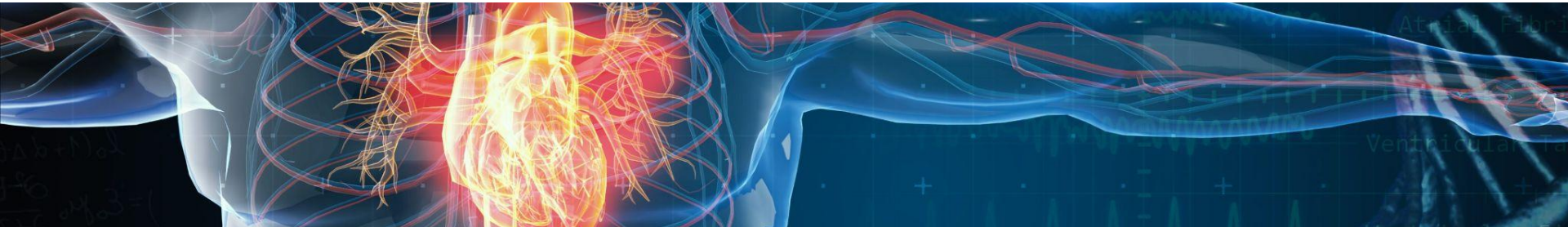
Predicting Heart Disease

Based on Physical Key Indicators.

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.



Questions to answer

1. Who is at higher risk of suffering Heart Failure based on sex/gender?
2. What is the correlation between age & heart failure?
3. Are BMI, Smoking, Alcohol drinking, and prior stroke associated to Heart disease?

Technologies Used

Technologies Used

Data Cleaning & Analysis:
Pandas will be used to clean the data and perform an exploratory analysis . Further analysis will be completed using Python.

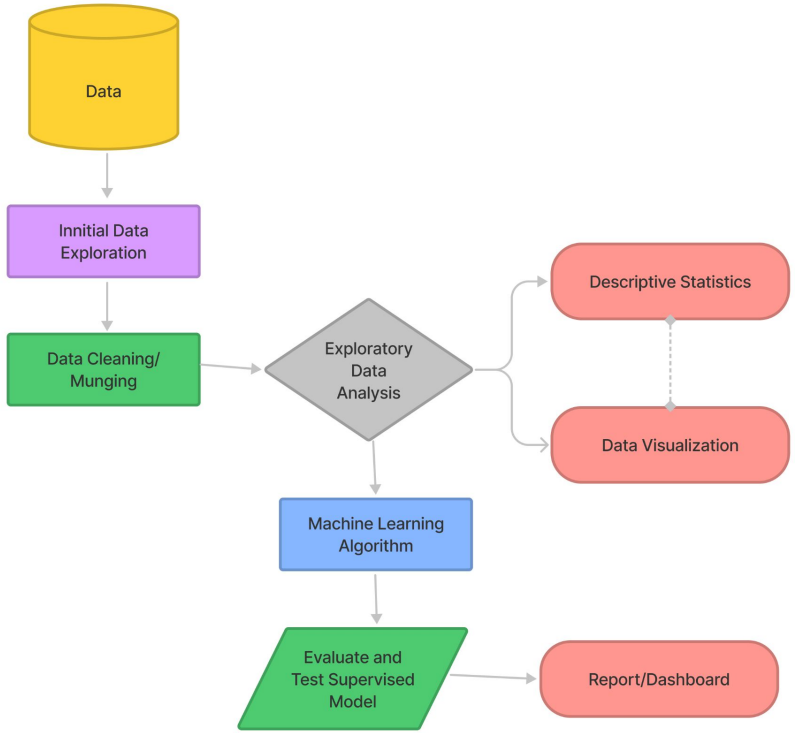
Database Storage:
Postgresql will be the intended database to use , and will be populated from jupyter notebook using SQLite.

Communication Protocols

- Team members will communicate using the #final-project Slack channel.
- Team members will meet at least 1x per week.
- Team members will plan next time they meet at the end of every meeting-- being flexible and open to each other's schedules.
- Team members will communicate with each other in the case they need internal deadline extensions or help with their part of the project.
- Team members will distribute work evenly amongst each other and be responsible for their distribution.
  Segment 1 Responsibilities: due date 4.6.2022
  All the members of the team are responsible of preparing the deliverable required by segment 1 and the majority of the work have been performed in group meetings but the final commit of the information have been splitted as follows:
  - Ivan: Performing the Github Repository for the project, creating the branches and setting up the Flow of the Machine Learning Model
  - Pavel: Uploading the jupyter notebook file and loading up the Database to be used until the end of the project
  - Gustavo: Preparing the final version of the Readme file to submit.
  - Jhonatan: Preparation of the slides presentation to be submitted.

Machine Learning Model

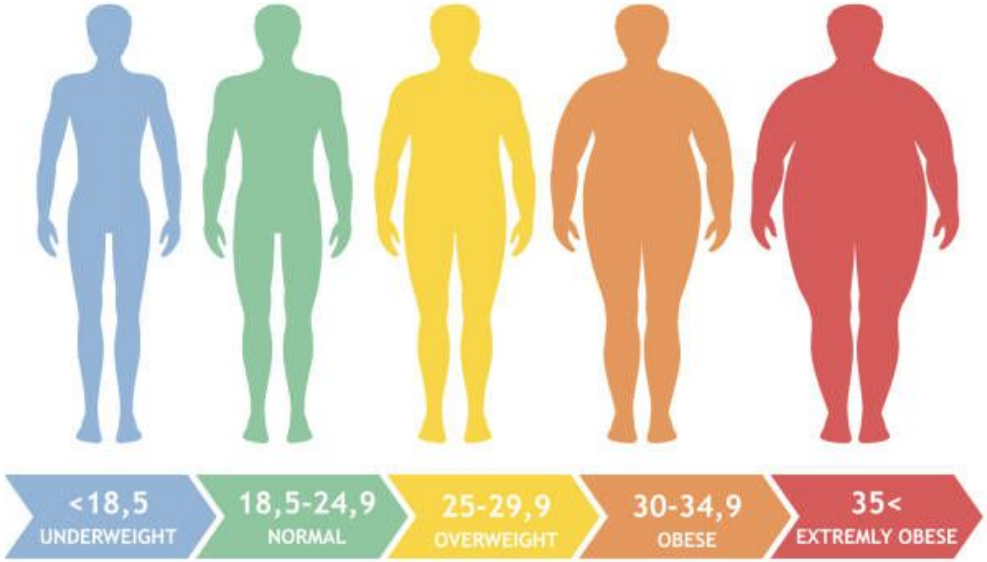
Data flow and processing




Data
Exploration
Process

Body Mass Index

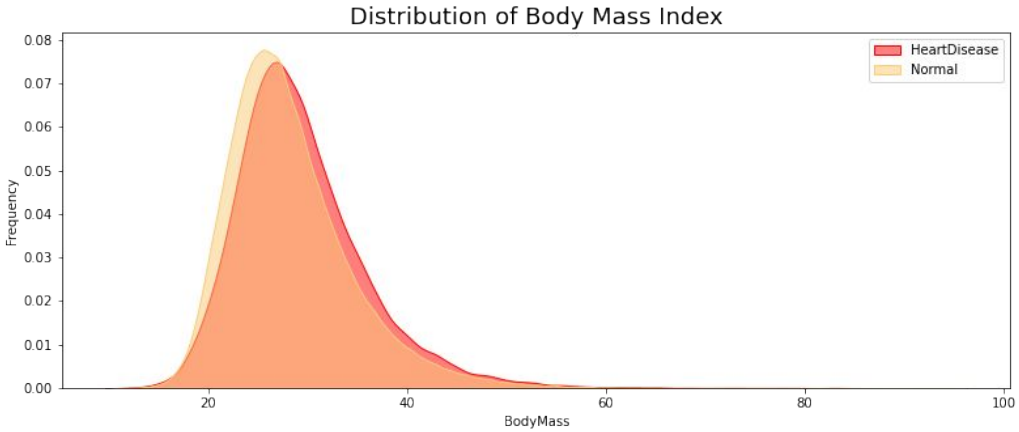
We have validated the average provided by the database with a external source (Stadista as per the graphs below) and the average BMI of our database is coherent with the information of the external source.



Distribution of BMI

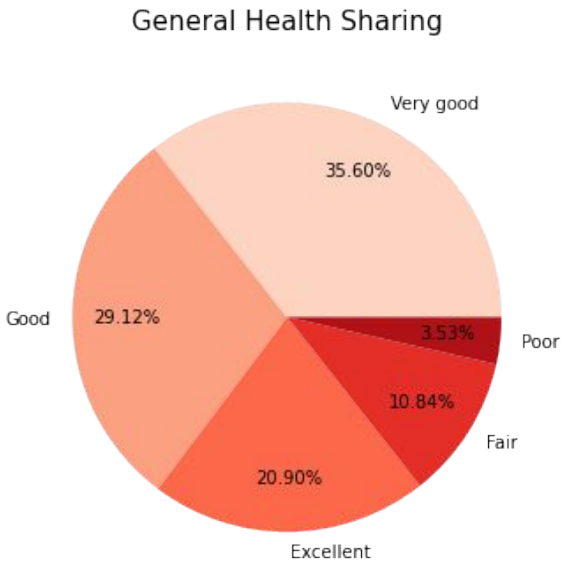
The BMI is defined as the body mass divided by the square of the body height, and is expressed in units of kg/m², resulting from mass in kilograms and height in metres. Higher BMI has a stronger association with incident heart failure.

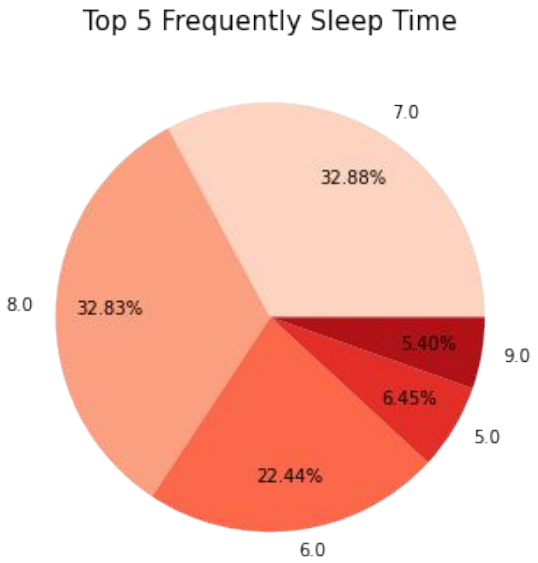
We can confirm that that the population that suffer a heartdisiease have on average a higher BMI, which means that there is a positive correlation between the high BMI and the risk of heart disiease.



Key-Categories Percentage

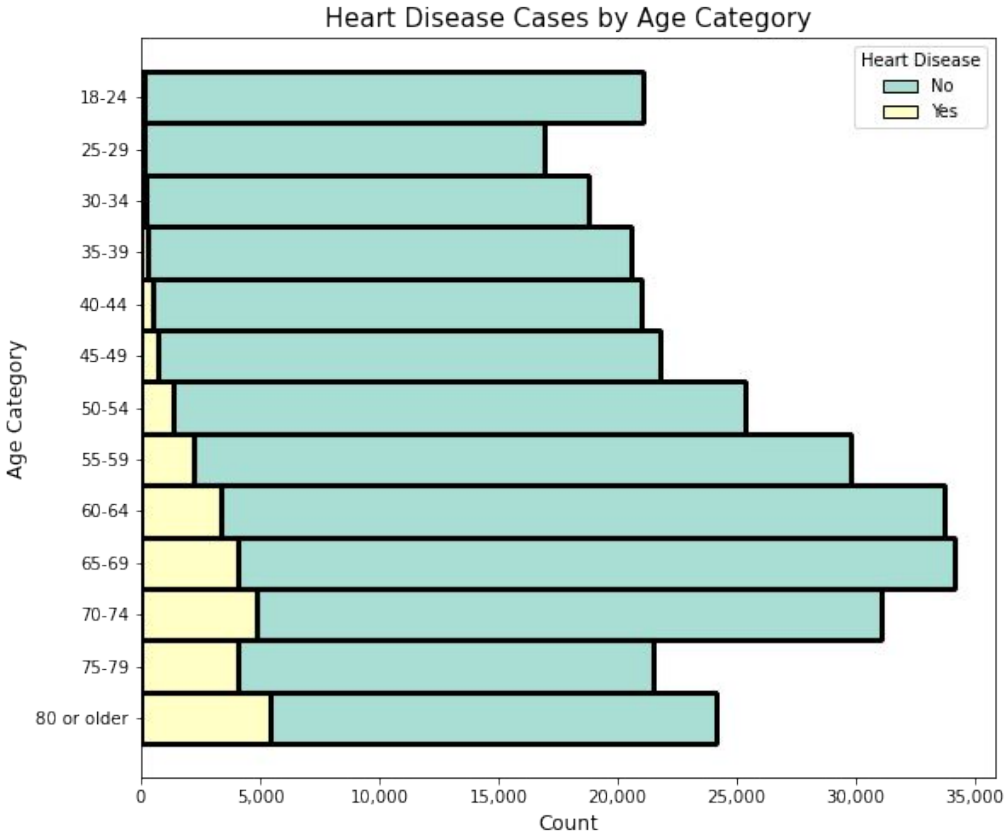
These Pie Graphs show us the percentage composition for two main categories: General Health Sharing and Frequently Sleep Time.





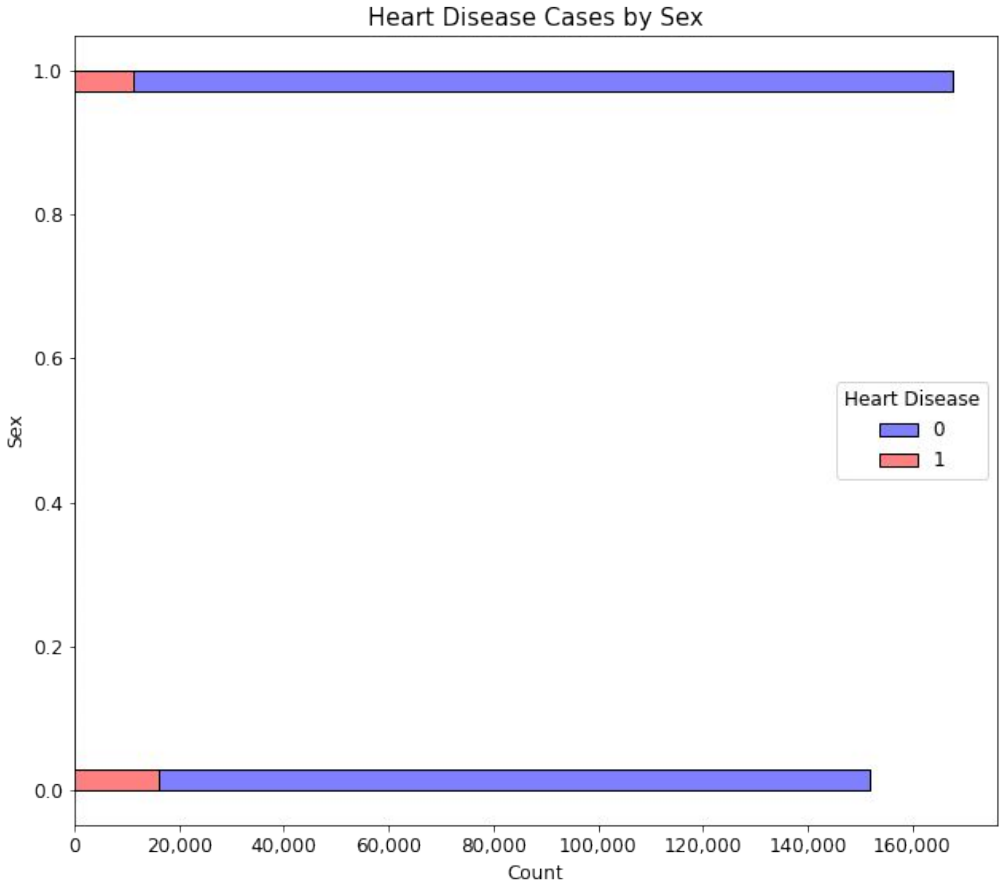
Heart Disease by Age Category

By using this Graph we can confirm what age group is heart disease most common. As we can visualize adults age 65 and older are more likely than younger people to suffer from cardiovascular disease.



Heart Disease Cases by Gender

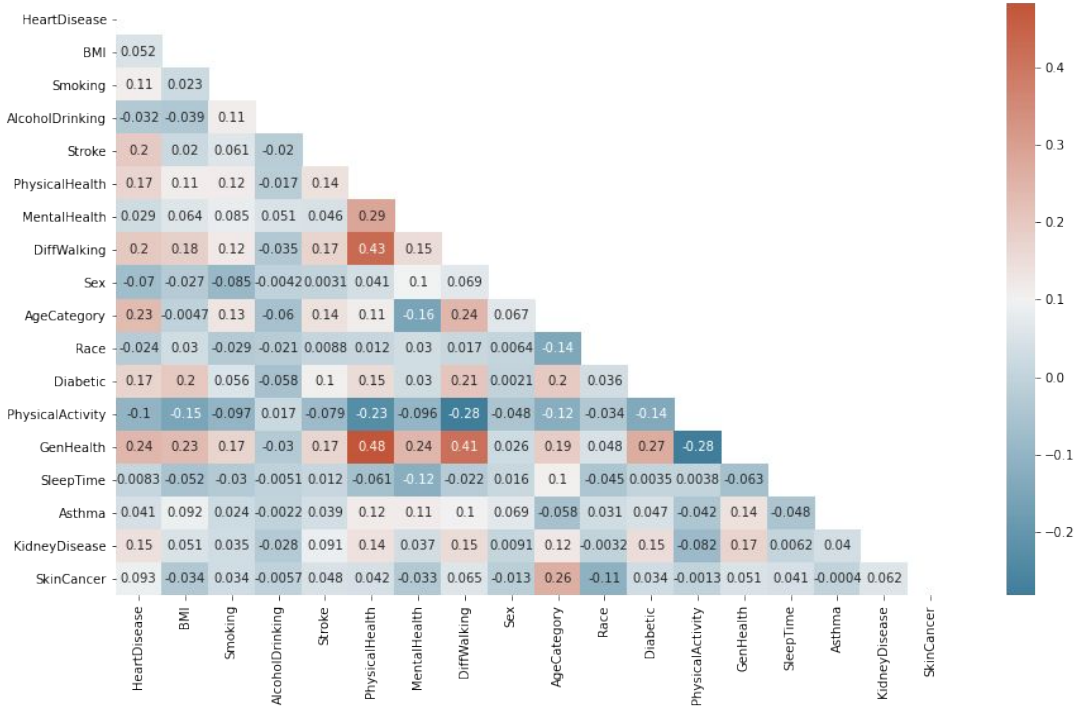
In this graph we can confirm that Males have a higher propensity of developing a Heart Disease compare with Females.



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 HeartDisease Variable, we notice as the variables with higher correlation the one listed below:

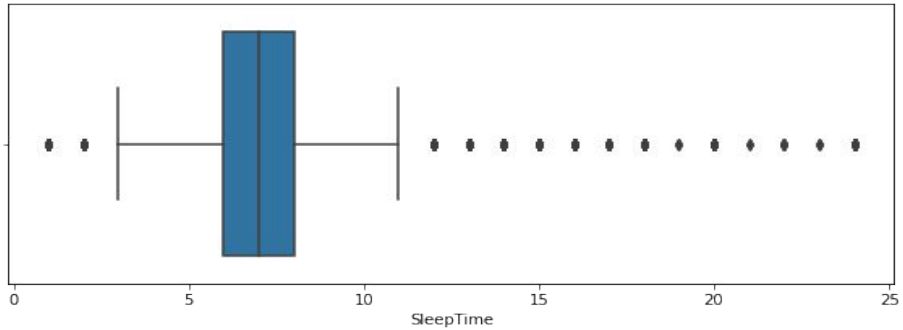
- GeneralHelath
- Age
- Stroke
- DiffWaking

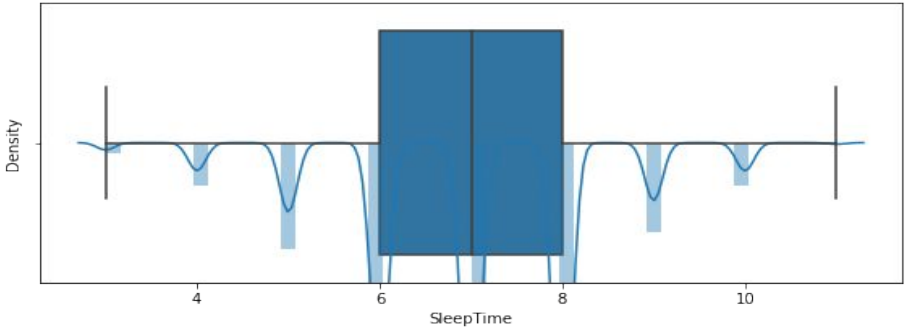


Sleeptime Variable

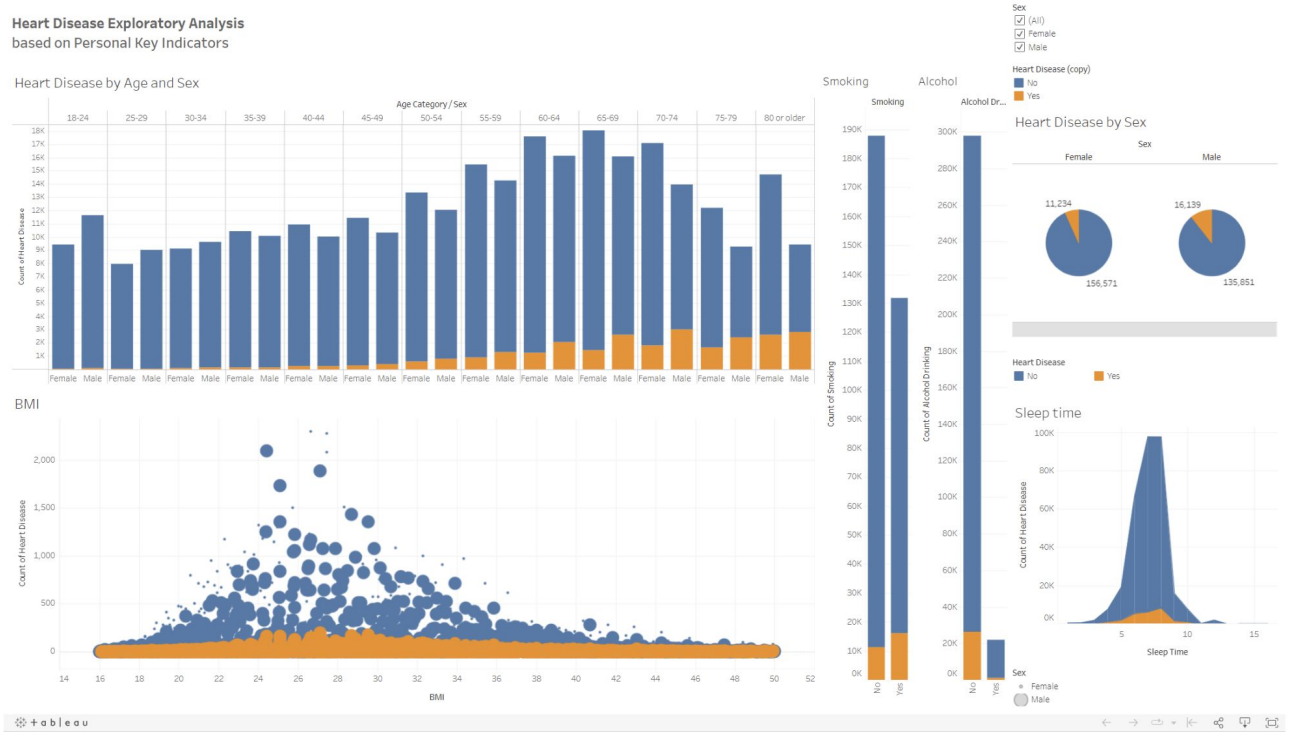
While analyzing variables, we discovered that on the Sleeptime column, there were outliers caused by information of patients sleeping too long.

When the column was cleaned there were only samples where patients sleep between 6 and 8 hours a day.



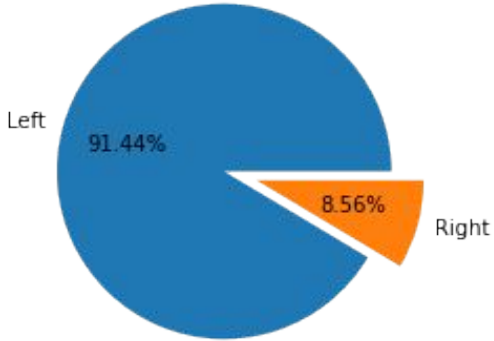


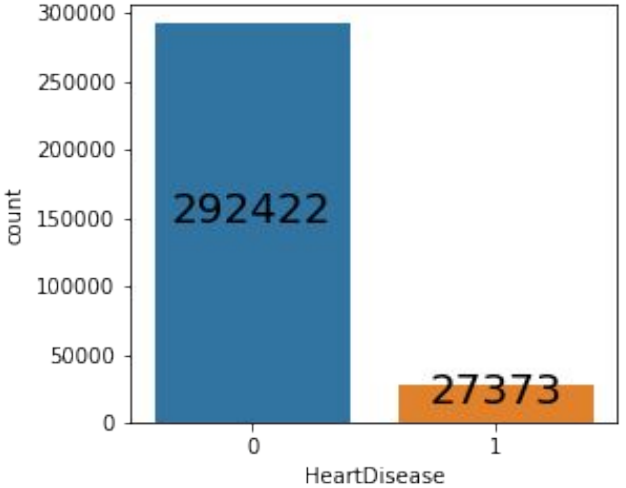
Interactive Dashboard



Sample Imbalance

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





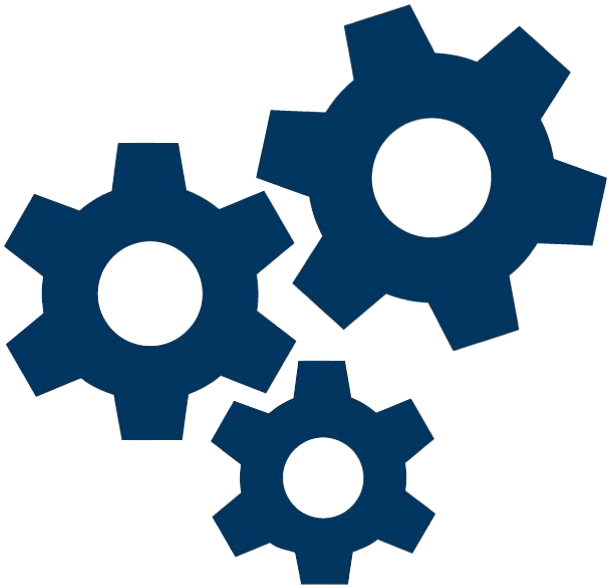
Machine Learning Model

Testing Machine Learning Models

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

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

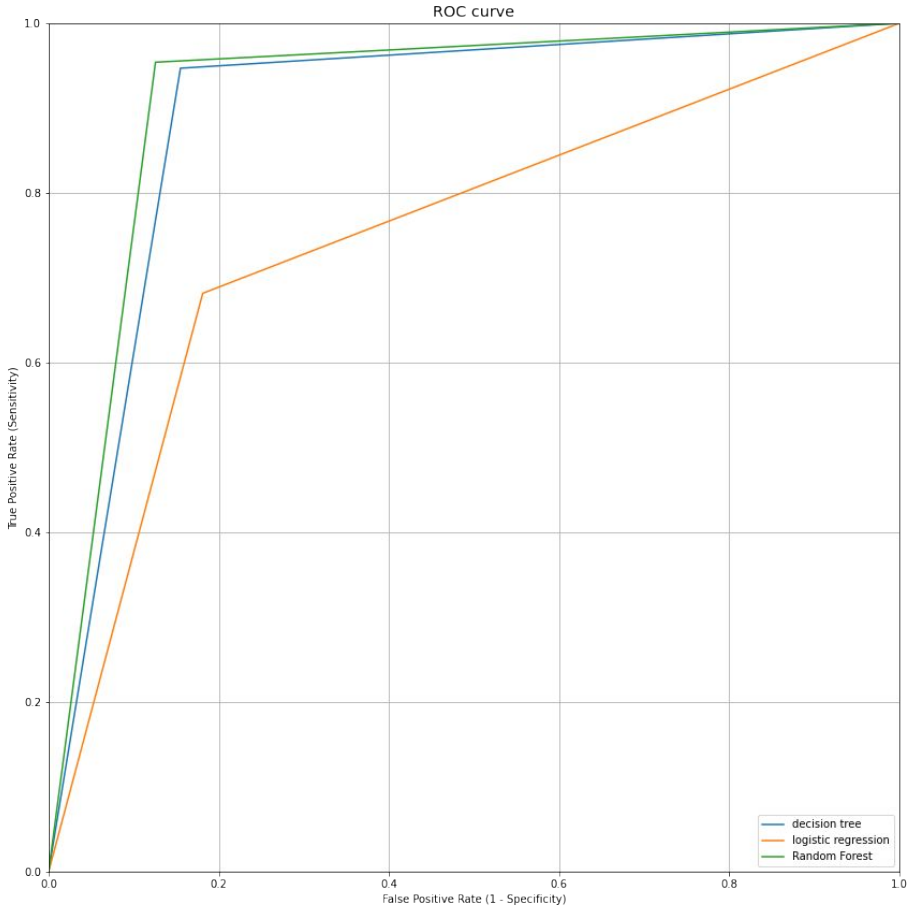
- Accuracy 0.90790
- Precision 0.84292
- Recall 0.95356



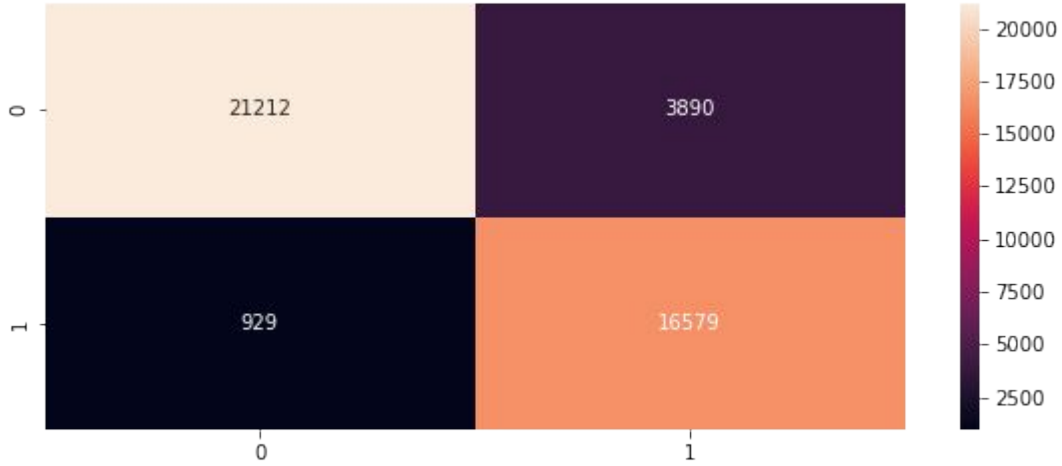
ROC Chart Models

Performance comparison to determine which model generated the best results.

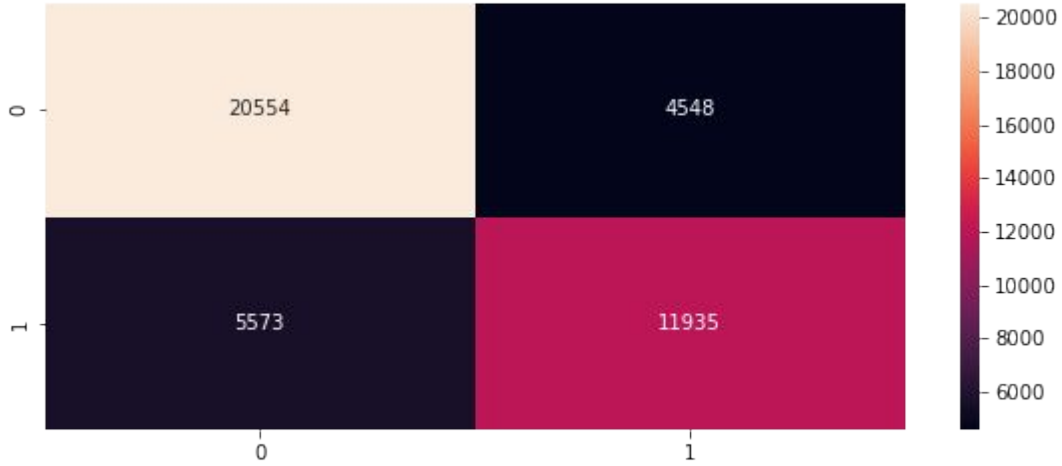
In this case the better-performing model is Random Forest.



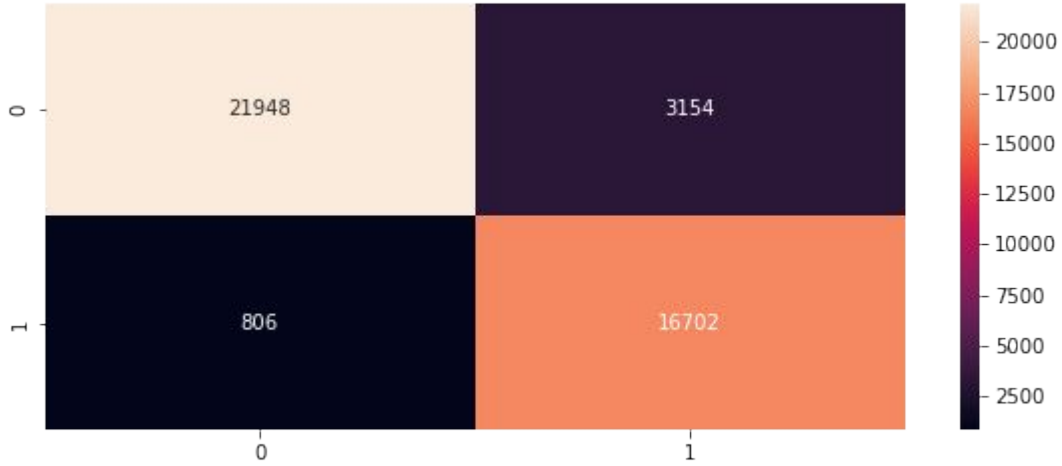
Confusion Matrix



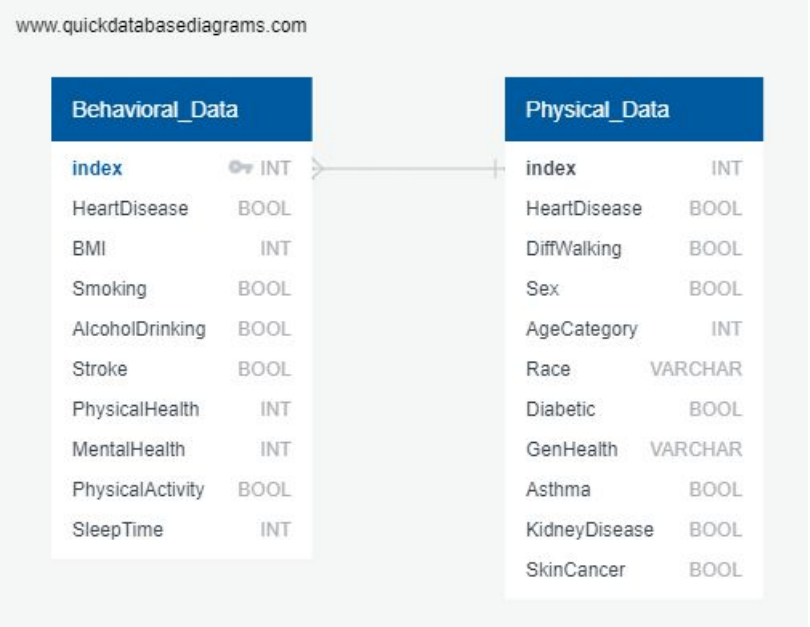
Logistic Regression



Random Forrest Clasifyer



Database Diagram




Thank you.