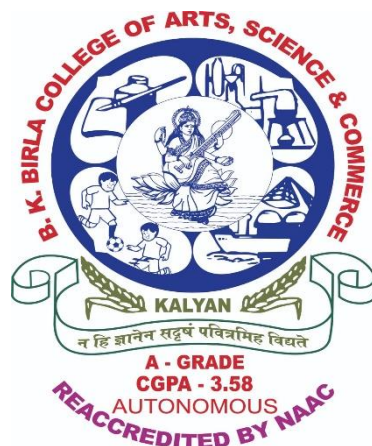


**B. K. Birla College of Arts, Science & Commerce
(AUTONOMOUS)**

Kalyan (W.)

Department of IT

**Master of Science in Data Science & Big Data Analytics
(M.Sc. DSBDA) – 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

Signature of External

Date: _____

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

Sr. No.	Topic	Pg. No.
1	Abstract	4
2	Introduction	5
3	Objective	6
4	Dataset	7-8
5	Methodology <ul style="list-style-type: none">• Data Cleaning• Data Exploration• Feature Engineering• Data Visualization• Power BI Dashboard• Stream lit Application.	9-22
6	Machine Learning Modelling and Prediction <ul style="list-style-type: none">• Training / Testing Model• Comparison Of Algorithms• Results of training and testing	23-29
7	Conclusion <ul style="list-style-type: none">• Data Analysis• Streamlit Application	30
8	Future Work	31
9	Reference	31

List Of Figures

Sr. No.	Figure Name	Page No.
1	Data frame view	5
2	View of Null values in dataset	6
3	Summary of the data	7
4	Data After Feature Engineering	8
5	Count and percentage of heart disease patients	9
6	Bar plot of BMI category wise Count of heart disease patients	9
7	Bar plots for heart disease wise count of Smoking , Alcohol Drinking , stroke ,Difficult in walking ,Sex and Skin cancer	10
8	Bar plots for heart disease wise count of Race , Diabetic ,Physical Activity and General health	11
9	Bar plots for heart disease wise count of Asthma ,Kidney Disease And Age Category	12
10	Bar plot for Sex wise count of Age Category	12
11	Kdeplot (Kernel Distribution Estimation Plot) for heart disease wise sleep time distribution.	13
12	Kdeplot (Kernel Distribution Estimation Plot) for heart disease wise Mental health distribution.	13
13	Kdeplot (Kernel Distribution Estimation Plot) for heart disease wise Physical Health distribution.	14
14	Kdeplot (Kernel Distribution Estimation Plot) for heart disease wise BMI distribution.	14
15	Image of PowerBI Dashboard	15
16	Image of Application	17
17	Image of pdf generated by application (Image of Health Report)	18

18	Image of Dashboard of application	19
19	Image of Confusion matrix of Logistic regression Algorithm	20
20	Image of ROC curves for Logistic regression Algorithm	21
21	Image of Confusion matrix of Decision Tree Algorithm	22
22	Image of Confusion matrix of Random Forest Algorithm	23
23	Image of Confusion matrix of SGD Algorithm	24
24	Bar Plot For recall comparison for Logistic regression, Decision Tree ,Random Forest And SGD Algorithms	25
25	Image of scores of Logistic regression, Decision Tree ,Random Forest And SGD Algorithms	26
26	Bar code For Application	29

Abstract

Heart disease is currently the leading cause of death in the world. According to the Center 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 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 [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
```

3. Feature Engineering

Features are distributed as continuous and categorical based on the class that is numeric or character, leading us to categorize the data and store them at respective levels using factors. The vectors of type string and integer for the unique values here are converted into numeric, with levels "Yes" or "No" having labels of 1 and 0. In the data, it is clearly seen that the variables AgeCategory has factors with 13 levels; Race has 6 levels, GenHealth has 5 levels; Diabetic has 4 levels; and all the other variables consist of factors with 2 levels. The correlation values of the numeric variables help to visualize the data further and predict the closeness of the truth value.

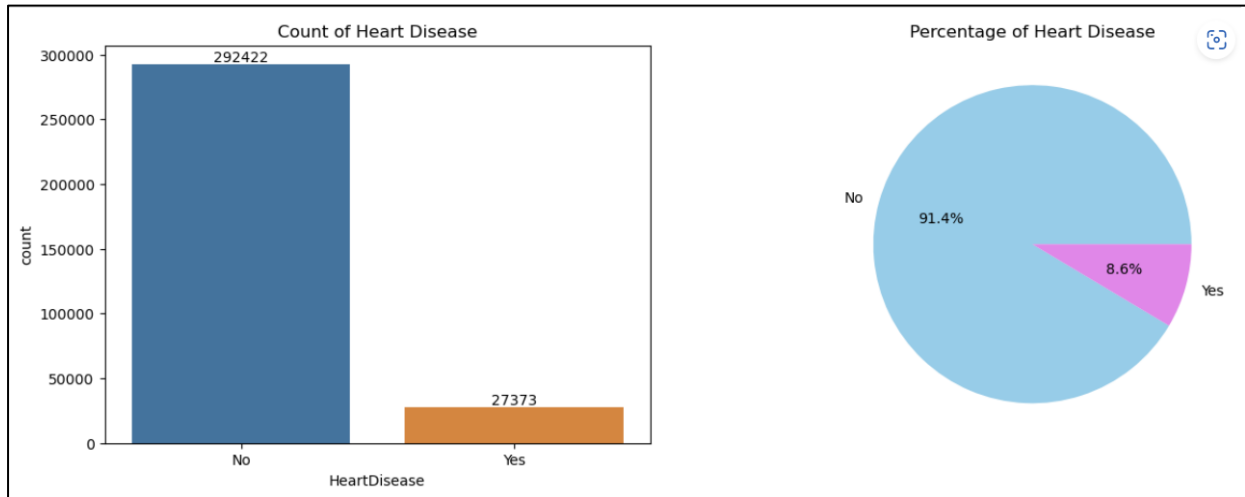
Data After Feature Engineering :

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

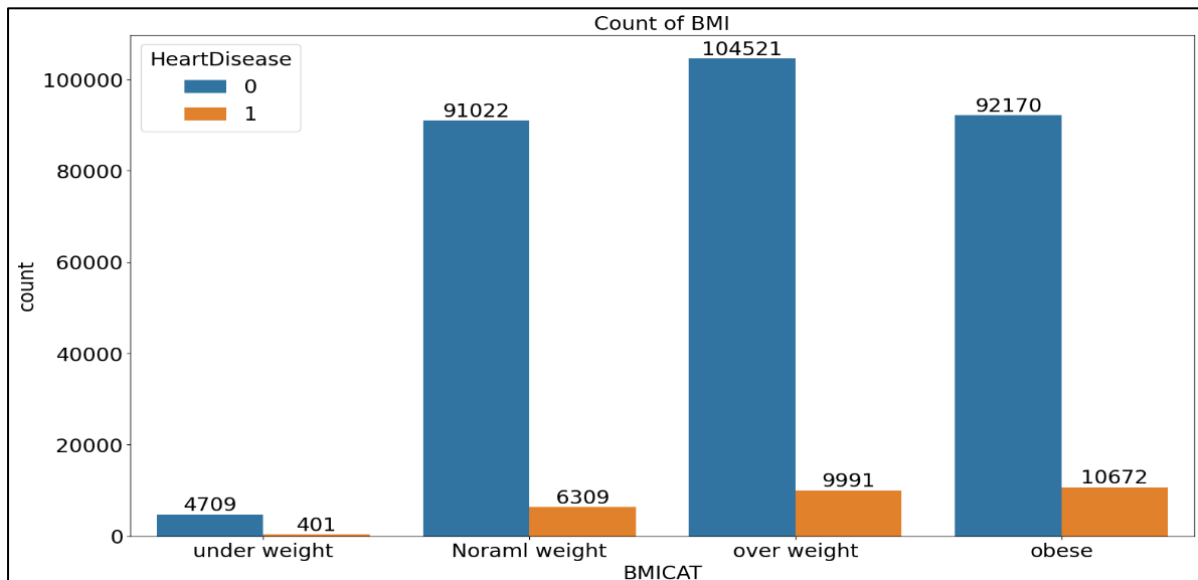
	HeartDisease	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer	BMICAT
0	0	1	0	0	3.0	30.0	0	0	7	3	1	3	5.0	1	0	1	0
1	0	0	0	1	0.0	0.0	0	0	12	0	1	3	7.0	0	0	0	1
2	0	1	0	0	20.0	30.0	0	1	9	3	1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

4. Data Visualization

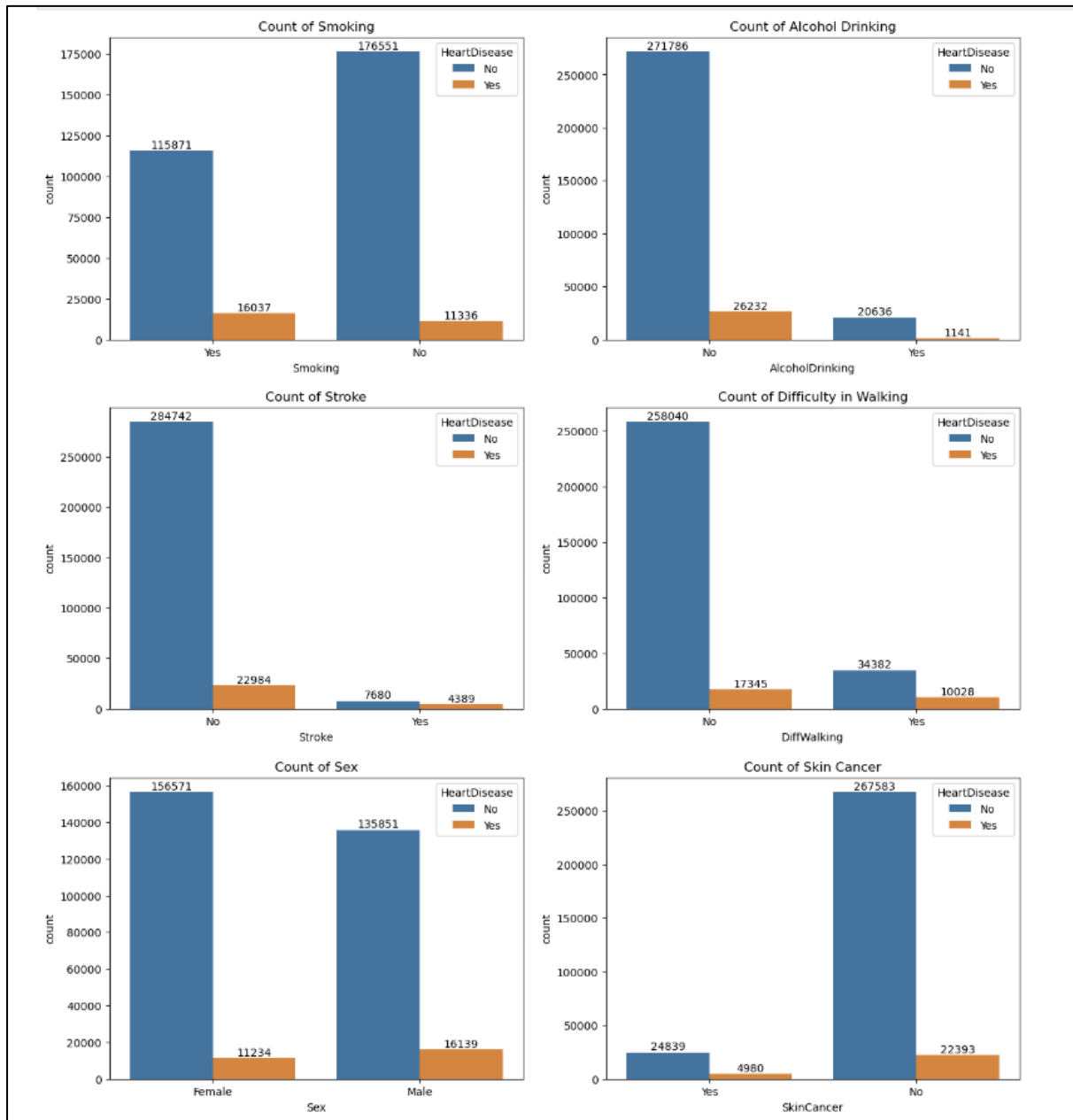
The first visualization aid I created was a pie chart. This diagram displayed the percentage of people with heart disease from the sample.



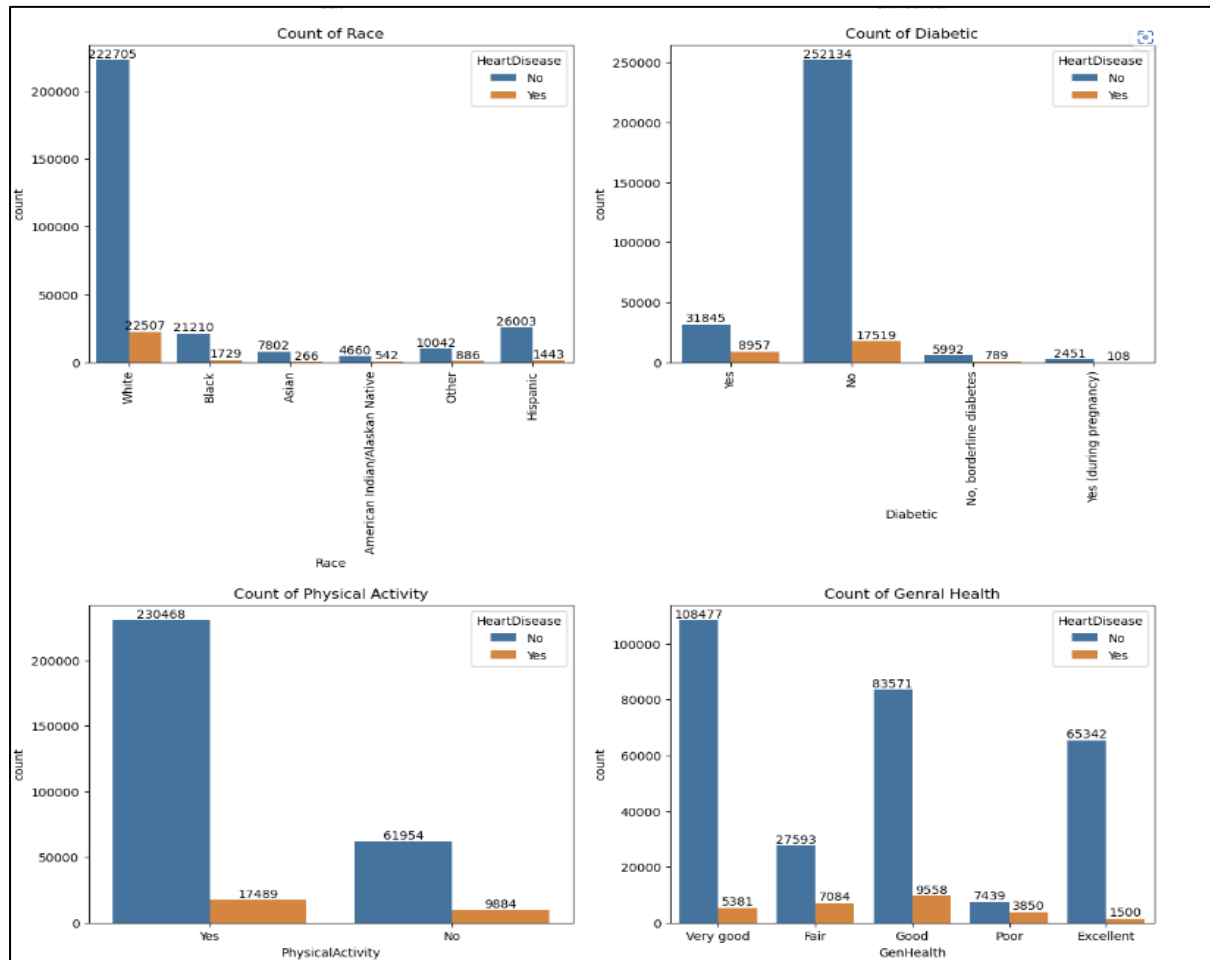
This pie chart reveals that only 8.56 % of people suffers from heart disease, while 91.4% do not.



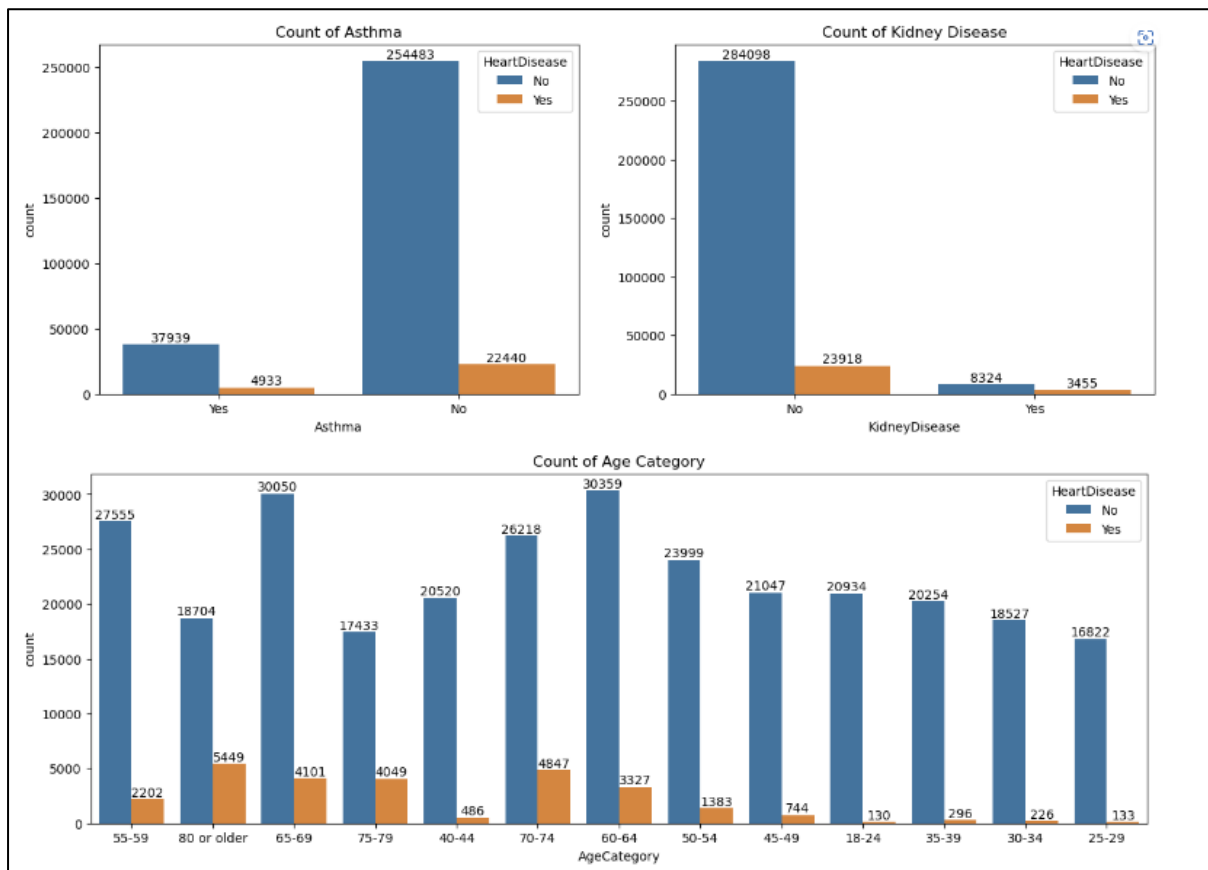
Heart disease is increase with BMI as above graph shows number of heart disease count increase with BMI , Obese has higher number of heart disease.



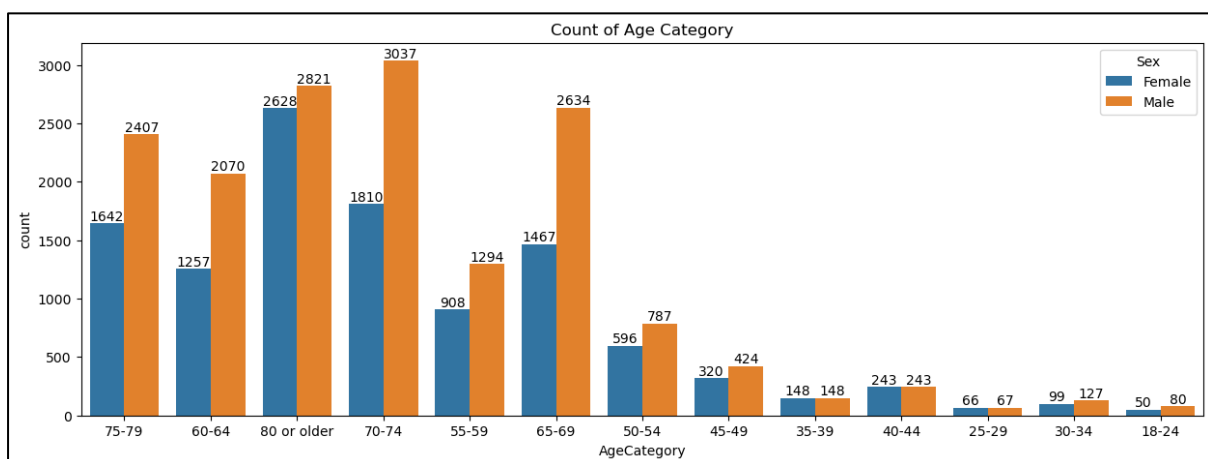
- observe that the people who are smoking are more susceptible to the heart disease.
- People who are not drinking alcohol, some of them have a heart disease.
- Stroke is highly correlated with heart disease.
- Heart disease is more commonly present in male.



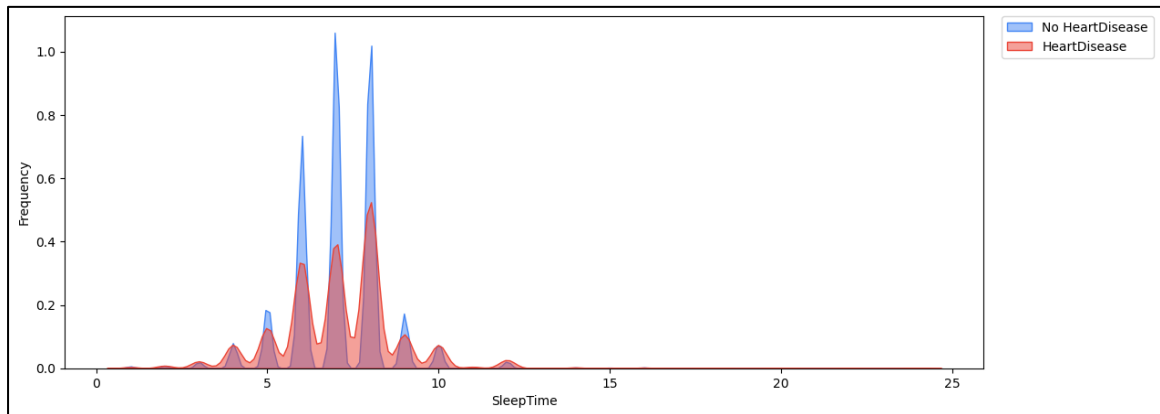
- Most people in the data are white and have no diabetic.
- Most people said that they have generally very good health. A few of people who said that they have generally a poor health.



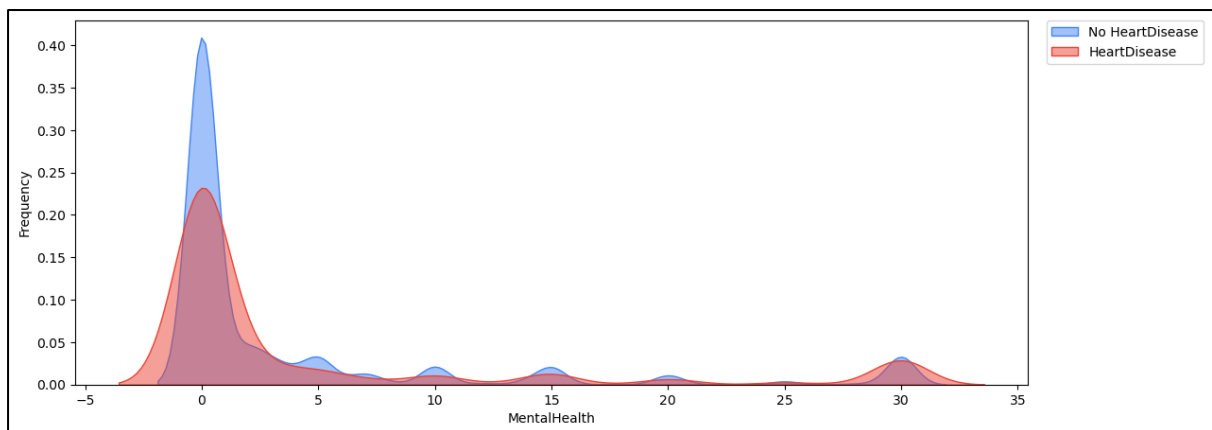
- A little of them who have asthma, kidney disease and skin cancer.
- Big factor in heart disease, as the amount of heart disease patients increases with age. The most susceptible people to the heart disease are people who are greater than 70 years old.



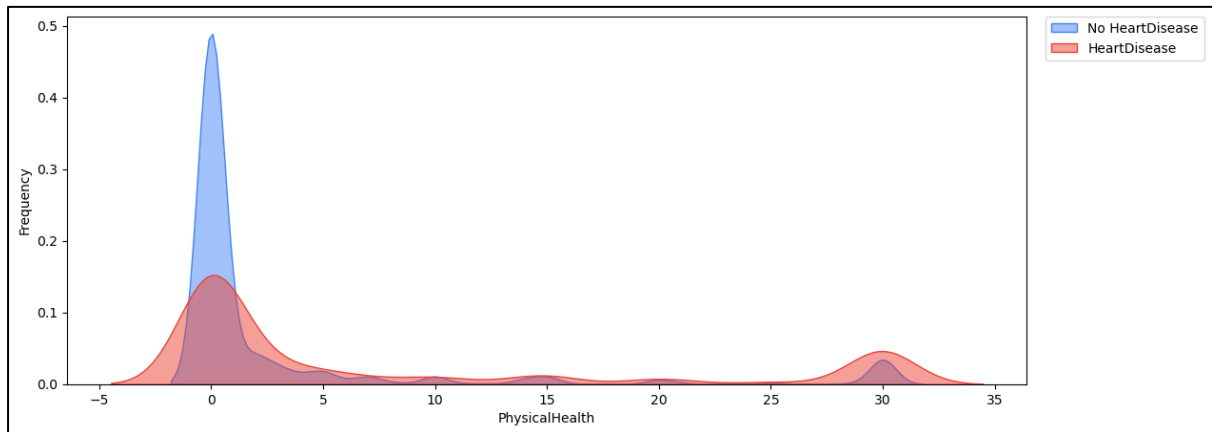
- From all the graphs presented, it can be concluded that alcohol consumption, smoking and age are the main factors in heart disease.
- Males are more susceptible to the heart disease.



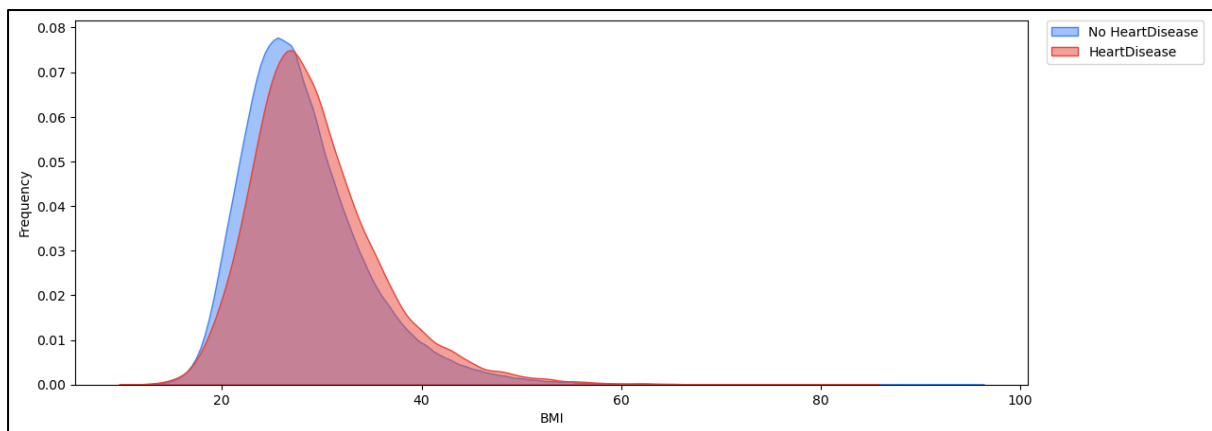
- Abnormal sleep duration is more prevalent in heart disease patients.



- Mental illness influences heart disease.



- Physical Health illness influences heart disease.



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

5. 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 as 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.


6. Streamlit 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:

The screenshot displays the 'Dr. Logistic's Heart Care' application interface. At the top, there is a navigation bar with links to Home, Dashboard, Data And Analysis, and Health tips. The main heading is 'Heart Health Checkup' with the subtext 'THE BEST WAY TO FIGHT A DISEASE IS TO PREVENT IT.' Below this, a paragraph explains the app's purpose: 'Are you wondering about the condition of your heart? This app will help you to diagnose it!'. It mentions that the app uses machine learning models to predict heart disease risk based on personal key indicators. A list of steps to follow is provided: 1. Enter the parameters that best describe you. 2. Press the 'Predict' button and wait for the result. A note states: 'Keep in mind that this results is not equivalent to a medical diagnosis! This model would never be adopted by health care facilities because of its less than perfect accuracy, so if you have any problems, consult a human doctor.' A link to the GitHub repository is also provided. On the right side, there is an illustration of a doctor in a white coat holding a stethoscope. Below the text, a form titled 'Fill The Following Form To Check Your Heart Health' is shown. The form contains various input fields and dropdown menus for personal information and health indicators. The fields include: Enter Your Name, Select Your Birth Date (2000/06/12), Enter Your City, Enter Your Phone Number, Age category (18-24), BMI category (Under_weight), How Many Hours On Average Do You Sleep? (7), How Can You Define Your General Health? (Very good), For How Many Days During The Past 30 Days Your Physical Health Not Good? (0), For How Many Days During The Past 30 Days Your Mental Health Not Good? (0), Sex (Female), Have You Played Any Sports (running, biking, etc.) In The Past Month? (No), Have You Smoked At Least 100 Cigarettes In Your Entire Life (approx. 5 packs)? (No), Do You Have More Than 14 Drinks Of Alcohol (jests Or More Than 7 ounces) In A Week? (No), Did You Have A Stroke? (No), Do You Face Difficulty Walking Or Climbing Stairs? (No), Have You Ever Had Diabetes? (Yes), Do You Have Asthma? (No), Do You Have Kidney Disease? (No), and Do You Have Skin Cancer? (No). A 'Predict' button is located at the bottom of the form.

- 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.
- The Data and Analysis menu of the Streamlit application provides information about surveys of data and insights from analytical studies, which could be useful for future work.
- The health tips tab displays health-related advice. It provides recommendations for preventing heart disease.

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

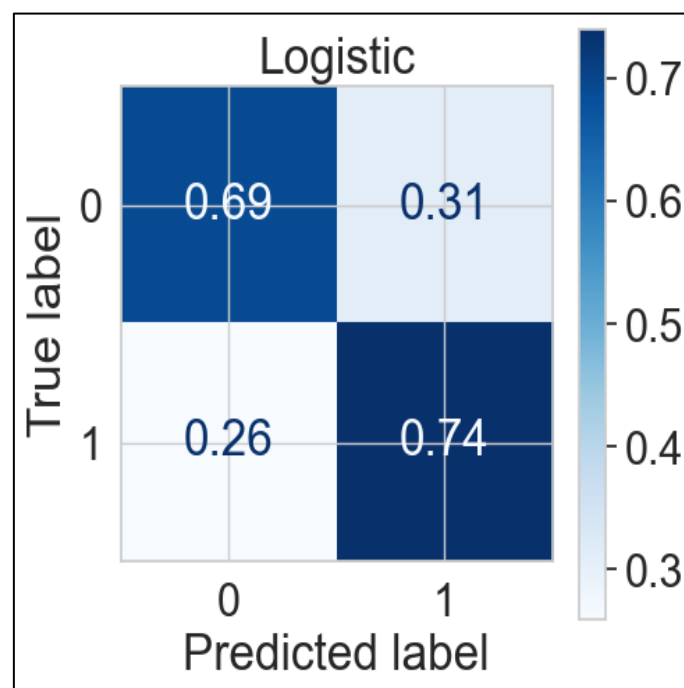
1. Training / Testing Model :

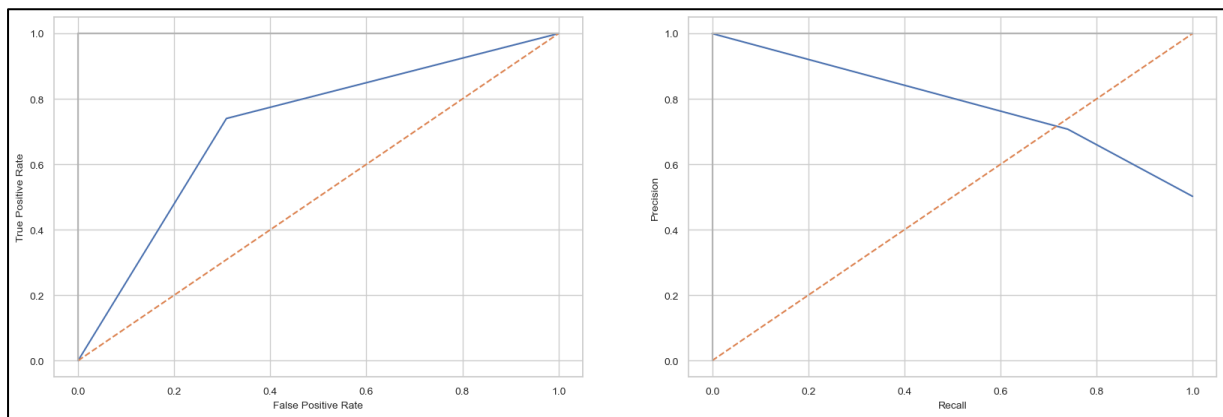
Using machine learning, I developed classification models, including logistic regression, random forest, decision tree, and SGD. I make non-sampled data in a sampled format, then divide the sampled dataset into 80% training data and 20% test data. After fitting the model to the training data, I examined 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.

1. 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.



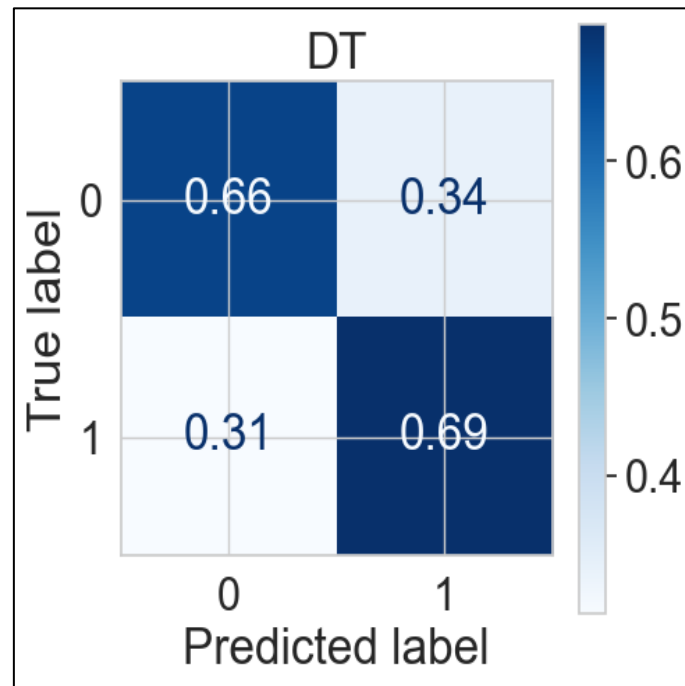


cutoff_df1				
	prob	accuracy	sensi	speci
0.0	0.0	0.502188	1.000000	0.000000
0.1	0.1	0.570712	0.986823	0.150943
0.2	0.2	0.642972	0.956429	0.326758
0.3	0.3	0.683958	0.913284	0.452616
0.4	0.4	0.712669	0.845484	0.578688
0.5	0.5	0.715871	0.740276	0.691252
0.6	0.6	0.701676	0.610414	0.793739
0.7	0.7	0.662184	0.440383	0.885935
0.8	0.8	0.592166	0.234006	0.953473
0.9	0.9	0.524602	0.061849	0.991424

I determined that the precision of logistic regression is 71.58% and the AUC is 0.7157.

2. Decision Tree:

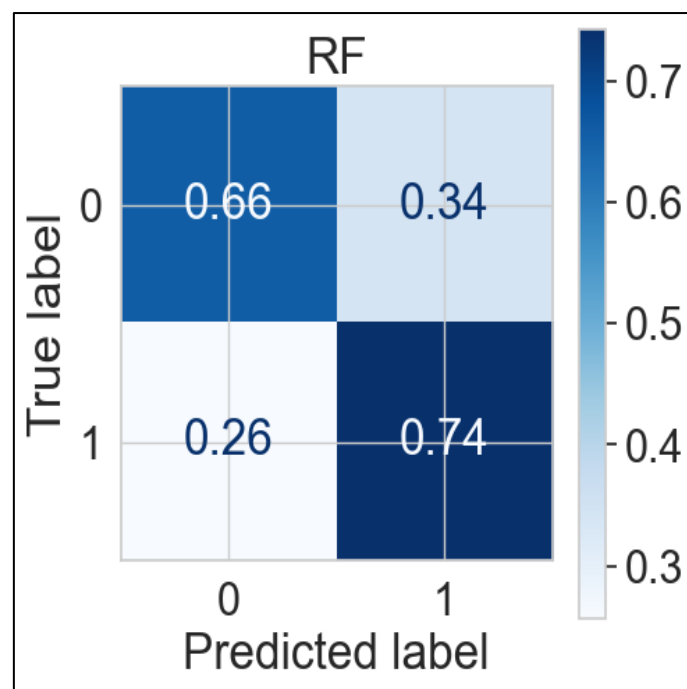
Compared to other machine learning models, Decision Tree models are relatively easy to comprehend and interpret, and the fact that there are nonlinear relationships between parameters does not affect the tree's performance. In contrast, this model contributes to the issue of overfitting and generates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



The DT model train score is 84.5 percent, while the test score is 67.3 %. This model shows overfitting because it performs well during training but not during testing.

3. 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.

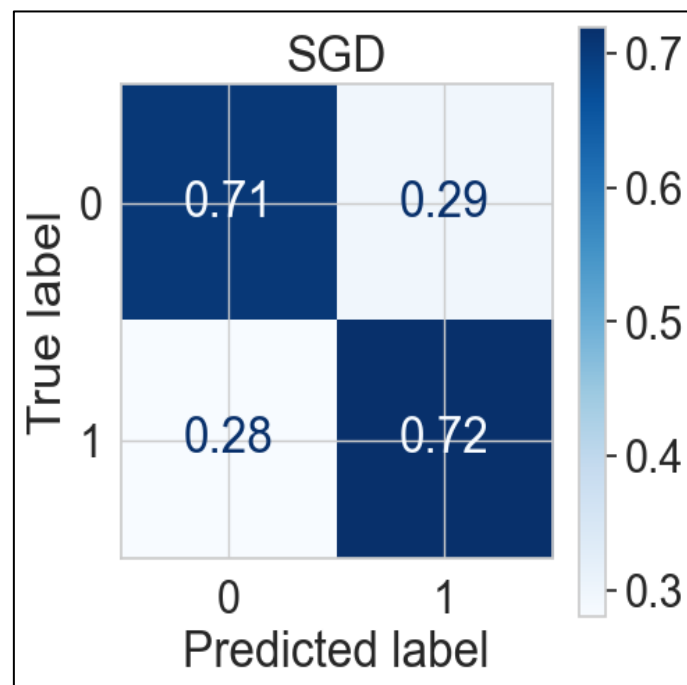


The model train score for Random Forest is 96.6%, while the test score is 70.1%. This model shows overfitting because it performs well during training but not during testing.

4. 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.

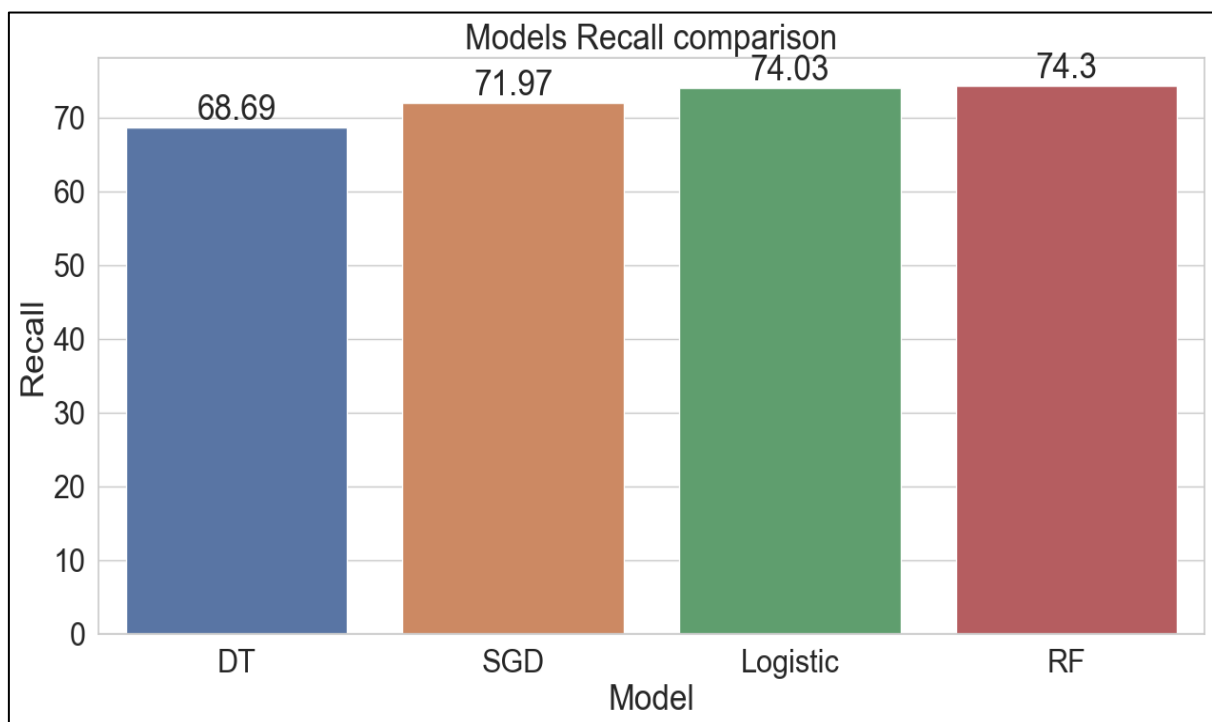


In this SGD model , train score is 72.3% while the test score is 71.3 %.

2. Comparison Of Algorithms :

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.



Form above bar graph Random Forest shows high recall but this model is already overfitted during testing. After Random Forest the highest Recall is Logistic regression. 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.

3. Results for training and testing :

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

results						
	Model	Train Score	Test Score	Recall	Precision	f1-score
0	DT	0.845	0.673	68.69	66.98	67.83
1	Logistic	0.725	0.716	74.03	70.75	72.35
2	SGD	0.723	0.713	71.97	71.13	71.55
3	RF	0.966	0.701	74.30	68.74	71.41

The above table shows the training score ,testing score , recall , Precision and F1 Score of all four models that trained. Decision Tree and Random Forest Model Shows overfitting problem. In other hand SGD and Logistic Regression has approximately equal training and testing score but Logistic Regression perform well as compared to SGD with Recall , Precision and F1 score.

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

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 -

