Analysis of Lifestyle Risk Factors for Heart Diseases

Group 66

Venkata Satyanarayana Murthy Iragavarapu
Ashritha Nayana Orampati

iragavarapu.v@northeastern.edu
orampati.a@northeastern.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Murthy Iragavarapu

Signature of Student 2: Ashritha Nayana O

Submission Date: 21 April, 2023

ANALYSIS OF LIFESTYLE RISK FACTORS FOR HEART DISEASE

I. Problem Setting

The term "heart disease" refers to a variety of heart conditions and is also known as "cardiovascular disease." Cardiovascular disease is characterized by constricted or blocked blood arteries, which can lead to a heart attack or stroke. Heart disease is the leading cause of death in both men and women. Someone in the United States has a heart attack every 40 seconds.Every year, approximately 8,05,000 people in the United States suffer from a heart attack; 6,05,000 of these are first-time heart attacks, while the remaining 2,00,000 are repeat heart attacks. One in every five heart attacks is silent; the victim is injured but is unaware of it. In general, a healthy lifestyle can prevent up to 80% of coronary artery disease and 50% of ischemic strokes.This analysis may be especially useful for people in their forties and fifties who do not have higher clinical risk factors like high blood pressure or high cholesterol but are still at high risk of developing cardiovascular disease due to their lifestyle.

II. Problem Definition

Though there are many different types of heart disease, we will focus on the two most common: coronary heart disease (CHD) and myocardial infarction (MI), as well as the risk factors for these heart diseases. The primary goal of this analysis is to visualize the distribution of traits that appear to be associated with heart disease using personal key indicators of heart disease. The goal is to learn how people's lifestyle choices, such as smoking, excessive alcohol consumption, being overweight, eating an unhealthy diet, and being physically inactive, put them at risk for heart disease.

III. Data Sources

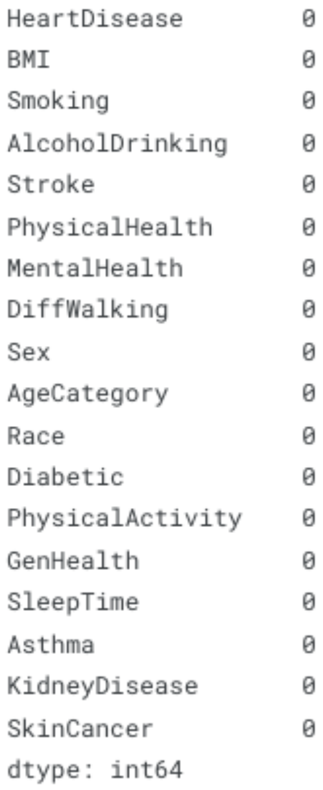
The dataset is collected by the Centers for Disease Control and Prevention (CDC) and is a key component of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather information on the health of Americans. The dataset has been taken from Kaggle (https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease).

IV. Data Description

The dataset contains 18 variables (9 booleans, 5 strings, and 4 decimals), including BMI, smoking and drinking habits, physical and mental health status. There are 18 attributes in total, with the target attribute "HeartDisease" being a binary ("Yes" - respondent has heart disease, "No" - respondent does not have heart disease). Other important indicators include diabetes status, obesity (high BMI), a lack of physical activity, and excessive alcohol consumption.

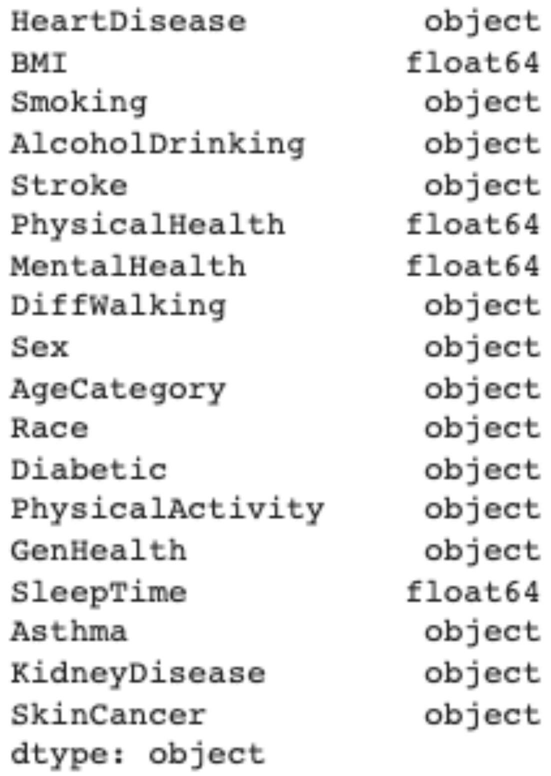
V. Data Cleaning

Missing data, also known as missing values, occur when data for certain variables or participants is not stored. Data might go missing for a variety of causes, including incorrect data entry, equipment faults, lost files, and many others.



```
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
```

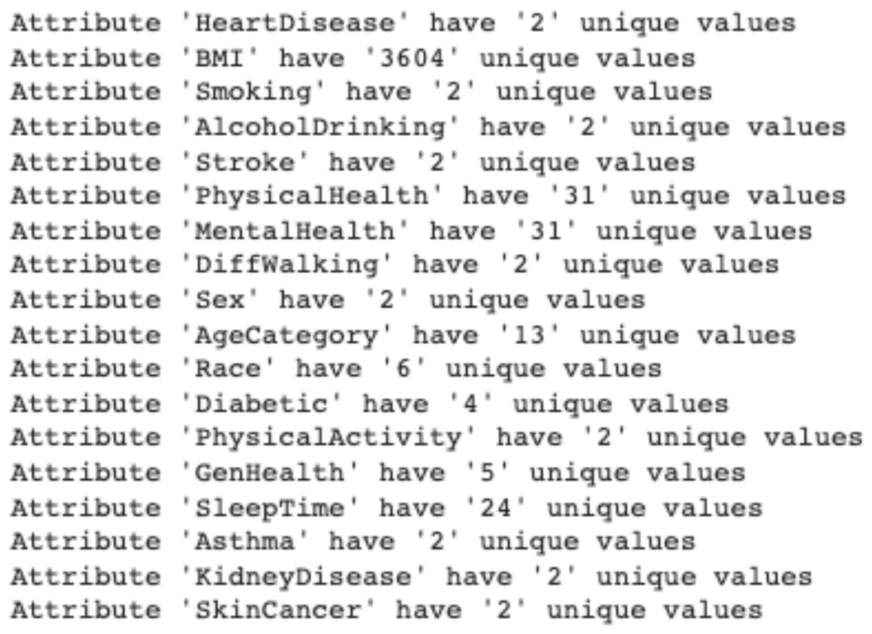
The Data Set has no null values. As a result, we may proceed with checking the values to see if they are valid or if any random meaningless data is provided.



```
HeartDisease         object
BMI                 float64
Smoking              object
AlcoholDrinking      object
Stroke               object
PhysicalHealth      float64
MentalHealth        float64
DiffWalking          object
Sex                  object
AgeCategory          object
Race                 object
Diabetic             object
PhysicalActivity     object
GenHealth            object
SleepTime           float64
Asthma               object
KidneyDisease        object
SkinCancer           object
dtype: object
```

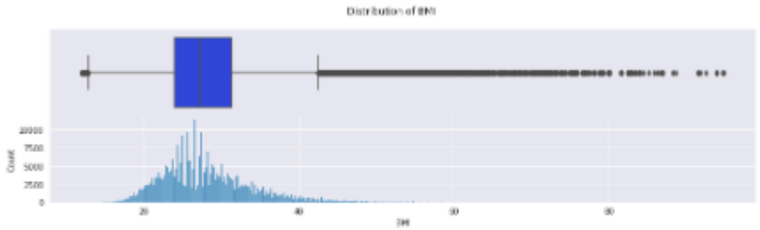
From the above analysis it can be said that we have are the PhysicalHealth, SleepTime, MentalHealth,BMI and rest of the variables we have are categorical.

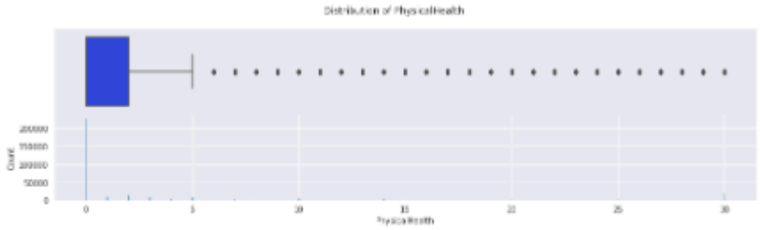
Analyzing the number of unique values for each attribute so that we get to know the essential material and wrong input values provided.

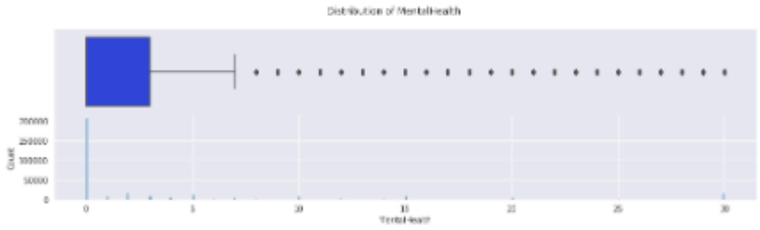


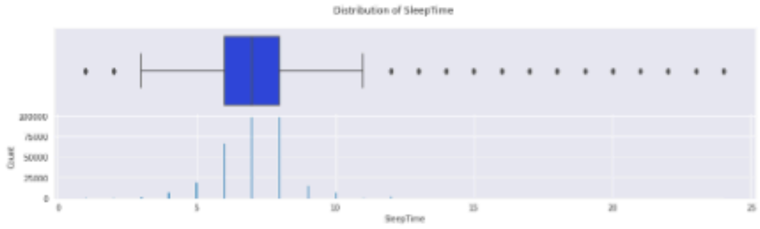
```
Attribute 'HeartDisease' have '2' unique values
Attribute 'BMI' have '3604' unique values
Attribute 'Smoking' have '2' unique values
Attribute 'AlcoholDrinking' have '2' unique values
Attribute 'Stroke' have '2' unique values
Attribute 'PhysicalHealth' have '31' unique values
Attribute 'MentalHealth' have '31' unique values
Attribute 'DiffWalking' have '2' unique values
Attribute 'Sex' have '2' unique values
Attribute 'AgeCategory' have '13' unique values
Attribute 'Race' have '6' unique values
Attribute 'Diabetic' have '4' unique values
Attribute 'PhysicalActivity' have '2' unique values
Attribute 'GenHealth' have '5' unique values
Attribute 'SleepTime' have '24' unique values
Attribute 'Asthma' have '2' unique values
Attribute 'KidneyDisease' have '2' unique values
Attribute 'SkinCancer' have '2' unique values
```

Outliers can provide useful information about the topic and the data collection process. Understanding how outliers form and whether they are likely to reoccur as a normal component of the process or study area is critical. As a result, we investigated the outliers of each numerical attribute and discovered that all numerical variables contain outliers.

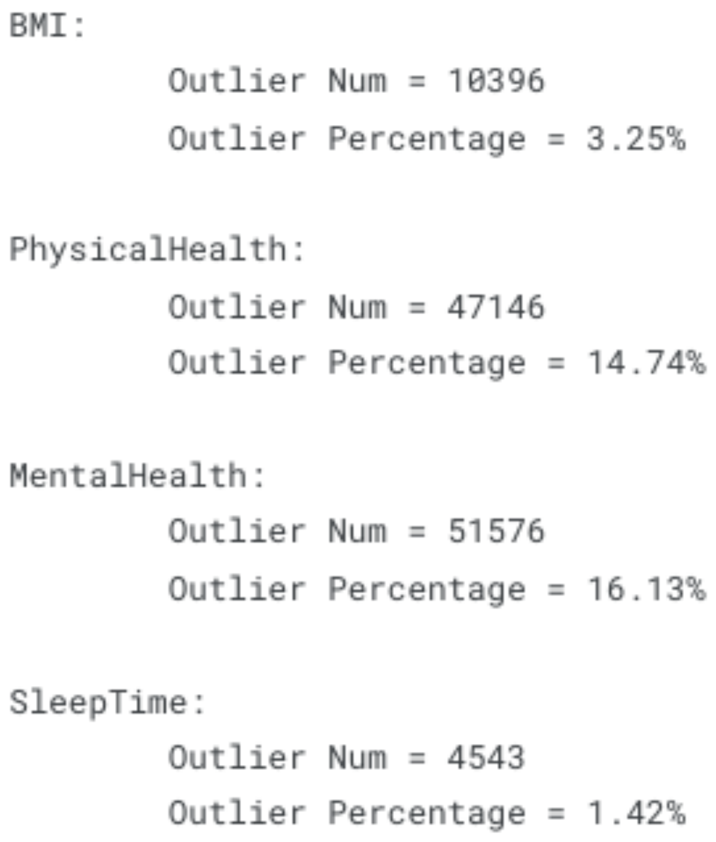

Distribution of BMI


Distribution of PhysicalHealth


Distribution of MentalHealth


Distribution of SleepTime

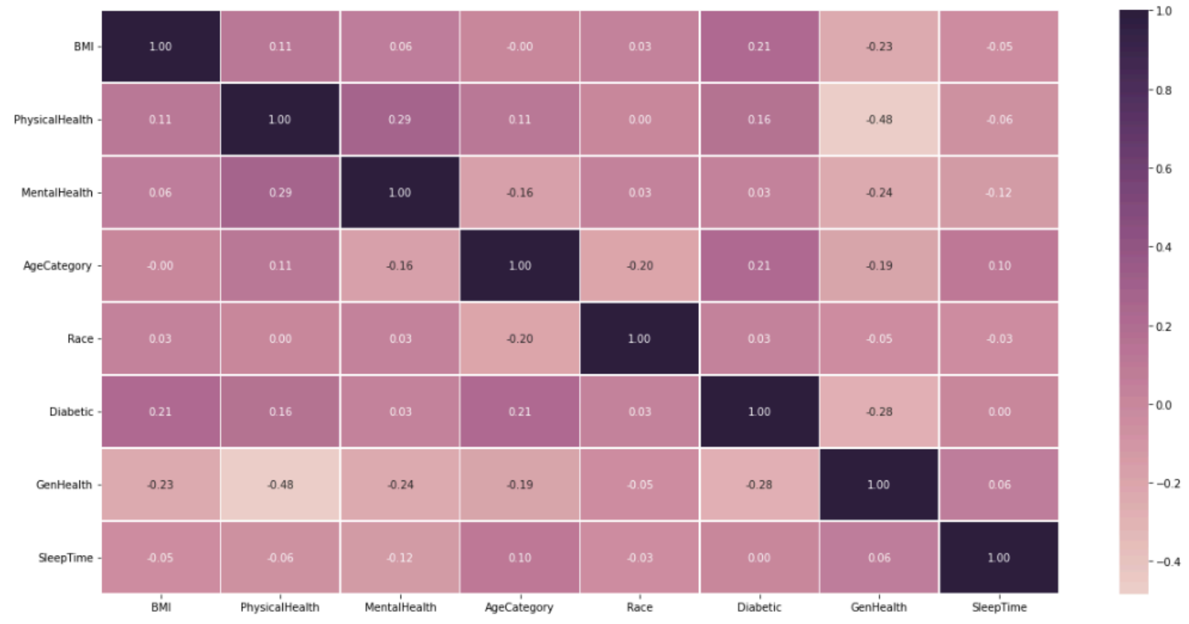
Hence we examined the outliers more closely by getting the count and percentage of outliers.



```
BMI:
        Outlier Num = 10396
        Outlier Percentage = 3.25%

PhysicalHealth:
        Outlier Num = 47146
        Outlier Percentage = 14.74%

MentalHealth:
        Outlier Num = 51576
        Outlier Percentage = 16.13%

SleepTime:
        Outlier Num = 4543
        Outlier Percentage = 1.42%
```

VI. **Data Preprocessing**

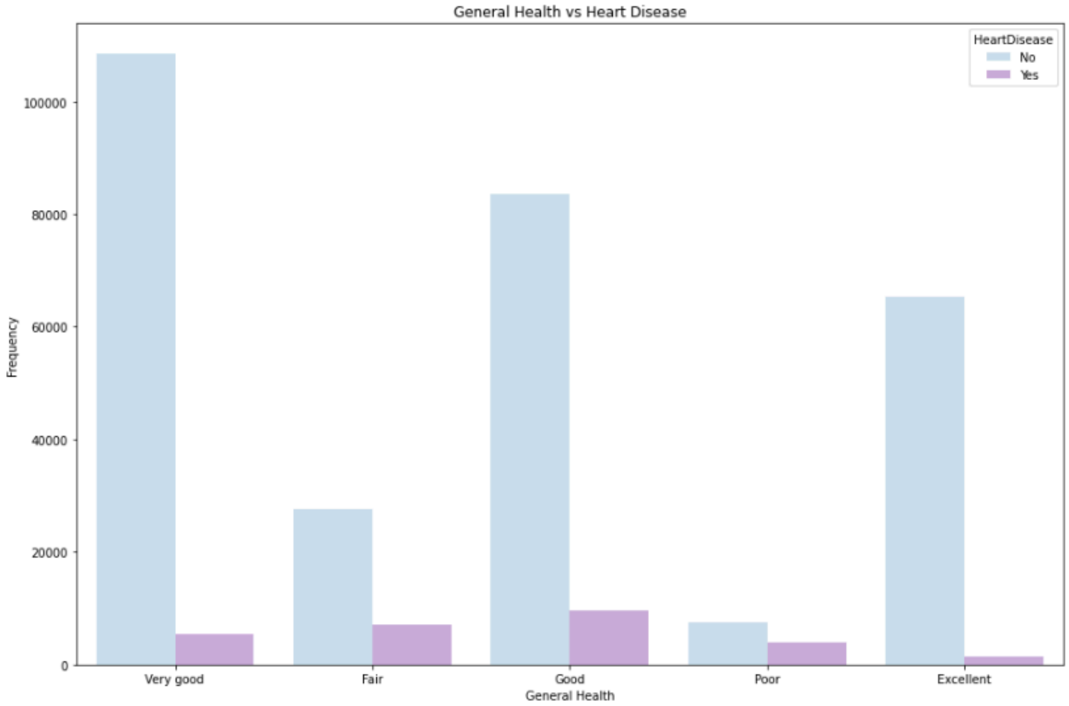
We started cleaning and preprocessing by converting binary yes/no data to numerical numbers. As a result, we converted the column attributes sex, heart disease, general health, diabetes, race, alcohol consumption, smoking diff_walking to numerical values and plotted the correlation between the parameters.



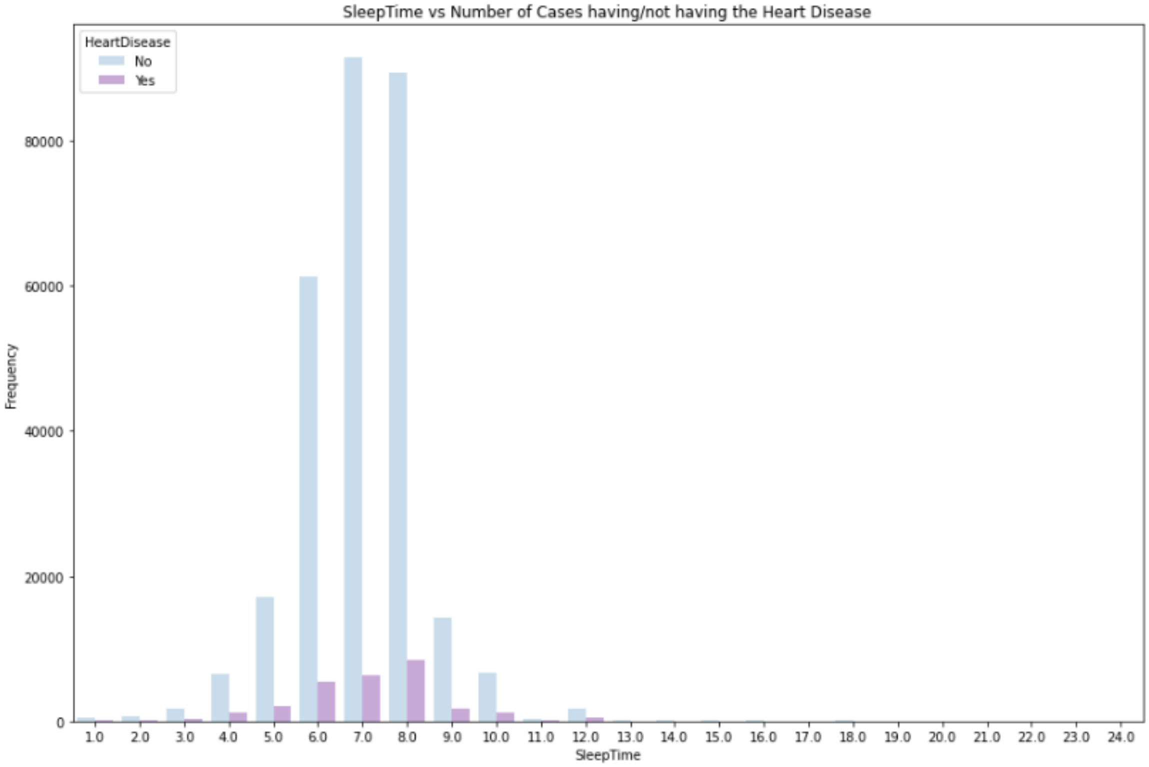
	BMI	PhysicalHealth	MentalHealth	AgeCategory	Race	Diabetic	GenHealth	SleepTime
BMI	1.00	0.11	0.06	-0.00	0.03	0.21	-0.23	-0.05
PhysicalHealth	0.11	1.00	0.29	0.11	0.00	0.16	-0.48	-0.06
MentalHealth	0.06	0.29	1.00	-0.16	0.03	0.03	-0.24	-0.12
AgeCategory	-0.00	0.11	-0.16	1.00	-0.20	0.21	-0.19	0.10
Race	0.03	0.00	0.03	-0.20	1.00	0.03	-0.05	-0.03
Diabetic	0.21	0.16	0.03	0.21	0.03	1.00	-0.28	0.00
GenHealth	-0.23	-0.48	-0.24	-0.19	-0.05	-0.28	1.00	0.06
SleepTime	-0.05	-0.06	-0.12	0.10	-0.03	0.00	0.06	1.00

VII. Data Visualization

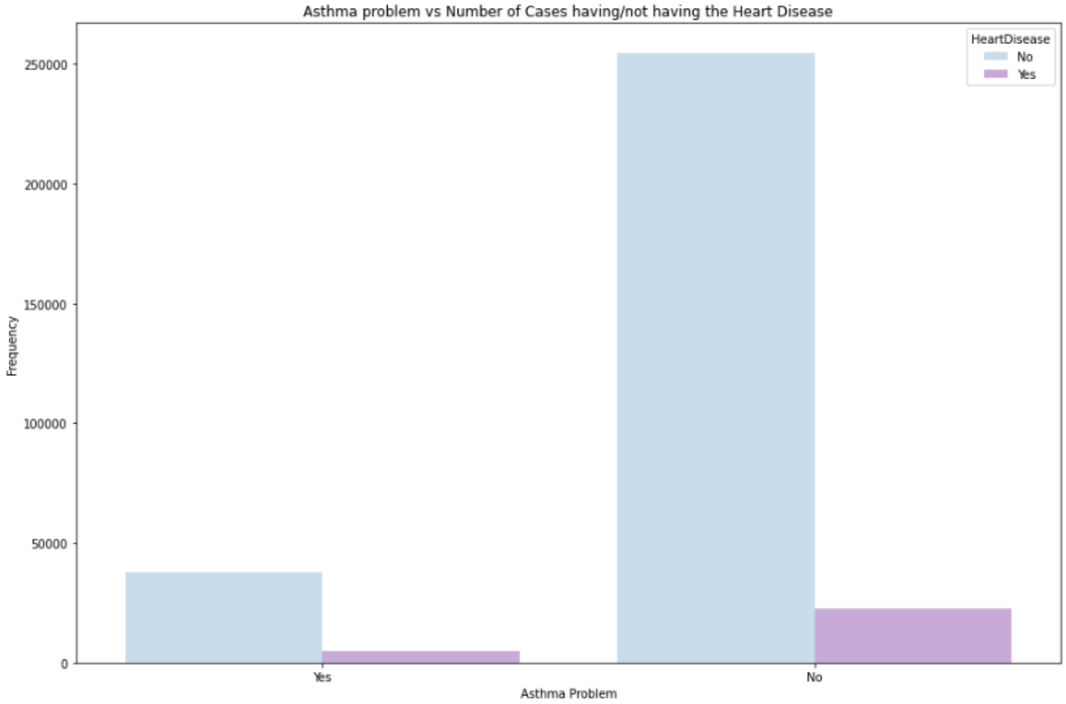
We used graphs to determine the relationship between the attributes and Heart Disease. According to the graph below, general health plays an important role in our lives, so heart disease is also affected by general health in some way.



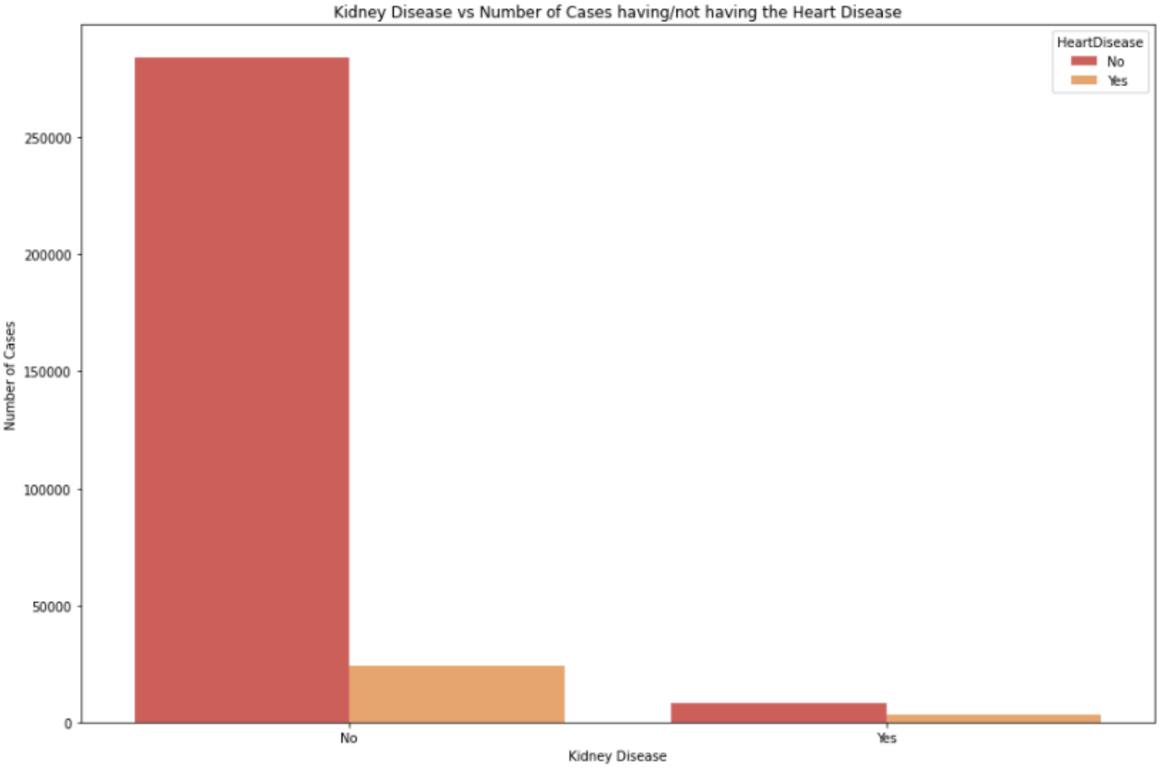
Sleeping for a longer period of time can also contribute to severe cardiac conditions, so sleep time should be regulated to 7 hours.



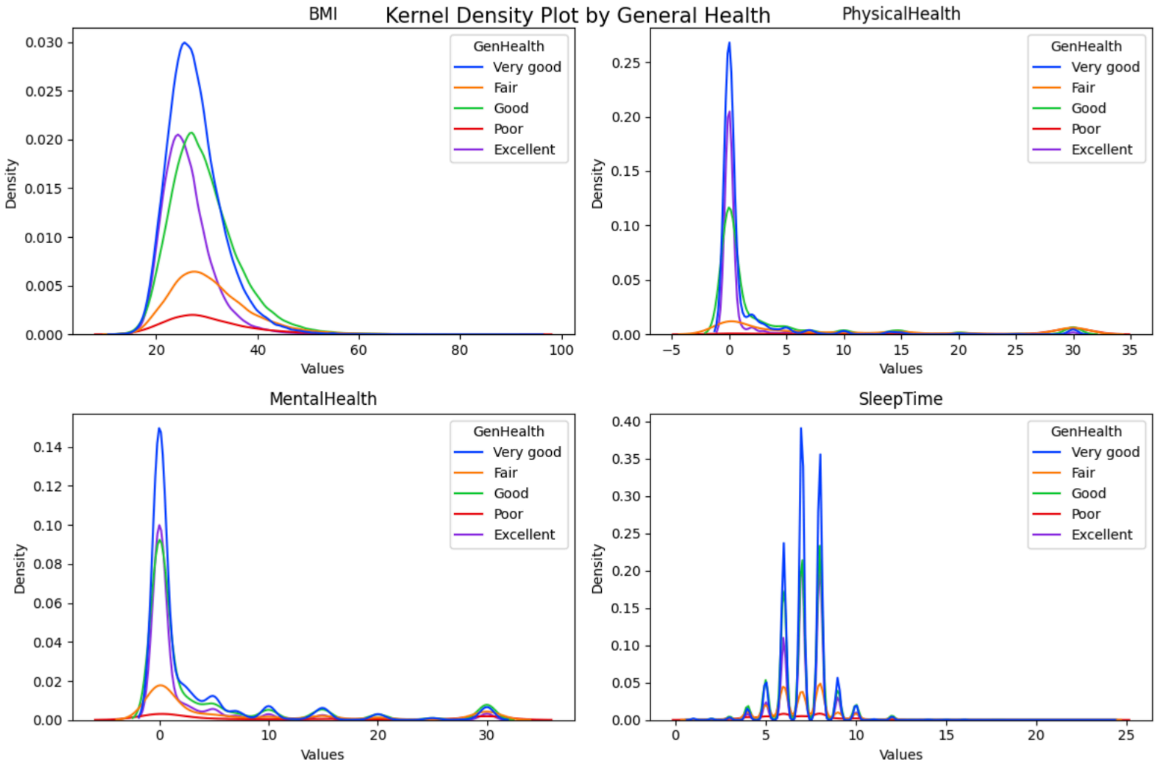
From the below graph it can be said that Asthma cannot be taken as the sole reason for the Heart Disease

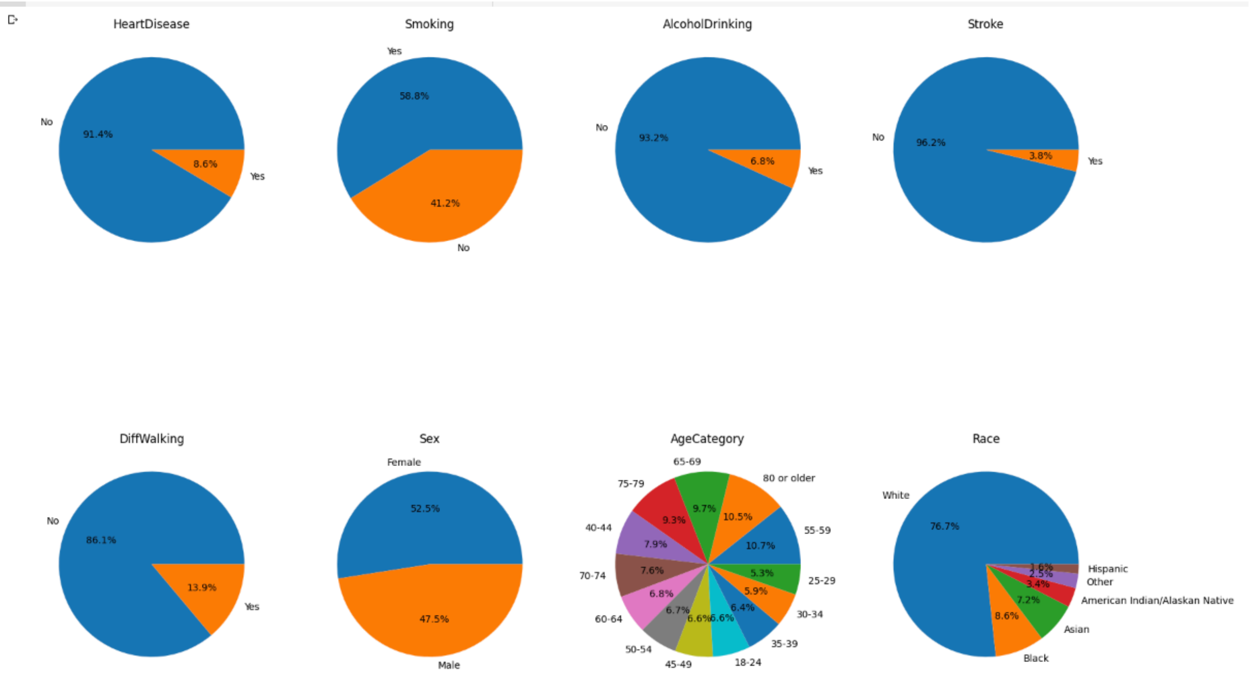


From the below graph it can be said that same as the Asthma, Kidney problem cannot be taken as the sole reason for the Heart Disease



The graphs below depict the distribution of each numerical variable in the dataset, with different colors representing the various levels of "GenHealth." Overall, this code is useful for visually exploring the relationships between numerical variables and categorical variables in a dataset.





The above visualization helps to give an overview of the distribution of categorical variables in a dataset, allowing for insights into any potential patterns or trends that may exist.

VIII. Data partitioning

The predictor variables were denoted by variable 'X,' and the target variable "Heart Disease" was denoted by variable 'y. 'X' denoted variables with indexes ranging from 0 to 32, while 'y' represented the target variable with index 24. The Hold out Method was used to partition the dataset, resulting in train and test sets with a 70:30 ratio. This means that 70% of the dataset was used to train the classification models, with the remaining 30% serving as a test set to evaluate the models' classification performance. X_ train contained 2,11,201 records and 23 variables, while X_ test contained 90,516 records and 23 variables. The y train had 2,11,201 records and one variable, while the y_ test had 90,516 records and one variable. During the model-building phase, the forward-selection variable selection method was used.

IX. Data Mining Models/ Methods

We ran five data mining classification models on the aforementioned training data, which are described below.

1. K - Nearest Neighbors (KNN)

The KNN classification algorithm is based on the use of records from the training data that are comparable to new records from the test data. The majority decision rule is used to classify the new record as a number of the k-neighbors' majority class, and the k-neighbors are used to determine related records.

Advantages: - Because K-NN is a non-parametric technique, it makes no assumptions about the underlying data distribution.
- It is capable of dealing with multi-class classification problems.- It is simple to implement and comprehend.

Disadvantages: - It is susceptible to irrelevant features or noisy data, resulting in poor classification performance.
- Significant computing costs during the prediction phase, especially when there are many training instances.
- Data with missing values cannot be handled.

Implementation:

Accuracy: 90.11555968005656 %
Precision: 37.81299524564184 %
Recall: 14.587918806554171 %
F1 Score: 21.053560398835263 %

2. Random Forest

The key idea behind the random forest classifier is to combine the predictions of multiple decision trees to produce a more accurate and robust classification model. Random forest regression is an ensemble learning technique that works by building multiple decision trees on different subsets of training data and then averaging their predictions. A random forest regression model trains each decision tree on a randomly selected subset of features and data points, which reduces overfitting and improves the model's generalization performance. When building each decision tree, a random subset of the features is used to increase the model's diversity and accuracy.

The basic idea behind random forest regression is to use ensemble learning to combine the predictions of many decision trees trained on different subsets of the training data to produce a more precise and reliable regression model.

Advantages: - The reliable random forest technique can handle outliers and missing data in input data without the need for preparation or data normalization.

- It is a fast algorithm capable of handling massive datasets and producing precise predictions.
- The random forest algorithm is non-parametric because it makes no assumptions about the distribution or relationships of the input data.

Disadvantages: Despite the fact that the algorithm favors the majority class, random forest models may underperform when there is a class imbalance in the input data.

- When working with extremely large datasets or a large number of features, training time and memory requirements can be extremely high.

Implementation:

Accuracy: 89.93658579698617 %

Precision: 33.77483443708609 %

Recall: 11.848862802641232 %

F1 Score: 17.54322440481579 %

3.Naive Bayes Bernoulli

The Naive Bayes Bernoulli algorithm is trained on a labeled dataset, and each example consists of a set of binary features and a matching binary label. Using the training data, it computes the probabilities of each feature and label, and then uses these probabilities to classify brand-new, unlabeled inputs. The algorithm is called "naive" because it assumes that each feature is independent of every other feature, which is rarely the case in real-world situations. Nonetheless, this simplification improves computation efficiency and frequently works well in practice, especially for high-dimensional data. Overall, Naive Bayes Bernoulli is a powerful and widely used classification technique, particularly in situations with a high proportion of binary variables.

The advantages: - Simple and easy to implement.

- Large datasets as well as data with multiple dimensions can be handled.
- - It has fast prediction and calculation times.

Disadvantages: - Because this is not always the case in real-world data, it assumes that all features are independent of one another.

- If the training data is unbalanced or contains missing values, the results may be skewed.

Implementation:

Accuracy: 84.65575146935348 %

Precision: 27.510435535953377 % Recall: 42.712154561017364 %

F1 Score: 33.465868263473055 %

4. Logistic Regression

The logistic regression statistical method can be used to model the relationship between a binary dependent variable and one or more independent variables. In the context of heart attack prediction, logistic regression can be used to investigate the relationship between various risk variables and the likelihood of a person having a heart attack. The model will calculate the likelihood that a person will have a heart attack based on the values of the independent variables. In this scenario, the dependent variable would be a binary variable indicating whether or not the subject had a heart attack. Identifying those at high risk of heart attacks and implementing preventative measures to reduce the risk of heart attacks

Advantages: - Simple and straightforward: Linear regression is a straightforward, simple, and straightforward strategy.

- It allows for a quick analysis of how the dependent variable and independent variables are related.
- Capable of dealing with continuous variables: The ability of linear regression to handle continuous variables may be useful for assessing variables such as blood pressure and cholesterol levels.

Disadvantages: - Logistic regression assumes a linear relationship between the input variables and the result, which may not be true for all datasets.

- Logistic regression may not perform well in datasets with imbalances, where one class is significantly more abundant than the other.
- Overfitting is an issue with logistic regression, particularly when the dataset is small and the number of input variables is large.

Implementation:

Accuracy: 91.04025807592028 %

Precision: 52.4390243902439 %

Recall: 8.938615798483736 %
F1 Score: 15.273715002089428 %

5. XGBoost

XGBoost works by iteratively building decision trees and aggregating their predictions. The method aims to optimize a loss function that measures how well the model works during each iteration. Because predicting heart attacks is a binary classification problem, XGBoost, which can handle both regression and classification tasks, may be appropriate. The training set is used to fit the model, and then the testing set is used to evaluate performance. A number of hyper-parameters must be adjusted to improve the model's performance. Some common hyper-parameters to tweak are the learning rate, the maximum depth of the decision trees, the number of trees, and the subsampling rate. After training the XGBoost model, we can use it to make predictions based on new data. To determine whether a person is likely to have a heart attack, you would enter relevant features about them into the trained model and look for the predicted label. Overall, XGBoost has the potential to be a powerful tool for predicting heart attacks; however, optimal performance necessitates careful tuning and feature selection. It is also critical to validate the model's performance on different datasets in order to confirm its generalizability.

Advantages: - High accuracy: XGBoost is frequently used in machine learning competitions and is well-known for its high accuracy.-
Handles missing values: In real-world datasets, XGBoost can manage missing values without the need for imputation.- The feature importance score provided by XGBoost can help you find the most important characteristics in predicting cardiac illnesses.
- Fast training: Because it can train on large datasets quickly, XGBoost is well suited for real-time applications.

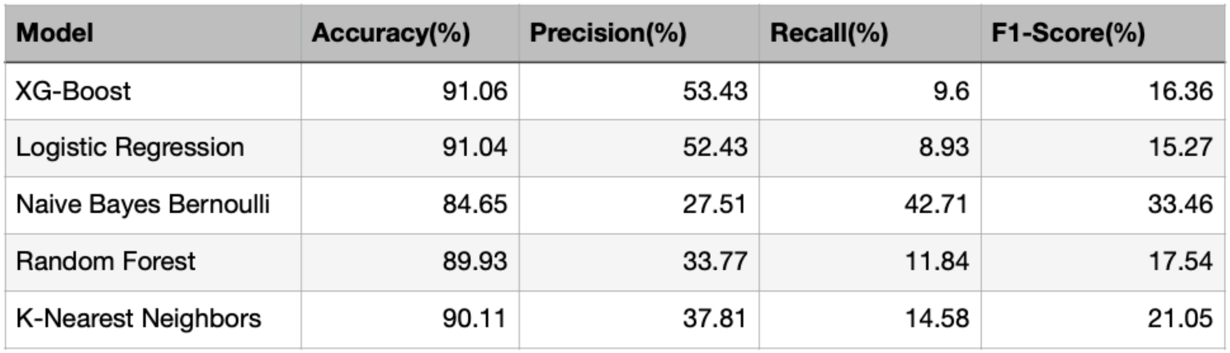
Disadvantages: - Overfitting: If XGBoost is not properly tuned, it can easily overfit the data, which has a negative impact on generalization performance.

- XGBoost is only effective with tabular data; it may not work as well with other types of data, such as image or text data.

Implementation:
Accuracy: 91.06787750232003 %
Precision: 53.12290127602418 %
Recall: 9.672291513817559 %
F1 Score: 16.364952932657495 %

After performing all these models on the data we have selected the XGBoost is classifying with good accuracy score and various other metrics.

X. Performance Evaluation :



Model	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
XG-Boost	91.06	53.43	9.6	16.36
Logistic Regression	91.04	52.43	8.93	15.27
Naive Bayes Bernoulli	84.65	27.51	42.71	33.46
Random Forest	89.93	33.77	11.84	17.54
K-Nearest Neighbors	90.11	37.81	14.58	21.05

Based on the performance metrics, XGBoost is the best model for this problem. It has the highest accuracy score of any model and outperforms the others in terms of precision, recall, and F1 score. While logistic regression and XGBoost have similar accuracy scores, XGBoost has higher precision and recall scores, indicating that it is more accurate at identifying positive cases. As a result, XGBoost is the best model for this particular issue.

Logistic regression has an extremely high accuracy score of 91.04%, which is nearly identical to the best model XGBoost. However, its precision and recall scores are lower than those of XGBoost, implying that it may struggle to correctly identify positive cases. Logistic regression is a simple and easy-to-understand model that works well. It is also computationally efficient and works well with large datasets. However, logistic regression assumes a linear relationship between the input features and the target variable, which may not be appropriate for complex datasets with nonlinear relationships.

KNN has a lower accuracy score than XGBoost and logistic regression, but a higher precision score than all of the other models. This implies that when KNN predicts a positive case, it is more likely than the other models to be correct. However, KNN has a very low recall score, which means it misses a lot of positive cases. KNN is effective for problems with a small number of input features and a nonlinear decision boundary. However, KNN can be computationally expensive for large datasets and requires the definition of a distance metric, which may not be appropriate for all datasets.

Random Forest has the lowest accuracy score of any model, as well as lower precision and recall scores than XGBoost and logistic regression. Random Forest is an ensemble learning method that combines multiple decision trees to create a more robust model. It is effective for problems involving nonlinear relationships between input features and the target variable, and it can handle missing data and irrelevant features. However, Random Forest is prone to overfitting and may require hyperparameter tuning to improve performance.

Although Naive Bayes has the lowest accuracy score of any model, it has the highest recall score. This shows that while Naive Bayes is good at identifying positive cases, it has a high false positive rate, resulting in lower precision. Naive Bayes is a probabilistic model that can handle missing data and works well for problems with categorical input features. It's also computationally efficient and requires little training data. Naive Bayes, on the other hand, assumes that the input features are independent, which may not be appropriate for some datasets with correlated input features.

In conclusion, XGBoost is the best model for this problem due to its higher accuracy, precision, recall, and F1 score. However, other models, such as logistic regression and KNN, can be useful for a variety of datasets and problems. Because of their lower accuracy and precision/recall scores, Random Forest and Naive Bayes may not be the best choices for this problem.

XI. <u>Accuracy Scores :</u>

Based on Accuracy of all the models, from the below plot we can say XG Boost Classifier is having the higher accuracy.

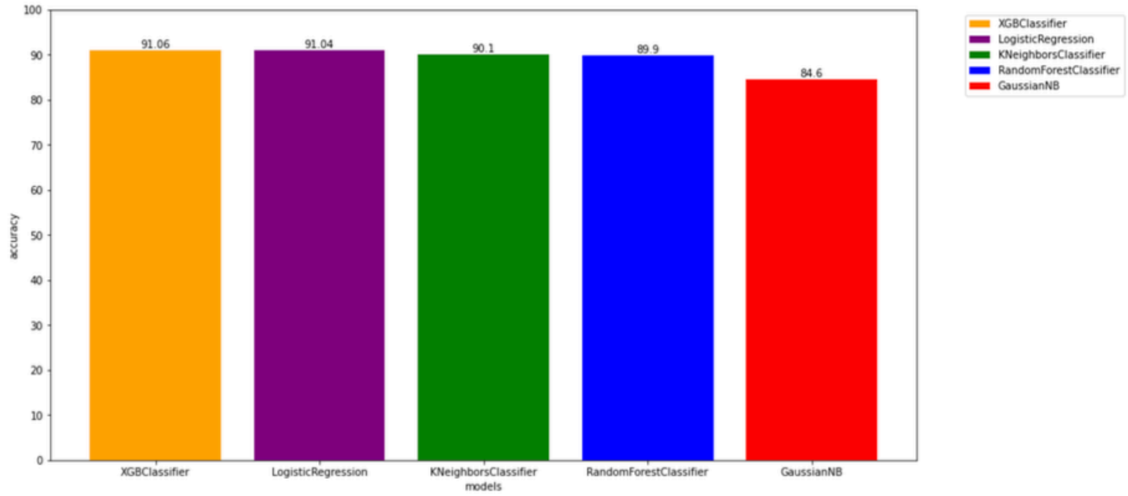
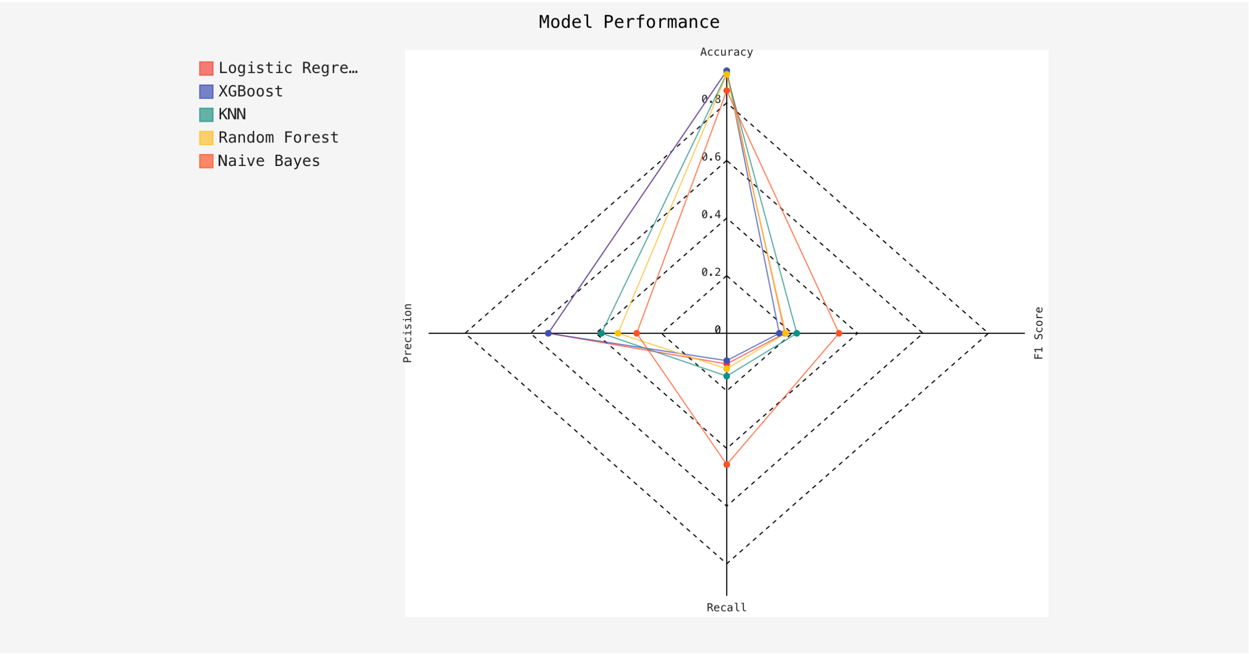


Figure: Plot of Accuracy for all models



The Area Under the ROC Curve (AUC) is a popular metric for assessing a classification model's performance. It assesses the model's ability to distinguish between positive and negative examples across all classification thresholds. AUC represents the likelihood that the model will correctly rank a randomly selected positive example higher than a randomly selected negative example. A higher AUC indicates that the model performed better in positive examples. classifying and negative examples.

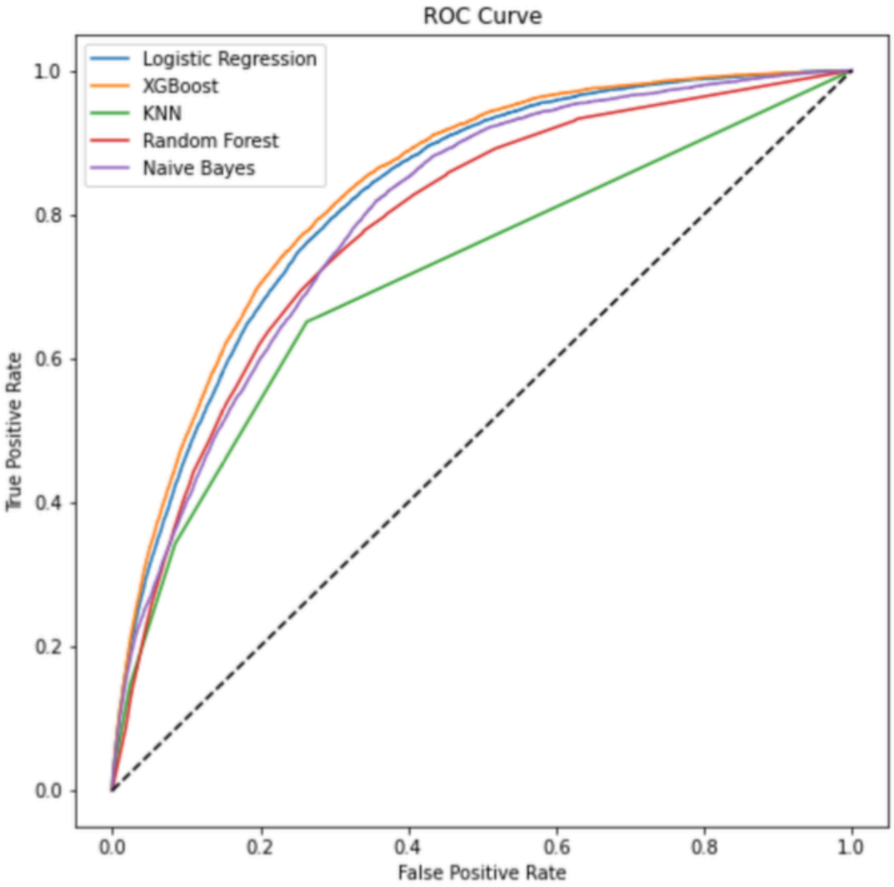


Figure: ROC curve for all models

The best performing model can be determined from the above ROC Curve figure by observing which model has an area under curve that is close to the one in this case, which is the XG Boost classification model.

XII. The Impact of the project outcomes:

The project's outcome could have a number of consequences. For starters, it can assist healthcare providers in identifying patients who are at risk of heart disease based on personal key indicators. This enables early interventions and lifestyle changes to reduce the risk of heart disease and improve health outcomes.

Second, the project's findings may aid researchers in better understanding the link between personal key indicators and heart disease. This could result in the development of new diagnostic tools and treatments for heart disease, as well as new insights into the disease's underlying mechanisms.

The project's outcome could have several ramifications. For starters, it can assist healthcare providers in identifying patients who may be at risk of heart disease based on their personal key indicators. This enables early interventions and lifestyle changes that can reduce the risk of heart disease and improve health outcomes.

Second, the project results can help researchers better understand the relationship between personal key indicators and heart disease. This can lead to the development of new diagnostic tools and treatments for heart disease, as well as new insights into the disease's underlying mechanisms.