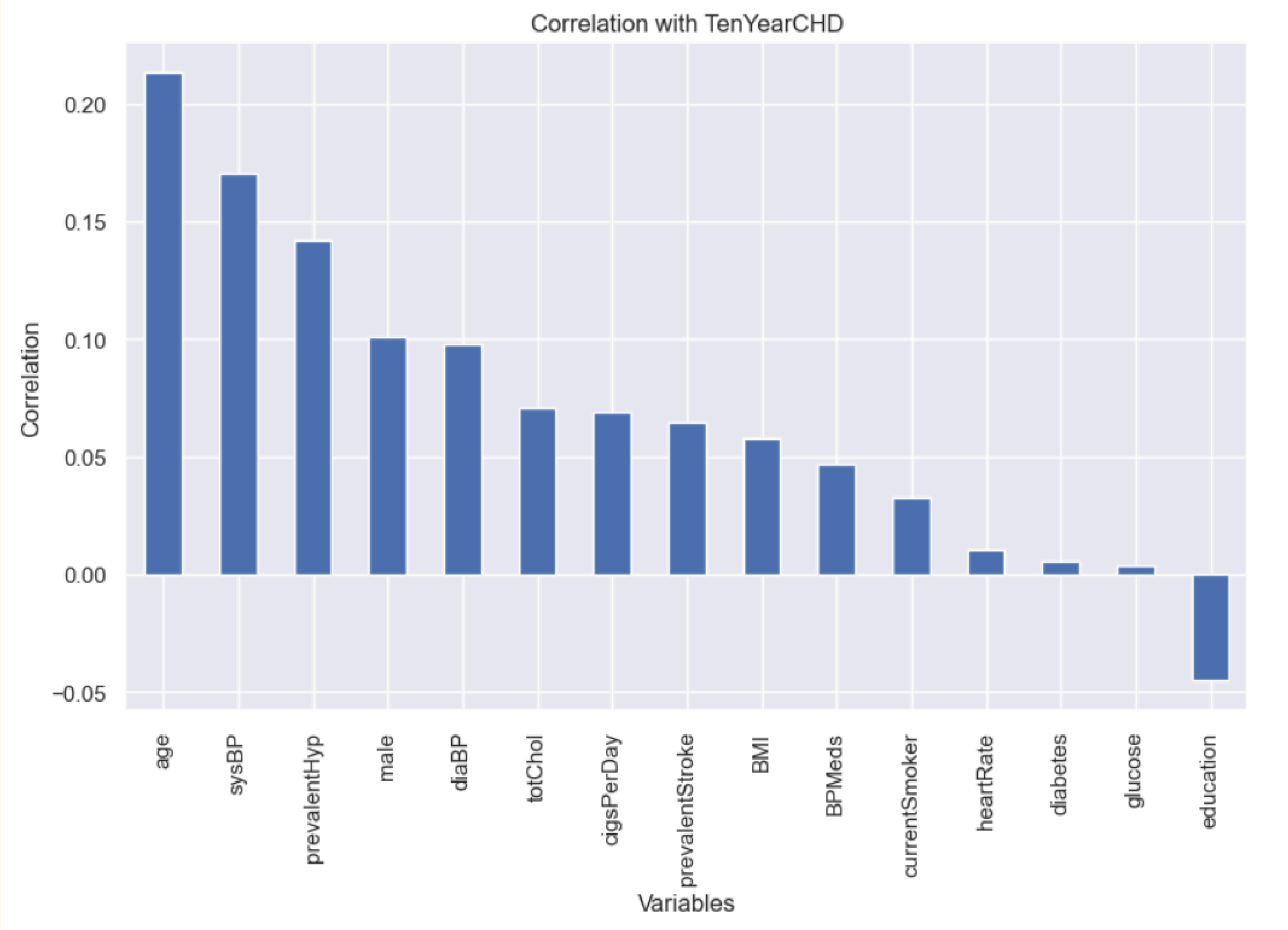
DATASET EXPLORATION

- First we looked through what the variables in the dataset can offer
  - Generate different graphs to see the frequency range of values
  - Observe how certain variables correlates with the end result

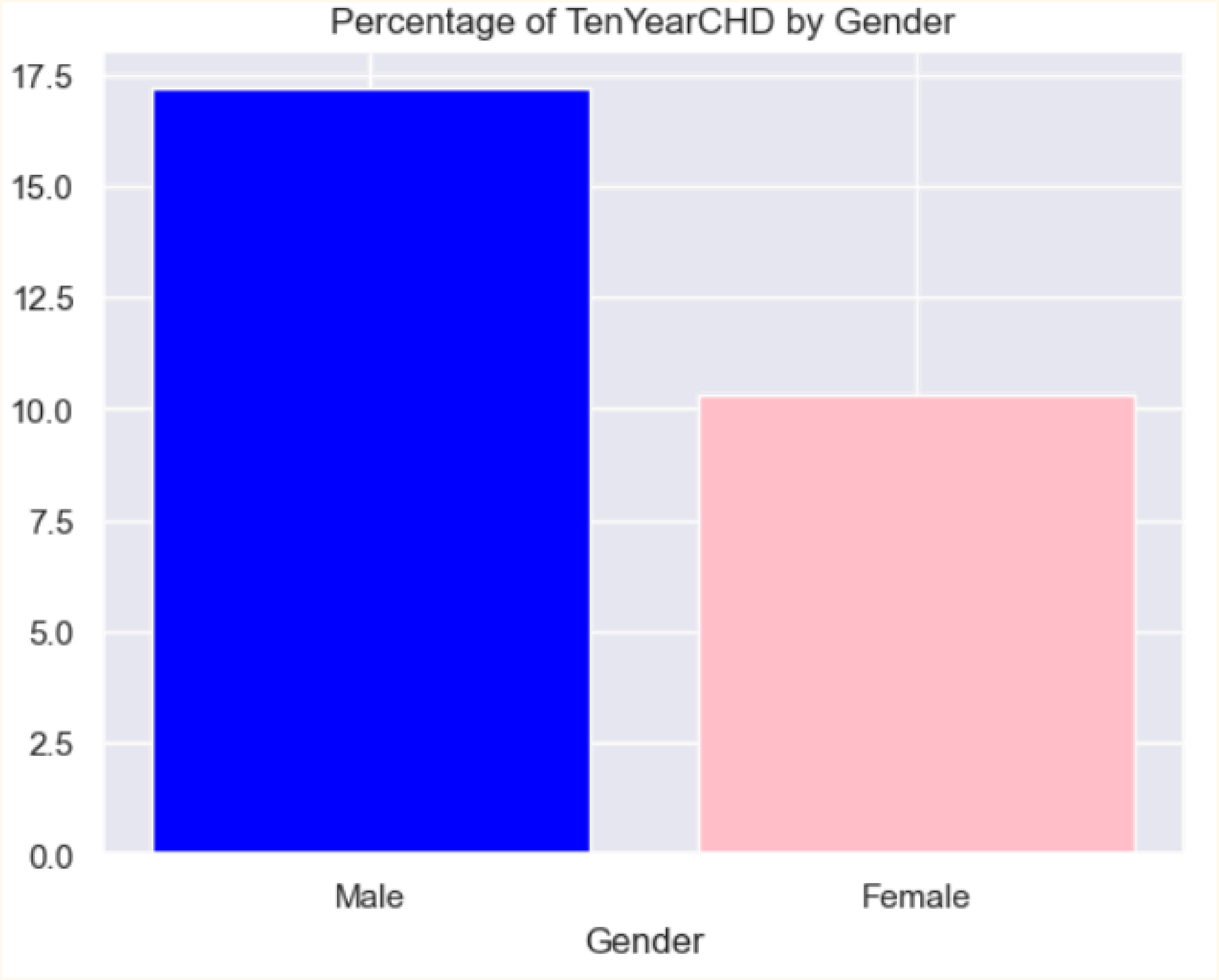
```
In [5]: # Information about the Variables
        heartData.info()
        heartData.shape

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4238 entries, 0 to 4237
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   male             4238 non-null   int64
 1   age              4238 non-null   int64
 2   education        4133 non-null   float64
 3   currentSmoker    4238 non-null   int64
 4   cigsPerDay       4209 non-null   float64
 5   BPMeds           4185 non-null   float64
 6   prevalentStroke  4238 non-null   int64
 7   prevalentHyp     4238 non-null   int64
 8   diabetes         4238 non-null   int64
 9   totChol          4188 non-null   float64
 10  sysBP            4238 non-null   float64
 11  diaBP            4238 non-null   float64
 12  BMI              4219 non-null   float64
 13  heartRate        4237 non-null   float64
 14  glucose          3850 non-null   float64
 15  TenYearCHD       4238 non-null   int64
dtypes: float64(9), int64(7)
```



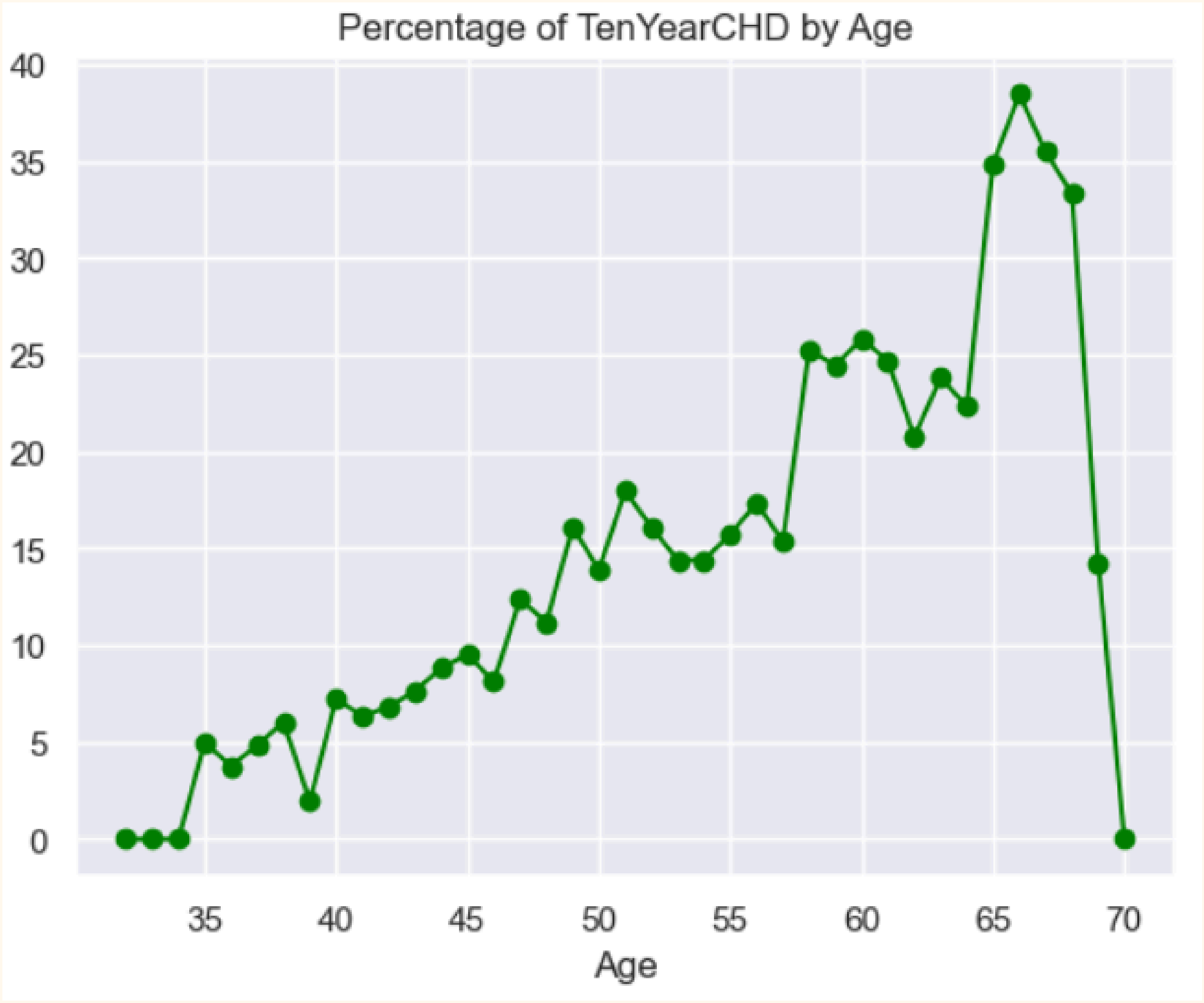
PREDOMINANT RISK FACTORS



RISK FACTORS - DAILY LIFE

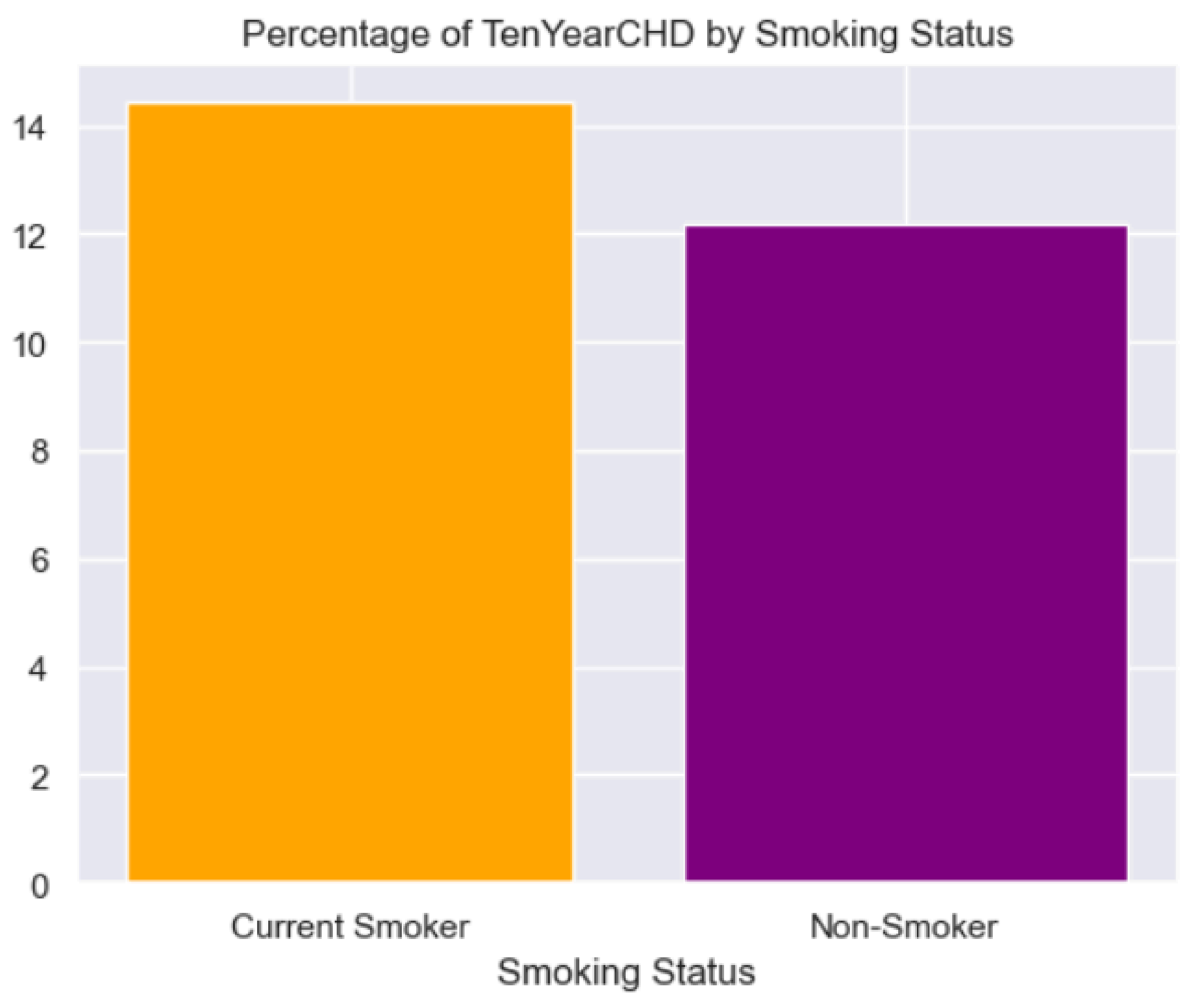


66.96% more likely for biological males

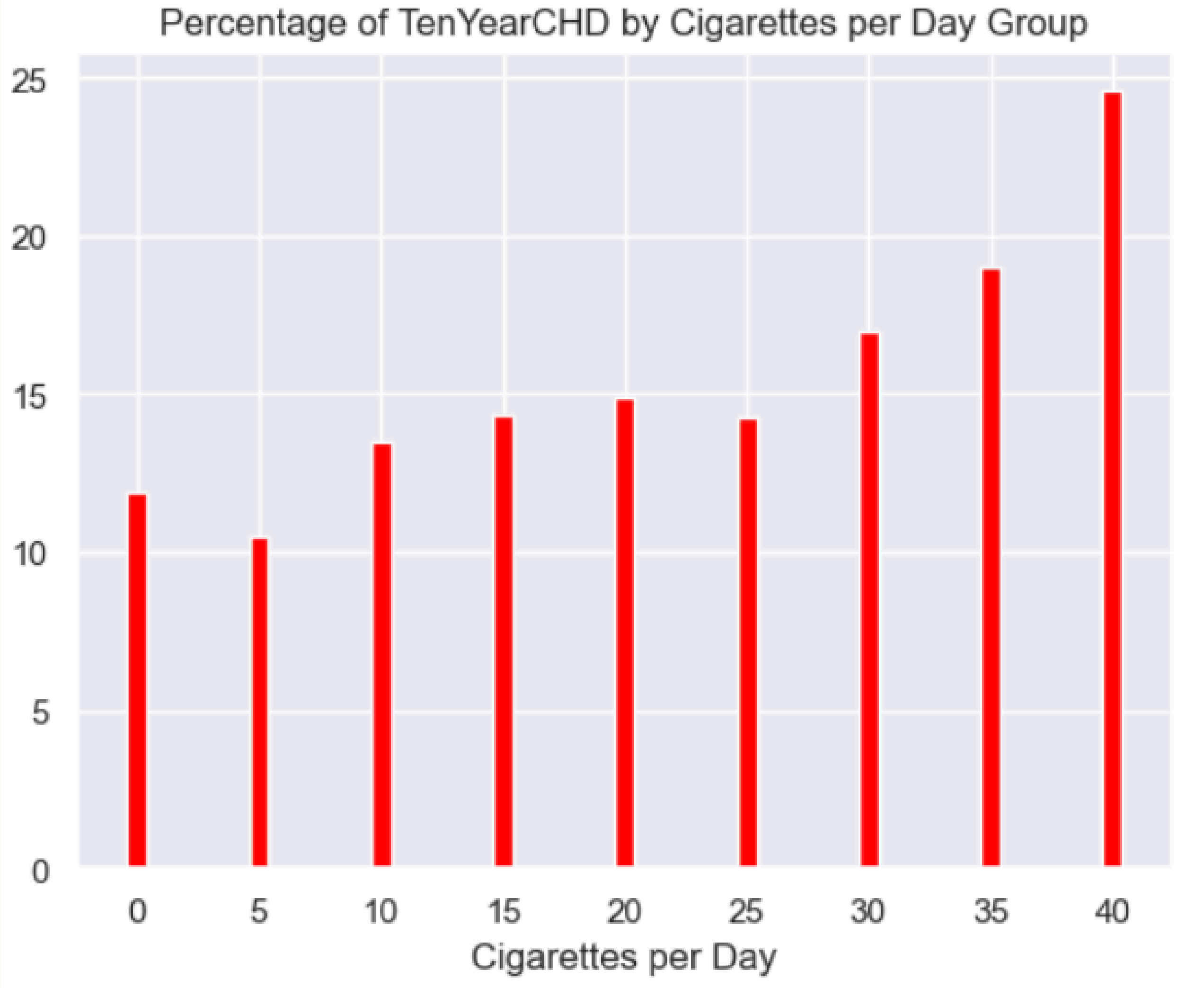


7.66% increase per year

RISK FACTORS - DAILY LIFE

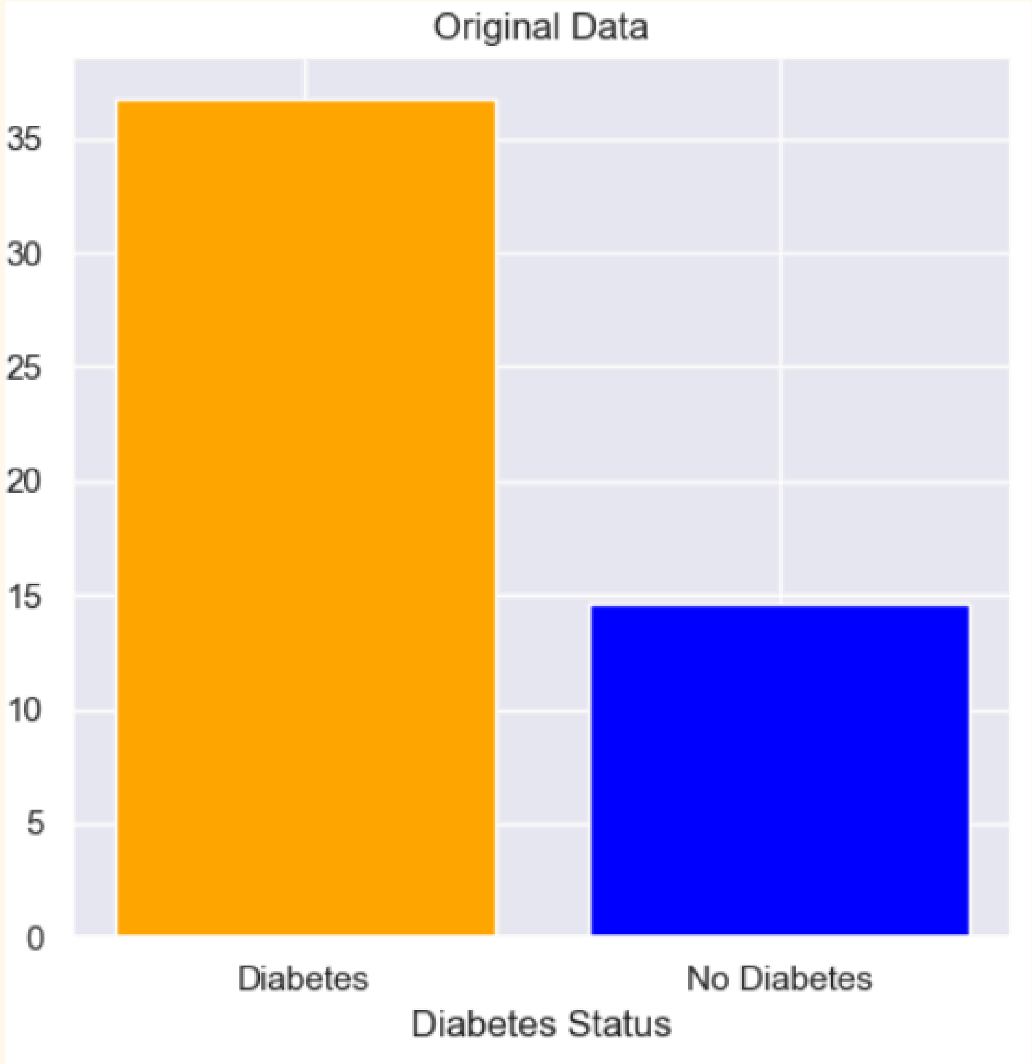


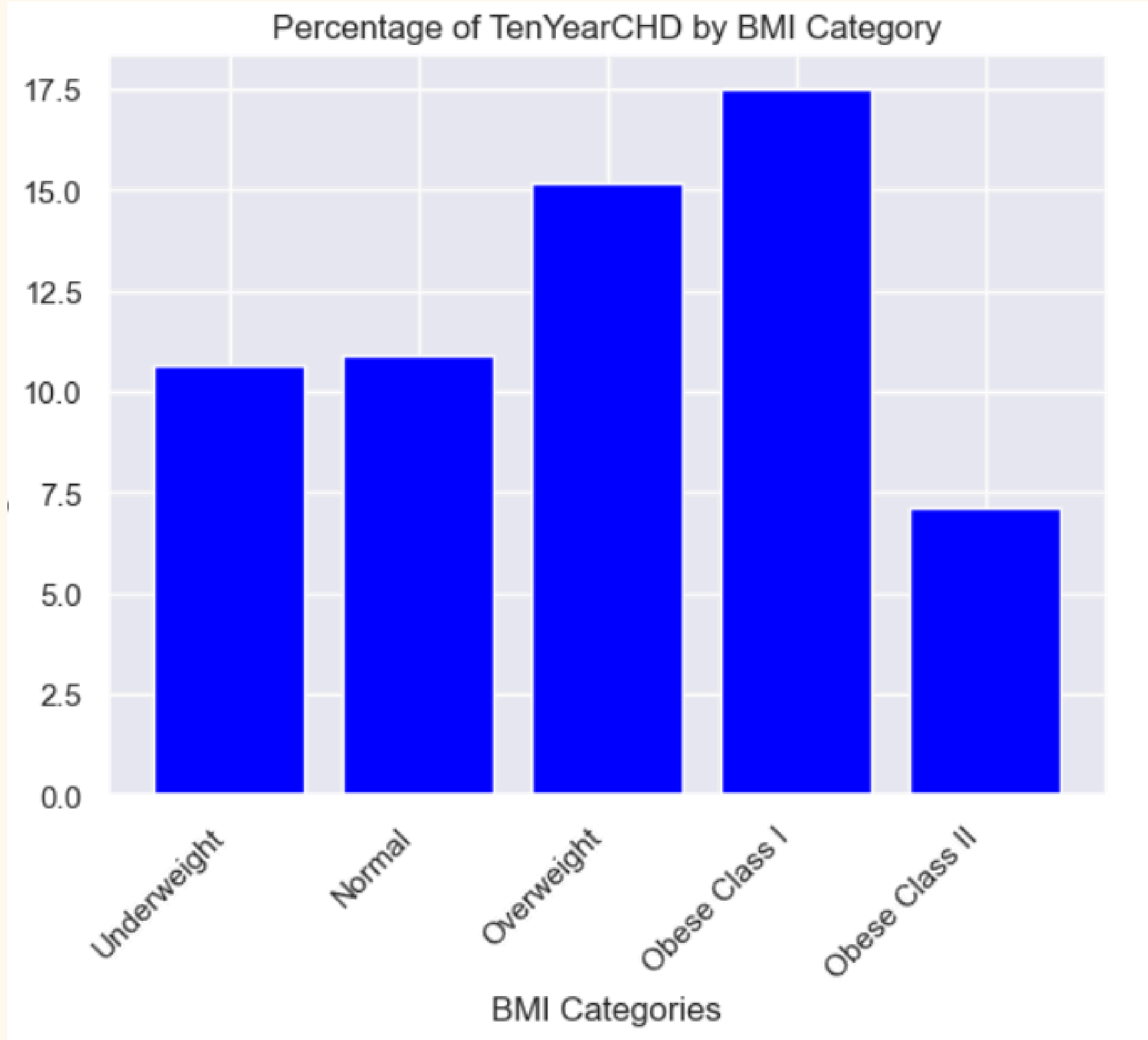
18.3% higher chance for smokers



1.64% increase per cigarette smoked a day.

RISK FACTORS - DAILY LIFE

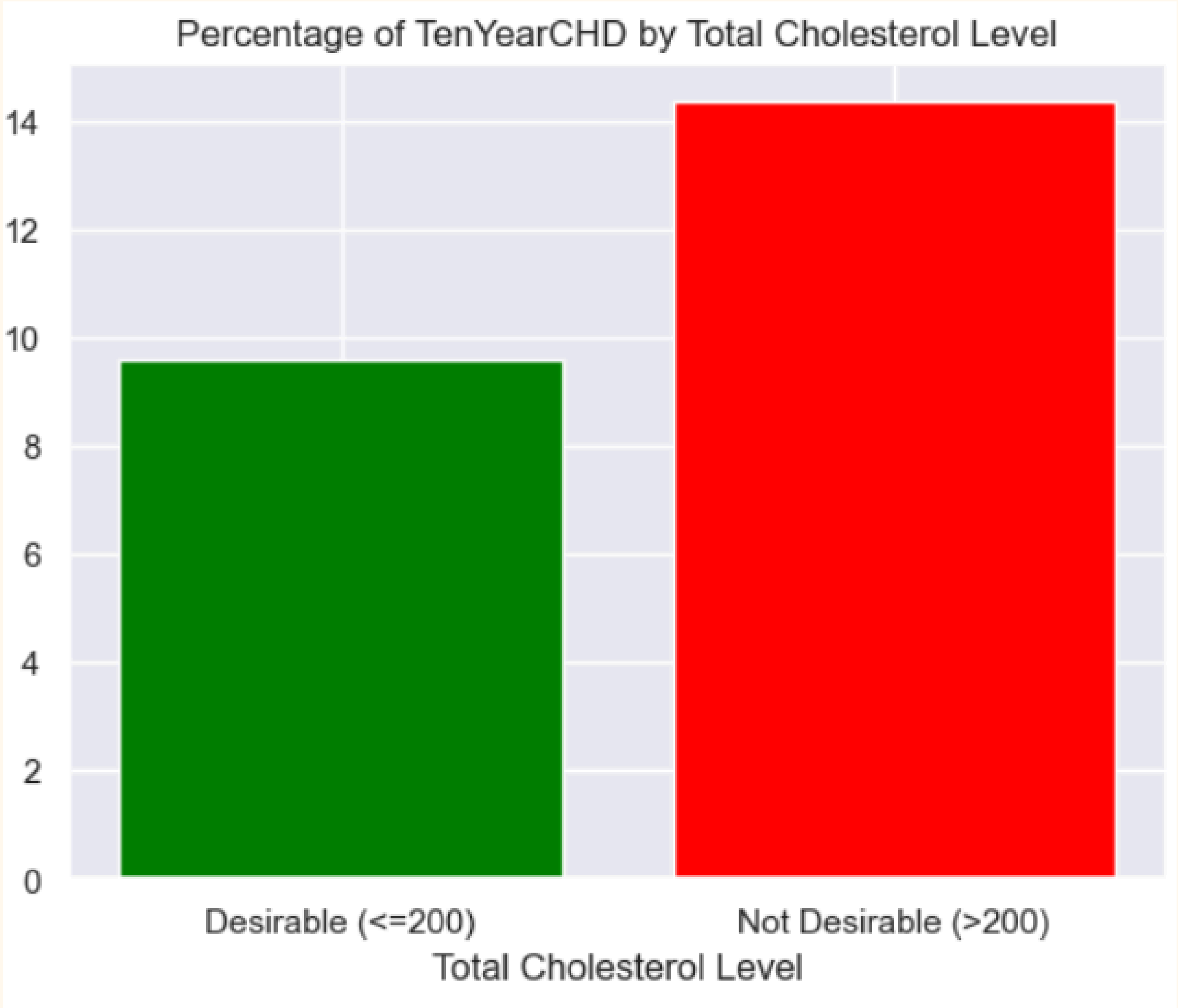


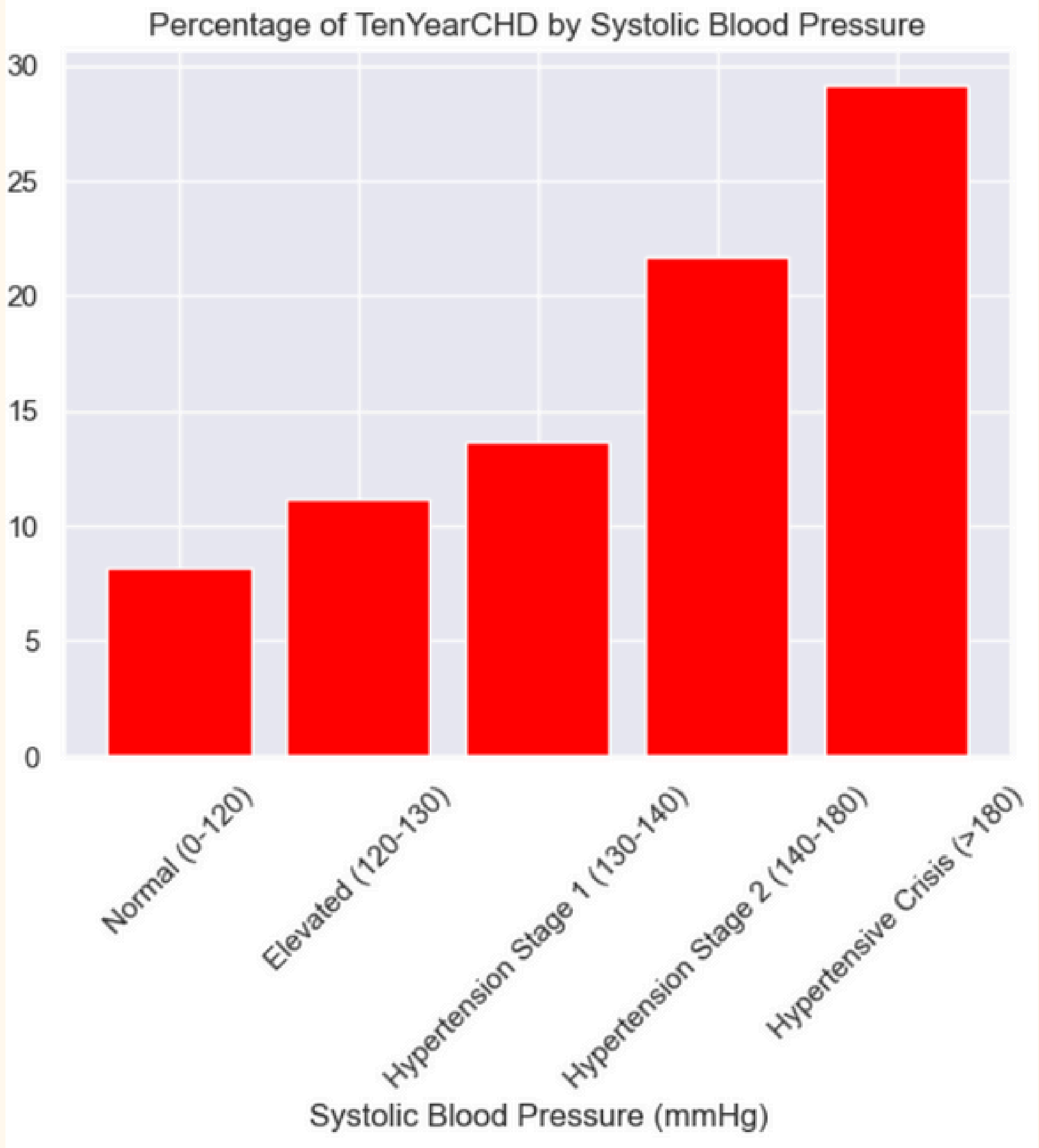


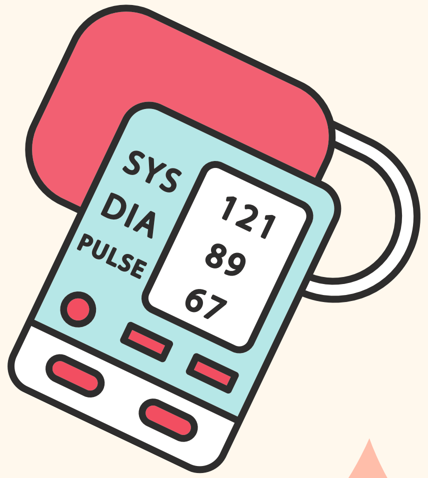
150.87% more likely for people with diabetes

4.96% increase per unit increase in BMI

RISK FACTORS - DAILY LIFE



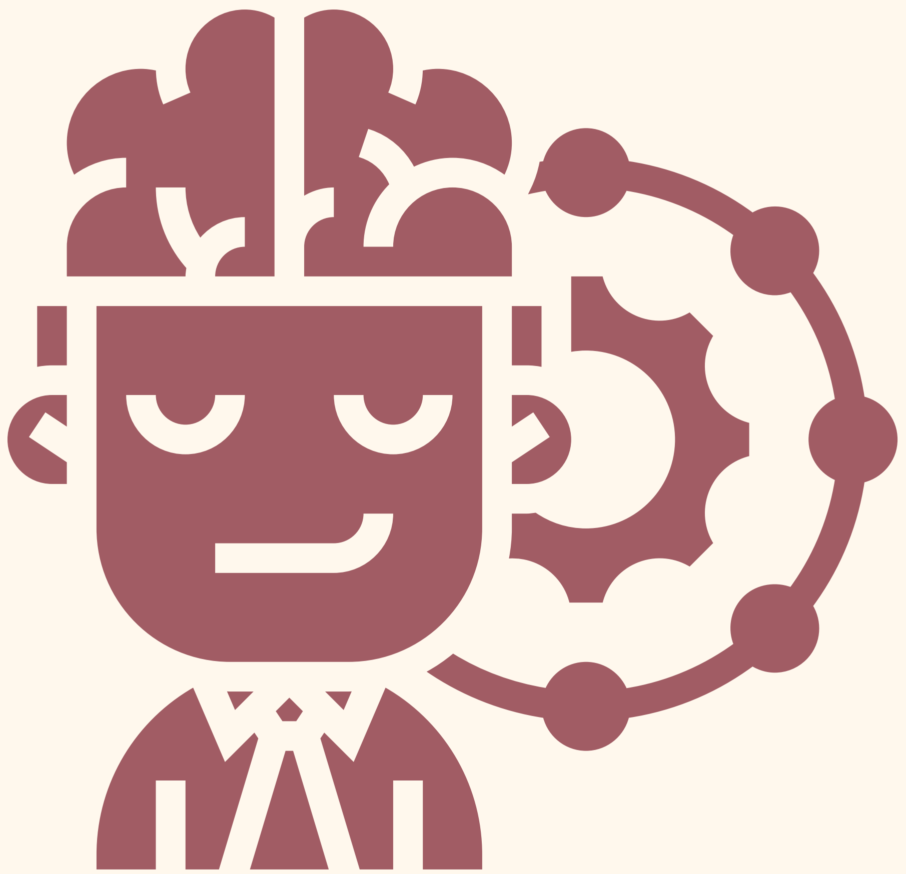




49.77% more likely for people with high cholesterol

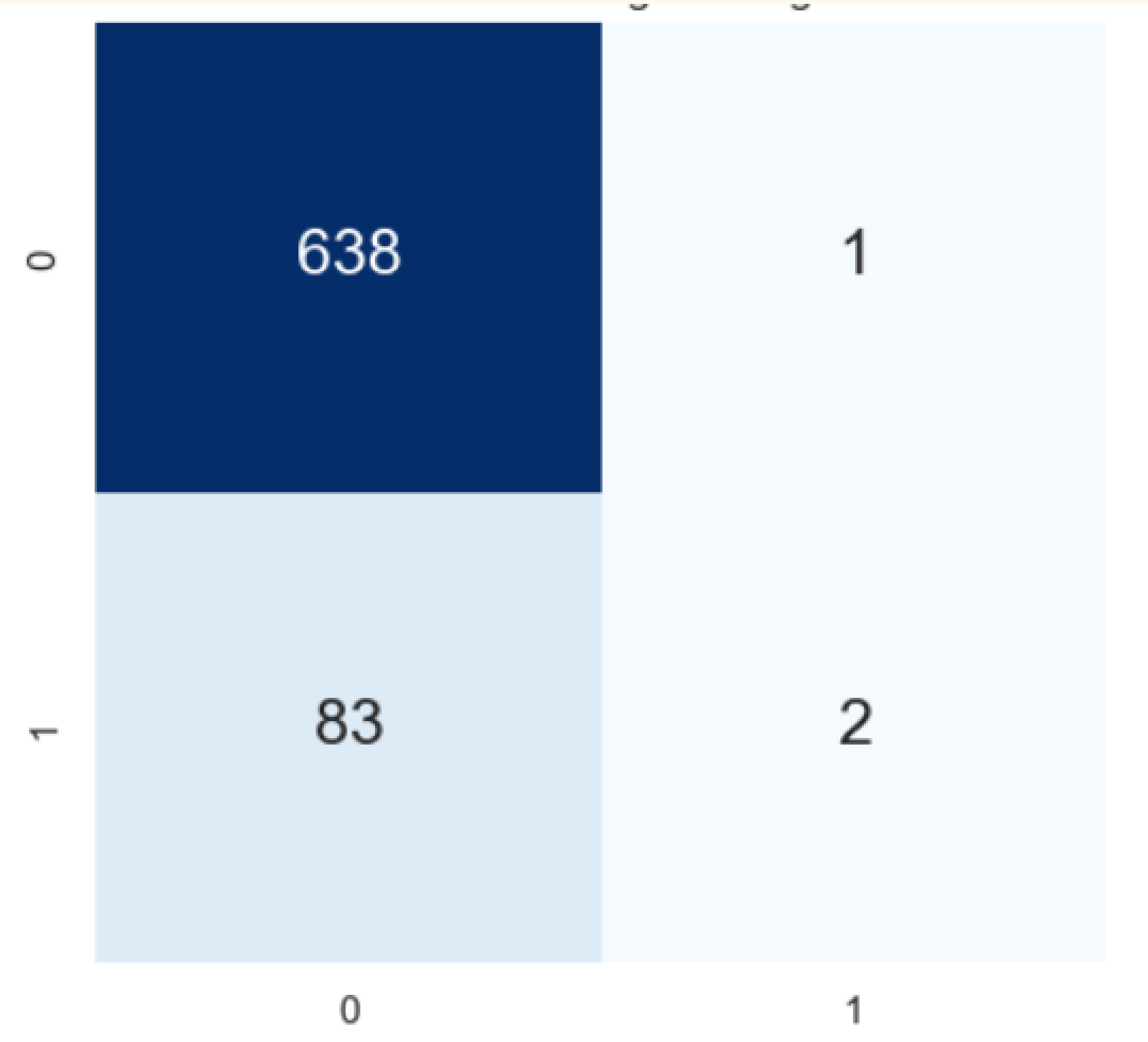
95.68% more likely for people with high blood pressure

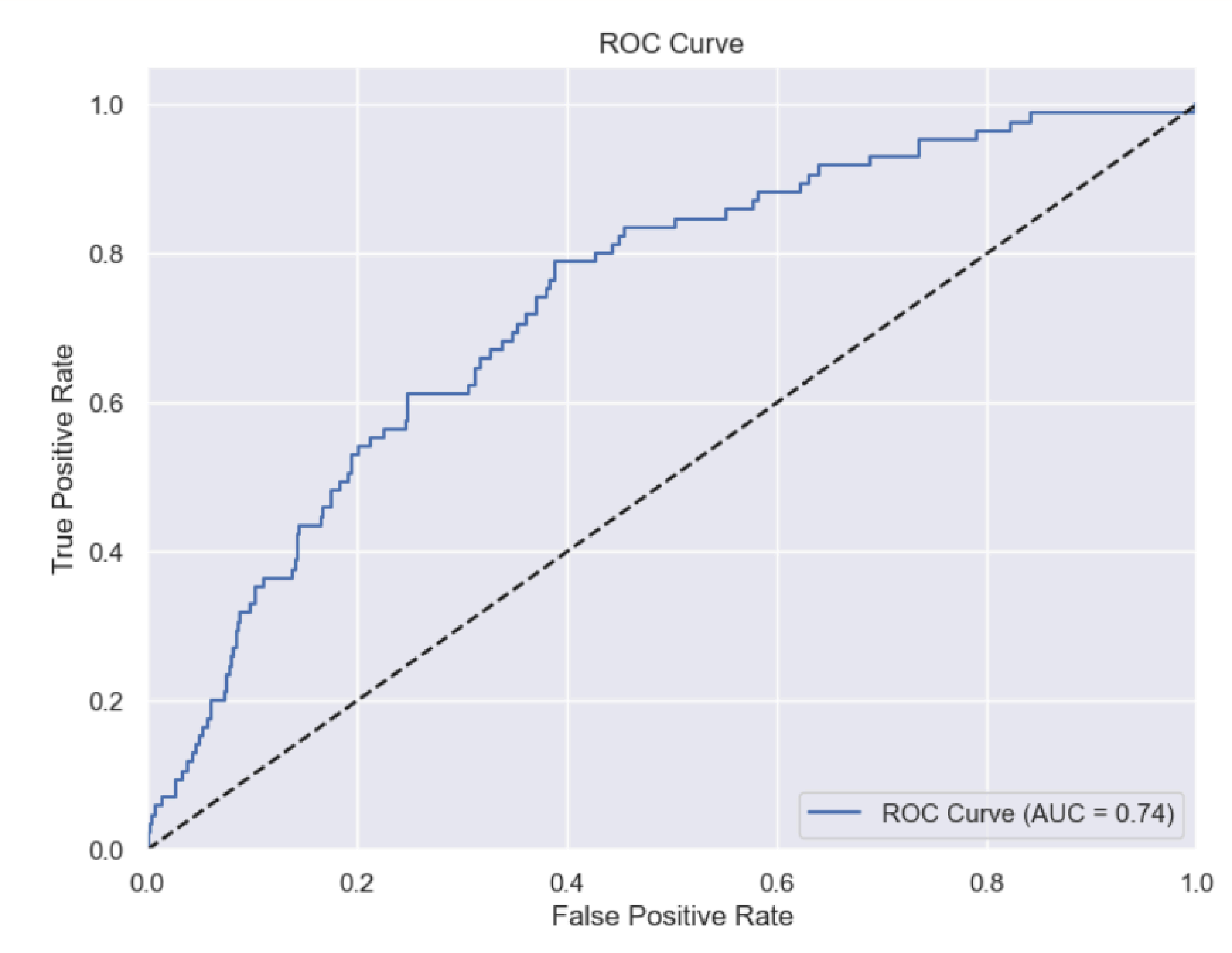
MACHINE LEARNING TECHNIQUES



LOGISTIC REGRESSION

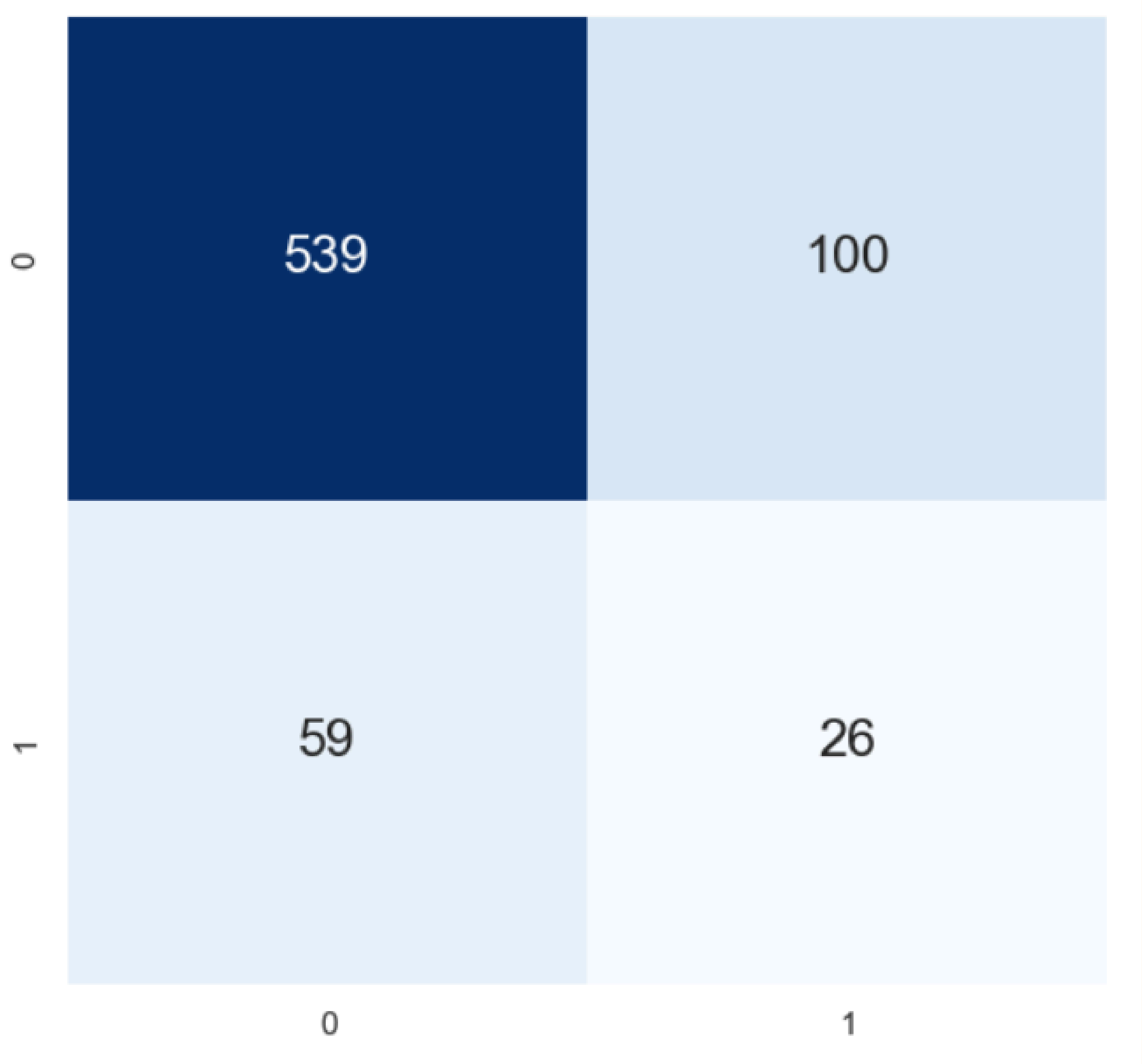
Accuracy: 88.39%
True Positive Rate: 2.35%
True Negative Rate: 99.8%

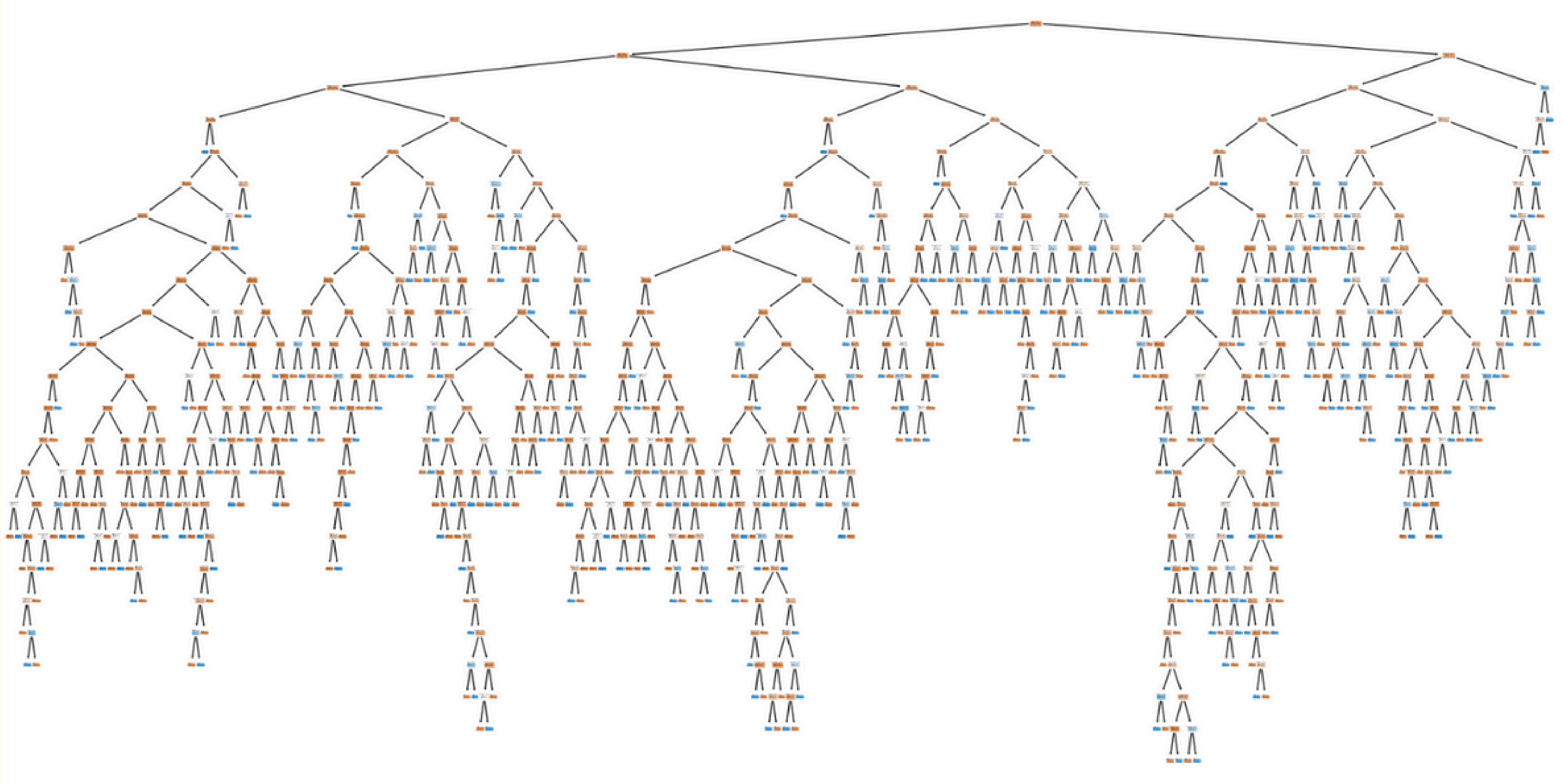




DECISION TREE

Accuracy: 78%
True Positive Rate: 30.5%
True Negative Rate: 84.3%





RANDOM FOREST

- Builds and merges multiple decision trees (more accurate and stable)
- Via bootstrap sampling, creating diverse trees less likely to overfit data

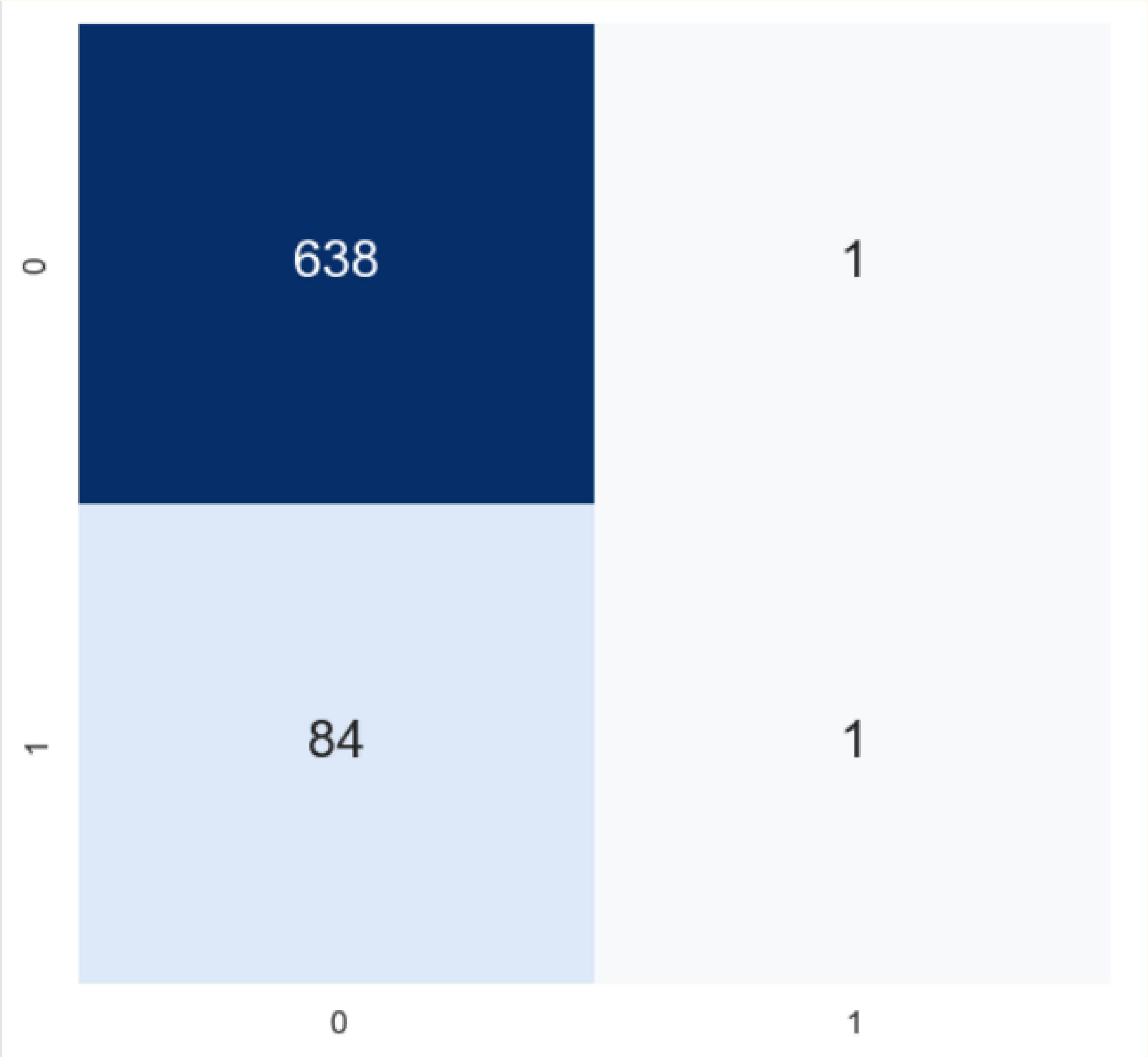
CLASSIFICATION:

- Each tree predicts class label of a new data point
- Class receives most "votes" chosen as final prediction

REGRESSION:

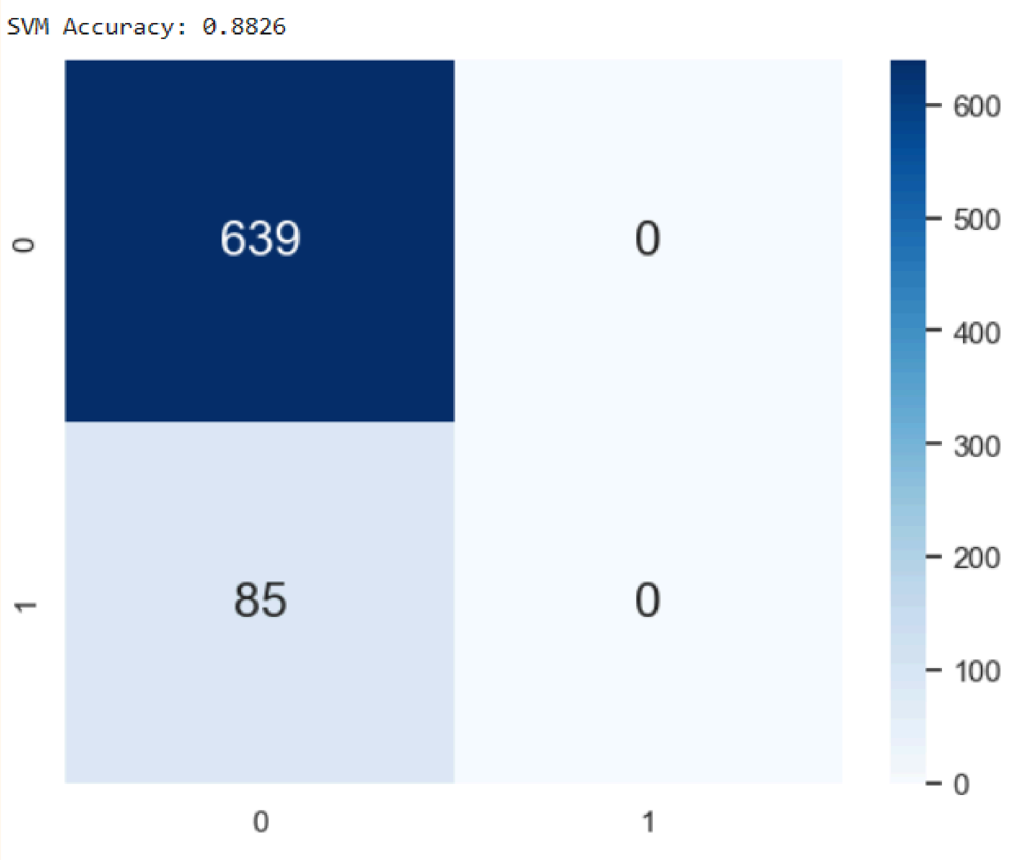
- Average prediction of all trees taken as the final prediction

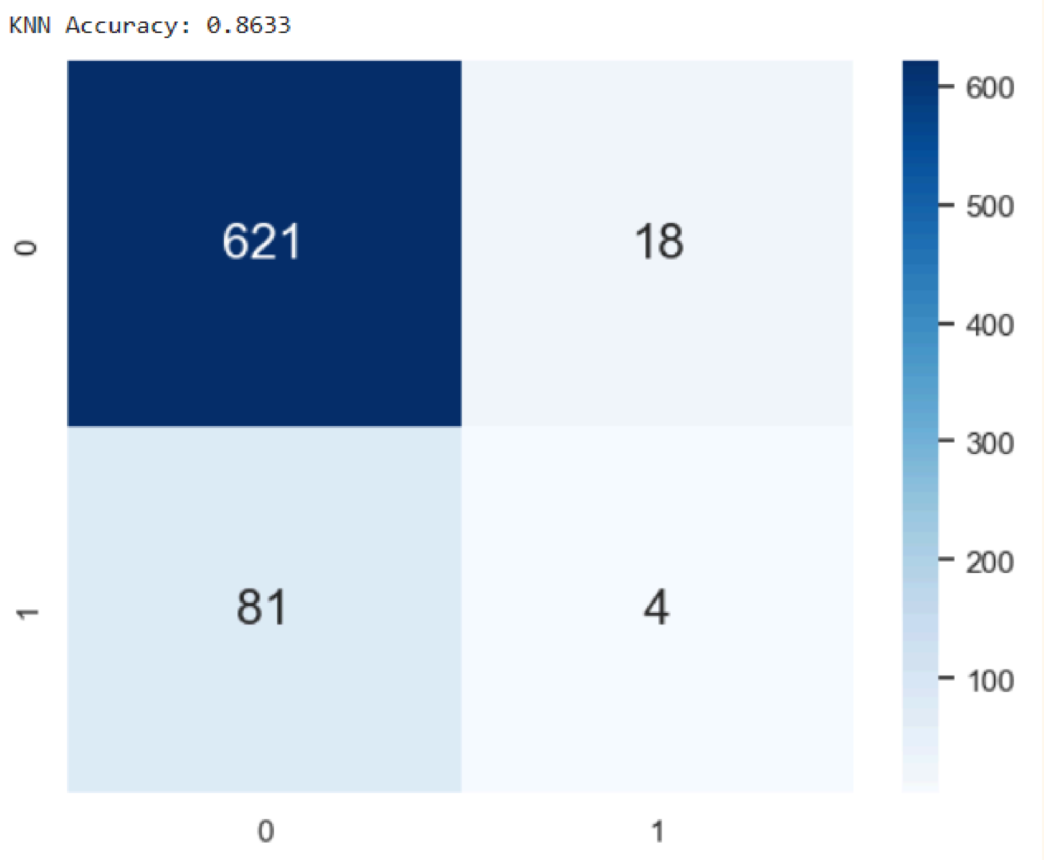
RANDOM FOREST

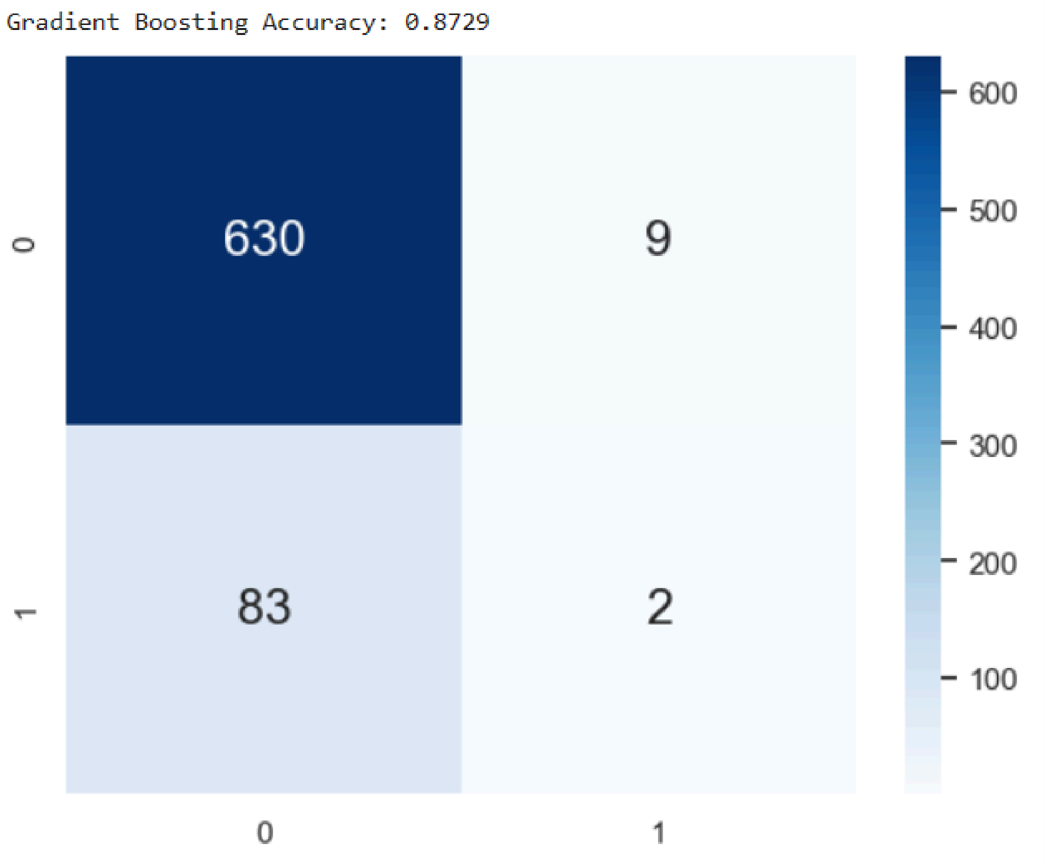


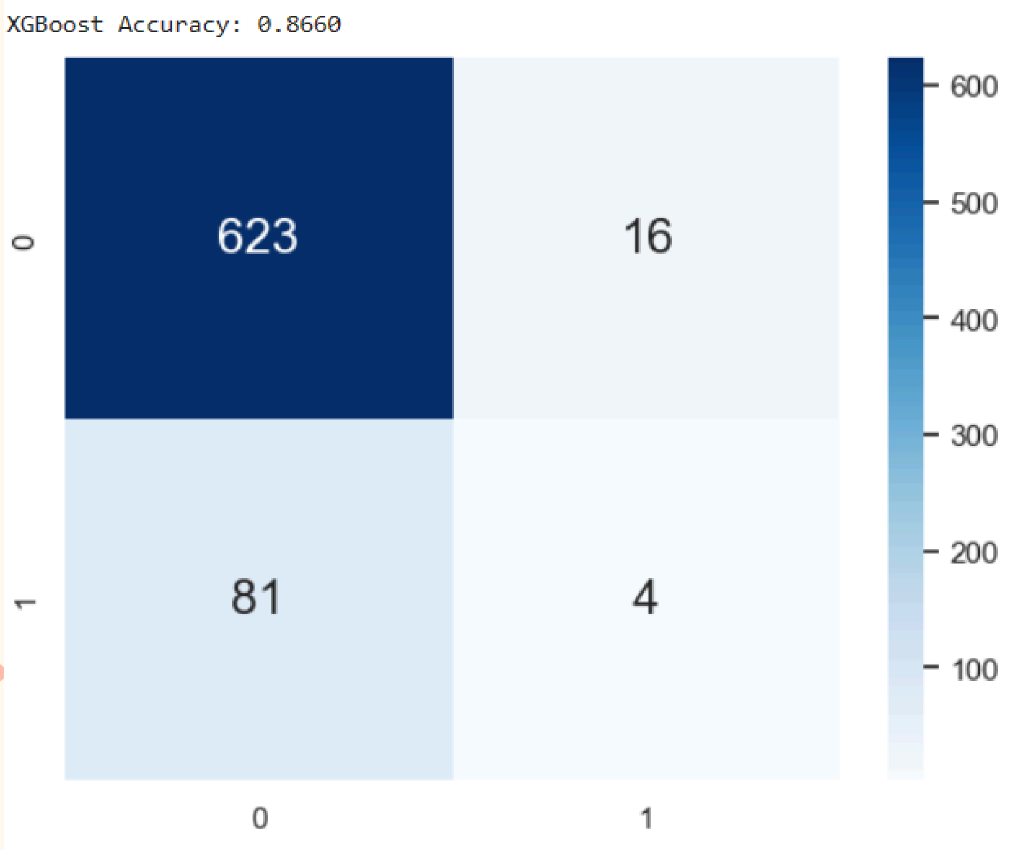
Accuracy: 88%
True Positive Rate: 1%
True Negative Rate: 99.8%

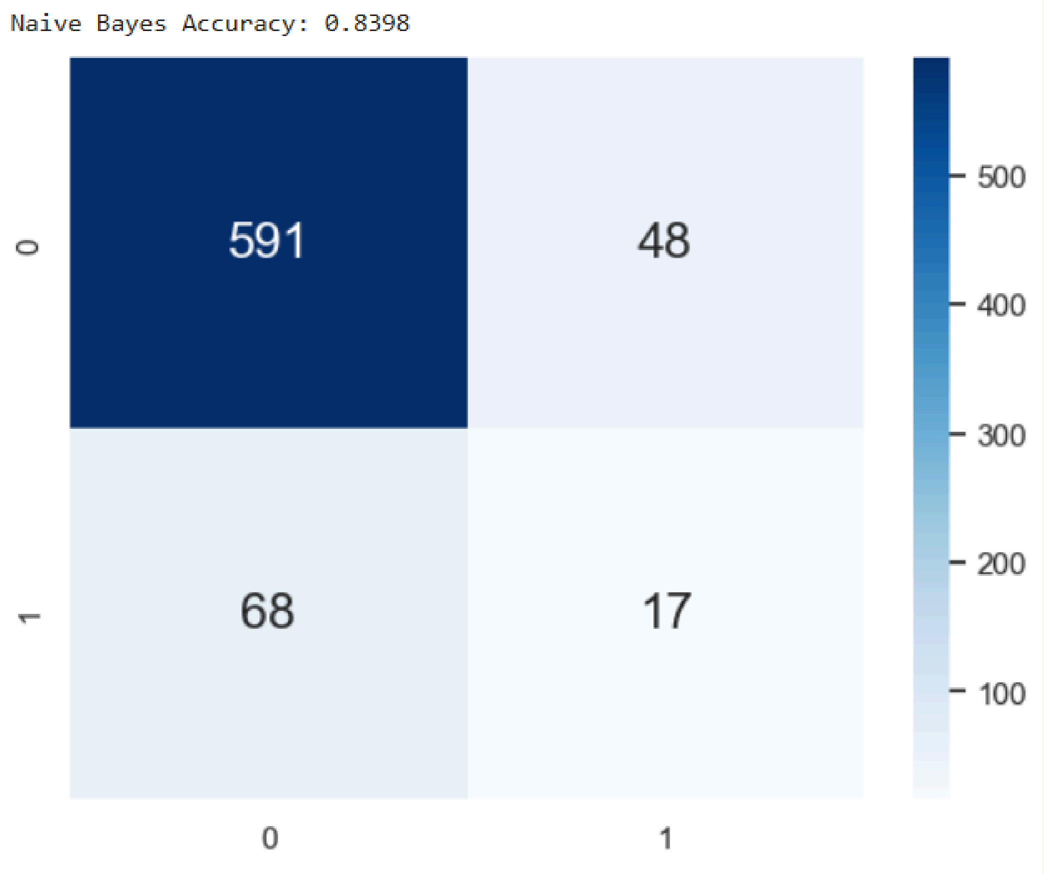
OTHER MODELS

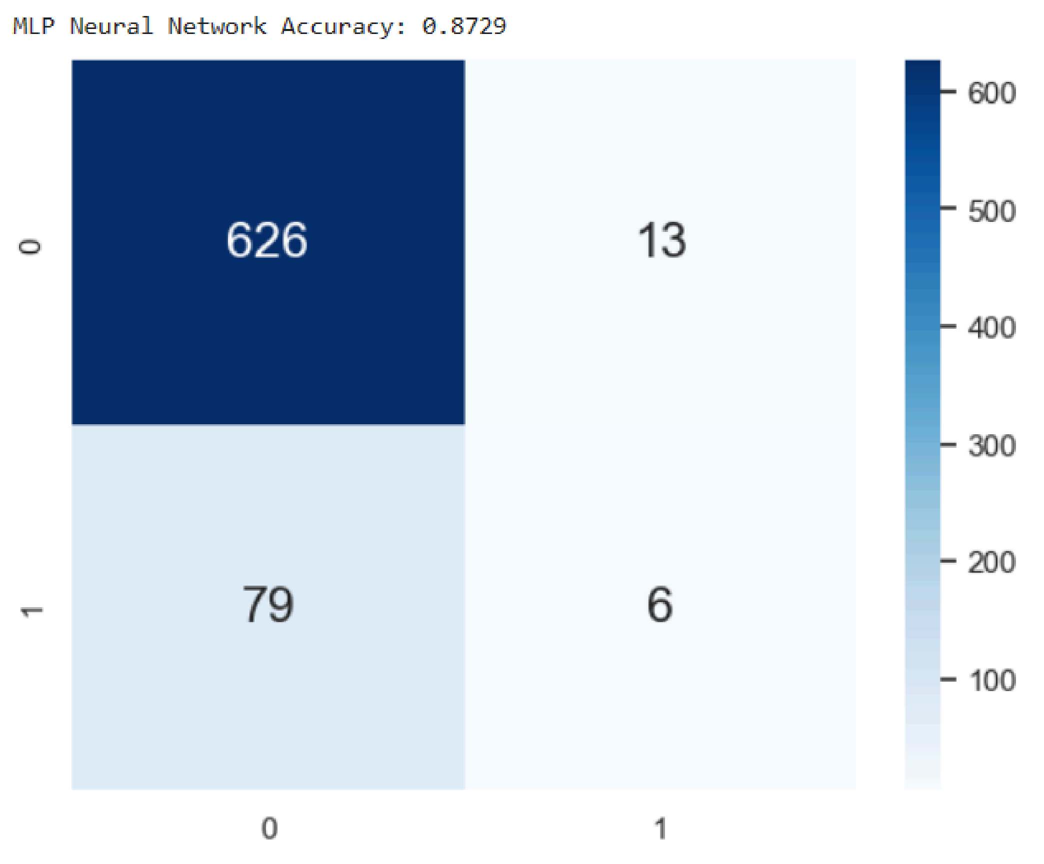








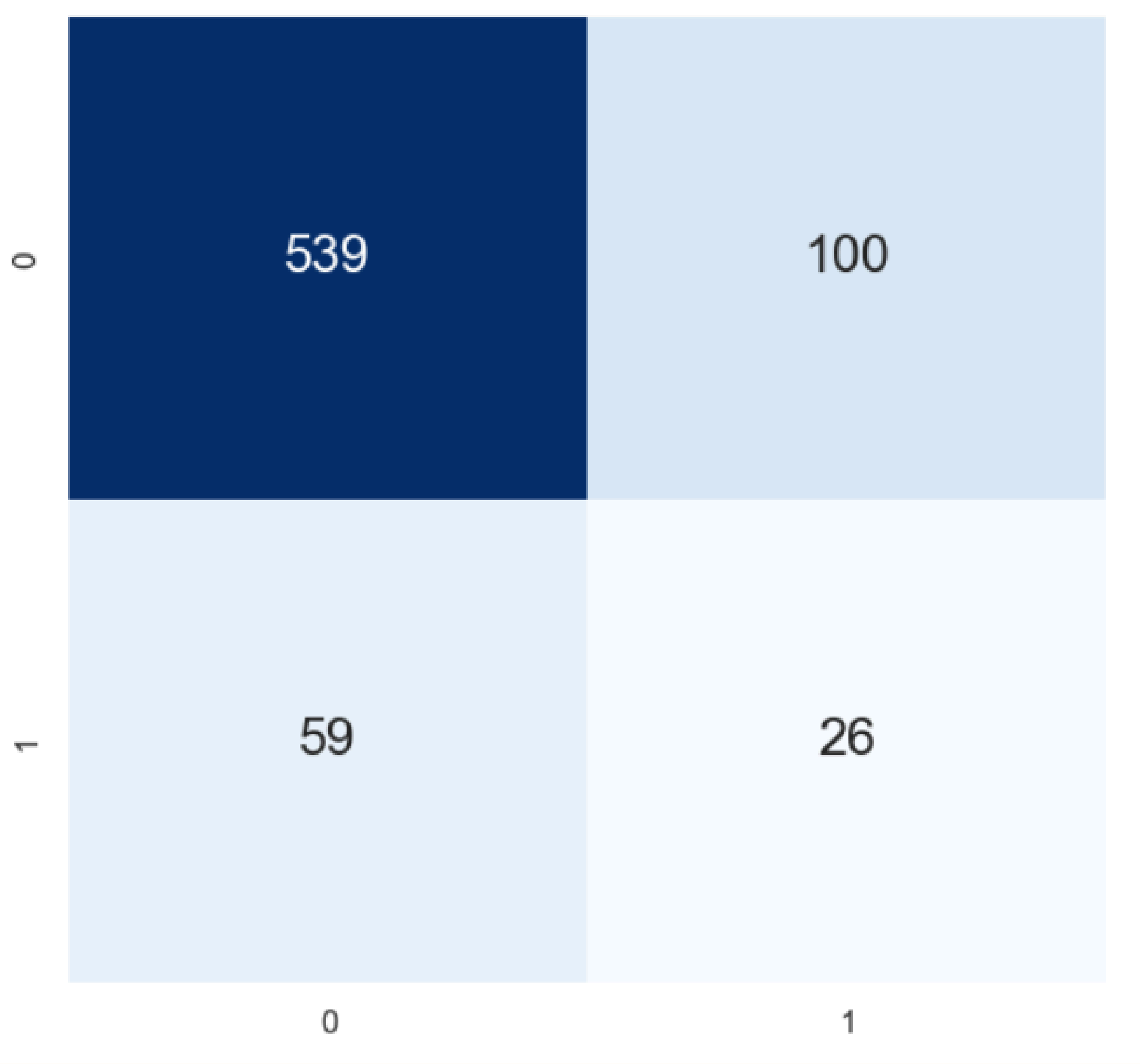




BEST MODEL - DECISION TREE

Accuracy: 78%
True Positive Rate: 30.5%
True Negative Rate: 84.3%

High true positive rate and low false negative is crucial



WHAT ELSE CAN WE TRY?

CONVOLUTIONAL NEURAL NETWORK

Cardiomegaly Disease Prediction Using CNN

Cardiomegaly Disease

kaggle.com

4438 train images
1114 test images

CONVOLUTIONAL NEURAL NETWORK

```python
model = Sequential()
model.add(Conv2D(32, (3, 3), activation='relu', input_shape=(img_width, img_height, 3)))
model.add(MaxPooling2D((2, 2)))
model.add(Conv2D(64, (3, 3), activation='relu'))
model.add(MaxPooling2D((2, 2)))
model.add(Conv2D(128, (3, 3), activation='relu'))
model.add(MaxPooling2D((2, 2)))
model.add(Conv2D(128, (3, 3), activation='relu'))
model.add(MaxPooling2D((2, 2)))
model.add(Flatten())
model.add(Dense(512, activation='relu'))
model.add(Dense(1, activation='sigmoid'))

# Compile the model
model.compile(loss='binary_crossentropy',
              optimizer='adam',
              metrics=['accuracy'])
```

```
1/1 [==============================] - 0s 17ms/step
Chances of having cardiomegaly: 78.73 %
```

```
1/1 [==============================] - 0s 16ms/step
Chances of having cardiomegaly: 15.09 %
```

Validation Accuray - 73.81%

DATA-DRIVEN INSIGHTS

RISK FACTORS RANKING

Based on the decision tree, the risk factors are ranked as such

1. Cholesterol
2. BMI
3. Blood Pressure
4. Glucose Level
5. Age
6. Heart Rate
7. Cigarettes smoked per day
8. Gender

PREVENTABLE RISK FACTORS

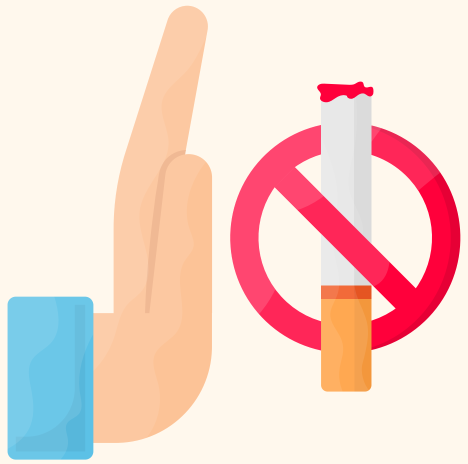
- Hypertension/Blood Pressure
- Heart rate
- Glucose/Diabetes
- Cholesterol Level
- BMI
- Cigarretes Smoked

NON-PREVENTABLE RISK FACTORS

- Age
- Gender

CONCLUSION + FUTURE POSSIBILITES

MAINTAIN A HEALTHY DIET,
EXERCISE REGULARLY,
MAINTAIN A HEALTHY WEIGHT,
LIMIT ALCOHOL COMSUMPTION
QUIT SMOKING



REGULAR SCREENING FOR EARLY DETECTION IS CRUCIAL, PARTICULARLY FOR INDIVIDUALS WITH MULTIPLE RISK FACTORS AND OLDER PEOPLE

TO OBTAIN BETTER RESULTS WHEN PREDICTING THE RISK OF CORONARY HEART DISEASE, A COMBINATION OF CT SCANS AND DEMOGRAPHICS WILL HELP TREMENDOUSLY

Thank you!