****

**(Affiliated to Tribhuvan University)**

**A**

**Lab Report**

**On**

**Predicting Heart Disease Using Machine Learning**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

**Submitted By:**

**Name:** Bivek Chaudhary

**Faculty:** BSc. CSIT

**Roll No:** 23871/076

**Batch:** 2076

**Year:** IV

**Semester:** VIII

**Submitted To:**

Department of CSIT

Orchid International College

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

This report tells discusses various Python-based ML and Data Science libraries in an attempt to build a Machine Learning model capable of predicting whether or not someone has heart disease based on their medical attributes. We're going to take the following approach:

1. Problem Definition
2. Understanding Features
3. Data Preparation and its tools
4. Exploratory Data Analysis
5. Modelling
6. Model Evaluation

## Problem Definition

Given clinical parameters about a patient, can we predict whether or not they have heart disease? Our aim is to reach the model accuracy of more than 85% . If the model scores better than 85%, we will select the model.

## Understanding Features

1. Heart\_disease: It is the target variable indicating the presence of heart disease (object type) into the value of:

0 = No

1 = Yes

1. BMI: It Stands for Body Mass Index, a measure of body fat based on height and weight (float) which is in numeric values.
2. Smoking: It indicates if the individual smokes (object) or not categorized into:

0=No

1= Yes

1. Alcohol\_drinking: It indicates if the individual consumes alcohol (object) or not categorized into:

0=No

1= Yes

1. Stroke: It indicates if the individual has had a stroke (object) ) or not categorized into:

0=No

1= Yes

1. Physical\_health: Number of days the individual experienced poor physical health in the past 30 days (float).
2. Mental\_health: Number of days the individual experienced poor mental health in the past 30 days (float).
3. Diff\_walking: It indicates if the individual has difficulty walking (object) whose format:

0=No

1= Yes

1. Sex: Gender of the individual (object)

0=Female

1= Male

1. Age\_category: It displays the age category of the individual (object) whose format:

18-24': 21,

'25-29': 27,

'30-34': 32,

'35-39': 37,

'40-44': 42,

'45-49': 47,

'50-54': 52,

'55-59': 57,

'60-64': 62,

'65-69': 67,

'70-74': 72,

'75-79': 77,

'80 or older': 85

1. Race: It displays the race of the individual (object), categorized into:

'White': 1,

'Black': 2,

'Asian': 3,

'American Indian/Alaskan Native': 4,

'Other': 5,

'Hispanic': 6

1. Diabetic: It indicates if the individual has diabetes (object) whose format:

0=No

1= Yes

1. Physical\_activity: Indicates if the individual engages in physical activity (object).
2. General\_health: Self-assessed general health status (object).
3. Sleep\_time: Average hours of sleep per night (float).
4. Asthma: Indicates if the individual has asthma (object) whose format:

0=No

1= Yes

1. Kidney\_disease: Indicates if the individual has kidney disease (object) whose format:

0=No

1= Yes

1. Skin\_cancer: Indicates if the individual has skin cancer (object) whose format:

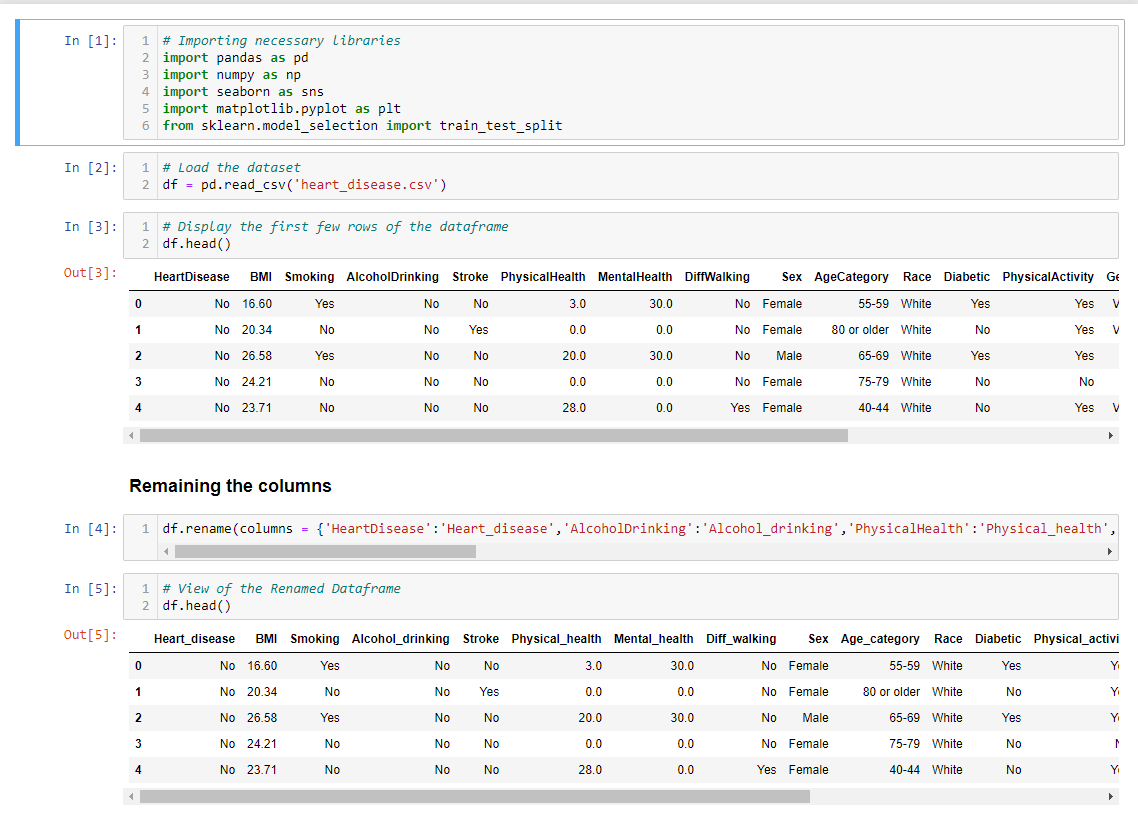
0=No

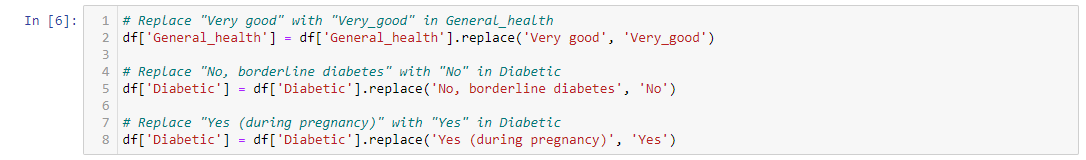
1= Yes

This dataset includes a mix of categorical and numerical features, capturing various health-related attributes and behaviors that can influence heart disease risk.

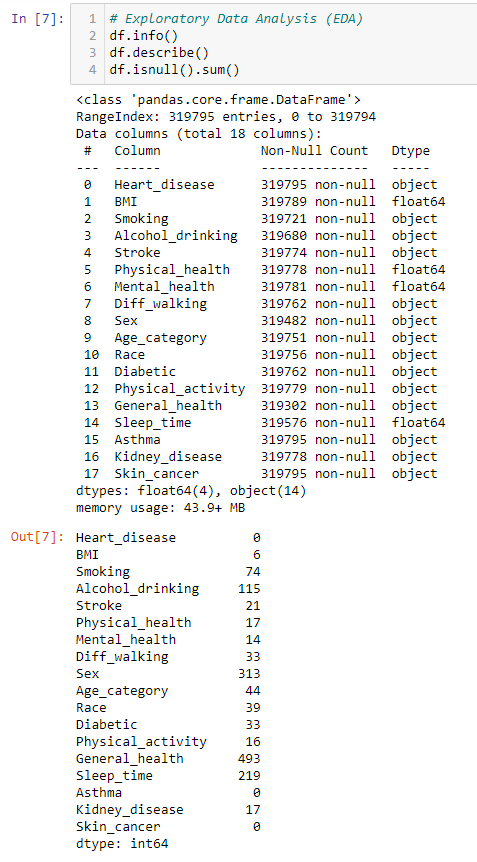
## Data preparation and its tools

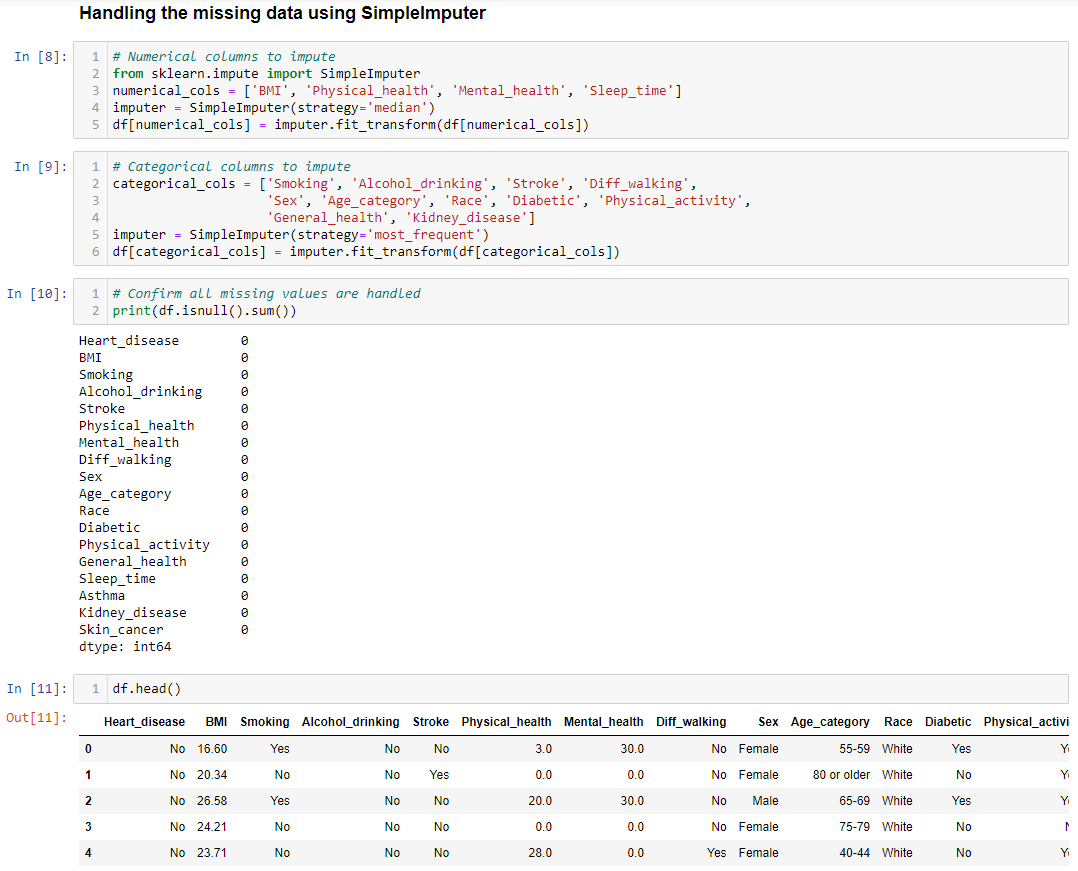
1. Pandas & NumPy: Pandas offers powerful data structures like Series and DataFrame for structured data manipulation, while NumPy provides support for numerical computing with multi-dimensional arrays and mathematical functions.
2. Matplotlib & Seaborn: Matplotlib is a versatile library for creating static, interactive, and animated visualizations, while Seaborn simplifies the process of generating complex statistical graphics on top of Matplotlib.
3. Scikit-Learn: Scikit-Learn is a comprehensive machine learning library with various algorithms for supervised and unsupervised learning tasks, along with utilities for model evaluation and selection.

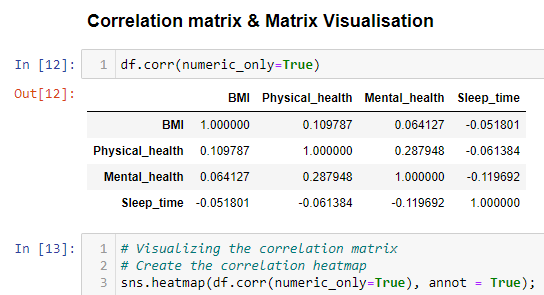


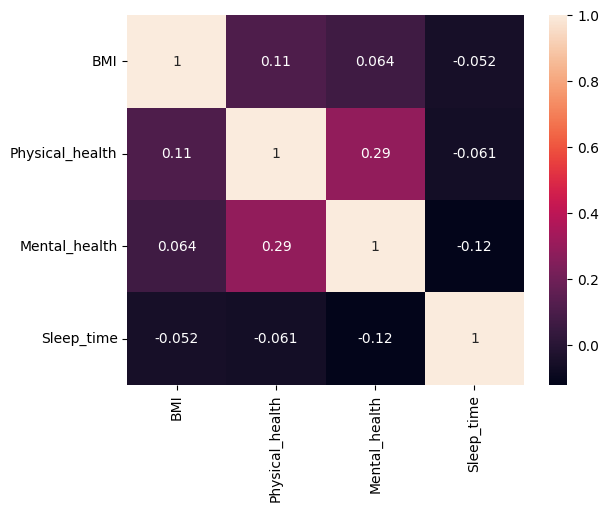


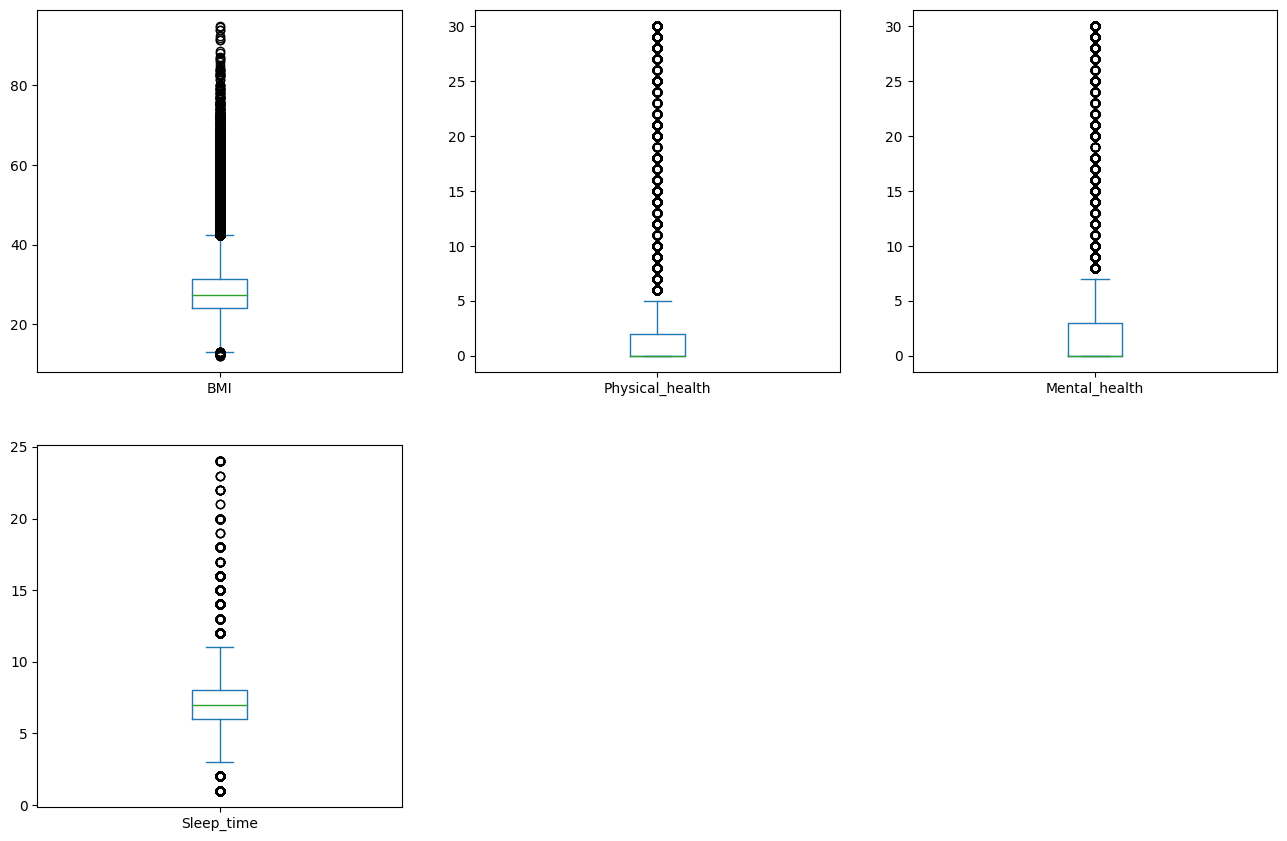
## Exploratory Data Analysis

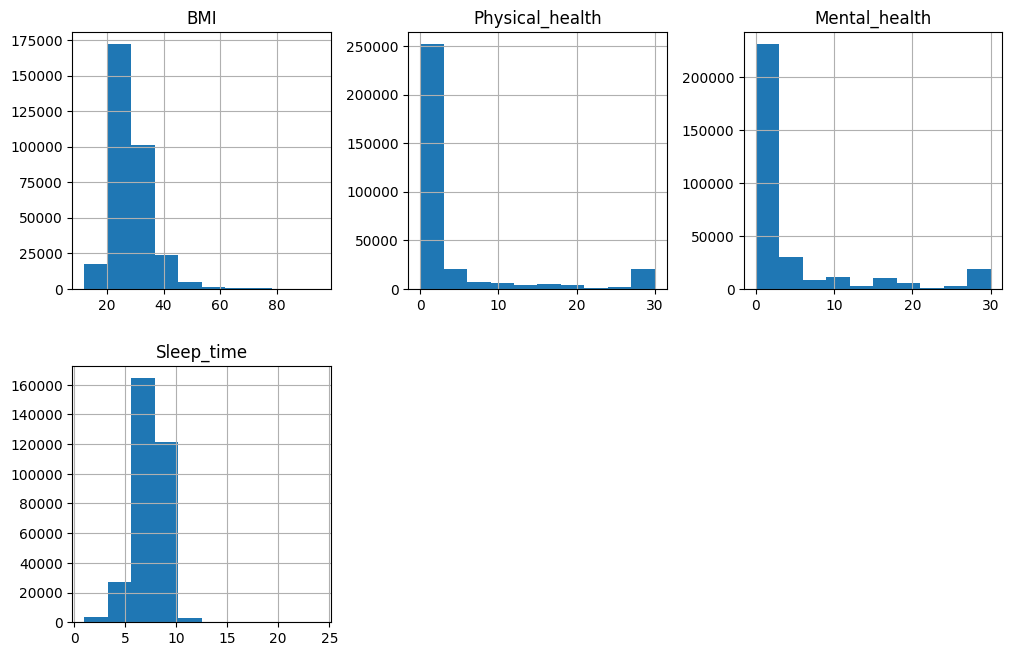


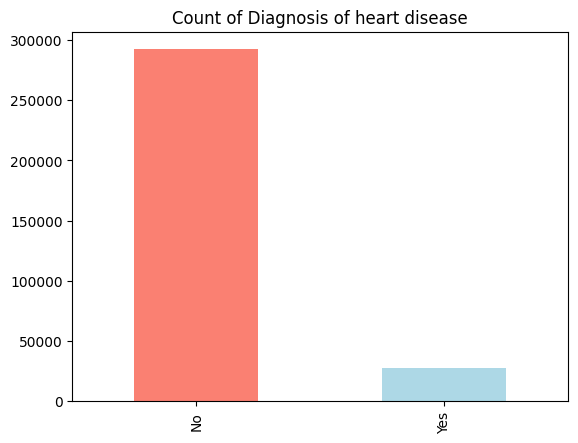


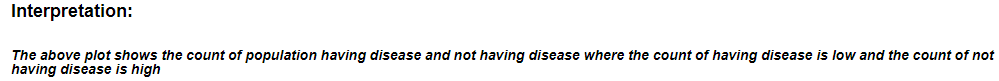




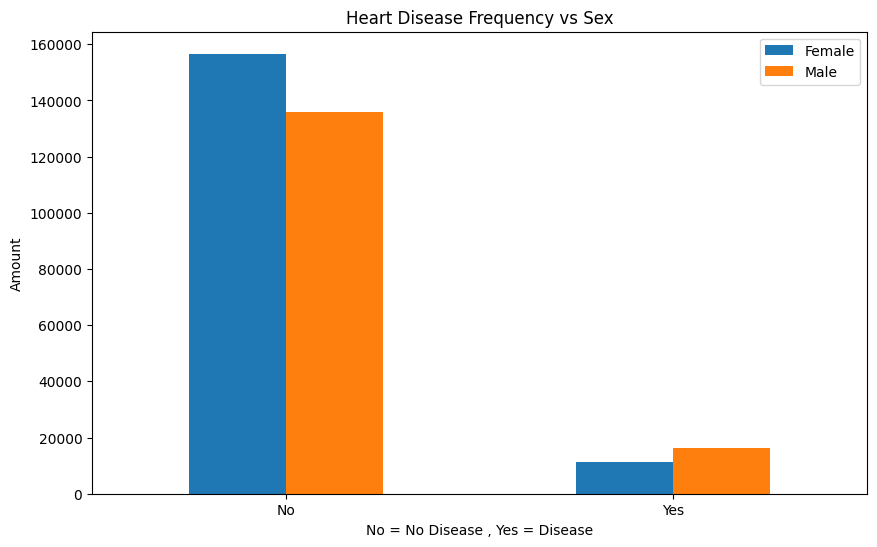


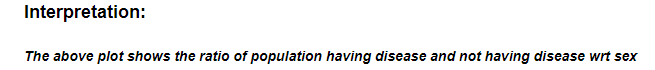


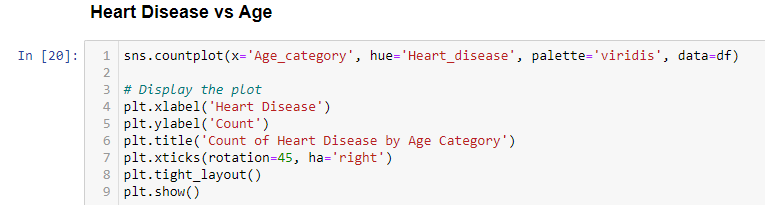


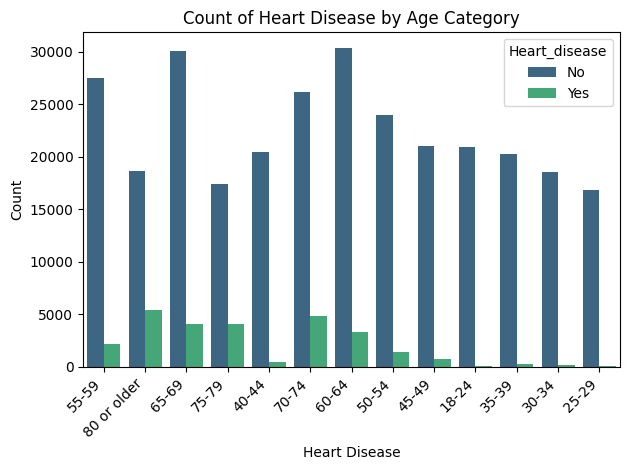




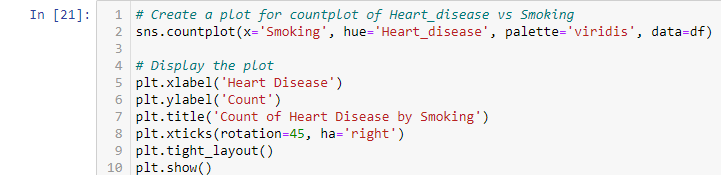


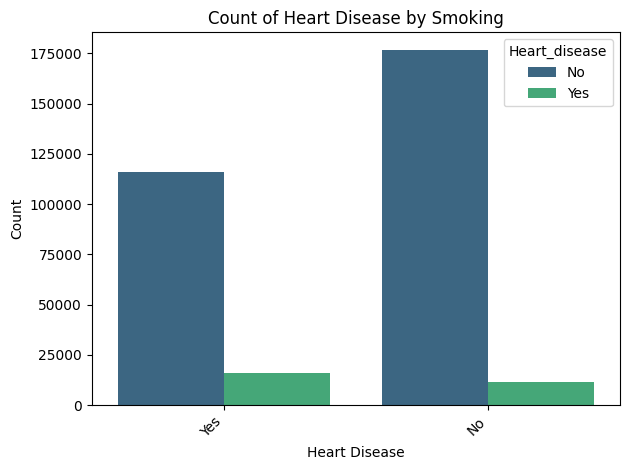


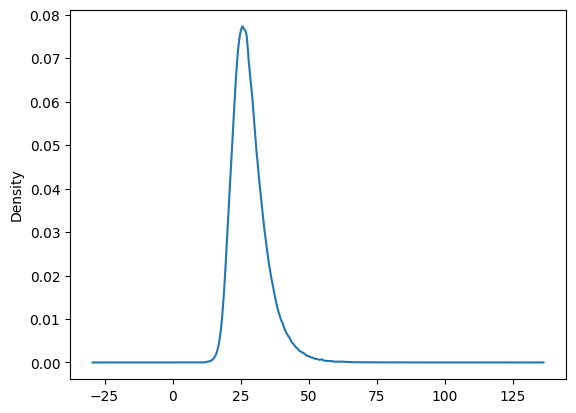




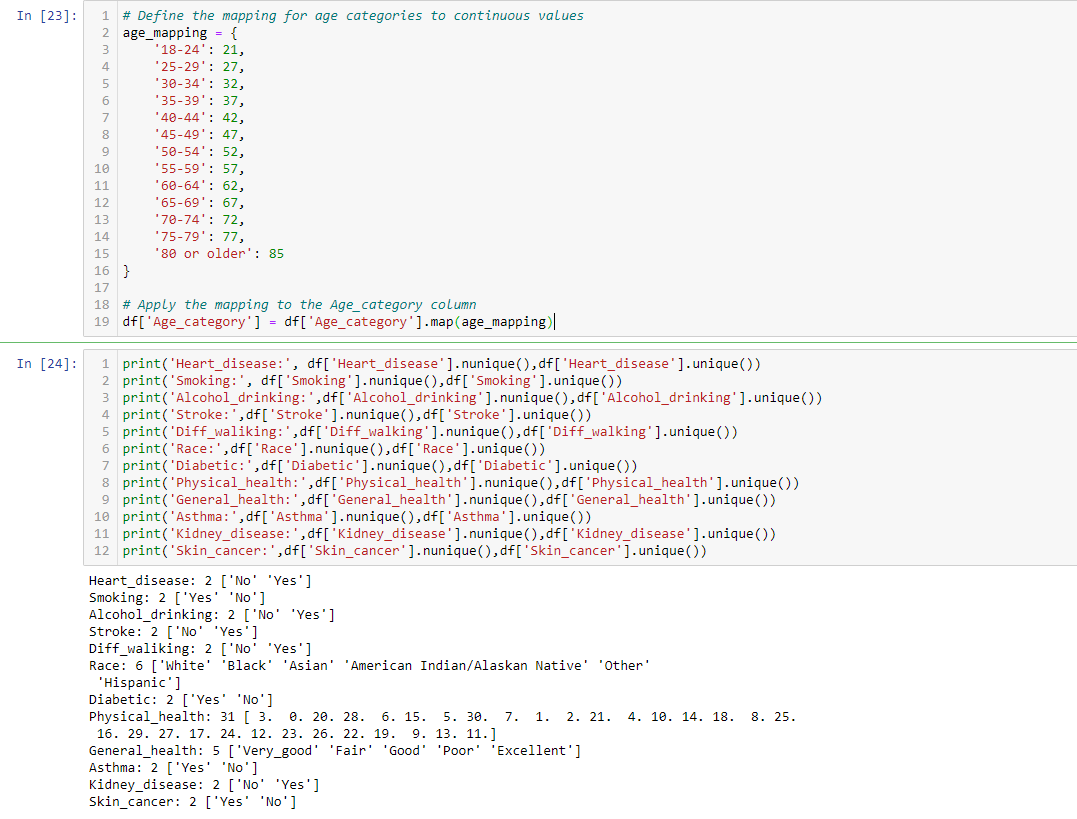
**Count plot of Heart Disease vs Smoking**



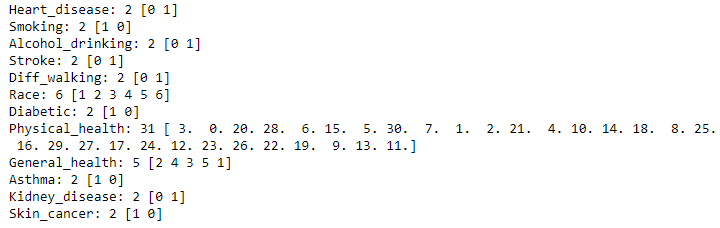


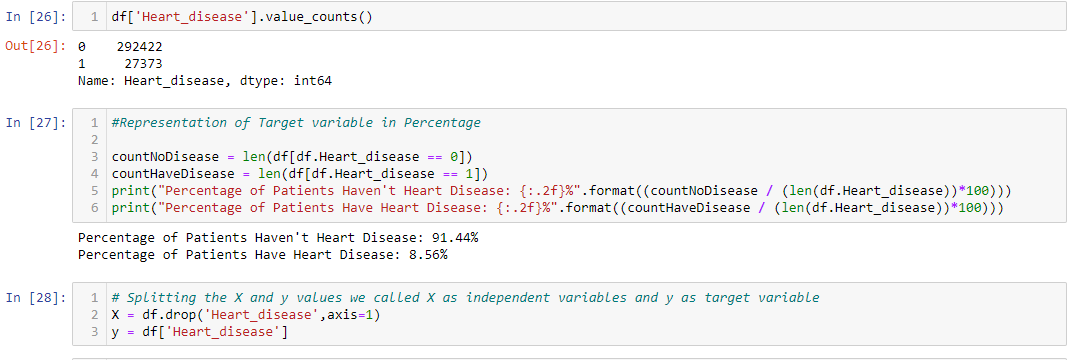








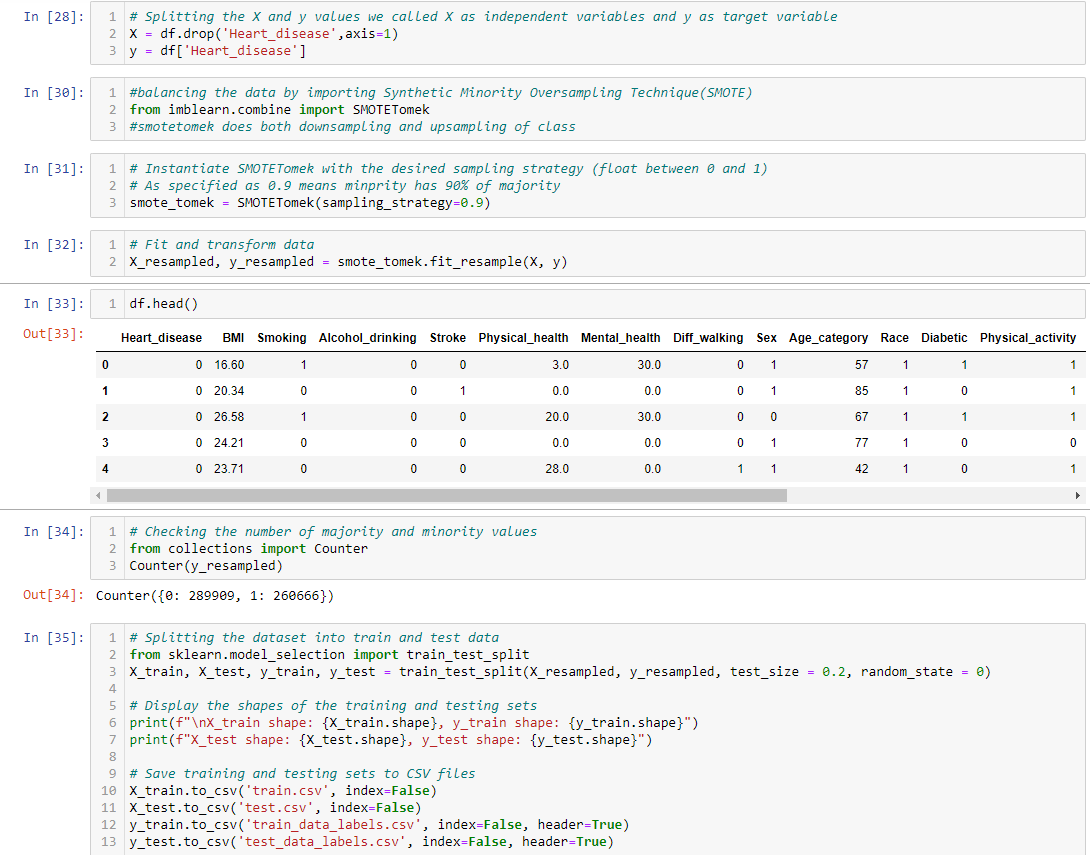




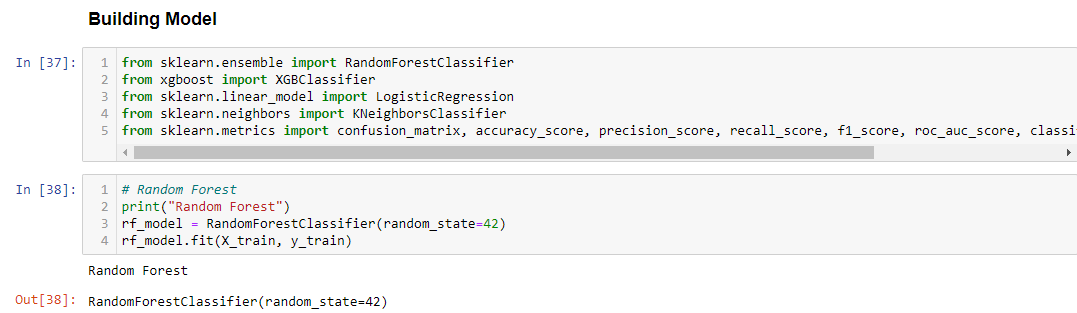
## Modelling

We will experiment with the models, trying the model and getting the results from them.

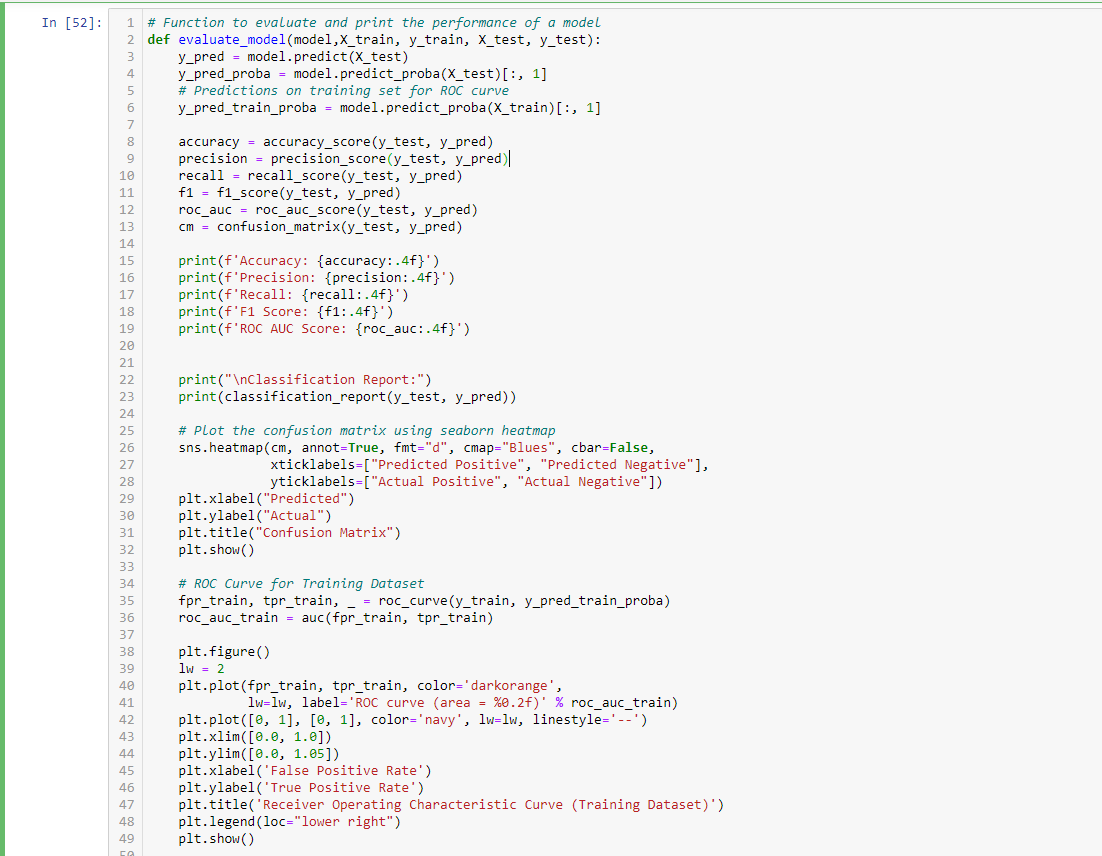
Since in the target variable, Percentage of Patients Haven't Heart Disease: 91.44% Percentage of Patients Have Heart Disease: 8.56% so to imbalance the ratio we use the SMOTETomek which helps in balancing the data by importing Synthetic Minority Oversampling Technique by doing both downsampling and upsampling of class.

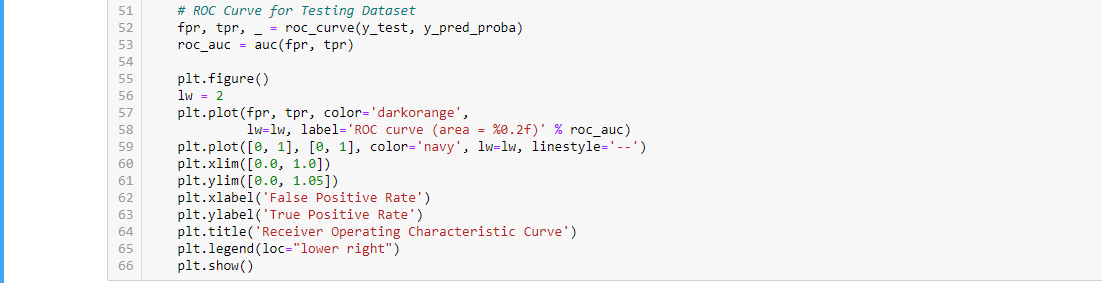


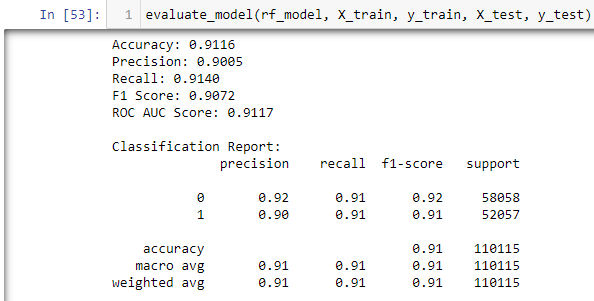
Now we have got our data split into training and test sets, it is time to build a Machine Learning model. We will train it (find the patterns) on the training set. And we will test it (use the patterns) on the test set.

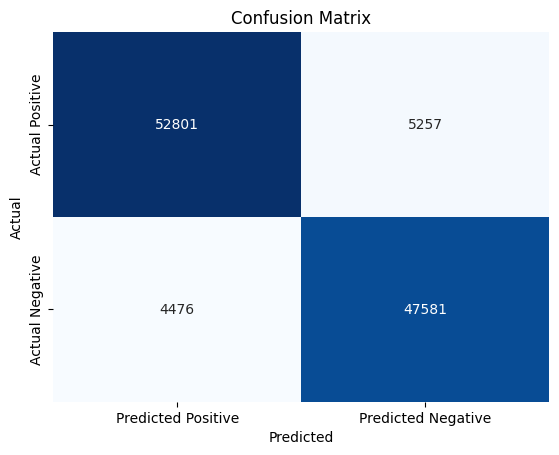


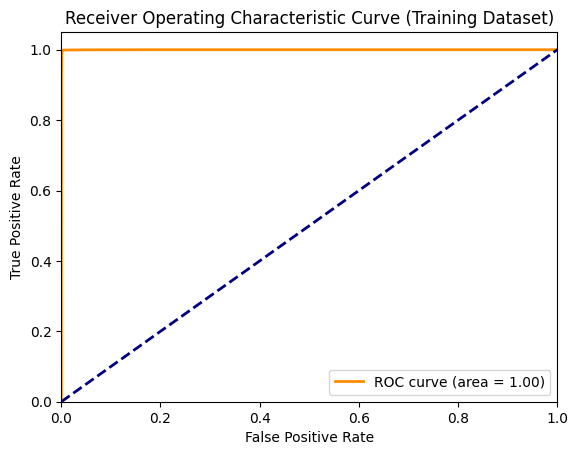
## Model Evaluation



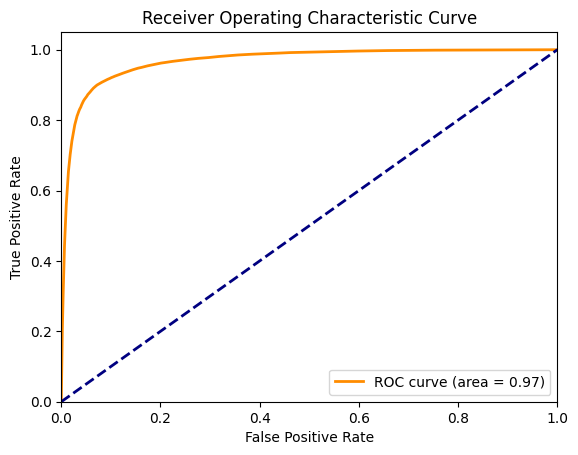






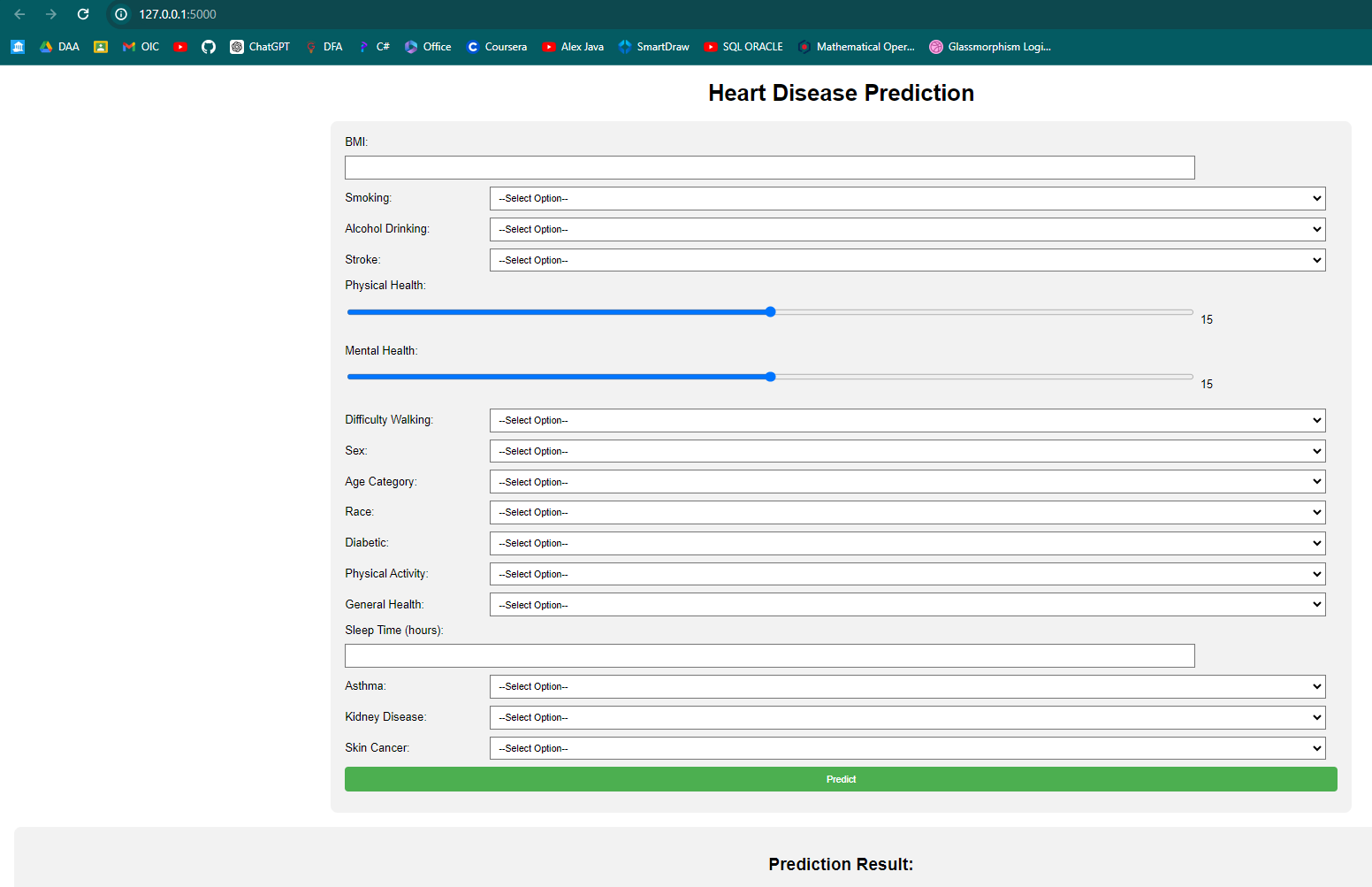


In training dataset. ROC Curve shows that area = 1which is equals to the AUC of 1 that indicates perfect performance.

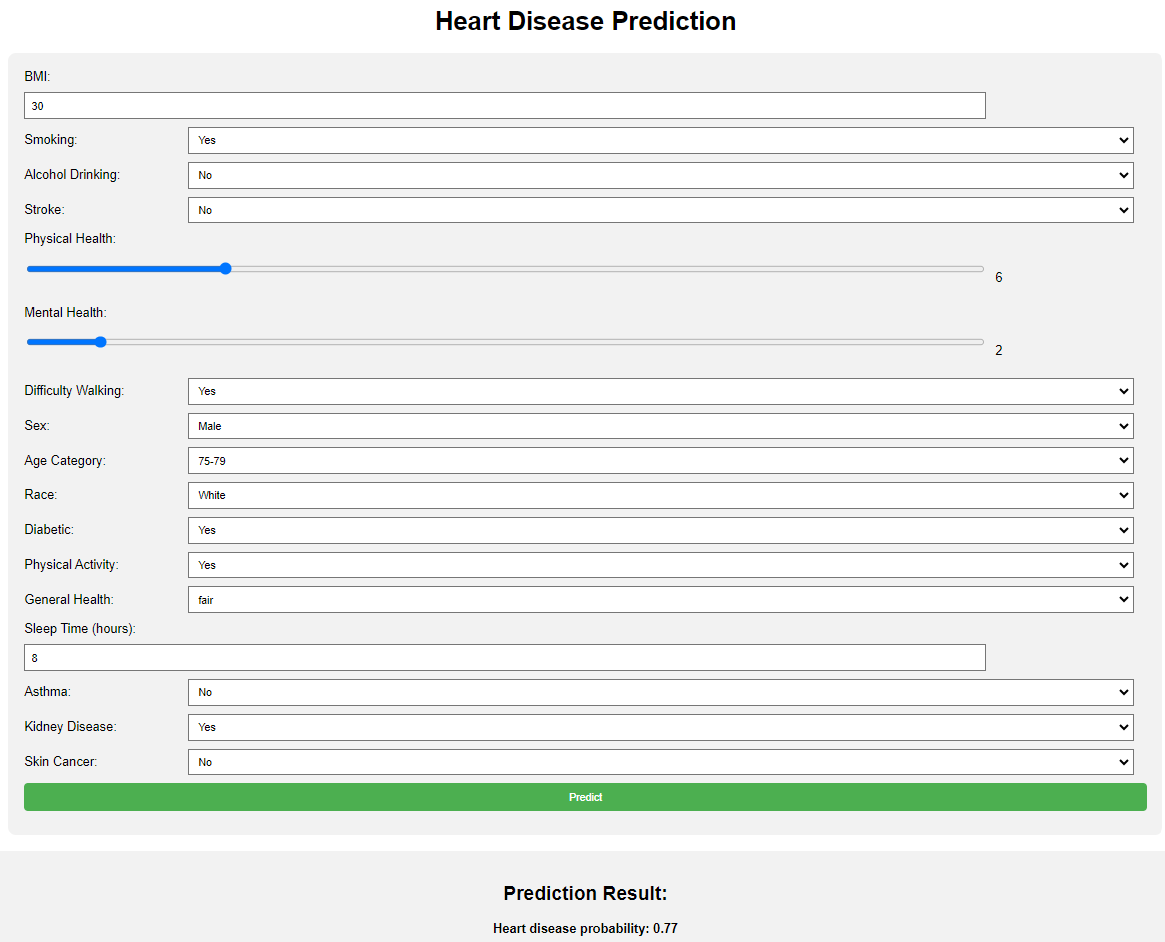


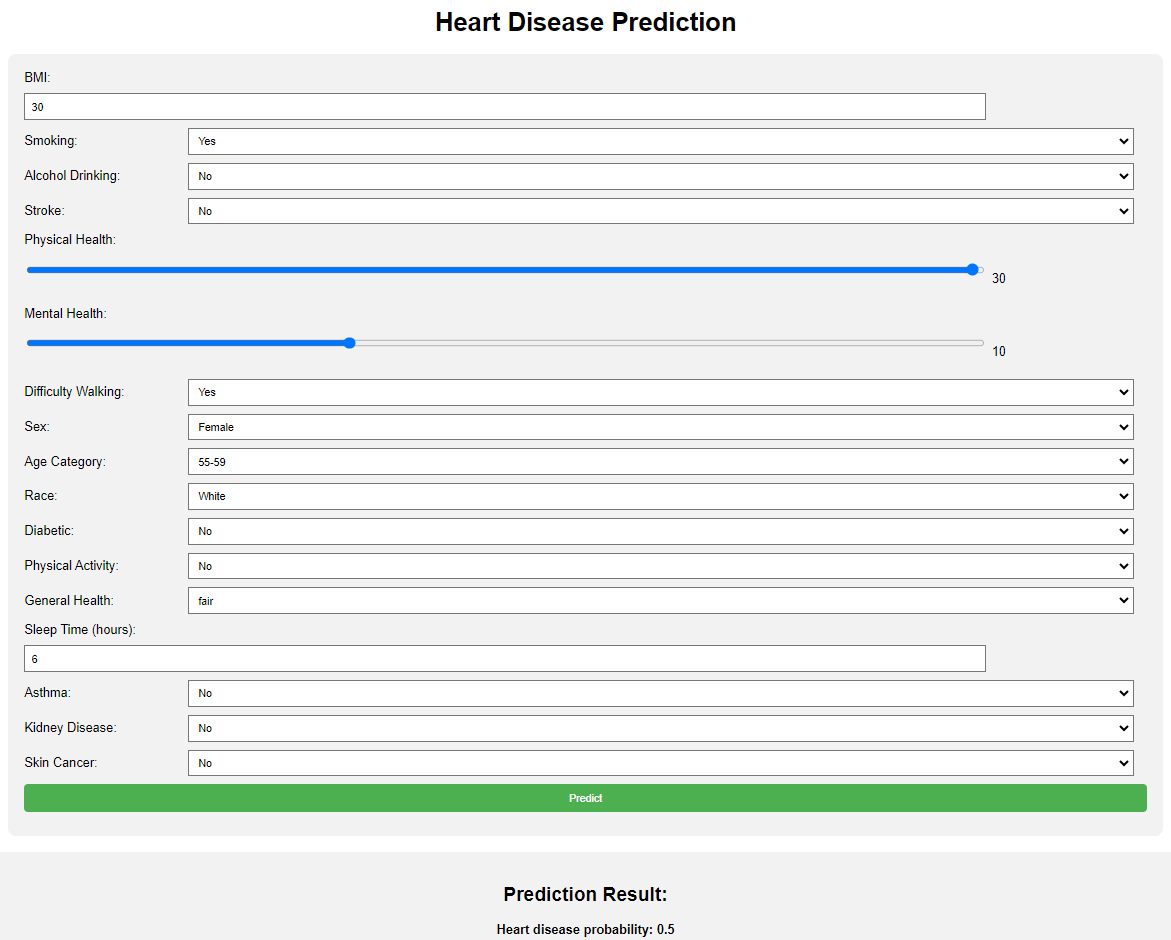
In testing dataset, ROC Curve shows that area = **0.97** which means nearer to the AUC of 1 that indicates better performance.

## Form For the Heart Disease Prediction



Result:





## Conclusion

From the above prediction analysis, we were able to predict the probability of the heart disease or not as in the result there is the heart disease probability is **0.77** which is suffering from heart disease where as there is also the heart disease probability is **0.5** which is not suffering from heart disease using Random Forest Classifier Model.