Heart Disease Analysis & Risk Prediction

A Project Report
Submitted in partial fulfilment of the
Requirements for the award of the degree of
Master Of Science
(Data Science & Big Data Analytics)

By:
Mr. Sanket Madan Bairagi

Under the esteemed guidance of

Ms. Esmita Gupta
(Vice Principle & HOD of IT Dept.)

Department Of Information Technology
B. K. Birla College of Arts, Science and Commerce
Kalyan - 421304.
2022-2023

B. K. Birla College of Arts, Science & Commerce Kalyan.

Department of Information Technology
Master of Science in Data Science & Big Data Analytics
(M.Sc. - Part II)

CERTIFICATE

This is to certify that data science project on **Analysis and Detection of Heart Disease Risk** entitled in healthcare sector submitted by **Mr. Sanket Madan Bairagi** Exam Seat No: _____ for the partial fulfilment of the requirement for award of degree Master of Science in Data Science And Big Data Analytics, to the University of Mumbai, is a bonafide work carried out during academic year 2022-23.

Place: Kalyan

Date: _____

Signature of External

_____

Signature of Principle

_____

Signature of HOD

_____

Declaration

I declare that this submission represents my ideas in my own words and where others idea or words have been declaring that I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date :- _____ _____

Sanket Madan Bairagi

INDEX

List Of Figures

Abstract

Heart disease is currently the leading cause of death in the world. According to the Canter for Disease Control (CDC), around 659,000 people die from heart related diseases every year in America, which is one in four deaths overall (CDC, 2021). From this number alone, we can see that this is a major problem causing the majority of deaths. Medically, heart disease arises when a layer of plaque blocks the arteries or blood vessels connected to the heart. This congests the arteries and does not allow the necessary nutrients and oxygen to reach the heart (Roth, 2018). Furthermore, there are many factors that make an individual more likely to suffer from heart disease. Some major risk factors include high blood pressure, smoking, obesity, and physical inactivity. While heart disease is very dangerous, many of the risk factors can be prevented with actions such as exercising and maintaining a healthy diet. That is why it is important to be able to predict possible heart disease when it is still preventable. In this project, I had analysed the data to determine the causes of heart disease. This analysis can be used to assist in the management of an individual's risk by identifying lifestyle choices and other heart disease-related health indicators.

Introduction

I chose to base my project on the key indicators of heart disease because it is leading cause of death in worldwide, and I am very interested in exploring the data. Heart disease is the leading cause of death in many countries, including the United States of America (Heart Disease, 2020). The term 'heart disease' can refer to numerous heart conditions such as a heart attack or coronary artery disease. A heart attack occurs when a section of the heart is not receiving enough blood, and therefore causes damage to the heart. Coronary disease is caused by a build-up of plaque on the walls of the arteries that pump blood to the heart and other parts of the body. Without the tireless effort of heart pumping blood, both conditions are highly likely to cause death.

The heart is one of the most important organs in the body and it is the main organ in the cardiovascular system. This system is made up of a network of blood vessels which pumps blood all around the body. The heart also controls the rhythm and speed of the heart rate in the body, along with maintaining blood pressure. Furthermore, undesirable carbon dioxide and waste products are carried away by the blood filled with nutrients and oxygen.

The heart is evidently important; therefore, it needs to be understood and taken care of. This dataset explores the indicators of heart disease, which sparks my interest as I feel this is vital information for everyone to know. Staying informed and being aware of the indicators keeps a person's risk of heart disease low. Some risk factors are uncontrollable, such as race or family background. However, many key risk factors are controllable and being aware of them can significantly reduce one's risk of heart disease. Additionally, the detection and prevention of heart disease is vital to healthcare.

In my project, I analysed data to determine the causes of heart disease. The project contains Streamlit application that includes a dashboard for the visualization of analysis and a predictive model that takes input from the user and shows how much the user is at risk of heart disease.

Objective

Here are some potential objectives for a machine learning project using the "Personal Key Indicators of Heart Disease" dataset:

1. Predicting heart disease: Build a classification model that predicts whether a patient has heart disease based on their personal key indicators such as age, sex, BMI, Smoking , and Sleep Time etc.
2. Feature importance: Identify which personal key indicators have the greatest impact on the likelihood of heart disease and use this information to develop targeted interventions or screening strategies.
3. Outcome prediction: Predict the likelihood of specific outcomes related to heart disease such as heart attacks based on personal key indicators.
4. Treatment effectiveness: Evaluate the effectiveness of different treatments or interventions for heart disease by comparing the personal key indicators of patients before and after treatment.

Overall, the goal of a project using this dataset would be to better understand the relationship between personal key indicators and heart disease and to develop predictive models that can be used to identify patients who are at risk or who would benefit from specific interventions.

Dataset

To analyse this problem, I utilized a data set found on Kaggle titled “Personal Key Indicators of Heart Disease” (https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease). This data was collected by the CDC as a part of the Behavioural Risk Factor Surveillance System (BRFSS). This is a large system that conducts telephone surveys of adults in the United States, and it is one of the most extensive health surveys in the country with about 400,000 surveys every year. For this project, we are using data collected from the 2020 survey. The original CDC dataset has about 400,000 entries and more than 300 columns containing survey questions on different demographic and health topics. This data was then reduced to approximately 320,000 entries and 18 columns by the creator of the Kaggle dataset. This was done to include only the data that is relevant to heart disease.

Columns in data -

- HeartDisease - Binary (Yes or No)
- BMI (Body mass Index) - (values)
- Smoking - Have you smoked at least 100 cigarettes in your entire life? (The answer is binary (Yes or No))
- AlcoholDrinking - Heavy drinkers : adult men having more than 14 drinks per week and adult women having more than 7 drinks per week (The answer is binary (Yes or No))
- Stroke - (Ever told) (you had) a stroke? (The answer is binary (Yes or No))
- PhysicalHealth - Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (0-30 days).
- MentalHealth - Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days).
- DiffWalking - Do you have serious difficulty walking or climbing stairs? (The answer is binary (Yes or No))
- Sex - Are you male or female? (The response is binary (Female or Male))
- AgeCategory - Fourteen-level age category.
- Race - Imputed race/ethnicity value.
- Diabetic - (Ever told) (you had) diabetes?
- PhysicalActivity - Adults who reported doing physical activity or exercise during the past 30 days other than their regular job.
- GenHealth - Would you say that in general your health is...
- SleepTime : On average, how many hours of sleep do you get in a 24-hour period?

- Asthma : (Ever told) (you had) asthma?
- KidneyDisease : Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
- SkinCancer : (Ever told) (you had) skin cancer?

Following Image Shows the data:

```
In [2]: df = pd.read_csv('https://raw.githubusercontent.com/SanketBairagi/FinalPro/main/heart_2020_cleaned.csv')
        df.head()
```

Out[2]:

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
0	No	16.60	Yes	No	No	3.0	30.0	No	Female	55-59	White	Yes	Yes	Very good	5.0	Yes	No	Yes
1	No	20.34	No	No	Yes	0.0	0.0	No	Female	80 or older	White	No	Yes	Very good	7.0	No	No	No
2	No	26.58	Yes	No	No	20.0	30.0	No	Male	65-69	White	Yes	Yes	Fair	8.0	Yes	No	No
3	No	24.21	No	No	No	0.0	0.0	No	Female	75-79	White	No	No	Good	6.0	No	No	Yes
4	No	23.71	No	No	No	28.0	0.0	Yes	Female	40-44	White	No	Yes	Very good	8.0	No	No	No

Methodology

1. Data Cleaning

Initially, I followed the cleaning of data using columns of the data frame after collecting from the dataset. Observed and found that there are no missing or null values to be removed.

```
In [5]: df.isnull().sum()

Out[5]: HeartDisease        0
        BMI                 0
        Smoking             0
        AlcoholDrinking     0
        Stroke              0
        PhysicalHealth      0
        MentalHealth        0
        DiffWalking         0
        Sex                 0
        AgeCategory         0
        Race                0
        Diabetic            0
        PhysicalActivity    0
        GenHealth           0
        SleepTime           0
        Asthma              0
        KidneyDisease       0
        SkinCancer          0
        dtype: int64

        There are no missing values!
```

2. Data Description

Pandas describe function shows the statistical information of numerical columns. Below images show the mean, standard deviation, count , percentiles, min and max details of numerical columns.

```
df.describe()
```

	BMI	PhysicalHealth	MentalHealth	SleepTime
count	305095.000000	305095.000000	305095.000000	305095.000000
mean	28.263328	3.079998	3.621089	7.157272
std	6.277987	7.561576	7.590524	1.098027
min	12.020000	0.000000	0.000000	5.000000
25%	24.020000	0.000000	0.000000	6.000000
50%	27.290000	0.000000	0.000000	7.000000
75%	31.320000	1.000000	3.000000	8.000000
max	94.850000	30.000000	30.000000	10.000000

3. Data Exploration

I organized the data based on the structure using exploratory analysis. Different types of data can be seen from the explanation of dataset variables. The summary of the data shown gives us the description of the common attributes. This tells us that there are 18 columns with 319795 values. In dataset 4 variables are float and 14 variables are object in data type.

```
In [6]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 319795 entries, 0 to 319794
Data columns (total 18 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   HeartDisease      319795 non-null  object
 1   BMI               319795 non-null  float64
 2   Smoking           319795 non-null  object
 3   AlcoholDrinking   319795 non-null  object
 4   Stroke            319795 non-null  object
 5   PhysicalHealth    319795 non-null  float64
 6   MentalHealth      319795 non-null  float64
 7   DiffWalking       319795 non-null  object
 8   Sex               319795 non-null  object
 9   AgeCategory       319795 non-null  object
 10  Race              319795 non-null  object
 11  Diabetic          319795 non-null  object
 12  PhysicalActivity  319795 non-null  object
 13  GenHealth         319795 non-null  object
 14  SleepTime         319795 non-null  float64
 15  Asthma            319795 non-null  object
 16  KidneyDisease     319795 non-null  object
 17  SkinCancer        319795 non-null  object
dtypes: float64(4), object(14)
memory usage: 43.9+ MB
```

4. Feature Engineering

In feature engineering, categorical columns are covered by numeric columns. In the data frame 'HeartDisease', 'Smoking', 'AlcoholDrinking', 'Stroke', 'DiffWalking', 'PhysicalActivity', 'Asthma', 'KidneyDisease', 'SkinCancer' are in "yes" and "No" format, so "Yes" is assigned with 1 and "No" is assigned with 0. In the "Sex" column, "Female" is assigned with number 0 and "Male" assigned with number 1. In the data frame, 'AgeCategory' column has 13 different categories, and each category is assigned with 0 to 12 numbers, respectively. 'Race' column has 6 different categories, and each category is assigned with 0 to 5 numbers, respectively. 'Diabetic' column has 4 different categories, and each category is assigned with 0 to 3, respectively. The "GenHealth' column has 5 different categories, and each category is assigned with 0 to 4 numbers respectively. 'BMICAT' column has 4 different categories, and each category is assigned with 0 to 3 numbers, respectively.

Data After Feature Engineering :

```
data=data.drop_duplicates()
data.head()
```

	HeartDisease	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer	BMICAT
0	0	1	0	0	3.0	30.0	0	0	7	3	1	3	5.0	1	0	1	0
1	0	0	0	1	0.0	0.0	0	0	12	0	1	3	7.0	0	0	0	1
2	0	1	0	0	20.0	30.0	0	1	9	3	1	1	8.0	1	0	0	2
3	0	0	0	0	0.0	0.0	0	0	11	0	0	2	6.0	0	0	1	1
4	0	0	0	0	28.0	0.0	1	0	4	0	1	3	8.0	0	0	0	1

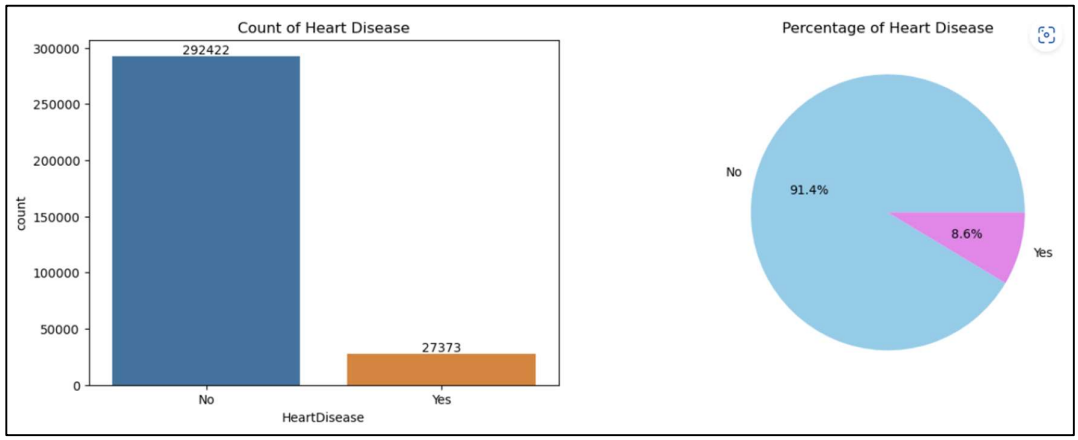
5. Feature Creation

Creating features involves creating new variables which will be most helpful for our model. This can be adding or removing some features.
In the dataset “BMICAT” column is made by “BMI” numerical column to get more insights form exploratory data analysis.in dataset those “BMI” is less than 18.5 are assigned with category "underweight". Those “BMI” between 18.5 to 25.5 are assigned with "Normal weight". Those are “BMI” between 25.0 to 30.0 are assigned with category “overweight". Remaining those “BMI” is more than 30.0 are assigned with category "Obese”.
The column “SleepTimeCat” is made up of column “SleepTime”, in which those are take less than 7-hour sleep are assigned with “Poor Sleep” category. Also, those take sleep 7-9 hours are assigned with “Good Sleep" category. Finally, those are take more than 9 hours of sleep are assigned with “Over Sleep" category.

6. Data Analysis

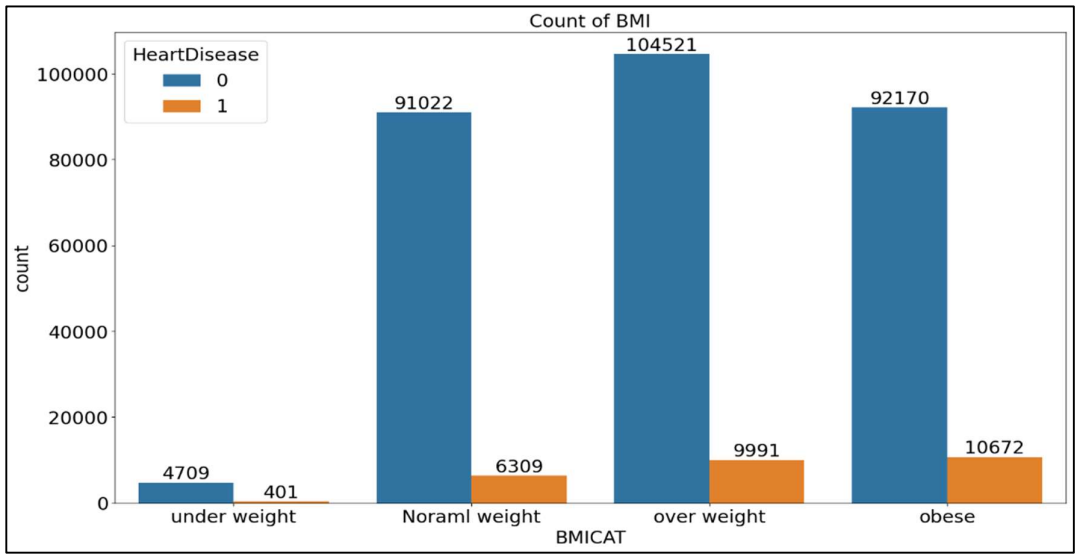
In the visualization I have created bar graphs, pie chart and area curves with respect to heart disease and gender to get better insights from data.

I. Count and percentage of heart disease patients



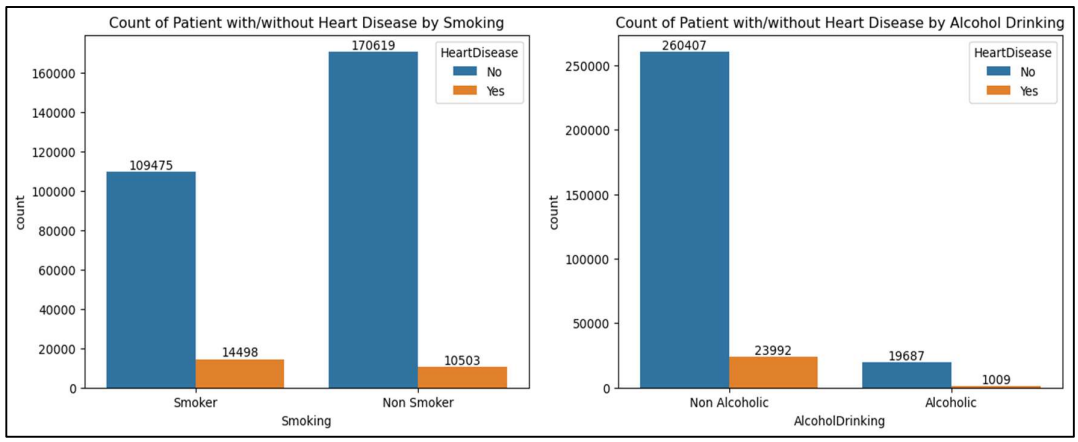
- This pie chart reveals that only 8.56 % of people suffers from heart disease, while 91.4% do not.

II. Bar plot of BMI category wise count of heart disease patients



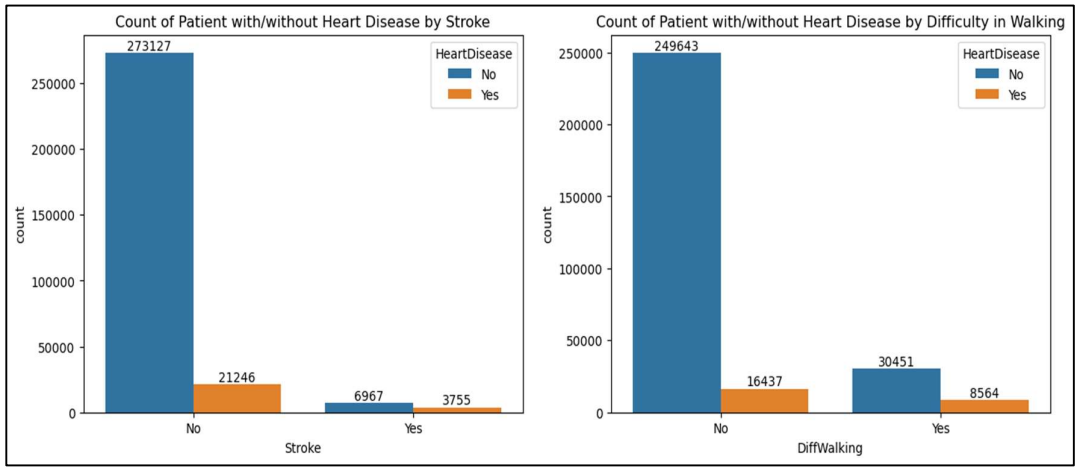
- Heart disease is increase with BMI as above graph shows number of heart disease count increase with BMI , Obese has higher number of heart disease.

III. Bar plots for heart disease wise count of Smoking and Alcohol Drinking



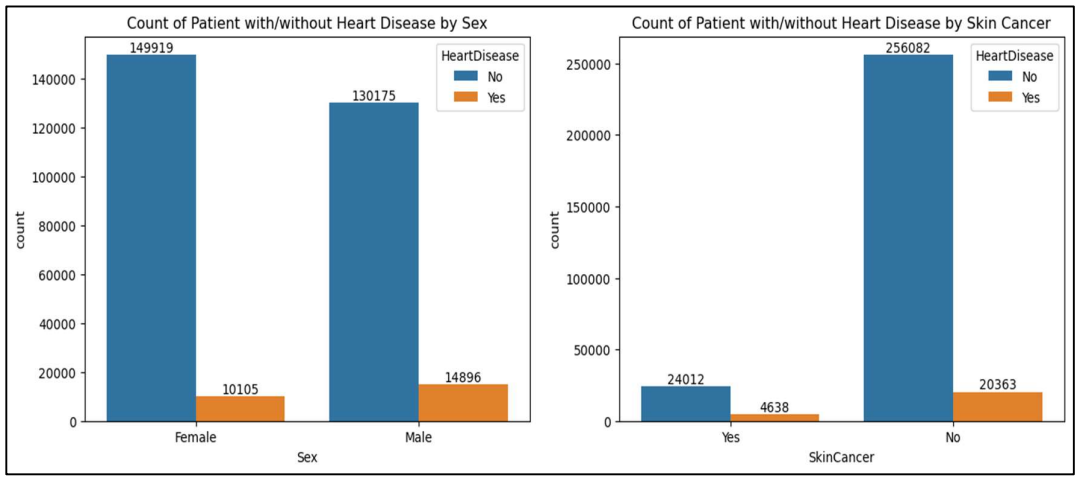
- observe that the people who are smoking are more susceptible to the heart disease.
- People who are not drinking alcohol, some of them have a heart disease.

IV. Bar plots for heart disease wise count of Stroke and Difficult in walking



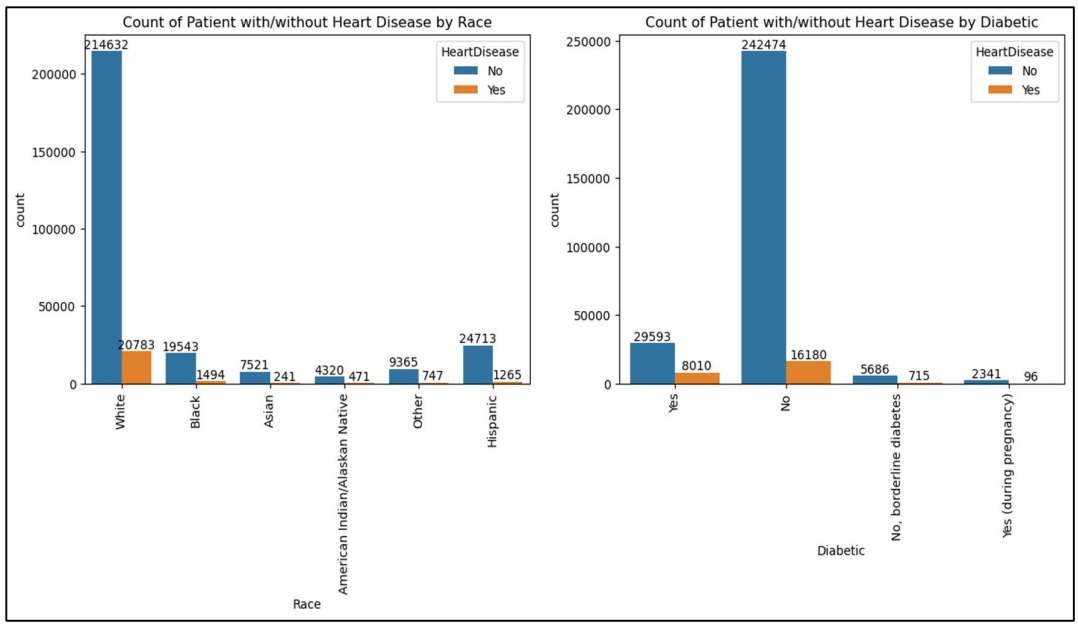
- Very few stroke patients have heart disease problem.
- Very few patients with difficulty in walk have heart disease.

V. Bar plots for heart disease wise count of Sex and Skin cancer



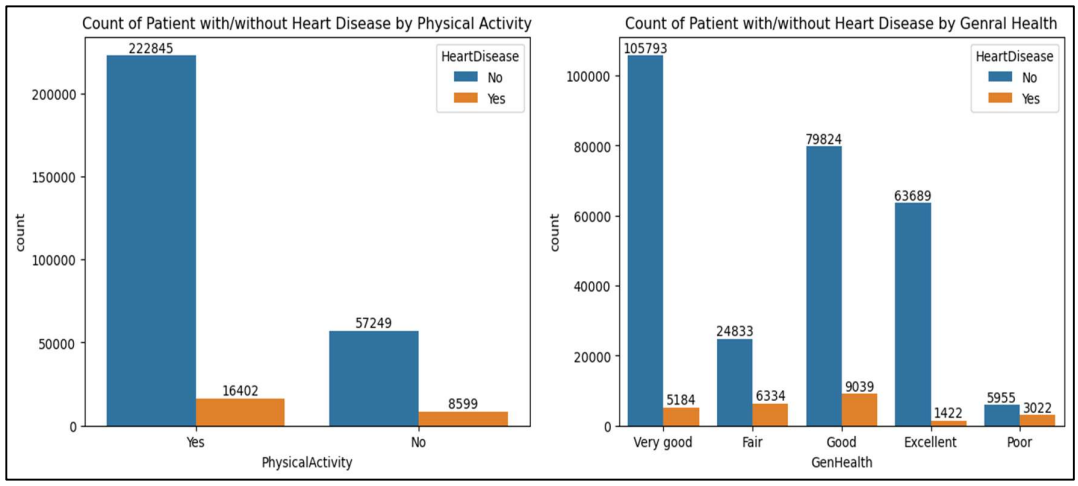
- Heart disease is more commonly present in male.
- Few skin cancer patients facing heart disease problem.

VI. Bar plots for heart disease wise count of Race and Diabetic



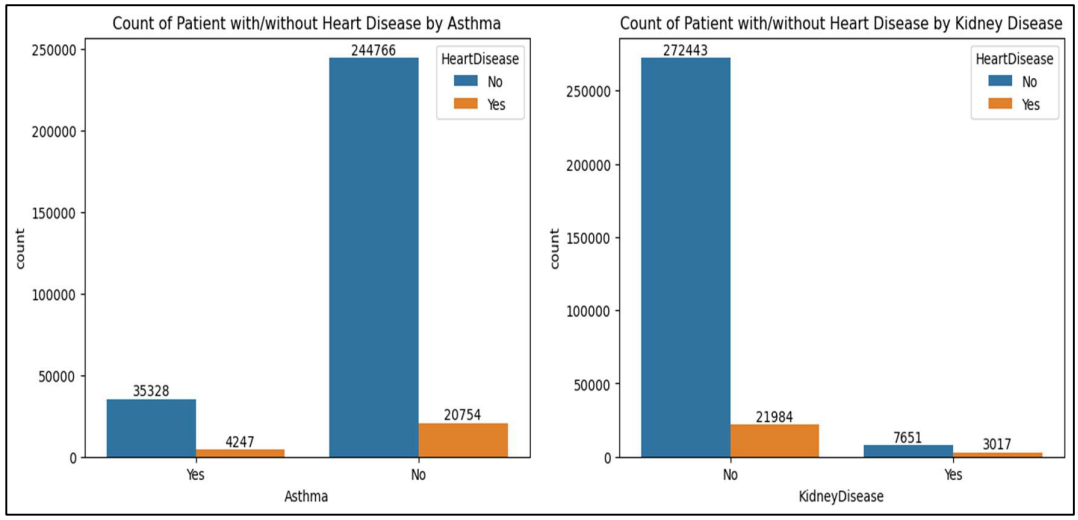
- Most people in the data are white and have no diabetic having heart disease.

VII. Bar plots for heart disease wise count of Physical Activity and General health



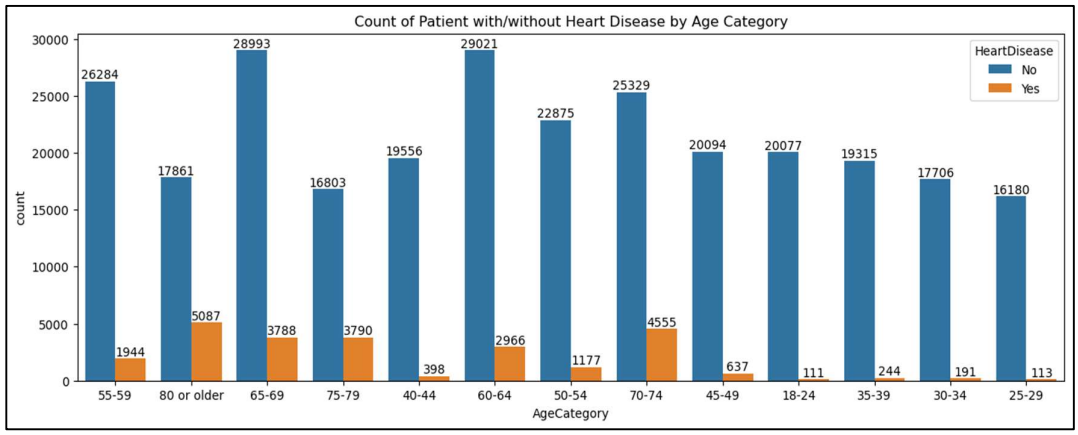
- Most people said that they have generally very good health. A few of people who said that they have generally a poor health.

VIII. Bar plots for heart disease wise count of Asthma and Kidney Disease And



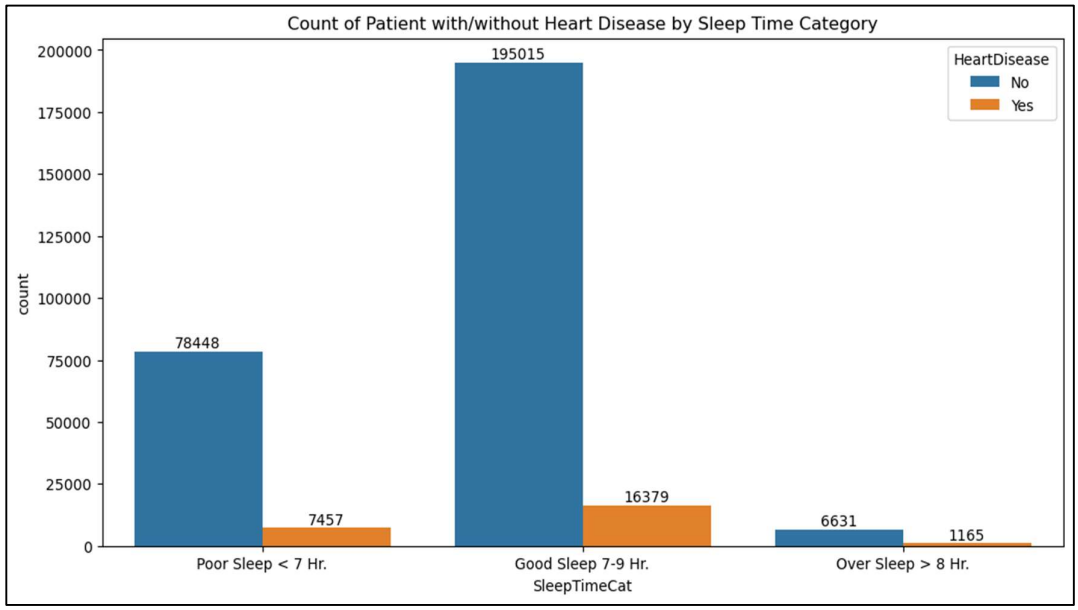
- A little of them who have asthma, kidney disease and skin cancer.

IX. Bar plots for heart disease wise count of Age Category



- Big factor in heart disease, as the amount of heart disease patients increases with age. The most susceptible people to the heart disease are people who are greater than 70 years old.

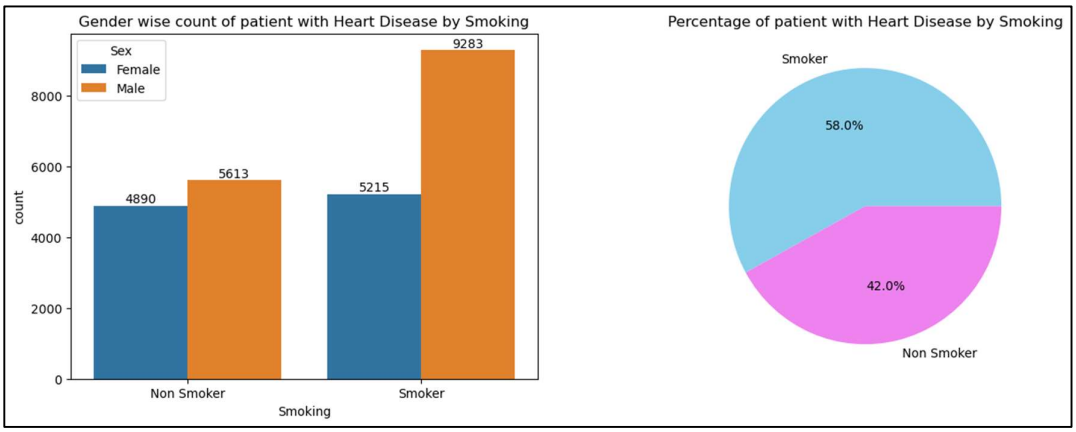
X. Bar plots for heart disease wise count of Sleep Time Category



- Oversleep reduces the chance of heart disease problem.

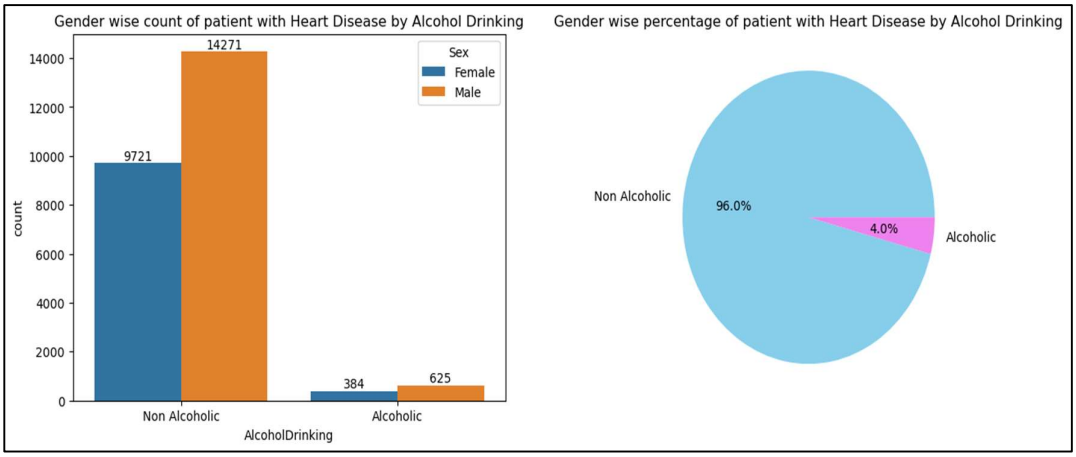
Analysing data of those , who has heart disease to understand what factors that effects on Heart Health

XI. Bar and pie chart for gender wise count and percentage of patients with heart disease by Smoking



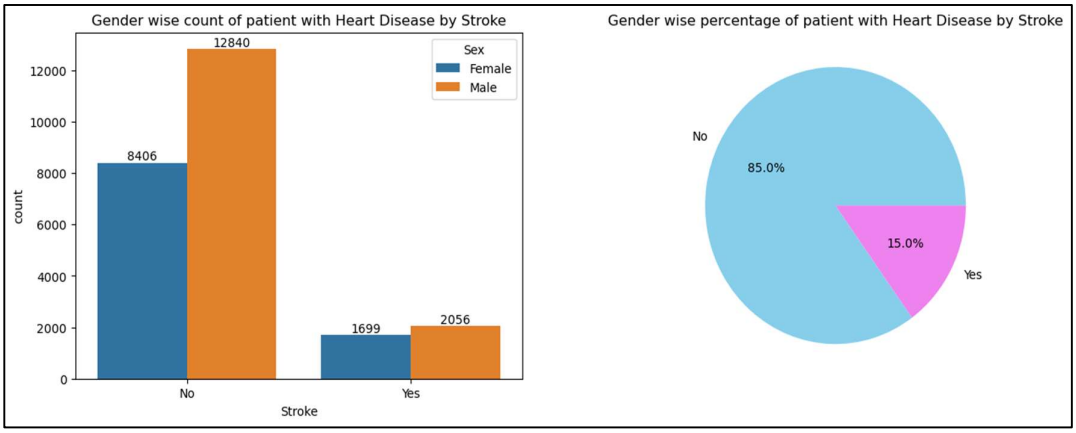
- observe that the people who are smoking are more susceptible to the heart disease.
- Heart disease problem observed more in Males as compared to Females who smoke.

XII. Bar and pie chart for gender wise count and percentage of patients with heart disease by Alcohol Drinking



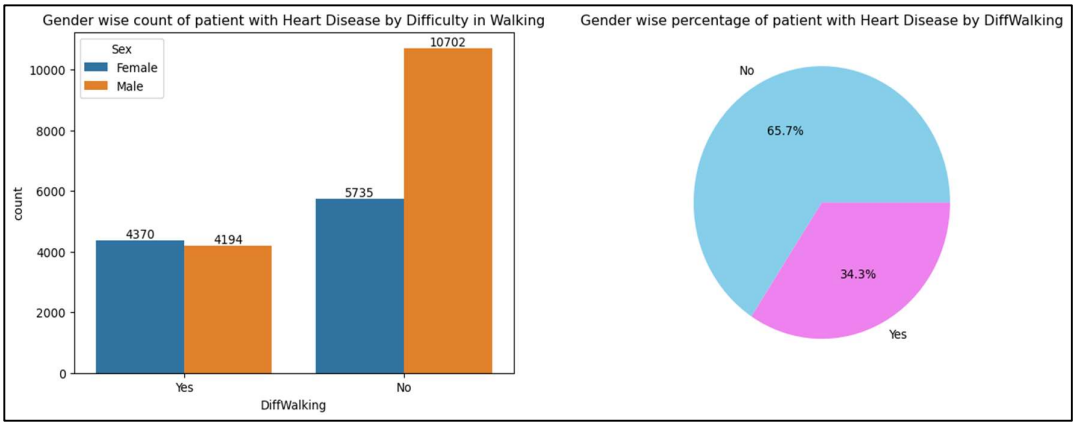
- Surprisingly those who consume alcohol have less chance of heart disease.
- Only 4% alcoholic people have heart disease.
- In alcohol category males has high count with heart disease.

XIII. Bar and pie chart for gender wise count and percentage of patients with heart disease by Stroke.



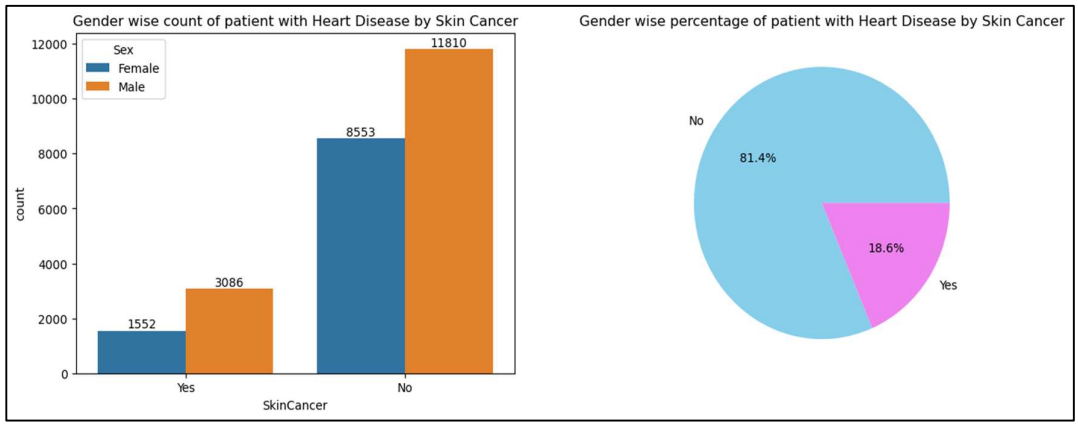
- Very few people have stroke and heart disease problem together.
- Count of Males are more than Females have stroke and heart disease Problem Together.

XIV. Bar and pie chart for gender wise count and percentage of patients with heart disease by difficulties in walking.



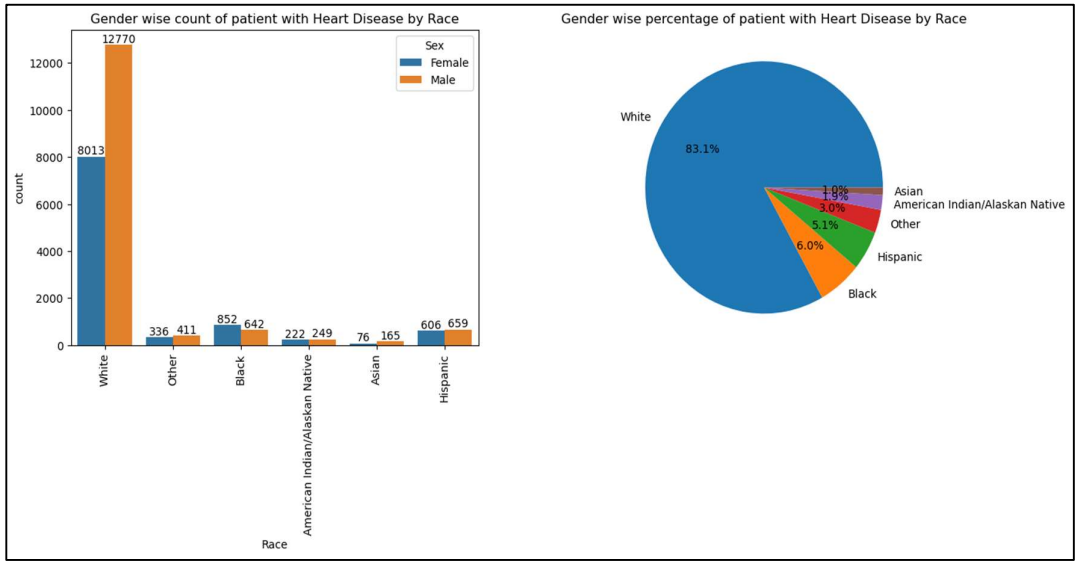
- Females with Difficulties in walking have heart disease problem and males with no difficulties in walking has high count of heart disease.

XV. Bar and pie chart for gender wise count and percentage of patients with heart disease by skin cancer



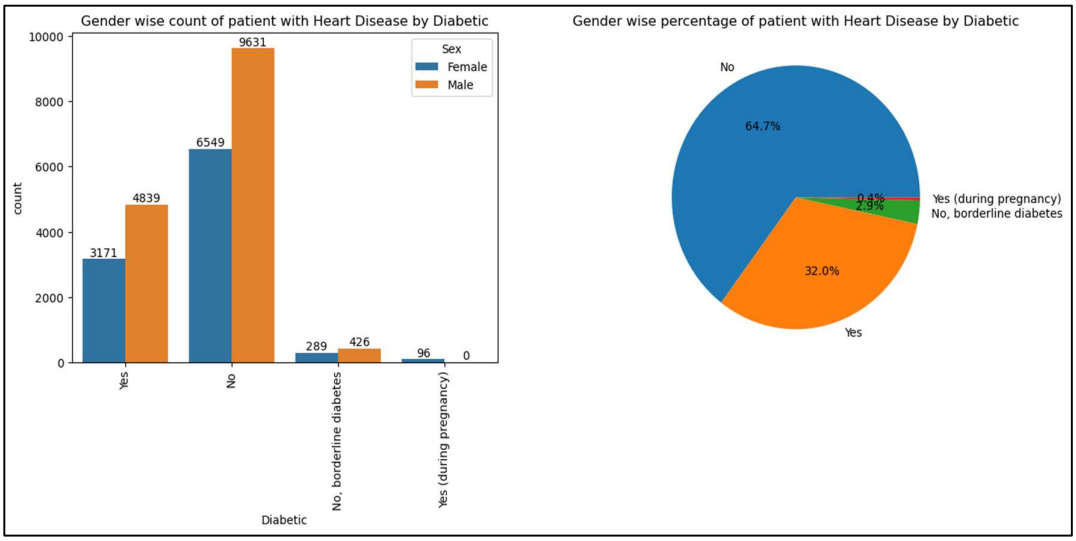
- 18.6% people have skin cancer and heart disease together.
- Males has high count of skin cancer.
- Above graph shows skin cancer slightly impacting on heart disease problem

XVI. Bar and pie chart for gender wise count and percentage of patients with heart disease by Race



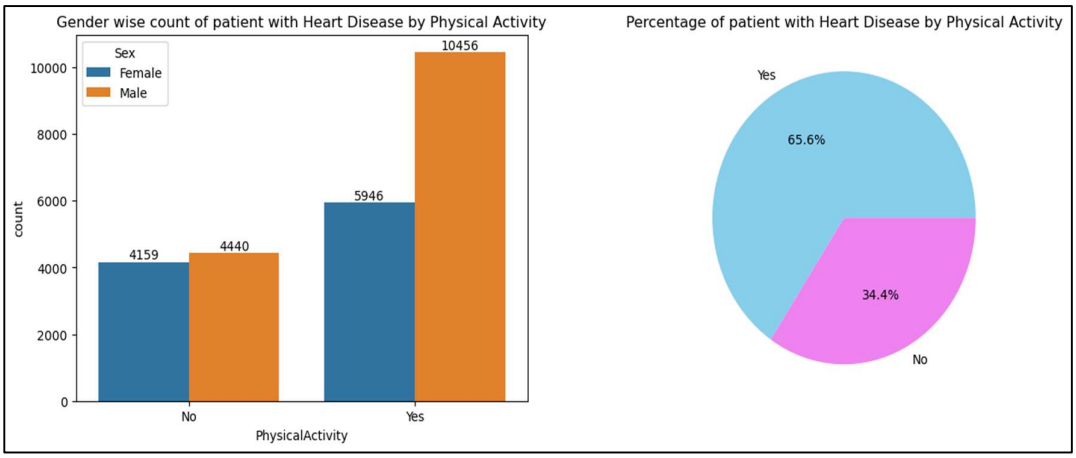
- 83.1% white people have high chances of heart disease.
- In the race category also males has high count with heart disease.
- Very few Asian has heart disease problem.

XVII. Bar and pie chart for gender wise count and percentage of patients with heart disease by Diabetic



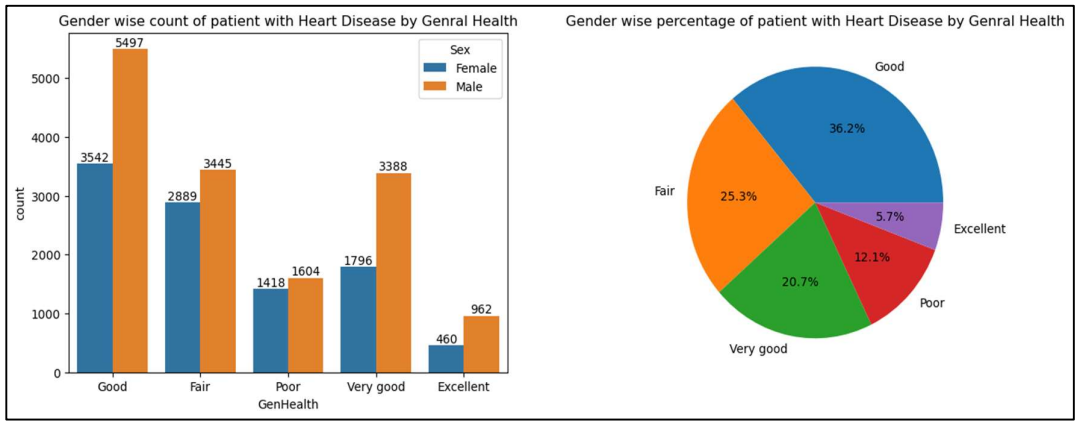
- 32% people have diabetics and facing heart disease problem.
- Males has high count of diabetic and heart disease problem together.
- Diabetics highly impact on heart disease problem.

XVIII. Bar and pie chart for gender wise count and percentage of patients with heart disease by physical activity.



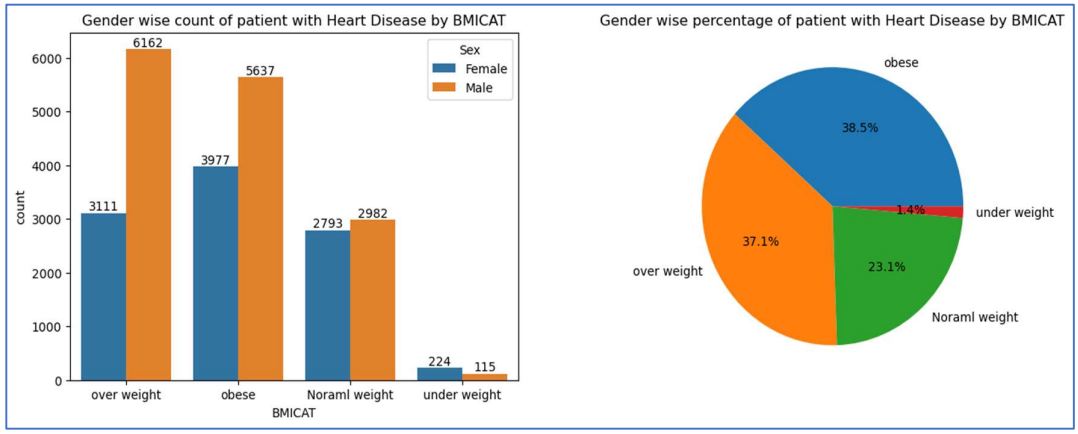
- Physical activity does not reduce the chance of heart disease problem.

XIX. Bar and pie chart for gender wise count and percentage of patients with heart disease by General Health



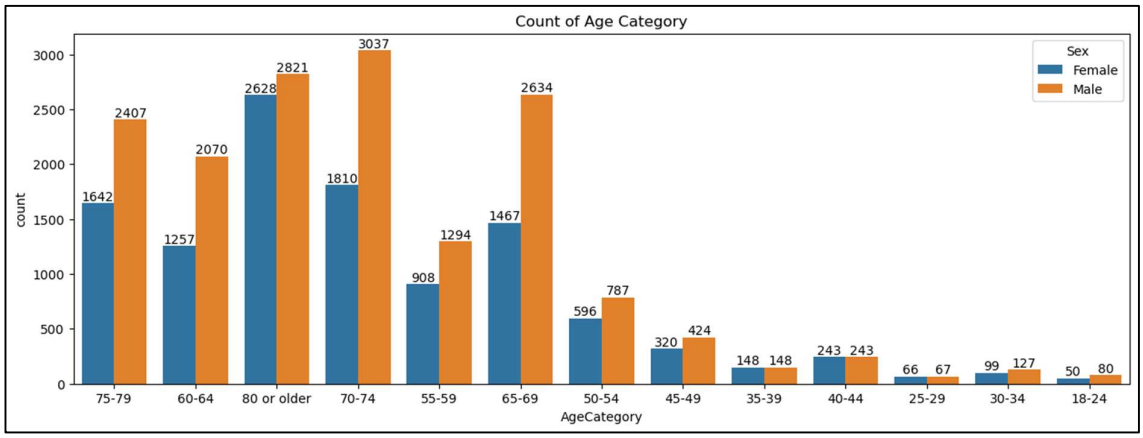
- Excellent General Health has low chances of heart disease.

XX. Bar and pie chart for gender wise count and percentage of patients with heart disease by BMI category



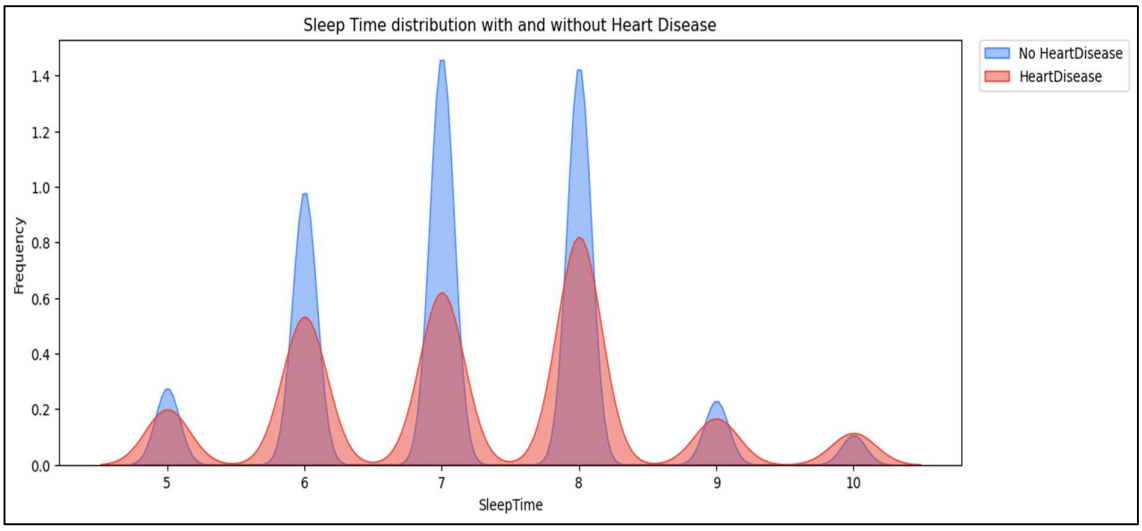
- Above plot shows BMI is highly impacting on Heart health
- Heart disease risk increases with BMI.
- Overweight has high numbers of heart disease count

XXI. Bar for gender wise count of patients with heart disease by Age Category



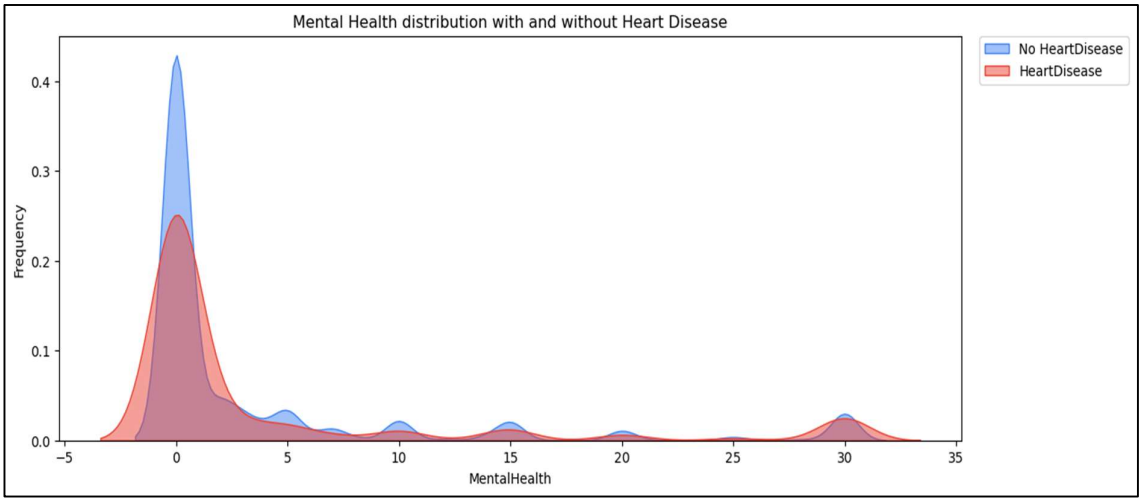
- From all the graphs presented, it can be concluded that alcohol consumption, smoking and age are the main factors in heart disease.
- Males are more susceptible to the heart disease.

XXII. Kdeplot (Kernel Distribution Estimation Plot) for Sleep time distribution with and without heart disease.



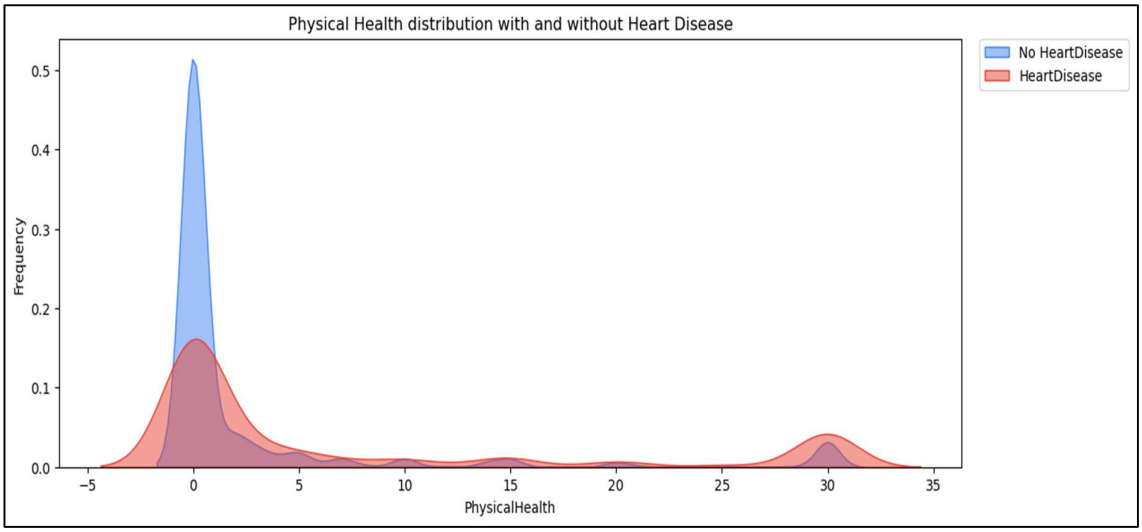
- Abnormal sleep duration is more prevalent in heart disease patients.

XXIII. Kdeplot (Kernel Distribution Estimation Plot) for Mental Health distribution with and without heart disease



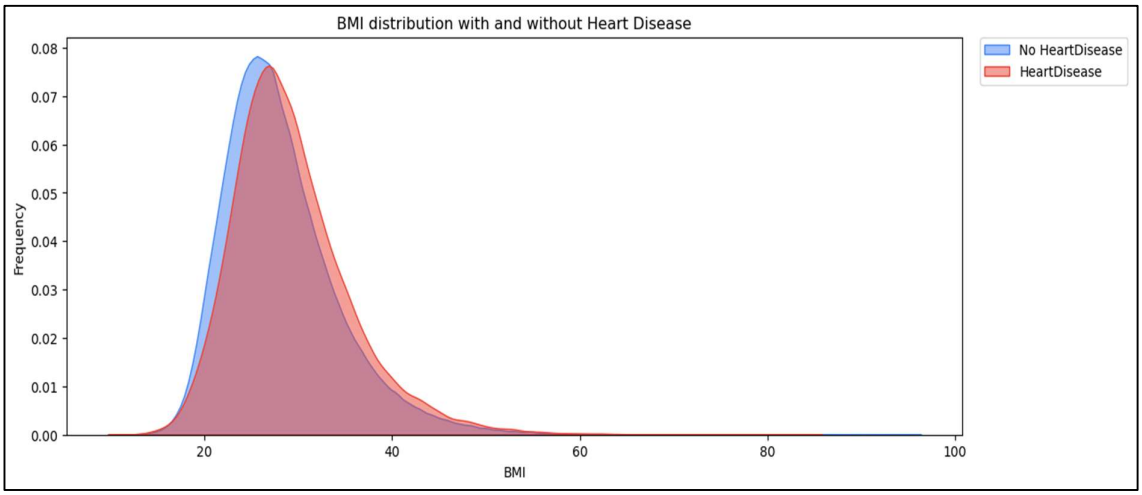
- Mental illness influences heart disease.

XXIV. Kdeplot (Kernel Distribution Estimation Plot) for Physical Health distribution with and without heart disease.



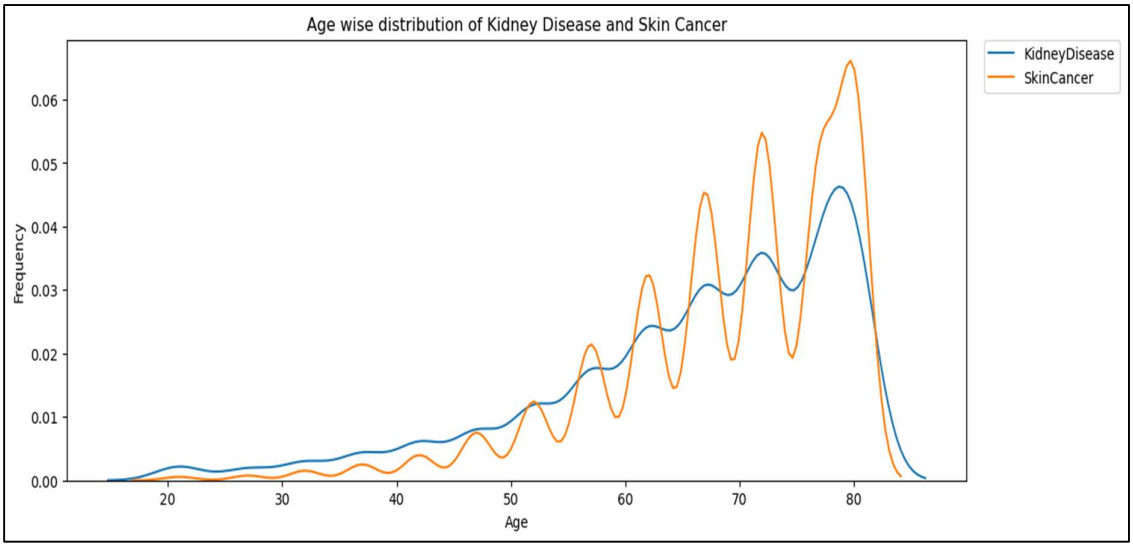
- Physical Health illness influences heart disease.

XXV. Kdeplot (Kernel Distribution Estimation Plot) for BMI distribution with and without heart disease



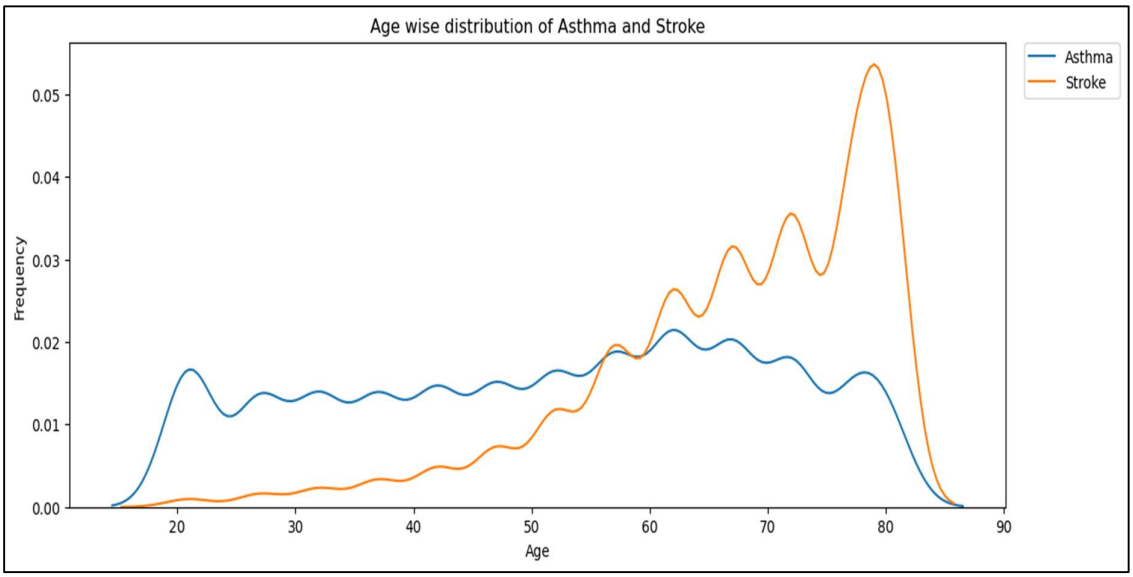
- Individuals with Heart Disease have a higher BMI than those do not have heart disease.

XXVI. Line plot for Age wise distribution of kidney disease and skin cancer.



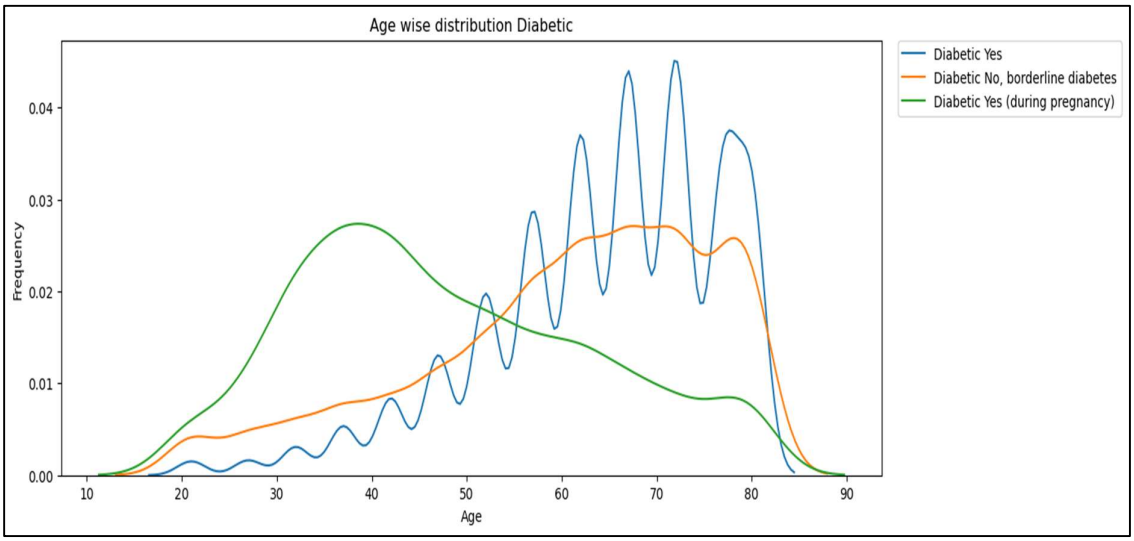
- Above graph shows kidney disease and skin cancer patients increases with increase in Age of an individual.

XXVII. Line plot for Age wise distribution of Asthma and Stroke



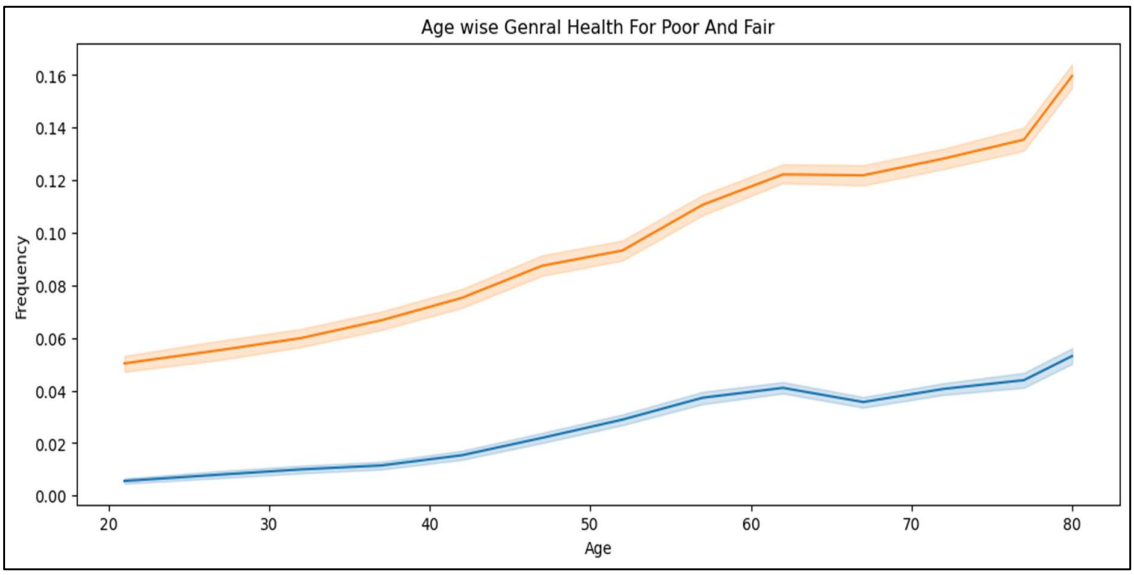
- Above graph shows Asthma and stroke patients increases with increase in Age of an individual

XXVIII. Line plot for Age wise distribution of Diabetes.



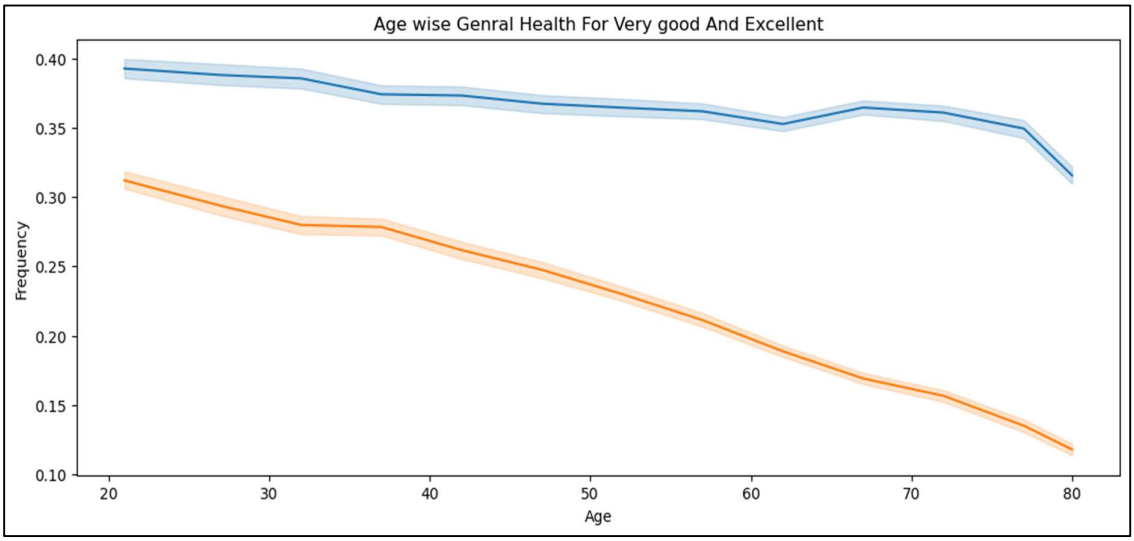
- Above graph shows Diabetic patients increases with increase in Age of an individual.

XXIX. Line plot for Age wise poor and fair General health



- Poor and fair health issues increase with age of an individual which is True.

XXX. Line plot for Age wise good and excellent general healths



- Above graph shows Good health and Excellent health decrease with age of an individual.

7. Power BI Dashboard

In this project, a Power BI dashboard is built for the "Personal Key Indicators of Heart Disease" dataset. This dashboard provides a visual representation of the data, allowing users to explore the relationships between various personal key indicators and heart disease. Here are some possible interpretations for the dashboard's diverse visualizations:

1. Bar chart: A bar used to compare the frequency of heart disease among different groups, such as males vs. females, heart disease (yes/no) This could help identify any significant differences between groups and inform targeted interventions or screening strategies.

2. Gauge: A gauge used to display a single personal key indicator, such average physical health , average mental health and indicate whether it falls within a healthy range or not. This could provide a quick and easy way to assess a patient's risk of heart disease.

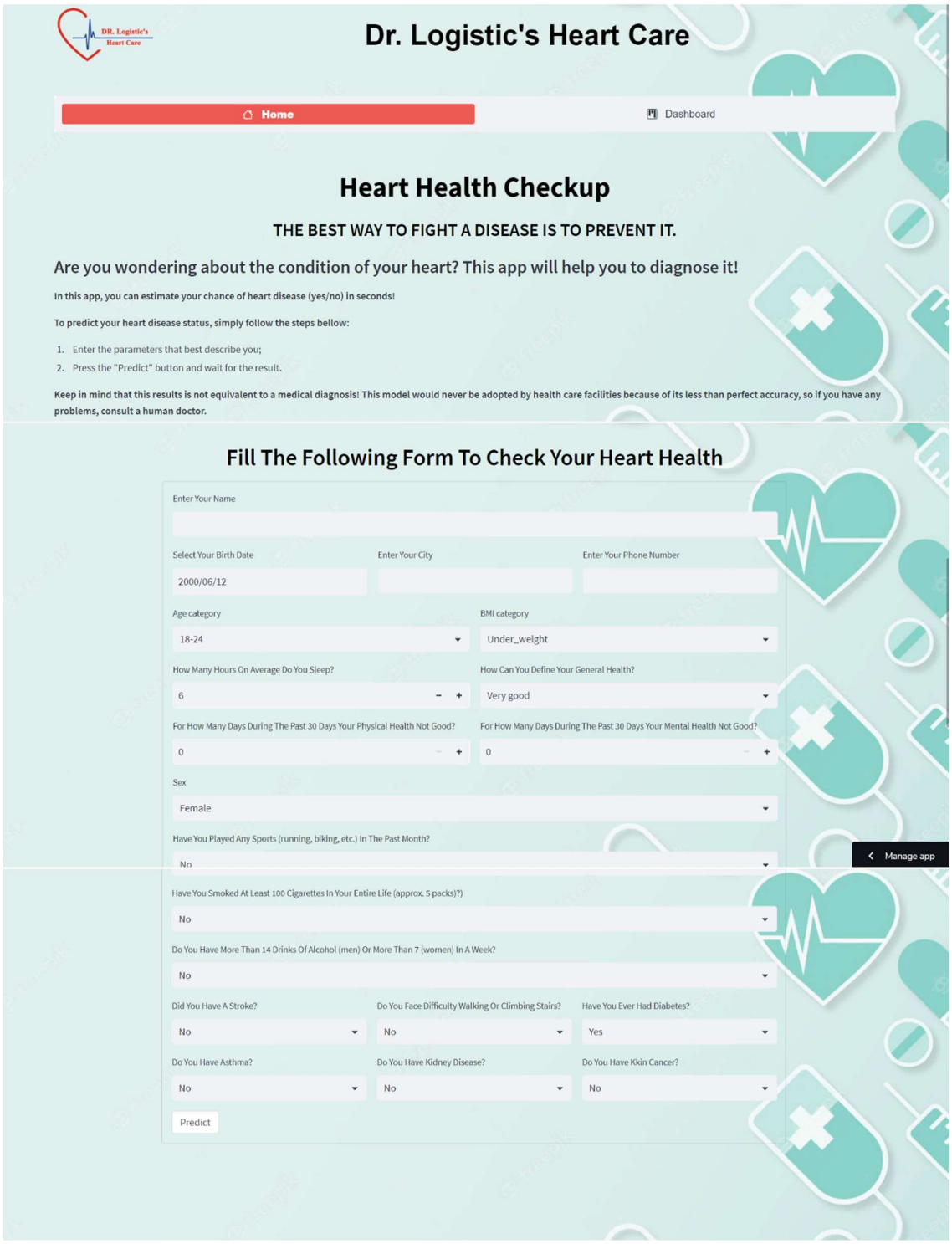
3. Cards : Cards used to display key summary statistics such as the total number of patients in the dataset, the number of patients with heart disease. cards provide a quick overview of the data and help contextualize the other visualizations in the dashboard.

4. Pie charts: Pie charts used to display the frequency of heart disease among different groups, such as males vs. females, smoker vs non-smoker with heart disease etc. The size of each slice is proportional to the number of patients in that group. This allows users to easily compare the prevalence of heart disease among different groups and identify any significant differences.

A Power BI dashboard provides a user-friendly interface for exploring the "Personal Key Indicators of Heart Disease" dataset and identifying patterns and relationships in the data. It is also used to track changes in key personal indicators over time and to inform the establishment of personalized treatment plans and lifestyle recommendations.

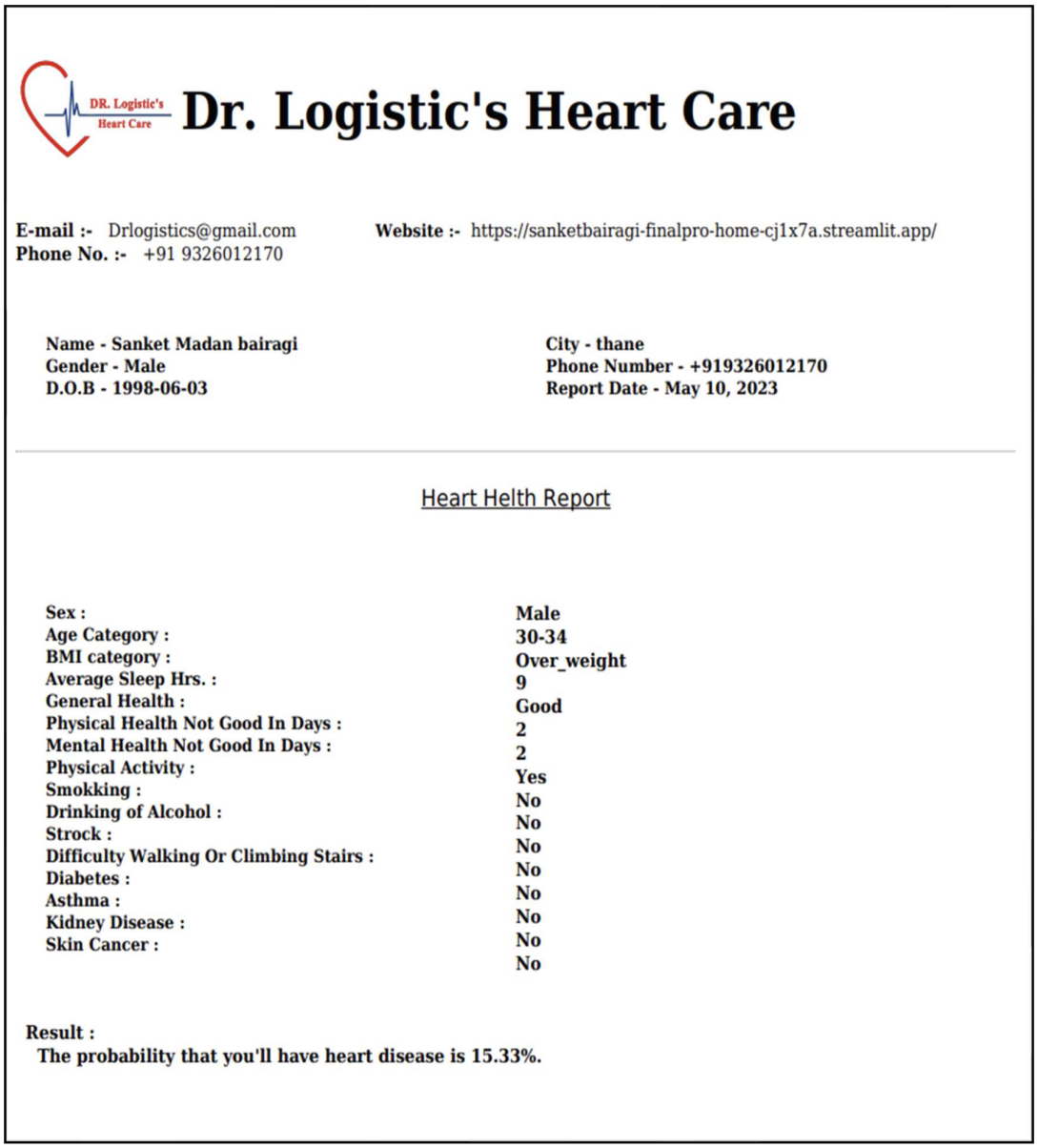
8. Stream lit Application.

In the project Streamlit application to predict heart disease using the "Personal Key Indicators of Heart Disease" dataset provides a user-friendly interface for healthcare professionals or patients to input personal key indicators and receive a risk prediction for heart disease. following are some potential interpretations for this type of application:



- Input fields: The application includes input fields for personal key indicators such as Age Category, sex, BMI category, Smoking habit , Alcohol Drinking habit, Physical Health, Mental Health , difficulty walking or climbing stairs, Diabetic, Physical Activity, General Health, Sleep Time, Asthma, Kidney Disease and Skin Cancer. These input fields given to saved machine learning model to get prediction in form of percentage.

- Prediction output: Once the user has entered their personal key indicators, the application generates a risk prediction for heart disease based on a machine learning model trained on the dataset. This prediction could be displayed as a details and percentage of risk of heart disease.



Dr. Logistic's Heart Care

E-mail :- Drlogistics@gmail.com **Website :-** https://sanketbairagi-finalpro-home-cj1x7a.streamlit.app/
Phone No. :- +91 9326012170

Name - Sanket Madan bairagi
Gender - Male
D.O.B - 1998-06-03

City - thane
Phone Number - +919326012170
Report Date - May 10, 2023

Heart Helth Report

Sex :	**Male**
Age Category :	**30-34**
BMI category :	**Over_weight**
Average Sleep Hrs. :	**9**
General Health :	**Good**
Physical Health Not Good In Days :	**2**
Mental Health Not Good In Days :	**2**
Physical Activity :	**Yes**
Smokking :	**No**
Drinking of Alcohol :	**No**
Strock :	**No**
Difficulty Walking Or Climbing Stairs :	**No**
Diabetes :	**No**
Asthma :	**No**
Kidney Disease :	**No**
Skin Cancer :	**No**

Result :
The probability that you'll have heart disease is 15.33%.

- This application also offers download option for reports in PDF format when the user clicks the predict button. This PDF contains the user information entered during prediction as well as the heart disease risk percentage that predicted during the prediction process.

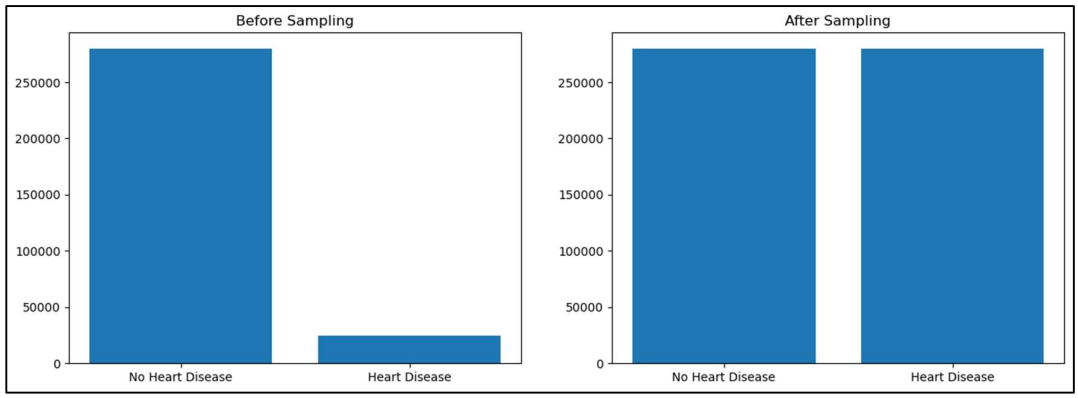
- This Streamlit application includes dashboard created on Power bi Application . dashboard could provide a user-friendly interface for exploring insights and help identify patterns or relationships in the analysis study so that user can understand what factors are important to keep heart healthy. It also used to monitor changes in personal key indicators over time and inform personalized treatment plans or lifestyle recommendations.

  This Streamlit application is able of predicting heart disease using the "Personal Key Indicators of Heart Disease" dataset and could be a useful tool for healthcare professionals or patients to assess their risk of heart disease and inform targeted interventions and lifestyle changes. The application might help users in understanding the data and factors that contribute to heart disease.

Machine Learning Modelling and Prediction

1. Sampling

Due to the imbalanced dataset, I have used SMOTE (Synthetic Minority Over-sampling Technique)sampling methods by considering the unequal distribution of the dependent variable. To improve the overall classification performance, oversampling was used to treat the imbalanced data. I have used the ROSE (Random Over Sampling Examples) package as it is a bootstrapping technique by random selection of the data to deal with the binary classification problems in the dataset.



Above graph shows before and after Sampling counts of dependent variable.

2. Scaling

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Before Scaling

PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma
560188.000000	560188.000000	560188.000000	560188.000000	560188.000000	560188.000000	560188.000000	560188.000000	560188.000000	560188.000000
4.858875	3.619772	0.143097	0.442703	7.668593	0.606152	0.647467	2.134944	7.192453	0.086003
9.510025	7.755592	0.350172	0.496707	3.241056	1.134487	0.477759	1.125881	1.063149	0.280369
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	5.000000	0.000000
0.000000	0.000000	0.000000	0.000000	6.000000	0.000000	0.000000	1.000000	6.298870	0.000000
0.000000	0.000000	0.000000	0.000000	8.000000	0.000000	1.000000	2.000000	7.000000	0.000000
3.509144	2.339668	0.000000	1.000000	10.000000	0.000000	1.000000	3.000000	8.000000	0.000000
30.000000	30.000000	1.000000	1.000000	12.000000	3.000000	1.000000	4.000000	10.000000	1.000000

Above image shows PhysicalHealth, MentalHealth, AgeCatgory, Diabetic, GenHealth, SleepTime is not equal range. To make this in range I have use Min-Max Scaling technique.

Min-max scaling, also known as normalization, is a data pre-processing technique used to rescale numeric features to a specific range. The purpose of this scaling is to bring all features to a similar scale, typically between 0 and 1.
The scaling equation for min-max scaling is as follows:
scaled_value = (original_value - min_value) / (max_value - min_value)

After applying min-max scaling, the transformed feature values will be within the range of 0 and 1. If the original value is equal to the minimum value, the scaled value will be 0.

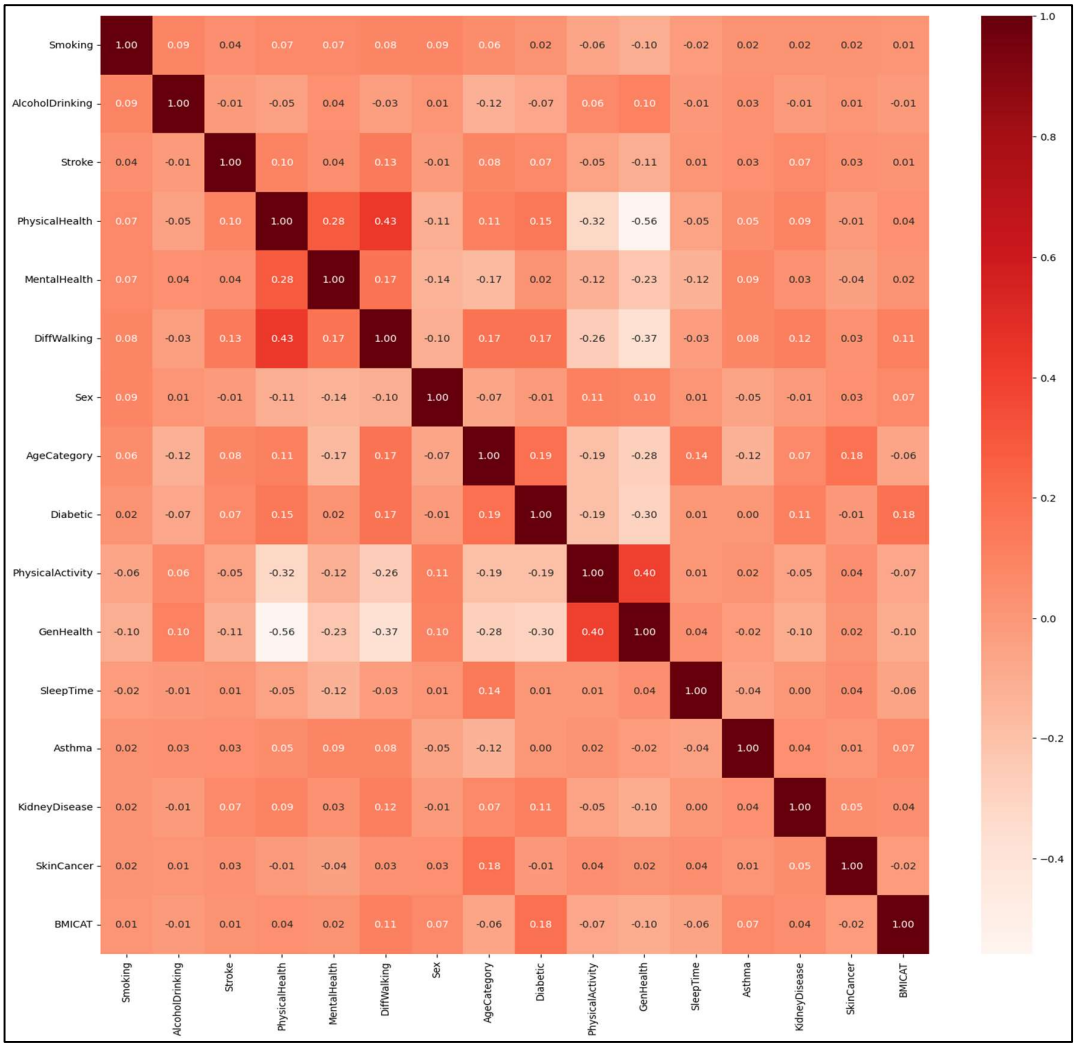
After Scaling -

PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma
448150.000000	448150.000000	448150.000000	448150.000000	448150.000000	448150.00000	448150.000000	448150.000000	448150.000000	448150.000000
0.162647	0.120983	0.142970	0.443135	0.639189	0.20172	0.647152	0.533461	0.438431	0.086054
0.317657	0.258975	0.350043	0.496756	0.269931	0.37785	0.477857	0.281674	0.212491	0.280444
0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.500000	0.00000	0.000000	0.250000	0.260245	0.000000
0.000000	0.000000	0.000000	0.000000	0.666667	0.00000	1.000000	0.500000	0.400000	0.000000
0.119666	0.078518	0.000000	1.000000	0.833333	0.00000	1.000000	0.750000	0.600000	0.000000
1.000000	1.000000	1.000000	1.000000	1.000000	1.00000	1.000000	1.000000	1.000000	1.000000

Above image shows PhysicalHealth, MentalHealth, AgeCatgory, Diabetic, GenHealth, SleepTime are in equal range.

3. Multicollinearity

Multicollinearity refers to a high correlation between two or more independent variables in a dataset. When variables are strongly correlated, it can pose challenges in interpreting the individual effects of the variables on the target variable. Moreover, multicollinearity can lead to unstable model coefficients and increase the uncertainty in the model's predictions. It is generally desirable to avoid multicollinearity or address it appropriately. Correlated independent variables can affect model performance in different ways. For example, in linear regression models, multicollinearity can inflate the standard errors of the coefficients, leading to imprecise estimates. In decision tree-based models, correlated variables may result in redundant splits, reducing the interpretability of the tree and potentially increasing overfitting. Additionally, correlated variables can impact feature importance rankings and the stability of the model's predictions. A correlation coefficient of 0 indicates no linear correlation, while values close to 1 or -1 indicate strong positive or negative correlation, respectively.



Above Heatmap shows, there is no Multicollinearity between independent variables.

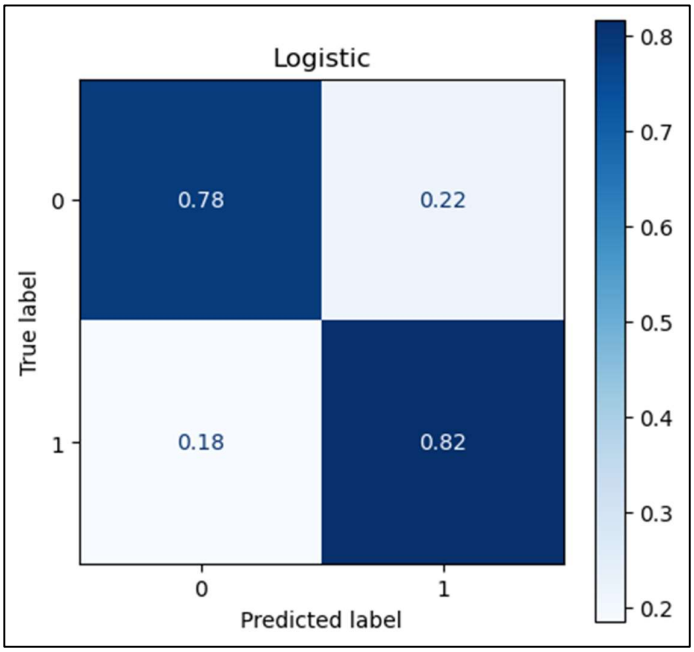
4. Training / Testing Model :

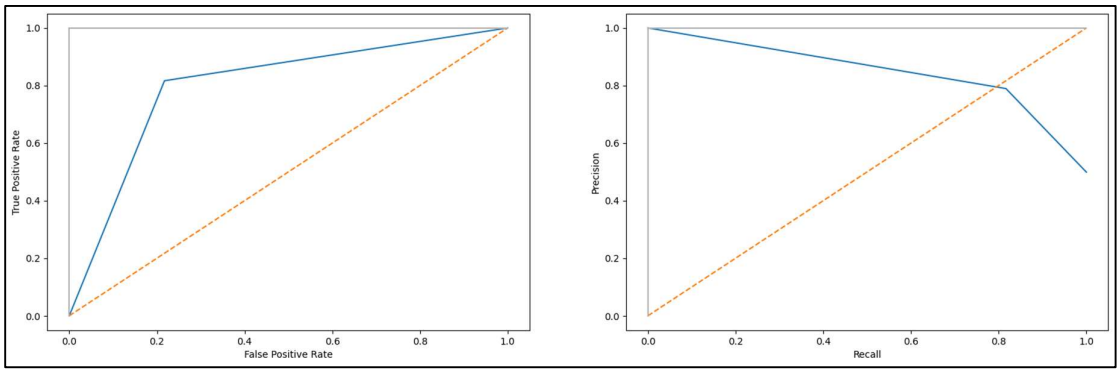
Using machine learning, I developed classification models, including logistic regression, random forest, decision tree, and SGD. I make non-sampled data in a sampled format, then divide the sampled dataset into 80% training data and 20% test data. After fitting the model to the training data, I examined the model's performance on the test data. Comparative performance metrics for these models have been calculated. The target variable is the heart disease variable, which indicates whether or not an individual has heart disease.

Logistic Regression:

Logistic regression allows for the analysis of dichotomous or binary outcomes with two mutually exclusive levels. The inclusion of continuous or categorical variables is permitted, and logistic regression allows for the adjustment of multiple factors. And after filtering and altering the equilibrium, this makes logistic regression particularly useful for the analysis of observational data, particularly when adjustments are needed to reduce the possibility of bias due to differences between the groups being compared.

Using the confusion matrix, I calculated the sensitivity, specificity, accuracy, and test error values shown in the results. then I have subsequently plotted ROC curve.





The AUC (Area Under the Curve) and ROC (Receiver Operating Characteristic) curve are commonly used evaluation metrics for binary classification models like logistic regression.

The ROC curve is a graphical representation of the performance of a binary classifier system. It illustrates the trade-off between the true positive rate (TPR or sensitivity) and the false positive rate (FPR or 1-specificity) as the classification threshold is varied. The curve is created by plotting TPR against FPR for various threshold values.
The x-axis represents the false positive rate (FPR), which is calculated as FPR = FP / (FP + TN), where FP is the number of false positives and TN is the number of true negatives.
The y-axis represents the true positive rate (TPR), also known as sensitivity or recall. It is calculated as TPR = TP / (TP + FN), where TP is the number of true positives and FN is the number of false negatives.
A good classifier has a higher TPR and a lower FPR, resulting in an ROC curve that hugs the top-left corner of the plot. The area under the ROC curve (AUC) is used as a summary statistic of the classifier's performance. It ranges from 0 to 1, with a higher value indicating better performance. An AUC of 0.5 indicates a random classifier, while an AUC of 1 represents a perfect classifier.

The AUC is a numerical value representing the area under the ROC curve. It provides a single measure of how well the classifier can distinguish between positive and negative instances. As mentioned earlier, it ranges from 0 to 1, with 0.5 indicating a random classifier and 1 representing a perfect classifier.

The AUC is advantageous because it is threshold independent. It considers the overall performance of the classifier across all possible classification thresholds. It is commonly used when comparing different classifiers or when optimizing the performance of a classifier by adjusting the classification threshold.

	prob	accuracy	sensi	speci
0.0	0.0	0.499054	1.000000	0.000000
0.1	0.1	0.667782	0.986175	0.350592
0.2	0.2	0.738687	0.961065	0.517149
0.3	0.3	0.776942	0.924955	0.629488
0.4	0.4	0.796382	0.878311	0.714762
0.5	0.5	0.799720	0.815302	0.784196
0.6	0.6	0.789027	0.733944	0.843902
0.7	0.7	0.762054	0.628351	0.895252
0.8	0.8	0.709679	0.479101	0.939385
0.9	0.9	0.623244	0.267362	0.977782

In logistic regression, sensitivity (also known as true positive rate or recall) and specificity are common evaluation metrics used to assess the performance of a binary classification model.

Sensitivity measures the proportion of true positive cases that are correctly identified by the model. It is calculated as follows:
Sensitivity = TP / (TP + FN)
where TP represents the number of true positives and FN represents the number of false negatives.
Sensitivity tells us how well the model identifies the positive cases out of all the actual positive cases. A higher sensitivity indicates a lower rate of false negatives, meaning the model is good at capturing positive instances.

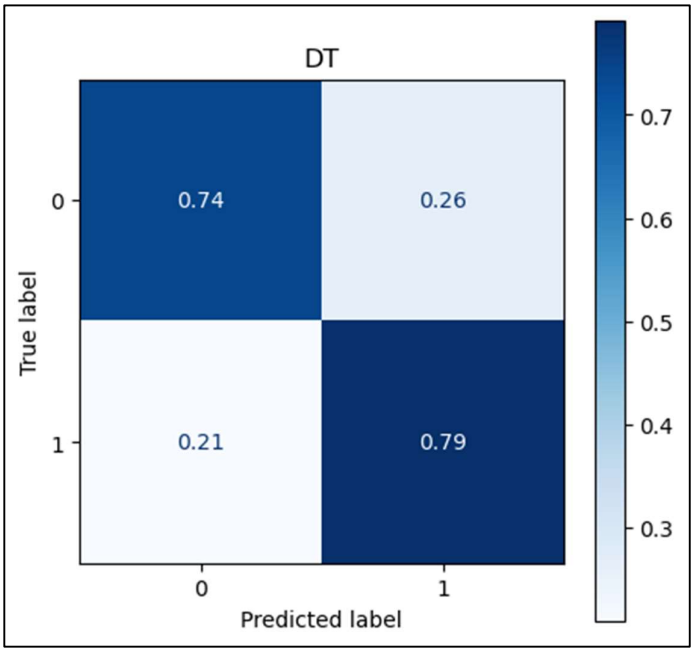
Specificity measures the proportion of true negative cases that are correctly identified by the model. It is calculated as follows:
Specificity = TN / (TN + FP)
where TN represents the number of true negatives and FP represents the number of false positives.
Specificity tells us how well the model identifies the negative cases out of all the actual negative cases. A higher specificity indicates a lower rate of false positives, meaning the model is good at correctly identifying negative instances.

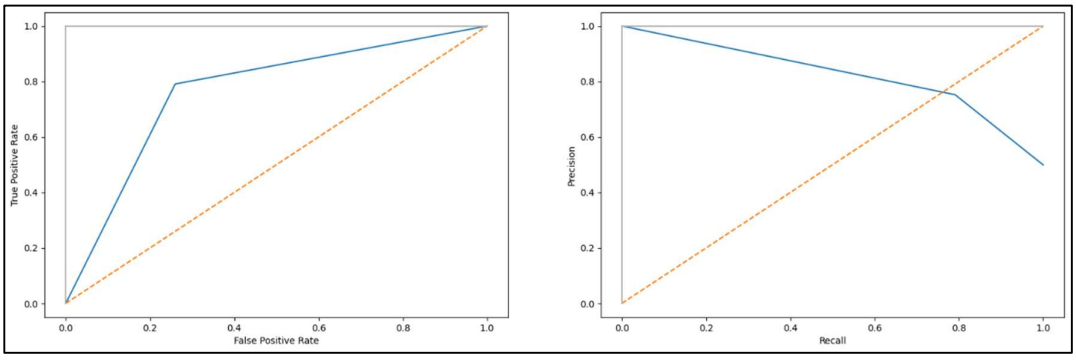
Both sensitivity and specificity provide valuable information about a logistic regression model's performance. However, it's important to note that they are often inversely related. Increasing sensitivity might lead to a decrease in specificity, and vice versa, depending on the classification threshold used. The choice of the threshold affects the trade-off between these two metrics. In the above image threshold set at 0.5 has sensitivity 81.53% and specificity is 78.41%.

Logistic regression gives accuracy 79.97%.

Decision Tree:

Compared to other machine learning models, Decision Tree models are relatively easy to comprehend and interpret, and the fact that there are nonlinear relationships between parameters does not affect the tree's performance. In contrast, this model contributes to the issue of overfitting and generates asymmetrical trees. A number of data variables have a particularly strong relationship with the dependent variable.

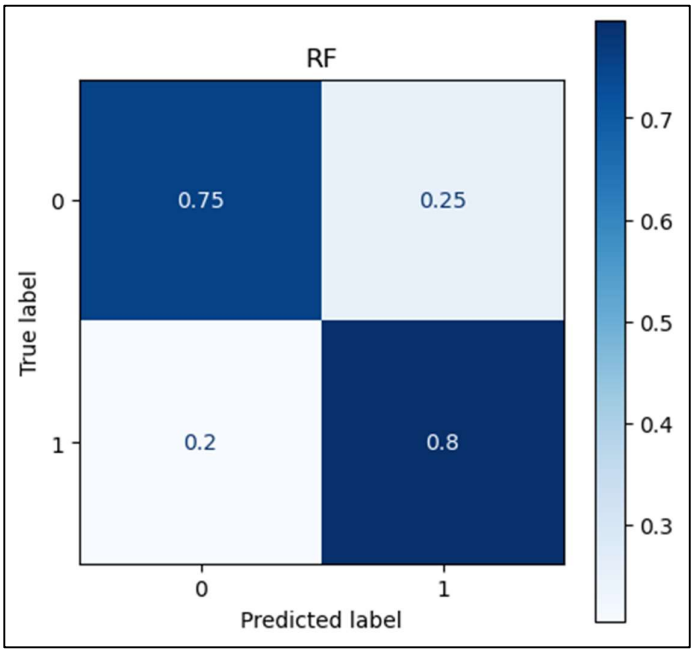


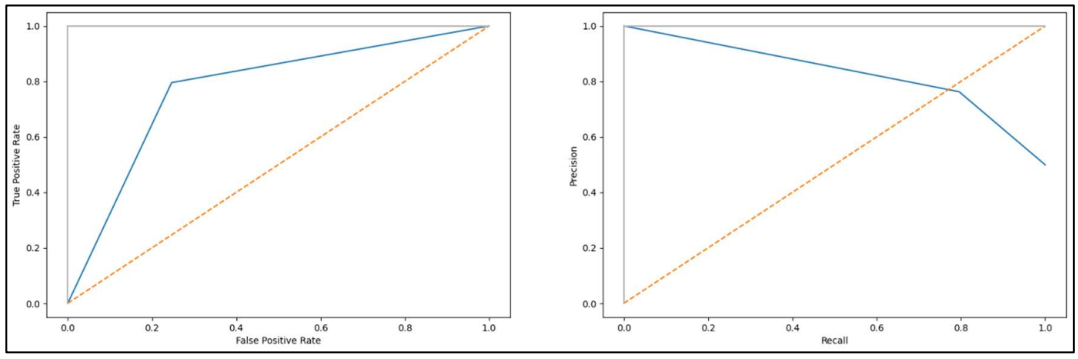


The DT model train score is 76.7% and the test score is 76.6%. accuracy of DT model is 76.56 %.

Random Forest:

Random Forest is an ensemble learning method that combines multiple decision trees to improve performance and reduce overfitting. It can be useful for datasets with a large number of features, such as the heart disease dataset. Random Forest could capture complex non-linear relationships between personal key indicators and heart disease. It could also provide feature importance measures to understand which personal key indicators have the strongest impact on the prediction.



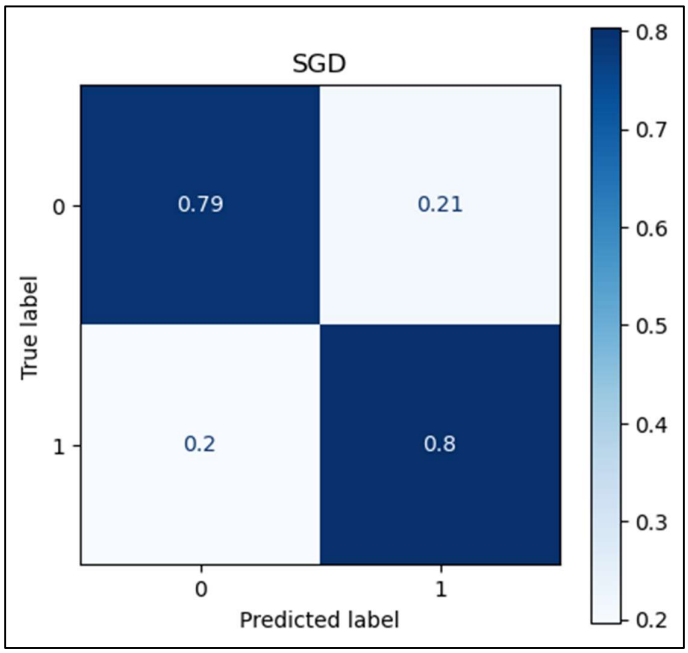


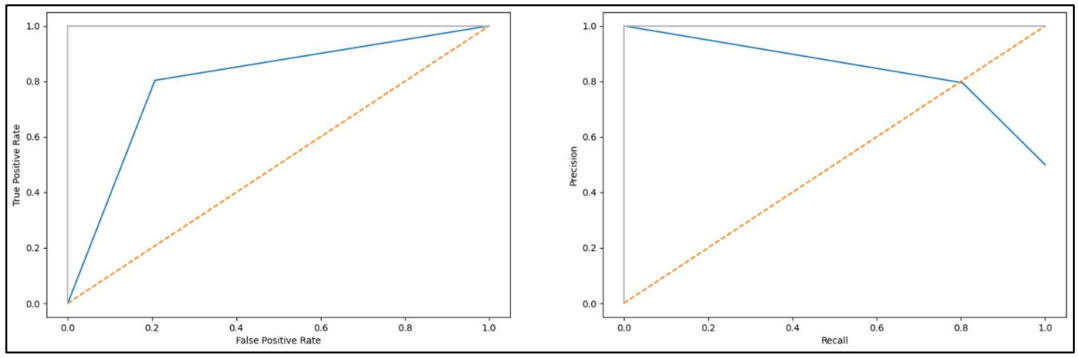
The model train score for Random Forest is 76.6% and the test score is 77.5%.
Accuracy of random Forest Model is 77.45%.

Stochastic Gradient Descent (SGD) :

Stochastic Gradient Descent (SGD) is a commonly used optimization algorithm for training linear models. It is useful for datasets with a large number of features, as it can train models quickly and efficiently. However, SGD can be sensitive to feature scaling and may require some pre-processing steps to ensure optimal performance.

In this project SGD algorithm perform well with good training and testing score.





In this SGD model , train score is 80.10% and the test score is 79.8 %.
Accuracy of SGD model is 79.85%.

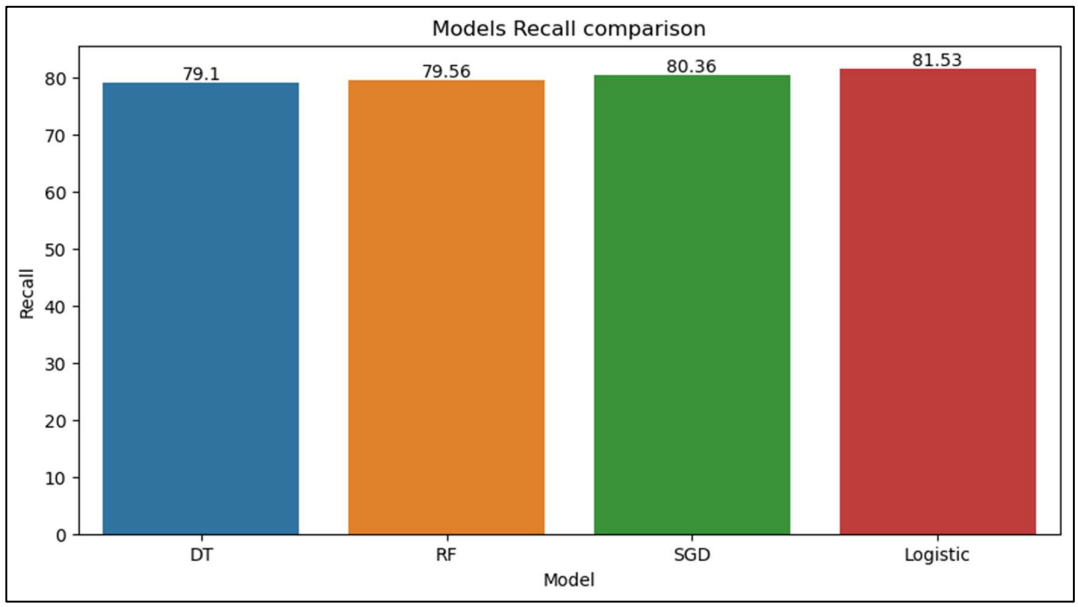
5. Comparison Of Algorithms :

Recall :

In the above training and testing model section Logistic Regression perform well compared to SGD , Random Forest and Decision Tree algorithm. Following bar graph sows more details and helps to compare all model Recall.
The recall is the measure of model correctly identifying True Positives. Thus, for all the patients who have heart disease, recall tells us how many correctly identified as having a heart disease. Therefore, recall is the best measure for comparing models.

A high recall value indicates a low rate of false negatives, meaning the model is effective at capturing most positive instances. On the other hand, a low recall value indicates a higher rate of false negatives, suggesting that the model is missing a significant number of positive instances.

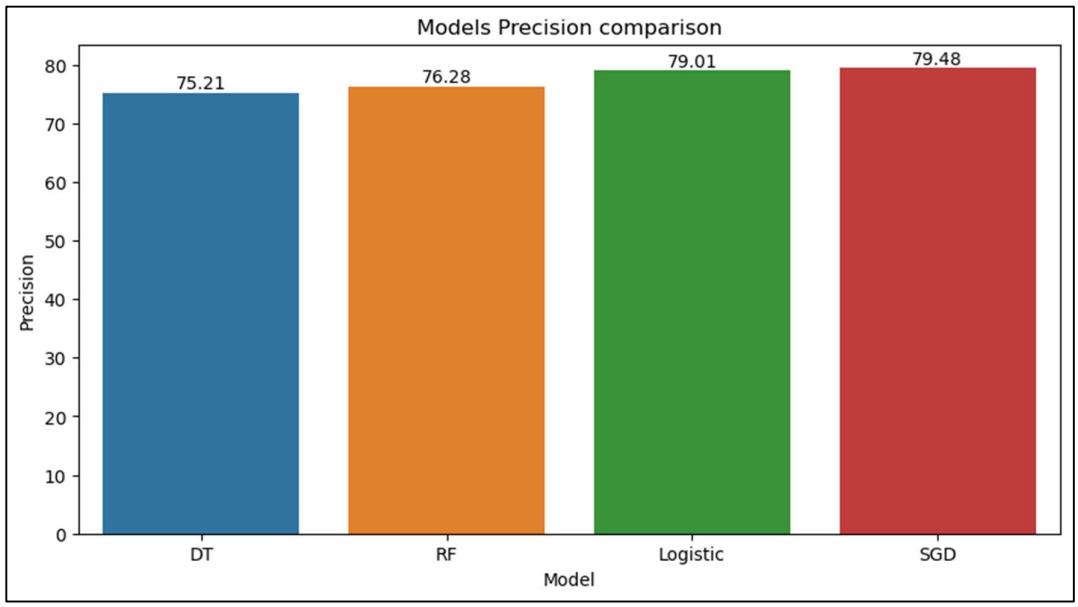
Recall is commonly used in situations where missing positive instances is considered more critical than incorrectly identifying negative instances as positive. For example, in medical diagnosis, it is important to have high recall to avoid missing potential cases of a disease, even if it leads to more false positives.



Form above bar graph Logistic Regression shows high recall . Therefore, I have chosen Logistic regression Model as final model with highest recall as well as good score of training and testing as compared to SGD and Decision Tree algorithm and Random Forest.

Precision :

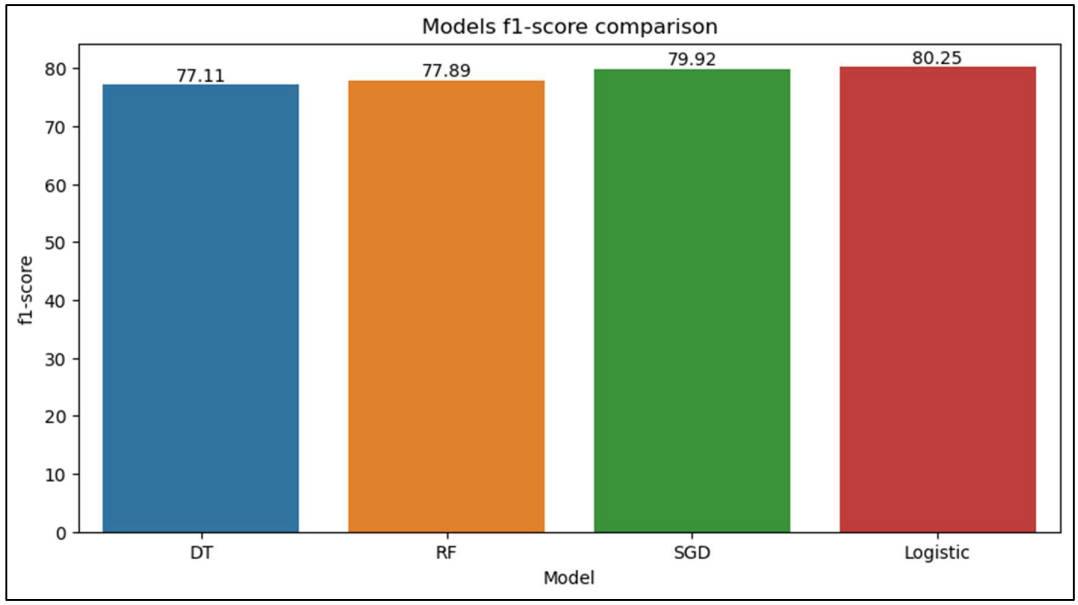
When comparing machine learning models, precision provides a measure of their performance in terms of positive prediction accuracy. A higher precision indicates better performance in accurately identifying positive instances.
Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive by the model. Precision is then calculated using the formula: Precision = TP / (TP + FP)
Precision quantifies the model's ability to make accurate positive predictions. It focuses on minimizing false positives, meaning instances that are predicted as positive but are actually negative.
A high precision value indicates a low rate of false positives, suggesting that the model is accurate and reliable in identifying positive instances. On the other hand, a low precision value indicates a higher rate of false positives, indicating that the model is incorrectly identifying negative instances as positive.



Above graph shows that SDG model has high precision followed by Logistic regression. However, precision should not be considered in isolation. It should be used in conjunction with other evaluation metrics such as recall, F1 score, or accuracy to get a comprehensive understanding of a model's performance.

F1-score :

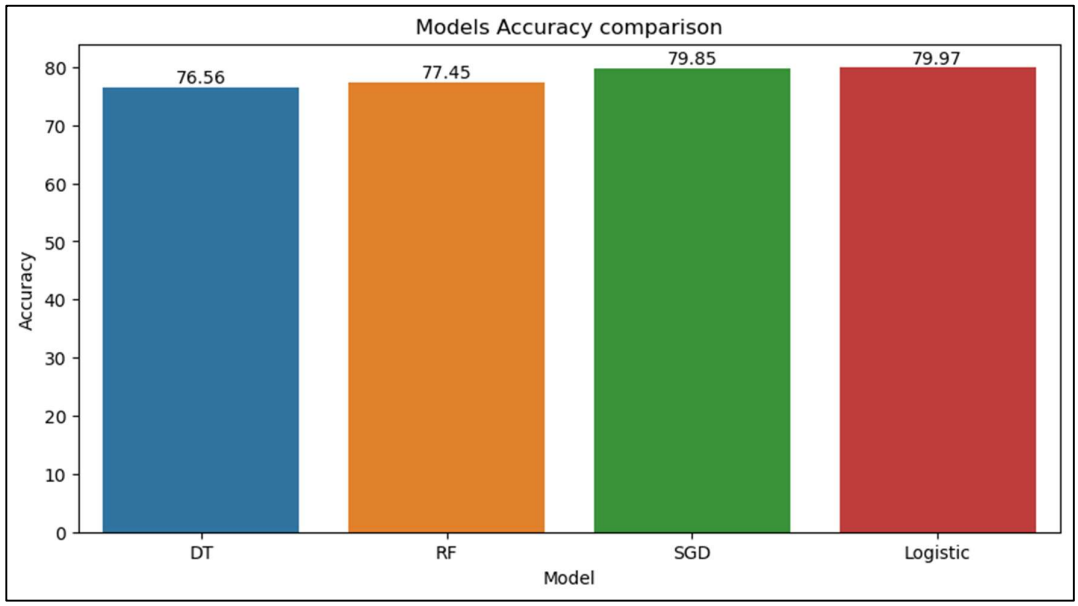
the F1 score is an evaluation metric that combines precision and recall to provide a single measure of a classification model's performance. It is particularly useful when dealing with imbalanced datasets or when both false positives and false negatives are important.
The F1 score is the harmonic mean of precision and recall, combining both measures into a single value. It provides a balanced assessment of the model's performance by considering both the model's ability to avoid false positives (precision) and its ability to avoid false negatives (recall). The F1 score ranges from 0 to 1, where 1 represents the best possible score (perfect precision and recall) and 0 represents the worst score (either precision or recall is 0).
The F1 score is commonly used in scenarios where false positives and false negatives have different costs or implications, such as in medical diagnosis, fraud detection, or information retrieval systems.



F1 score is only suitable for binary classification problems. Above graph shows that Logistic Regression has high F1-score as compared to other 3 model.

Accuracy :

Accuracy is a commonly used evaluation metric in machine learning to assess the performance of a classification model. It measures the proportion of correctly predicted instances out of the total number of instances in the dataset. Accuracy provides an overall measure of how well the model performs in terms of both positive and negative predictions. It considers the correct predictions for both classes and is suitable for balanced datasets where the classes are represented equally. A high accuracy value indicates that the model is making correct predictions for a large proportion of the instances in the dataset. However, accuracy may not be a reliable metric when dealing with imbalanced datasets, where one class significantly outnumbers the other. In such cases, the model might achieve high accuracy by simply predicting the majority class for most instances while performing poorly on the minority class.



When comparing different machine learning models, accuracy can provide a useful initial indication of their overall predictive performance.in above graph Logistic regression has high accuracy as compared to other models.

6. Results for training and testing :

After training and testing of all models , following results are generated which is shown in image of table below.

	Model	Train Score	Test Score	Recall	Precision	f1-score	Accuracy
0	DT	0.767	0.766	79.10	75.21	77.11	76.56
1	Logistic	0.802	0.800	81.53	79.01	80.25	79.97
2	SGD	0.801	0.798	80.36	79.48	79.92	79.85
3	RF	0.776	0.775	79.56	76.28	77.89	77.45

The above table shows the training score ,testing score , recall , Precision and F1 Score of all four models that trained. In logistic regression performs best amongst all other algorithms.

So, this is enough evaluation for to choose final model for the deploy machine learning application .

7. Deployment of Streamlit application –

Streamlit is a popular open-source framework used for model deployment by machine learning and data science teams. And the best part is it's free of cost and purely in python.

To deploy a machine learning application in Streamlit, I followed the following general steps:

machine learning model - After training and testing, logistic regression shows the best accuracy, recall, and F1 score, so I have saved the logistics regression model in a Pickle file with the extension ".sav".

Streamlit app file: I have created a Python file (app.py) where I did basic and advanced coding for the Streamlit application. This application starts with code for taking input from the user. Then input values are converted into numerical format with if-else conditions. This Python file also includes data pre-processing, feature extraction, or any other necessary steps before feeding the data to your model.

Loading a trained model: after pre-processing the input data, I have loaded a trained machine learning model into the Streamlit app file. all inputs given to the model to get predictions.

Deploying the Streamlit app on the Streamlit cloud: Streamlit works by reading code directly from a public GitHub repository. So to make the application work, I have made a repository that includes an app.py file, a train model, and a requirements.txt file that specifies the Python packages Streamlit needs to install for the app to run. After logging in to the steramlit cloud server, I have given access and the path of the repository.

Conclusion

This project on the "Personal Key Indicators of Heart Disease" dataset has provided valuable insights into the factors that contribute to the risk of heart disease. The project has included exploratory data analysis, machine learning classification models, interactive visualizations using Power BI and Streamlit application that predicts the risk of heart disease.

Exploratory data analysis has following points that found in as a conclusion of risk of heart disease.

1. heart disease affects roughly 8 out of every 100 people.
2. Patients with heart disease have a marginally higher BMI than healthy individuals.
3. Individuals who are older are more susceptible to heart disease.
4. A lot more people who suffer from heart disease say they have poor or fair health compared to those who don't.
5. Abnormal sleep duration is more prevalent in heart disease patients.
6. Diabetes increases the risk of heart disease by 25%.
7. People with asthma have a modestly increased risk of heart disease.
8. People with a history of skin cancer have a moderately increased risk of heart disease.
9. The mental health, sleep duration, and physical wellness of individuals with different diseases are equivalent.
10. Observe that those who smoke are more susceptible to heart disease.

On the basis of personal key indicators, machine learning models including Logistic Regression, Random Forest, Decision Tree, and SGD were trained to predict the presence of heart disease. The logistic regression model performed well as compared to other machine learning models.

The Power BI dashboard provided a visually appealing and interactive interface for exploring the data and gaining insight into the relationships between various personal key indicators and heart disease. The dashboard included a card visualization to display key metrics, such as the number of individuals with heart disease, and a pie chart to show the distribution of heart disease cases. bar charts for counts of people with heart disease according to different key indicators.

The Streamlit application demonstrated how the machine learning model could be used in a real-world scenario to provide risk predictions for heart disease based on personal key indicators. The application could be modified to include various input fields or models based on the user's specific requirements. The Streamlit application to predict heart disease could provide a user-friendly interface for healthcare professionals or patients to input personal key indicators and receive a heart disease risk prediction.

Overall, the initiative has provided significant insights into the risk factors for heart disease and demonstrated the potential of machine learning models to accurately predict heart disease risk based on personal key indicators. Individuals can enhance their heart health and reduce their risk of heart disease by implementing the interventions and lifestyle modifications recommended by this project.

Future Work

Following areas of future work can enhance the accuracy and effectiveness of the project in predicting the risk of heart disease based on personal key indicators and provide more comprehensive and personalized healthcare solutions.

1. Model Optimization: The current machine learning models can be optimized further by fine-tuning hyperparameters and exploring other model architectures, such as neural networks, to improve their performance.

2. Additional Data: More data can be collected to supplement the existing dataset, which can increase the accuracy and reliability of the models in predicting heart disease risk.

3. Domain-Specific Models: Different models can be trained for specific subgroups of the population, such as males and females, or people of different age ranges, to capture differences in the relationships between personal key indicators and heart disease risk.

4. Real-Time Monitoring: A real-time monitoring system can be developed using the models trained in this project to continuously monitor and predict the risk of heart disease in individuals, which can facilitate early interventions and improve health outcomes.

Reference

- Data Set : https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

- Additional Data & Information –
    1. https://www.cdc.gov/heartdisease/risk_factors.htm
    2. https://towardsdatascience.com/heart-disease-prediction-73468d630cfc

- GitHub Repository Link - https://github.com/SanketBairagi/FinalPro

- Application Link - https://sanketbairagi-finalpro-home-cj1x7a.streamlit.app/

- Application QR Code -