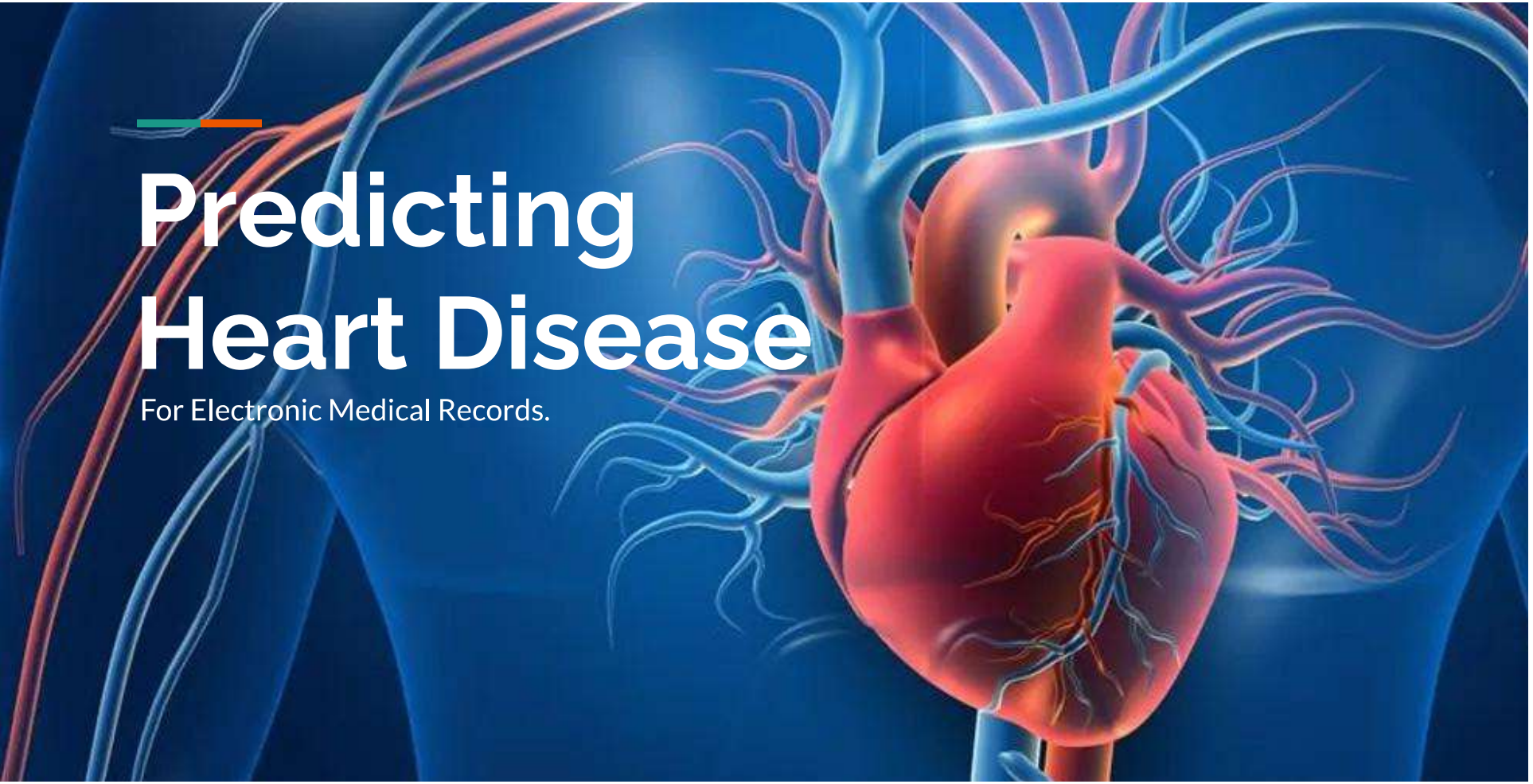


Predicting Heart Disease

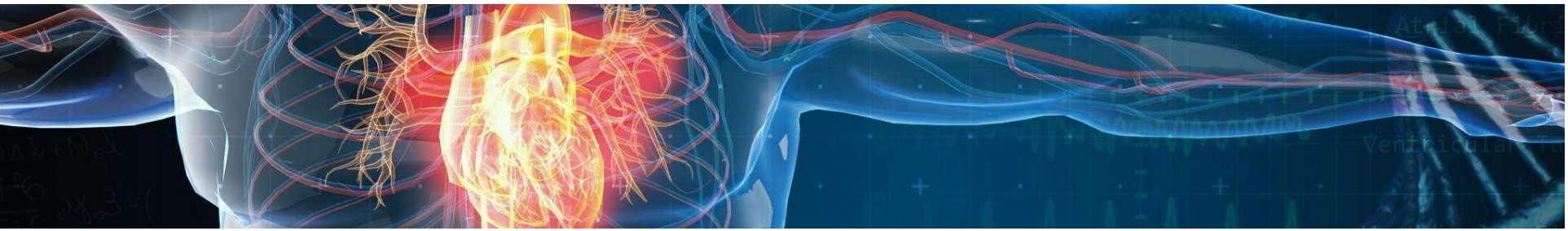
For Electronic Medical Records.

Significance

Several health conditions, your lifestyle, and your age and family history can increase your risk for heart disease. These are called risk factors. About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Some risk factors for heart disease cannot be controlled, such as your age or family history. But you can take steps to lower your risk by changing the factors you can control

Detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare. Computational developments, in turn, allow the application of machine learning methods to detect "patterns" from the data that can predict a patient's condition.

The purpose of this model is to predict a patient’s Heart Disease Risk



Questions to answer

1. Does Sex/Gender have an association with **Heart Disease**?
2. Is there an association between age & **Heart Disease**?
3. Are BMI, Smoking, Alcohol drinking, and prior stroke associated with **Heart Disease**?

Technologies Used

Technologies Used

Data Cleaning & Analysis:
The Pandas library in Python was used to clean the data and perform an exploratory analysis . Machine Learning was completed using the SKLearn and IMBLearn libraries.

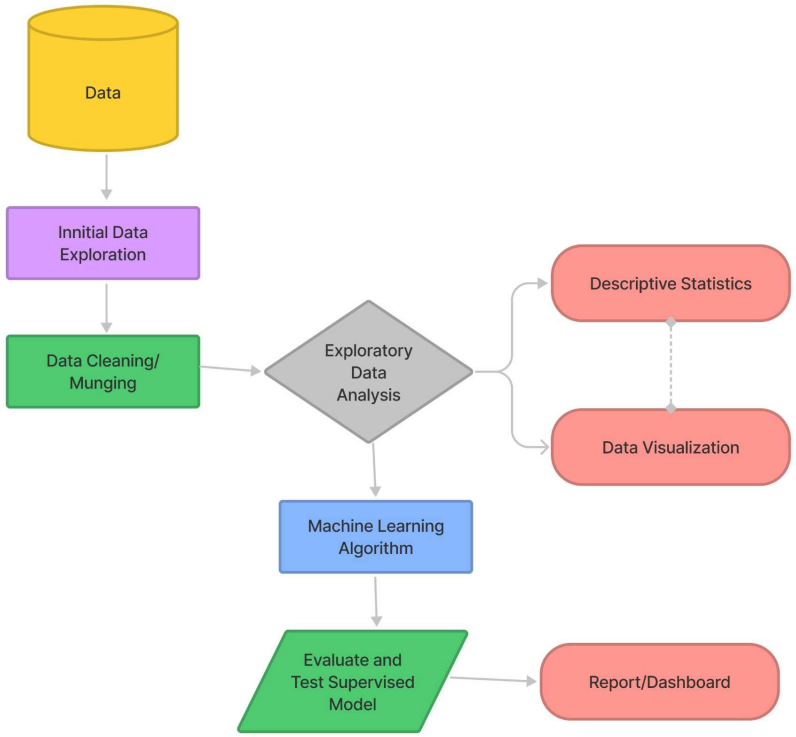
Database Storage:
PostgreSQL and PGAdmin were used to create the Database, and was populated from Jupyter Notebook using SQLite.

Dashboard:
Tableau was used to create our Interactive Dashboard visualizing our exploratory analysis.

Database Diagram

Machine Learning Model

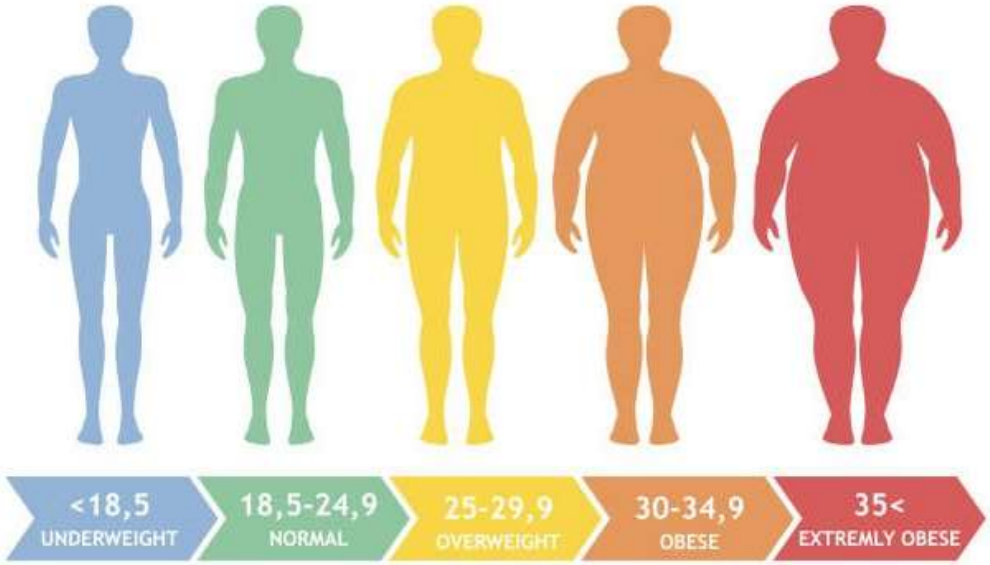
Data flow and processing



Data Exploration Process

Body Mass Index (BMI)

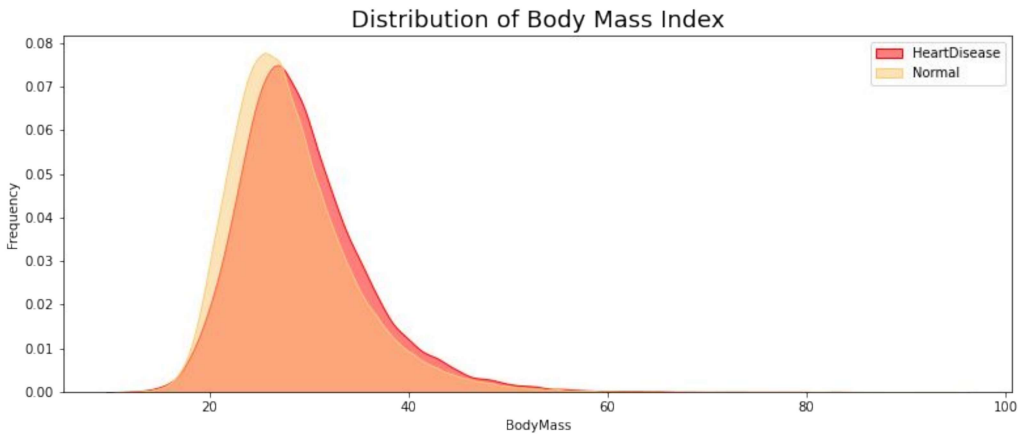
We validated the average **BMI** provided by the database with a external source (Statista as per the graphs below) and the average **BMI** of our database is validated by the information of the external source.



Distribution of BMI

The **BMI** is defined as the body mass divided by the square of the body height, and is expressed in units of kg/m², resulting from mass in kilograms and height in meters. Higher **BMI** has a stronger association with **Heart Disease**.

We can confirm that that the population that suffer from **Heart Disease** have a higher **BMI** on average, when compared to those that do not suffer from **Heart Disease.** This means that there is a positive correlation between the high **BMI** and the risk of **Heart Disease**.

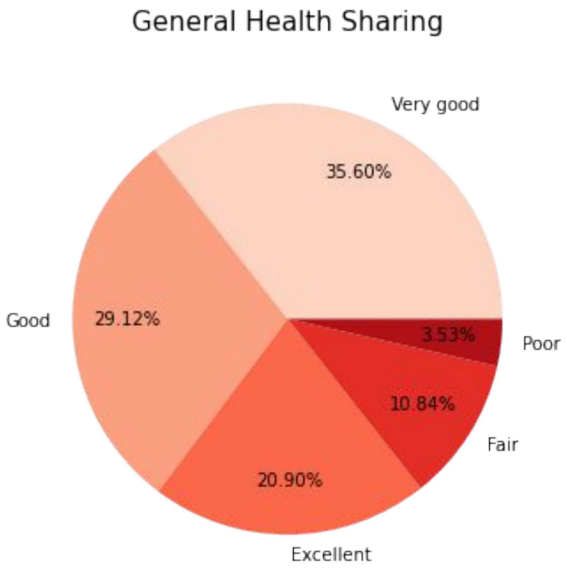


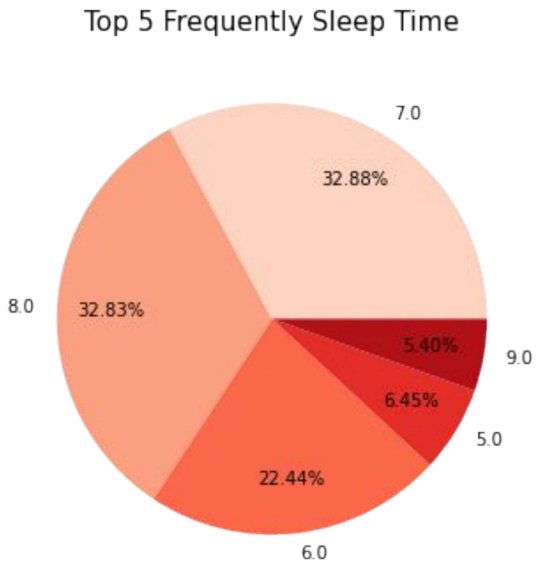
Key-Categories Percentage

These Pie Graphs show us the percentage composition for two main categories: General Health Sharing and Top 5 Frequently Sleep Time.

64.7% of the Sample population report **Good** or **Better** Sleep.

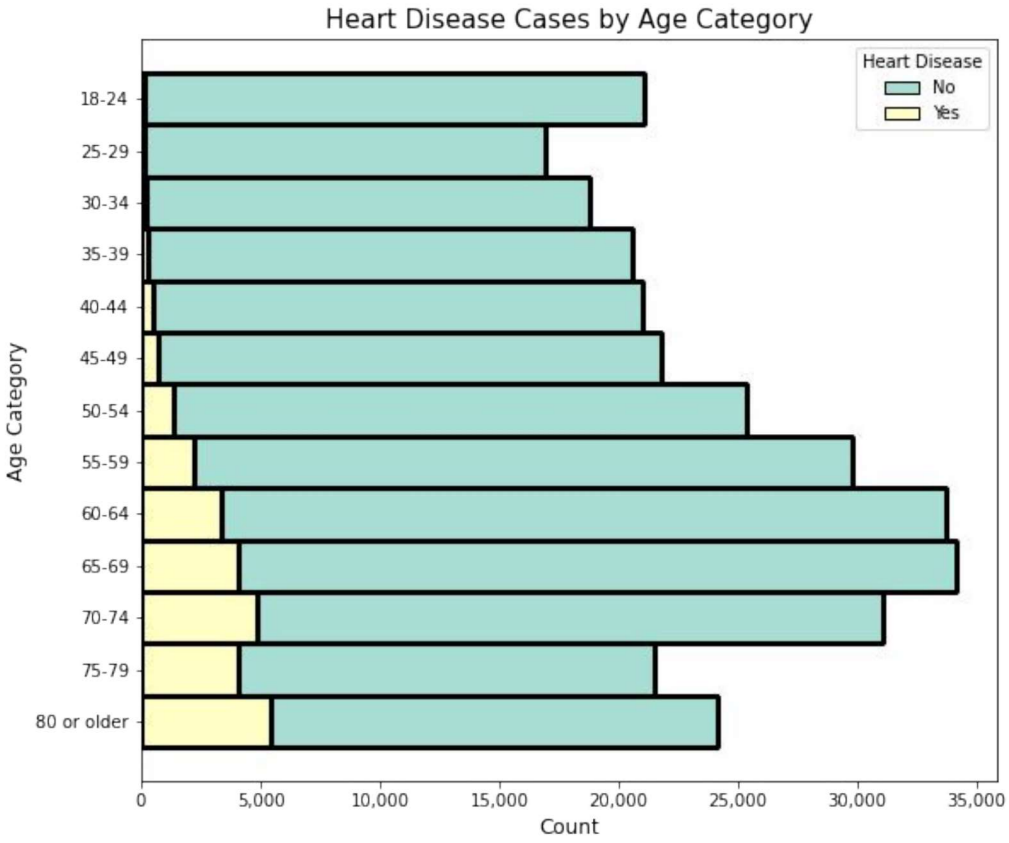
Out of those who report sleeping more than 5 hours, 38.2% sleep 8 hours or more.





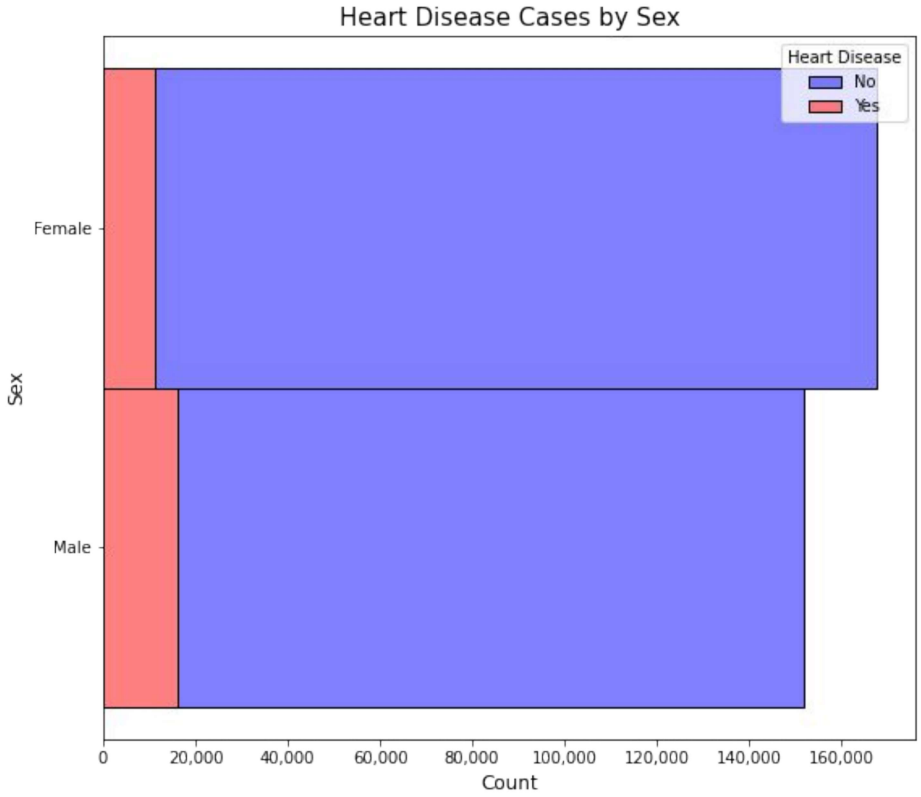
Heart Disease by Age Category

From this graph we can conclude that **Heart Disease** is more prevalent on those 55 years and older.

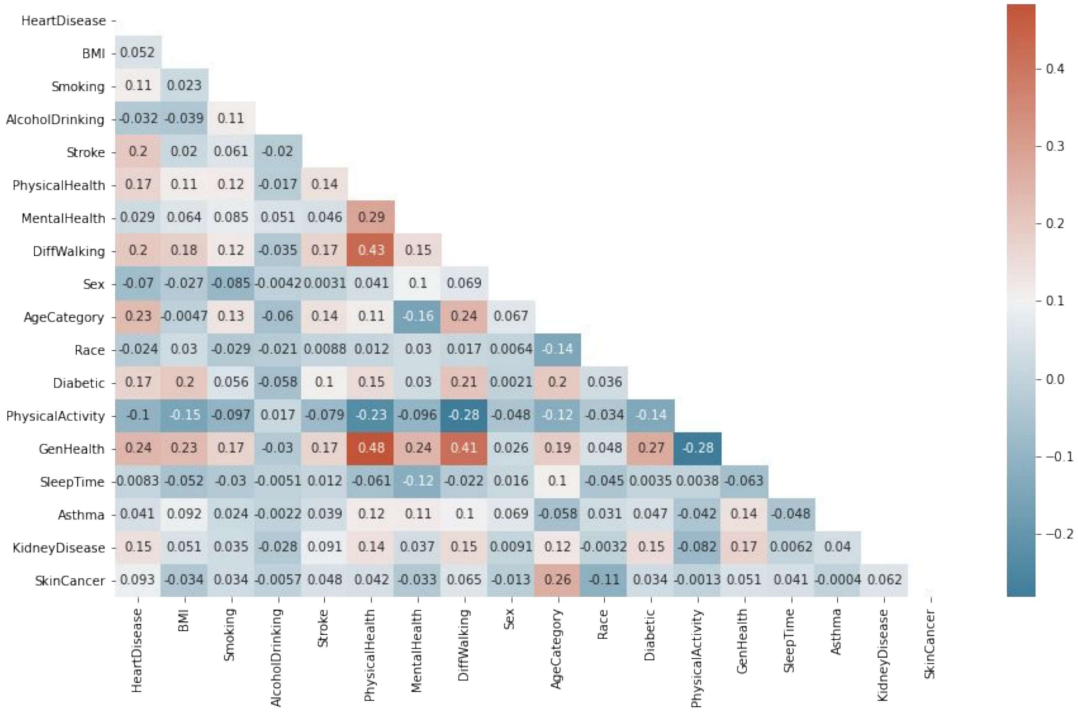


Heart Disease Cases by Gender

In this graph we can confirm that Males have a higher propensity of developing a **Heart Disease** compared to Females.



Correlation Heatmap



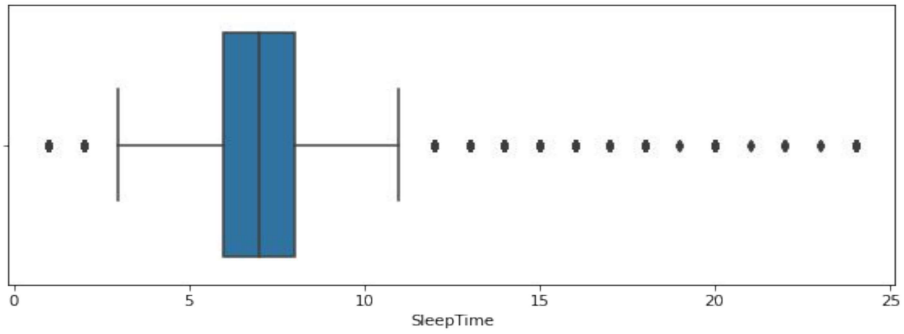
The Heat Map show us the correlation between two variables. If the value is 1, it is said to be a positive correlation between the two variables. Regarding the correlation factor against the **Heart Disease** Variable, the variables with higher correlation are the ones listed below:

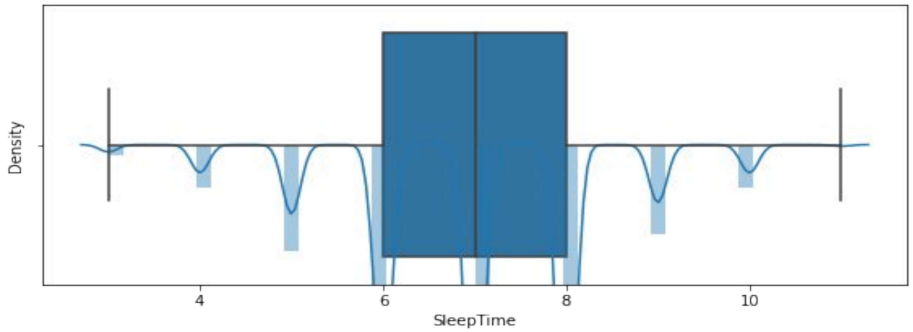
- GenHealth
- Age
- Stroke
- DiffWaking

Sleeptime Variable

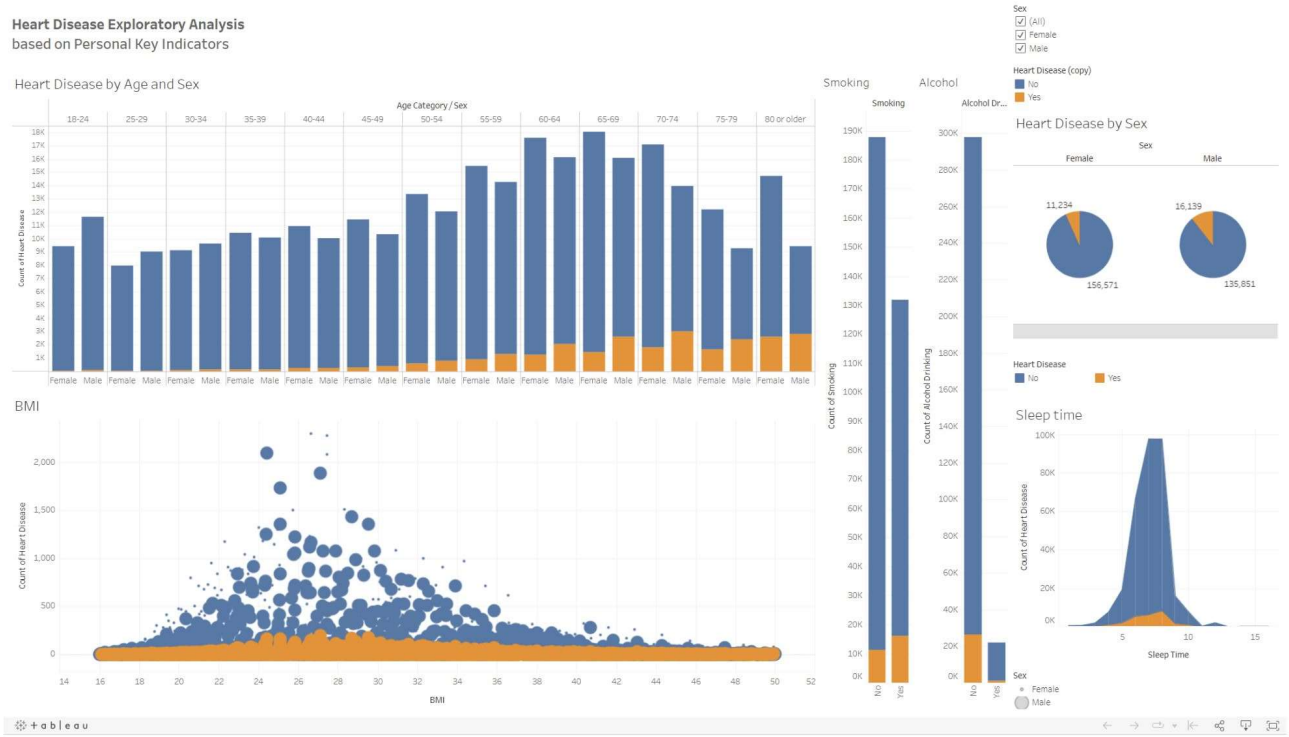
While analyzing variables, we discovered that on the **SleepTime** variable had outliers caused by information of patients sleeping too long (i.e. 24hrs).

When we removed the 4543 outliers, the distribution of **SleepTime** was more central from 6 to 8 hours.





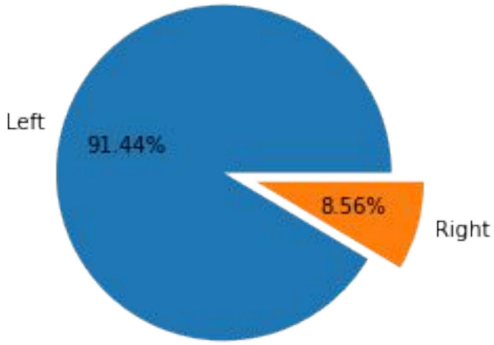
Interactive Dashboard

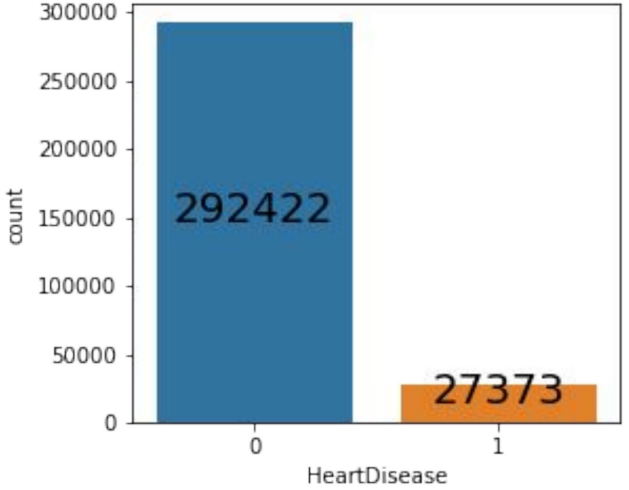


Machine Learning Model

Sample Imbalance

After importing dependencies, an unbalanced dataset with two classes is artificially created and plotted, as shown in the resulting charts.





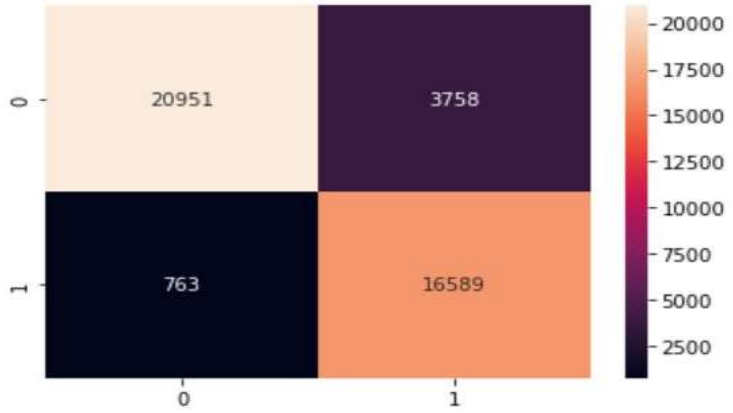
Testing Machine Learning Models

We decided to test three different Machine Learning Model Methods: Confusion Matrix, Logistic Regression and Random Forest Classifier.

The best results were achieved with the Random Forest Classifier method. Achieving the following efficiency values:

- **Accuracy** 0.9211621
- **Precision** 0.863853
- **Recall** 0.9602351

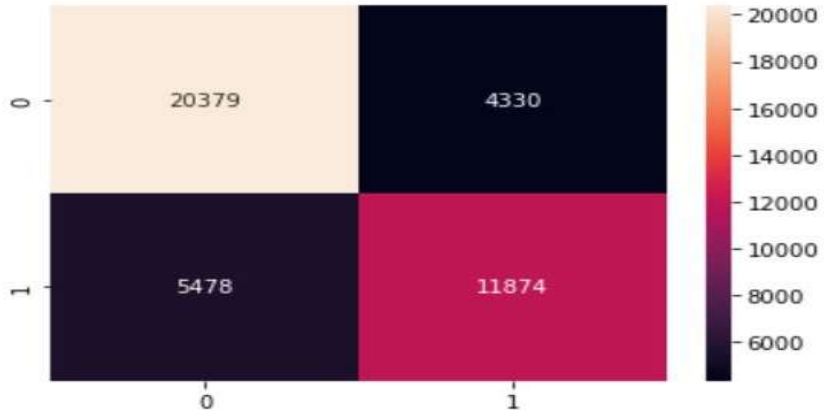
Confusion Matrix




```
]: print('accuracy',accuracy_score((y_test), y_predict_dt))
   print('precision',precision_score(y_test, y_predict_dt))
   print('recall',recall_score(y_test, y_predict_dt))
```

```
accuracy 0.892513254558855
precision 0.8153044674890647
recall 0.9560281235592439
```

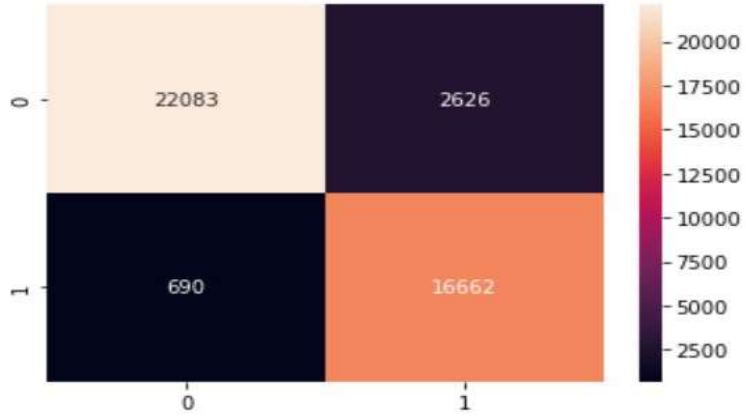
Logistic Regression




```
print('accuracy',accuracy_score((y_test), y_predict_lr))
print('precision',precision_score(y_test, y_predict_lr))
print('recall',recall_score(y_test, y_predict_lr))
```

```
accuracy 0.7668148641259124
precision 0.7327820291286102
recall 0.6843015214384509
```

Random Forest Classifier



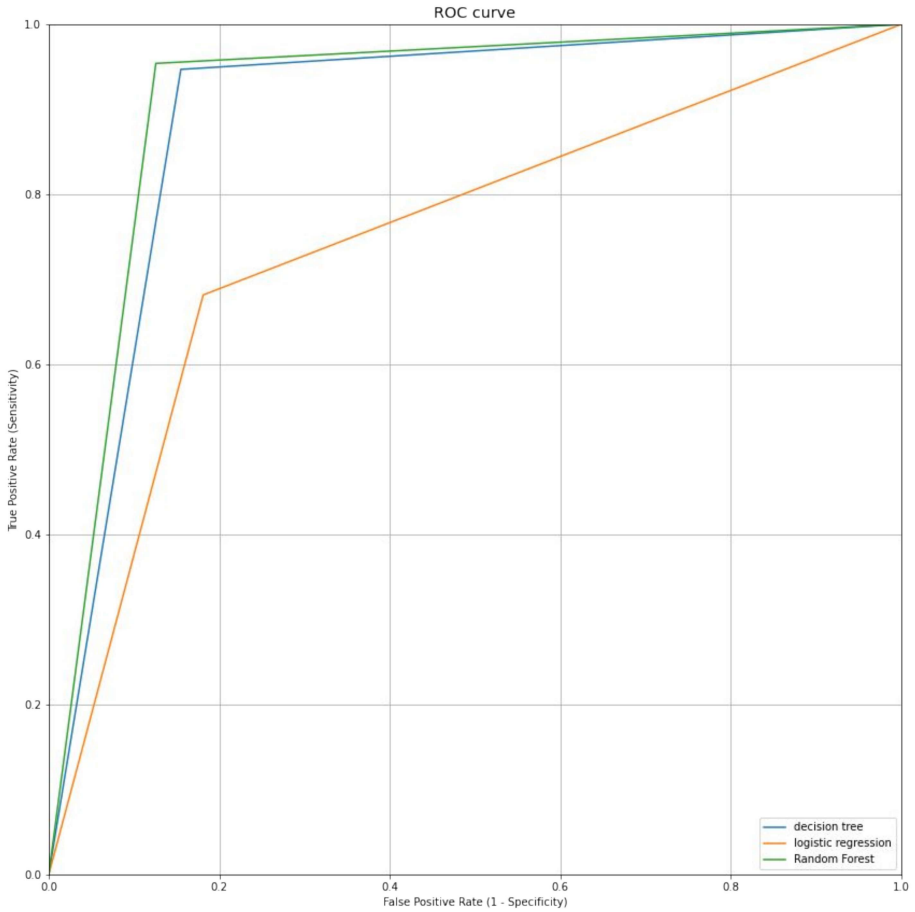

```
]: print('accuracy',accuracy_score((y_test), y_predict_rf))
   print('precision',precision_score(y_test, y_predict_rf))
   print('recall',recall_score(y_test, y_predict_rf))
```

```
accuracy 0.9211621216804166
precision 0.8638531729572791
recall 0.9602351313969572
```

ROC
Chart Models

Performance comparison to determine which model generated the best results.

In this case the better-performing model is Random Forest.



Conclusion

1. We performed a chi-square analysis to test if there is an association between gender and heart disease. We hypothesized that the distribution of heart disease is equal amongst the genders.The p-value (p=0.0) of our chi-square analysis is below the 0.05 significance level, this we reject our null hypothesis. **There is a difference in the distribution of heart disease amongst the genders.**
2. We performed a chi-square analysis to test if there is an association between age category and heart disease. We hypothesized that the distribution of heart disease is equal amongst the age groups. The p-value (p=0.0) of our chi-square analysis is below the 0.05 significance level, this we reject our null hypothesis. **There is a difference in the distribution of heart disease amongst the different age groups.**


Thank you.