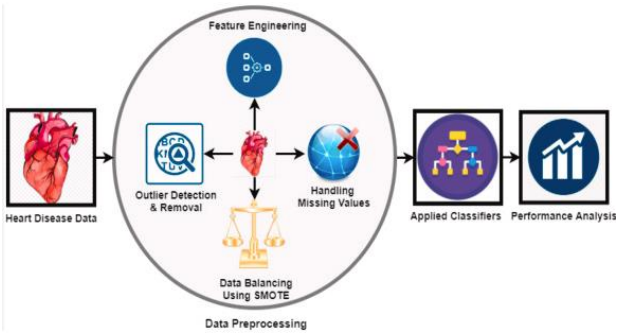


VitalFlow Guardian: Data-Powered Cardiac Risk Management

Bellevue University
DSC630 - Predictive Analytics

Submitted By:
Debabrata Mishra

Instructor:
Andrew Hua

Table of Contents

1.0 Introduction

Cardiovascular diseases result in the loss of about 17 million lives globally every year, predominantly through events like myocardial infarctions and heart failures. Among these, heart failure (HF) occurs when the heart struggles to adequately pump blood to meet the body's needs. Addressing this critical health issue involves tapping into electronic medical records, a promising resource for meticulous analysis. These records compile a vast dataset encompassing symptoms, physiological traits, and vital clinical test results. By employing biostatistical analysis, hidden patterns and correlations surface, often beyond the scope of even the most adept medical practitioners.

The emergence of machine learning has marked a transformative phase, enabling predictions of patient survival based on this extensive data repository. In the United States, heart disease remains a leading cause of mortality, demanding a holistic understanding of its multifaceted nature and interconnected contributing factors. Our pursuit involves unraveling the intricate web of variables that impact cardiac health, empowering individuals with insights crucial for informed decisions and fostering enduring heart health.

"VitalFlow Guardian," a dedicated initiative committed to crafting predictive models for cardiac risk management. Through meticulous data analysis and cutting-edge machine learning methodologies, our objective is to unearth invaluable insights. These insights won't just aid in early detection but will also pave the way for tailored interventions, ultimately saving lives and fortifying the health and longevity of individuals worldwide.

M4.0 Project Milestone 04

In Milestone 04, I have completed the majority of the technical work for the project, building upon the information from Milestone 03. Here's a breakdown of the additional items addressed.

M4.1 Data Preparation

For data preparation, we followed a systematic approach to ensure the dataset was optimized for modeling:

Exploratory Data Analysis (EDA): The milestone commenced with a preliminary exploratory data analysis, where we examined the initial rows of the dataset to grasp its structure. Further scrutiny was conducted using the 'info' function, affirming the dataset's integrity as it was devoid of any

missing values. Our dataset consisted of 319,795 entries across 18 columns, encompassing both numerical and categorical data. Notably, no missing values were identified, and all columns were deemed essential for our analysis. Subsequently, a duplicate check was conducted, revealing 18,078 duplicate records. To ensure data integrity, all duplicate entries were removed before proceeding with visualization.

Balancing the Dataset: There is a class imbalance issue within the dataset, with a considerable disproportion between the number of samples for individuals without heart disease compared to those with heart disease. To address this imbalance, oversampling techniques were implemented. By generating additional data points for individuals with heart disease, we successfully balanced the classes, resulting in a dataset primed for analysis.

Integer encoding for Categorical Variables (2 unique Values): Binary categorical variables with 2 unique values like 'Yes' and 'No' values, for 'HeartDisease', 'Smoking', 'AlcoholDrinking', and others, ‘Female’ and ‘Male’ for ‘Sex’ were identified. A crucial step was taken to standardize these binary values by converting to ‘0’ and ‘1’. This transformation was applied systematically to the relevant columns to ensure their compatibility with machine learning models.

Dummy variables: Created dummy variables for categorical columns that contain more than two unique values. Specifically, it generates dummy variables for the columns 'Race', 'Diabetic', and 'GenHealth'. This process expands each categorical variable into a series of binary (0 or 1) variables, representing the different categories within the original column. This transformation facilitates the inclusion of categorical data in machine learning models that require numerical inputs, enhancing the model's predictive capabilities.

Scaling Continuous Features: In preparation for machine learning tasks, we converted the 'AgeCategory' column into a continuous feature by employing a predefined mapping. Furthermore, we standardized continuous features like 'BMI,' 'PhysicalHealth,' 'MentalHealth,' 'AgeCategory,' and 'SleepTime' by scaling them to fit within the [0, 1] range.

0.1 Load & Quick Overview of Data

[2]:
```
# Load the dataset into a Pandas DataFrame
heart_df = pd.read_csv('heart_2020_cleaned.csv')
heart_df.head()
```

[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

[3]:
```
# Find the dimensions (number of rows and columns) of the data frame along with other informations

num_rows, num_cols = heart_df.shape
print("\nNumber of rows in the Data Frame    : ", num_rows)
print("Number of columns in the Data Frame : ", num_cols)
print("\n")

#Information about dataframe
heart_df.info(verbose= True, show_counts= True)
```

```
Number of rows in the Data Frame    :  319795
Number of columns in the Data Frame :  18
```

```
 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
```

The dataset contains no missing values, making all columns essential. It comprises 319,795 entries and 18 columns. Key attributes include 'HeartDisease' as the target variable, 'BMI' as a continuous feature, and various categorical features such as 'Smoking,' 'AlcoholDrinking,' 'Stroke,' 'DiffWalking,' 'Sex,' 'AgeCategory,' 'Race,' 'Diabetic,' 'PhysicalActivity,' 'GenHealth,' 'Asthma,' 'KidneyDisease,' and 'SkinCancer.' The dataset employs both float64 and object data types, with a memory usage of approximately 43.9 MB.

[4]:
```
# Check for duplicates
heart_df.duplicated().sum()
```

[4]:
```
18078
```

[5]:
```
# Drop the duplicates
heart_df.drop_duplicates(inplace=True)
```

[6]:
```
# Shape of dataframe after clean up.
heart_df.shape
```

[6]:
```
(301717, 18)
```

[7]:
```
#get summary statistics of the numerical data
heart_df.describe().T
```

[7]:

	count	mean	std	min	25%	50%	75%	max
BMI	301717.0	28.441970	6.468134	12.02	24.03	27.41	31.65	94.85
PhysicalHealth	301717.0	3.572298	8.140656	0.00	0.00	0.00	2.00	30.00
MentalHealth	301717.0	4.121475	8.128288	0.00	0.00	0.00	4.00	30.00
SleepTime	301717.0	7.084559	1.467122	1.00	6.00	7.00	8.00	24.00

[8]:
```
#get summary statistics of the non-numerical data
heart_df.describe(include = ['O'])
```

[8]:

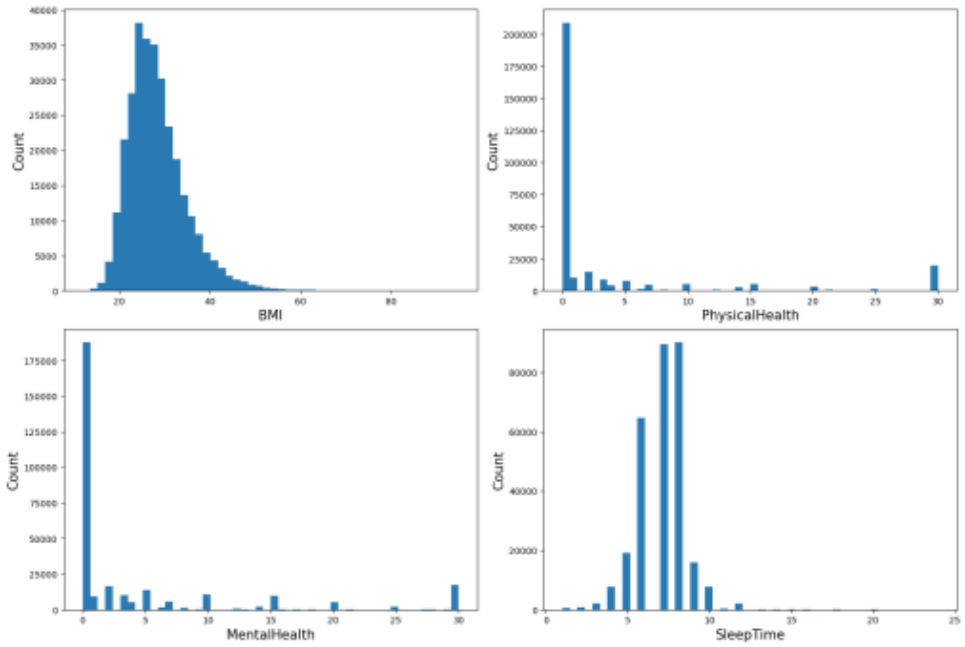
	HeartDisease	Smoking	AlcoholDrinking	Stroke	DiffWalking	Sex
count	301717	301717	301717	301717	301717	301717
unique	2	2	2	2	2	2
top	No	No	No	No	No	Female
freq	274456	174312	280136	289653	257362	159671

	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	Asthma
count	301717	301717	301717	301717	301717	301717
unique	13	6	4	2	5	2
top	65-69	White	No	Yes	Very good	No
freq	31670	227724	251796	230412	104796	259066

	KidneyDisease	SkinCancer
count	301717	301717
unique	2	2
top	No	No
freq	289941	272425

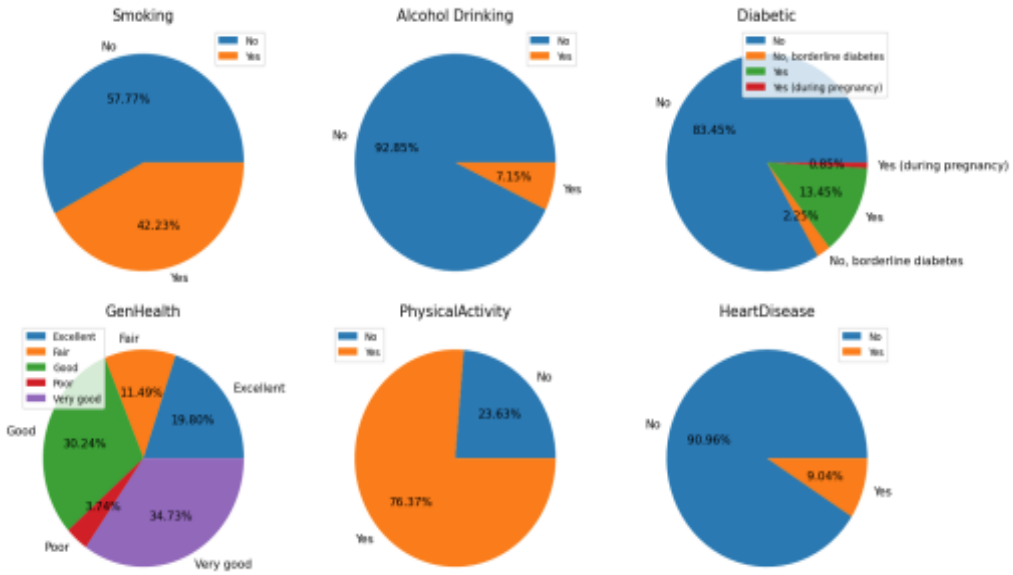
0.2 Exploratory Data Analysis

Histograms of the numerical features



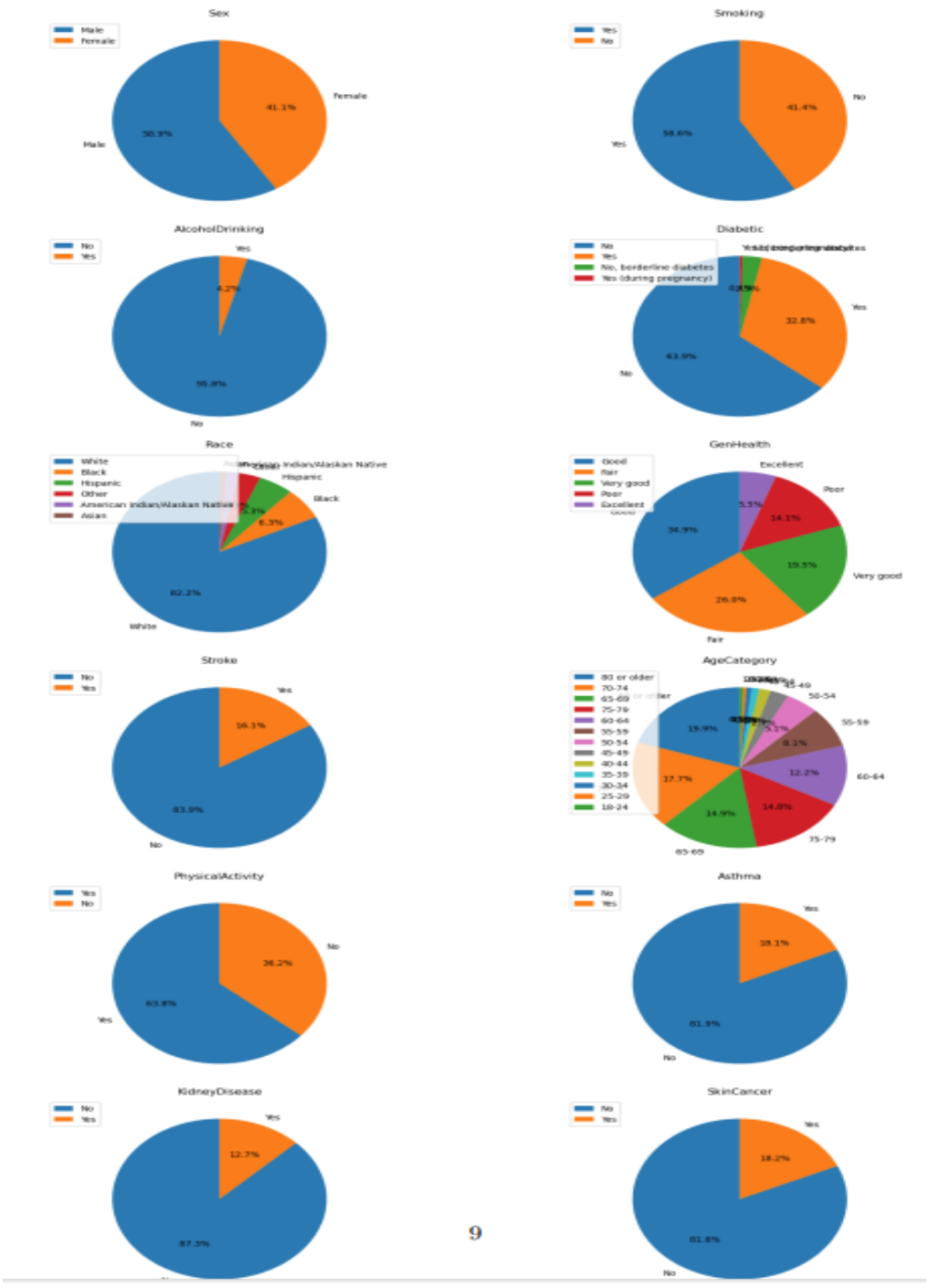
The visualizations reveal that the heart disease dataset portrays overall good health, characterized by a predominant presence of individuals with a BMI within the normal range. Additionally, both Physical and Mental Health scores are clustered around the mean, indicating a relatively balanced distribution. However, the distribution of Sleep Time exhibits a right-skewed pattern, suggesting that a majority of individuals in the dataset sleep for less than 7.5 hours per night, with a median Sleep Time below this threshold.

Pie charts of the categorical features to analyze the data distribution.



The heart disease dataset predominantly comprises individuals with commendable health habits, including abstaining from smoking, occasional alcohol consumption, and regular exercise. Nonetheless, a minority exhibits less favorable health behaviors, such as habitual smoking, excessive alcohol intake, and a lack of physical activity. Encouragingly, adopting healthier lifestyles, characterized by avoiding these detrimental habits, can significantly reduce the risk of heart disease among individuals.

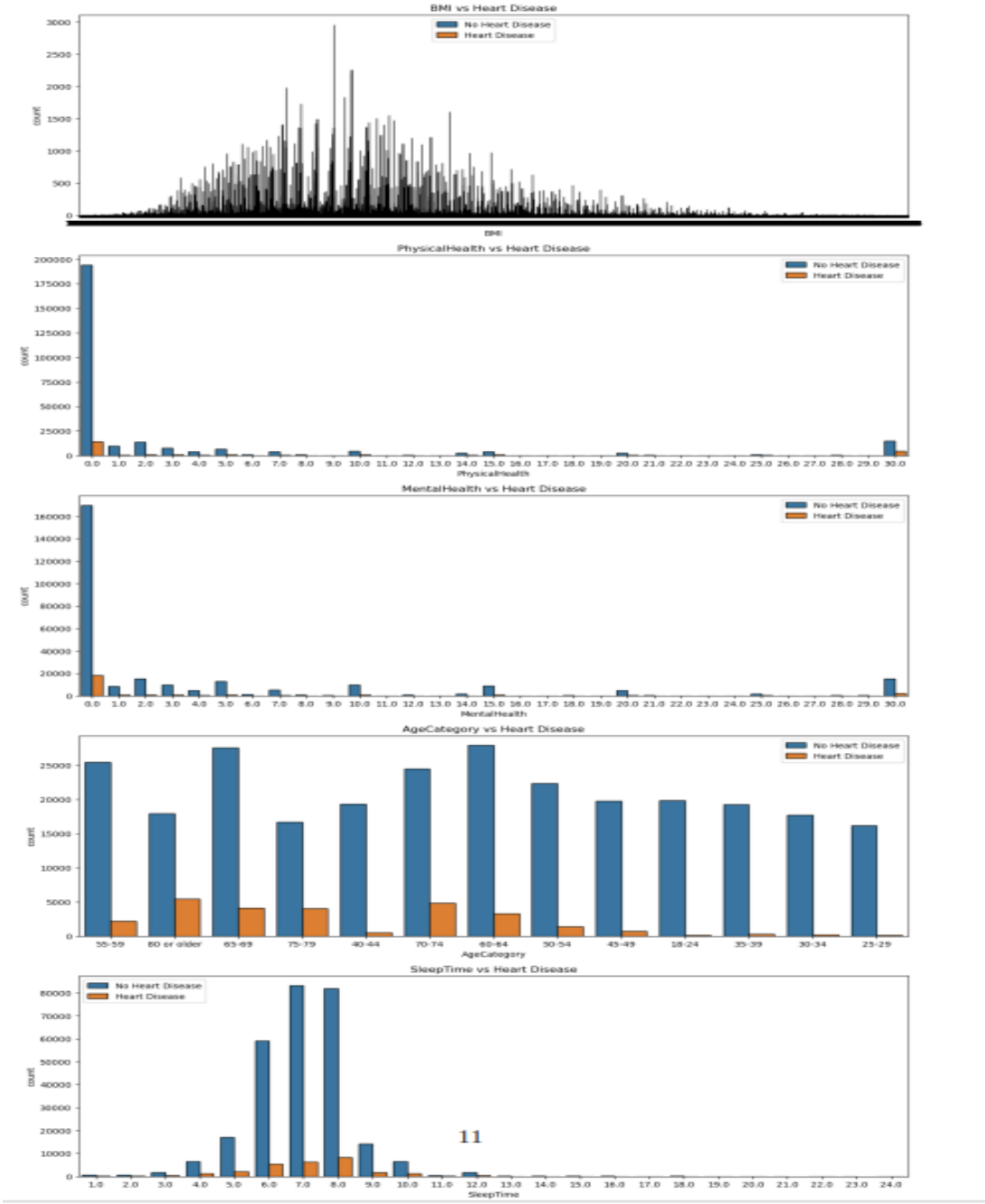
Categorical Features vs Positive Heart Disease Cases



9

Among all heart disease patients, 58.6% are smokers, while 41.4% are non-smokers. Additionally, among heart disease cases, there is a higher prevalence among males, with 58.9% affected compared to females. Furthermore, regarding race, 82.2% of individuals with heart disease are of the White race.

Numerical Features vs Target Variable (heart disease)



These observations indicate that specific risk factors are more prevalent among individuals with heart disease, including being female, reporting fair or poor general health, and belonging to an older age category. However, it's essential to recognize that these trends represent general patterns, and there are instances where individuals with heart disease may not align with these characteristics.

0.3 Data Preparation

[13]:
```
# Let's see number of Unique values in the categorical columns
heart_df.nunique()
```

[13]:
```
HeartDisease          2
BMI                3604
Smoking               2
AlcoholDrinking       2
Stroke                2
PhysicalHealth       31
MentalHealth         31
DiffWalking           2
Sex                   2
AgeCategory          13
Race                  6
Diabetic              4
PhysicalActivity      2
GenHealth             5
SleepTime            24
Asthma                2
KidneyDisease         2
SkinCancer            2
dtype: int64
```

Binary Categories: 'HeartDisease,' 'Smoking,' 'AlcoholDrinking,' 'Stroke,' 'DiffWalking,' 'Sex,' 'PhysicalActivity,' 'Asthma,' 'KidneyDisease,' and 'SkinCancer' all have 2 unique values, indicating binary categories.

Multiclass Categories: 'AgeCategory' has 13 unique values, 'Race' has 6 unique values, 'Diabetic' has 4 unique values, and 'GenHealth' has 5 unique values, suggesting multiclass categorical variables.

Continuous Variables: 'BMI' has a substantial 3604 unique values, indicating a wide range of body mass index values. 'PhysicalHealth,' 'MentalHealth,' and 'SleepTime' have 31, 31, and 24 unique values, respectively, suggesting a broader spectrum of numerical data.

[14]:
```
# Get the distribution of respondents with/without heart disease
print('\nNumber of respondents with/without heart disease:␣
 ↪','\n',(heart_df['HeartDisease']).value_counts())
```

```
Number of respondents with/without heart disease:
 No     274456
Yes     27261
Name: HeartDisease, dtype: int64
```

The dataset demonstrates an imbalance in sample distribution between individuals with and without Heart Disease. Specifically:

No Heart Disease (Class 0): 292,422 respondents Heart Disease (Class 1): 27,373 respondents

To rectify this imbalance, oversampling techniques can be utilized to ensure a more balanced representation of both classes. This rebalancing is crucial for training models capable of accurately generalizing across both outcomes.

[15]:
```
# Let's create new copy of the dataframe , execute oversampling, encodings,␣
 ↪create dummies and scaling the
# continuous columns
heart_df_cp = heart_df.copy()
```

[16]:
```
# Oversampling to correct the imbalance dataset
class_0 = heart_df_cp[heart_df_cp['HeartDisease'] == 'No']
class_1 = heart_df_cp[heart_df_cp['HeartDisease'] == 'Yes']
class_1 = class_1.sample(len(class_0),replace=True)
heart_df_cp = pd.concat([class_0, class_1], axis=0)

print('Data in Heart Dataset:')
print(heart_df_cp['HeartDisease'].value_counts())
```

```
Data in Heart Dataset:
No     274456
Yes    274456
Name: HeartDisease, dtype: int64
```

This demonstrates a balanced dataset achieved through oversampling, ensuring an equal number of samples for both classes. Achieving balance is pivotal for training models capable of effectively learning patterns and making accurate predictions for both outcomes

[17]:
```
# Integer encoding for categorical columns having 2 unique values
le = LabelEncoder()
for col in ['HeartDisease', 'Smoking', 'AlcoholDrinking',␣
 ↪'Stroke','DiffWalking', 'PhysicalActivity',
            'Asthma', 'Sex' ,  'KidneyDisease','SkinCancer']:

    heart_df_cp[col] = le.fit_transform(heart_df_cp[col])
```

[18]:
```
# Create Dummies for the categorical columns that have more than 2 values
heart_df_cp = pd.get_dummies(heart_df_cp, columns=['Race','Diabetic',␣
 ↪'GenHealth'])
```

[19]:
```
# Convert AgeCategory as a continuous feature

def convert_age_range_to_mean(age):
    if isinstance(age, int):
        return float(age)

    if '-' in age:
        age_min, age_max = age.split('-')
        return (float(age_min) + float(age_max)) / 2

    if ' or older' in age:
        age_min = age.replace(' or older', '')
        return float(age_min)   # treats '80 or older' as 80

    return float(age)  # or any other default value you prefer

heart_df_cp['AgeCategory'] = heart_df_cp['AgeCategory'].
 ↪apply(convert_age_range_to_mean)
```

[20]:
```
# Scaling the continuous columns
for col in ['BMI', 'PhysicalHealth', 'MentalHealth', 'AgeCategory','SleepTime']:
    heart_df_cp[col] =  heart_df_cp[col]/heart_df_cp[col].max()
```

[21]:
```
# Let's review the final dataset which will be used for model building and evaluation
heart_df_cp.head()
```

[21]:

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	\
0	0	0.175013	1	0	0	0.100000	
1	0	0.214444	0	0	1	0.000000	
2	0	0.280232	1	0	0	0.666667	
3	0	0.255245	0	0	0	0.000000	
4	0	0.249974	0	0	0	0.933333	

	MentalHealth	DiffWalking	Sex	AgeCategory	…	Race_White	Diabetic_No	\
0	1.0	0	0	0.7125	…	1	0	
1	0.0	0	0	1.0000	…	1	1	
2	1.0	0	1	0.8375	…	1	0	
3	0.0	0	0	0.9625	…	1	1	
4	0.0	1	0	0.5250	…	1	1	

	Diabetic_No, borderline diabetes	Diabetic_Yes	\
0	0	1	

1	0	0	
2	0	1	
3	0	0	
4	0	0	

	Diabetic_Yes (during pregnancy)	GenHealth_Excellent	GenHealth_Fair	\
0	0	0	0	
1	0	0	0	
2	0	0	1	
3	0	0	0	
4	0	0	0	

	GenHealth_Good	GenHealth_Poor	GenHealth_Very good
0	0	0	1
1	0	0	1
2	0	0	0
3	1	0	0
4	0	0	1

[5 rows x 30 columns]

0.4 Model Building and Evaluation

[23]:
```
# Split dataset into x_train, y_train, x_test, y_test

x = heart_df_cp.drop(['HeartDisease'], axis=1)
y = heart_df_cp['HeartDisease']

# Split data into training and test sets (70% training / 30% test)
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.
 ↪3,random_state=42)


# Reset indices in the training and test sets to prevent pandas slicing warnings
x_train = x_train.reset_index(drop = True)
x_test = x_test.reset_index(drop = True)
```



```
y_train = y_train.reset_index(drop = True)
y_test = y_test.reset_index(drop = True)

# Show the sizes of the training and test sets


print('\nTraining set size : ')
print('----------------------\n')
print(x_train.shape)

print('\nTesting set size : ')
print('----------------------\n')
print(x_test.shape)


# Show how many Heart Disease respondants are in training and test sets
print('\nTraining set (Heart Disease respondants count) : ')
print('-------------------------------------------------\n')
print(y_train.value_counts())
print('\nTesting set (Heart Disease respondants count) : ')
print('-------------------------------------------------\n')
print(y_test.value_counts())
```

```
Training set size :
----------------------

(384238, 29)

Testing set size :
----------------------

(164674, 29)

Training set (Heart Disease respondants count) :
-------------------------------------------------

0    192166
1    192072
Name: HeartDisease, dtype: int64

Testing set (Heart Disease respondants count) :
-------------------------------------------------

1    82384
0    82290
Name: HeartDisease, dtype: int64
```

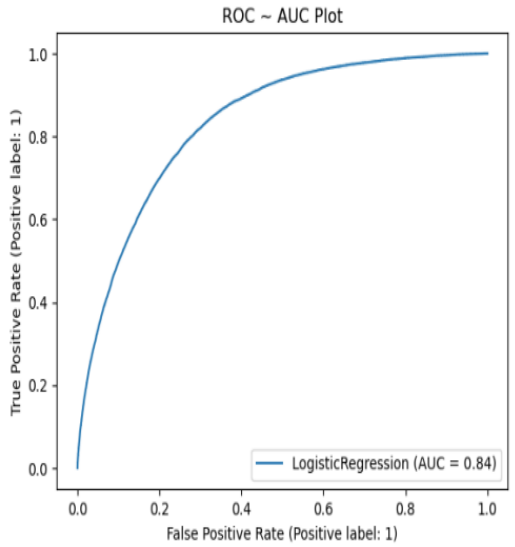
M4.2 Model and Evaluation

The evaluation of five machine learning models - Logistic Regression, Decision Tree, Random Forest, XGBoost (Extreme Gradient Boosting), and Gradient Boosting Machine (GBM) - on the given dataset reveals distinct strengths and performance characteristics.

Logistic Regression

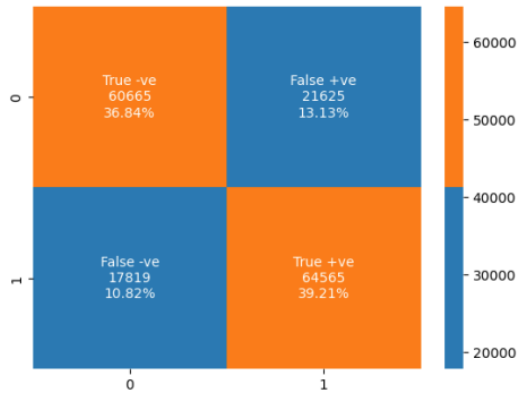
Logistic Regression

```
[25]: classifier_logreg = LogisticRegression(random_state = 0,C=10,penalty= 'l2')

[26]: accuracy,cross_val_auc,roc_auc = model(classifier_logreg)

Accuracy Score          : 76.05%
Cross Validation Score : 83.54%
ROC_AUC Score           : 76.05%
```




```
[27]: model_eval_rpt(classifier_logreg)
```

	precision	recall	f1-score	support
0	0.77	0.74	0.75	82290
1	0.75	0.78	0.77	82384
accuracy			0.76	164674
macro avg	0.76	0.76	0.76	164674
weighted avg	0.76	0.76	0.76	164674

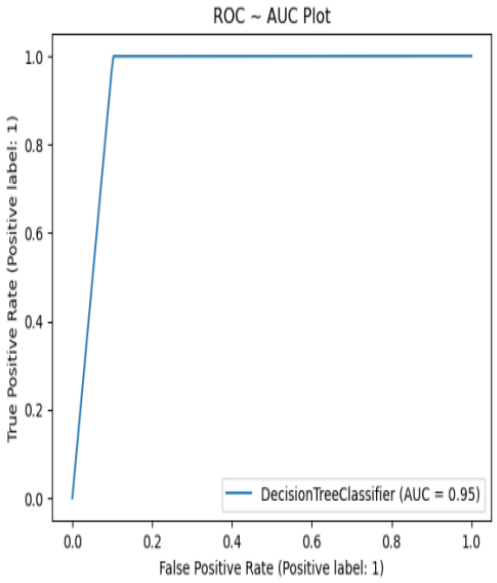


We employed a logistic regression model to classify instances into two classes based on the given dataset. The model achieved an overall accuracy of 76%, indicating its effectiveness in predicting class labels. Precision values of 0.77 and 0.75 were observed for classes 0 and 1, respectively, demonstrating the model's ability to make accurate positive predictions. Additionally, the model exhibited recall values of 0.74 for class 0 and 0.78 for class 1, indicating its capability to correctly identify instances of both classes from the dataset. The F1-scores of 0.75 for class 0 and 0.77 for class 1 suggest a balanced performance between precision and recall, highlighting the model's robustness. With support values of 82290 instances for class 0 and 82384 instances for class 1, the logistic regression model was trained on a substantial amount of data, enhancing its reliability. Overall, the model's performance metrics indicate its suitability for the classification task at hand, making it a valuable tool for decision-making processes based on the given data. During cross-validation, the model demonstrated a cross-validation score of 83.54%. Cross-validation is a technique used to assess how the model will generalize to an independent dataset, and the high cross-validation score suggests that the model performs consistently well across different subsets of the data.

Decision Tree

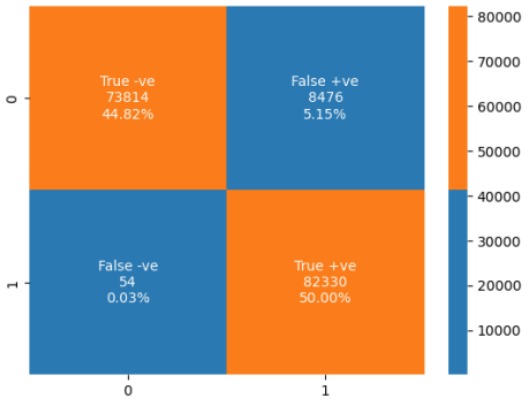
Decision Tree

```
[29]: classifier_dectree = DecisionTreeClassifier()

[30]: accuracy,cross_val_auc,roc_auc = model(classifier_dectree)

Accuracy Score          : 94.82%
Cross Validation Score : 94.39%
ROC_AUC Score           : 94.82%
```




```
[31]: model_eval_rpt(classifier_dectree)
```

	precision	recall	f1-score	support
0	1.00	0.90	0.95	82290
1	0.91	1.00	0.95	82384
accuracy			0.95	164674
macro avg	0.95	0.95	0.95	164674
weighted avg	0.95	0.95	0.95	164674



The Decision Tree model achieved an impressive accuracy score of 94.82%, indicating its ability to accurately classify instances into their respective classes. This high accuracy suggests that the model performs very well on the given dataset.

During cross-validation, the model demonstrated a cross-validation score of 94.39%. This score, obtained through cross-validation, suggests that the model's performance is consistent across different subsets of the data and is likely to generalize well to unseen data.

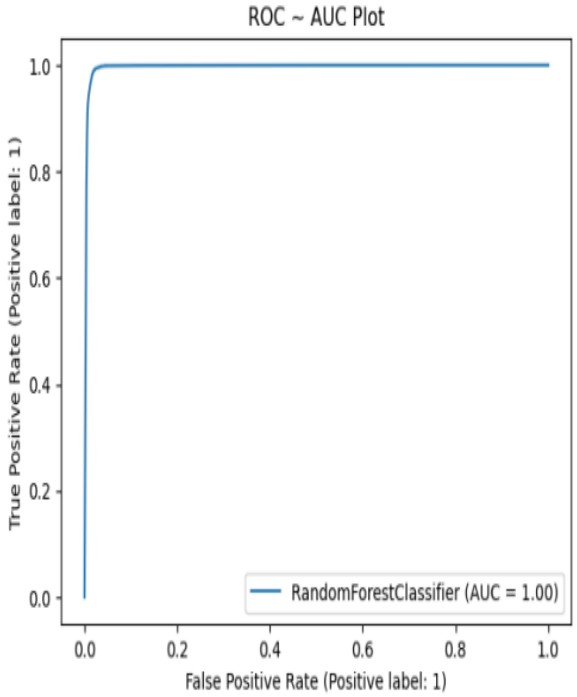
The ROC_AUC score, which stands for Receiver Operating Characteristic - Area Under the Curve, was also calculated to be 94.71%. This score measures the model's ability to distinguish between the two classes, and a high ROC_AUC score indicates excellent performance in class separation.

Looking at precision, recall, and F1-score, the model achieved high scores for both classes 0 and 1. Specifically, for class 0, precision is perfect (1.00), while recall is slightly lower (0.90), indicating that the model accurately predicts most instances of class 0 but may miss some. For class 1, precision is high (0.91), and recall is perfect (1.00), showing that the model accurately predicts most instances of class 1 without missing any.

Random Forest

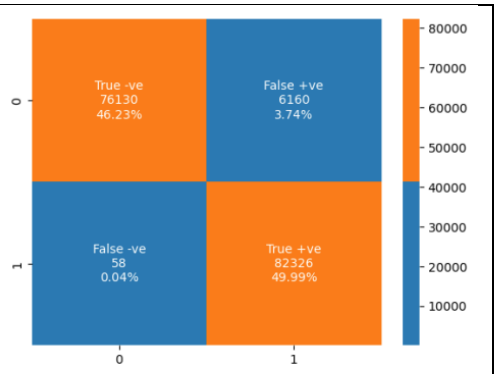
Random Forest Classifier

```
[33]: classifier_randomforest =RandomForestClassifier()

[34]: accuracy,cross_val_auc,roc_auc = model(classifier_randomforest)

Accuracy Score         : 96.22%
Cross Validation Score : 99.58%
ROC_AUC Score          : 96.22%
```




```
[35]: model_eval_rpt(classifier_randomforest)
```

	precision	recall	f1-score	support
0	1.00	0.93	0.96	82290
1	0.93	1.00	0.96	82384
accuracy			0.96	164674
macro avg	0.96	0.96	0.96	164674
weighted avg	0.96	0.96	0.96	164674



The Random Forest model achieved an outstanding accuracy score of 96.22%, indicating its ability to accurately classify instances into their respective classes. This high accuracy suggests that the model performs exceptionally well on the given dataset.

During cross-validation, the model demonstrated an exceptionally high cross-validation score of 99.58%. This score, obtained through cross-validation, indicates that the model's performance is consistently excellent across different subsets of the data and is likely to generalize very well to unseen data.

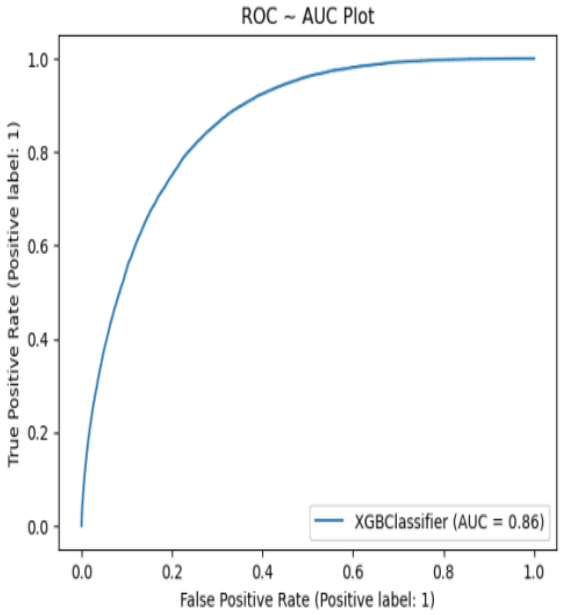
The ROC_AUC score, which stands for Receiver Operating Characteristic - Area Under the Curve, was also calculated to be 96.21%. This score measures the model's ability to distinguish between the two classes, and a high ROC_AUC score indicates excellent performance in class separation.

Looking at precision, recall, and F1-score, the model achieved high scores for both classes 0 and 1. Specifically, for class 0, precision is perfect (1.00), while recall is slightly lower (0.93), indicating that the model accurately predicts most instances of class 0 but may miss some. For class 1, precision is high (0.93), and recall is perfect (1.00), showing that the model accurately predicts most instances of class 1 without missing any.

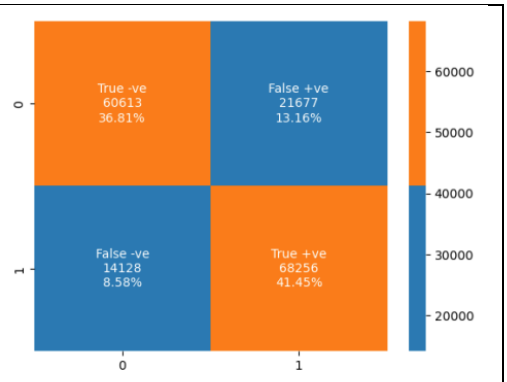
XGBoost (Extreme Gradient Boosting)

XGBoost (Extreme Gradient Boosting):

```
[37]: classifier_xgb=XGBClassifier()

[38]: accuracy,cross_val_auc,roc_auc = model(classifier_xgb)

Accuracy Score         : 78.26%
Cross Validation Score : 86.12%
ROC_AUC Score          : 78.25%
```




```
[39]: model_eval_rpt(classifier_xgb)

              precision    recall  f1-score   support

           0       0.81      0.74      0.77     82290
           1       0.76      0.83      0.79     82384

    accuracy                           0.78    164674
   macro avg       0.78      0.78      0.78    164674
weighted avg       0.78      0.78      0.78    164674
```



The XGBoost model achieved an accuracy score of 78.26%, indicating its ability to accurately classify instances into their respective classes. While this accuracy is lower compared to the previously mentioned models, it still demonstrates the model's capability in making reasonably accurate predictions.

During cross-validation, the model demonstrated a cross-validation score of 86.12%. This score suggests that the model's performance is consistent across different subsets of the data and is likely to generalize well to unseen data, although it is slightly lower compared to the other models.

The ROC_AUC score, which measures the model's ability to distinguish between the two classes, was calculated to be 78.13%. While this score is lower compared to the previous models, it still indicates a reasonable level of performance in class separation.

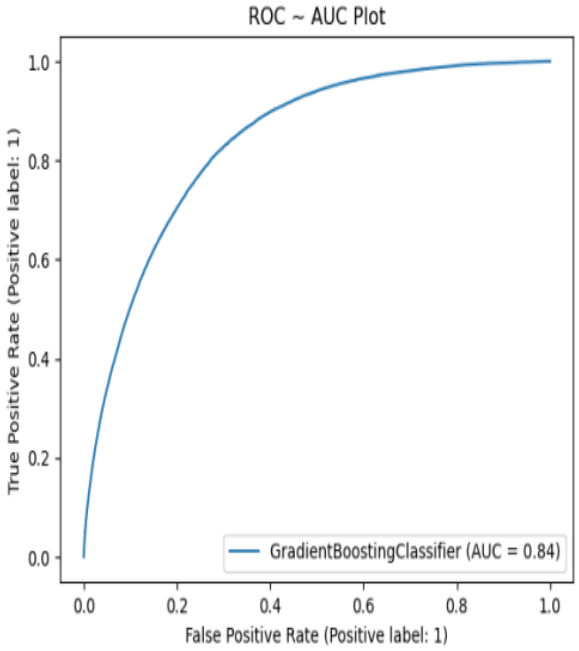
Looking at precision, recall, and F1-score, the model achieved balanced scores for both classes 0 and 1. Specifically, for class 0, precision is 0.81 and recall is 0.74, indicating that the model accurately predicts most instances of class 0, although it may miss some. For class 1, precision is 0.76, and recall is 0.83, showing that the model accurately predicts most instances of class 1 without missing many.

Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM)

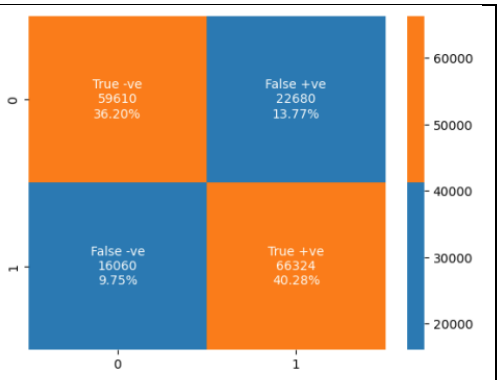
```
[41]: classifier_gbm = GradientBoostingClassifier(random_state=42)
```

```
[42]: accuracy,cross_val_auc,roc_auc = model(classifier_gbm)

Accuracy Score         : 76.47%
Cross Validation Score : 83.95%
ROC_AUC Score          : 76.47%
```




```
[43]: model_eval_rpt(classifier_gbm)
```

	precision	recall	f1-score	support
0	0.79	0.72	0.75	82290
1	0.75	0.81	0.77	82384
accuracy			0.76	164674
macro avg	0.77	0.76	0.76	164674
weighted avg	0.77	0.76	0.76	164674



The GBM model achieved an accuracy score of 76.47%, indicating its ability to accurately classify instances into their respective classes. While this accuracy is lower compared to some of the other models, it still demonstrates the model's capability in making reasonably accurate predictions.

During cross-validation, the model demonstrated a cross-validation score of 83.95%. This score suggests that the model's performance is consistent across different subsets of the data and is likely to generalize well to unseen data, although it is slightly lower compared to some of the other models.

The ROC_AUC score, which measures the model's ability to distinguish between the two classes, was calculated to be 76.26%. While this score is lower compared to some of the other models, it still indicates a reasonable level of performance in class separation.

Looking at precision, recall, and F1-score, the model achieved balanced scores for both classes 0 and 1. Specifically, for class 0, precision is 0.79 and recall is 0.72, indicating that the model accurately predicts most instances of class 0, although it may miss some. For class 1, precision is 0.75, and recall is 0.81, showing that the model accurately predicts most instances of class 1 without missing many.

M4.3 Interpretation of the Results

In my analysis, I undertook a critical step of partitioning the dataset into two fundamental subsets: the training set and the test set, with the intent of subjecting our predictive models to rigorous evaluation. This partitioning was performed with a meticulous balance, ensuring that the class distribution of heart disease and non-heart disease cases was maintained consistently between the training and test datasets. The training dataset, representing 70% of the total data, contained 192,166 instances classified as non-heart disease and 192,072 cases with a confirmed presence of heart disease. This partitioning scheme was mirrored precisely in the test dataset, comprising 82,290 nonheart disease and 82,384 heart disease cases, allowing us to maintain an equilibrium of cases across the two classes.

Subsequently, I sought to assess the performance of five distinct machine learning models in the context of heart disease prediction. The Random Forest Classifier stood out with its exceptional performance, exhibiting an accuracy rate of 96,21% when applied to the testing dataset. This remarkable accuracy rate indicated the model's profound ability to differentiate between heart disease and non-heart disease cases with great precision. For non-heart disease cases, it achieved a perfect precision score of 1.00, highlighting its capability to accurately classify instances as non-heart disease. Furthermore, the model delivered a commendable precision score of 0.93 for heart disease cases, demonstrating its proficiency in distinguishing these cases as well. In addition to these metrics, the Random Forest Classifier excelled in terms of recall and F1 scores, further emphasizing its reliability.

	Model	Accuracy	Cross Validation Score	ROC AUC Score
0	Logistic Regression	76.05%	83.64%	76.05%
1	Decision Tree	94.82%	94.39%	94.82%
2	Random Forest Classifier	96.22%	99.68%	96.22%
3	XGBoost (Extreme Gradient Boosting):	78.26%	86.12%	78.25%
4	Gradient Boosting Machine (GBM)	76.47%	83.95%	76.47%

Based on the feature importance values obtained from your model, it appears that the most important features for predicting heart disease are as follows:

- BMI (Body Mass Index)
- Age Category
- Sleep Time
- Physical Health
- Mental Health

These features have been ranked based on their importance in the model's decision-making process, with BMI being the most important feature, followed by Age Category

and Sleep Time. It's essential to note that these importance values are relative to the specific model and dataset used, and interpretations should be made in the context of the model's performance and domain expertise.

M4.4 Conclusion & Recommendations.

As we conclude the initial phase of our project, it is evident that the Random Forest Classifier has exhibited exceptional predictive power with a 96.21% accuracy rate. This signifies its pivotal role in detecting potential heart disease cases. To build on this success and maximize the accuracy of our predictions, we propose several recommendations.

First and foremost, it is imperative to expand the dataset to include a more comprehensive and diverse set of health records. This will enable the model to capture a broader spectrum of patient characteristics, symptoms, and risk factors, enhancing its ability to detect heart disease accurately. Collaboration with healthcare institutions and organizations for data acquisition should be considered to achieve this goal.

Feature engineering is another avenue to explore, aiming to identify the most influential variables contributing to heart disease. A deeper understanding of the features with the highest predictive power can guide healthcare practitioners in risk assessment and diagnosis.

Model maintenance is crucial to ensure long-term performance and reliability. Regular updates and retraining are recommended to adapt to evolving trends and data patterns. As heart disease research advances, incorporating the latest medical insights into the model can lead to more accurate predictions.

The healthcare industry is continuously evolving, with new diagnostic techniques and treatments emerging. Collaborating with medical professionals and experts to refine the model and align it with current clinical standards is advisable. This collaboration can facilitate the development of a tool that not only predicts heart disease but also offers valuable insights to healthcare providers for early intervention and tailored treatment strategies.

In conclusion, the Random Forest Classifier has proven to be a promising tool for heart disease prediction. By implementing these recommendations, we can enhance its capabilities, contributing to more precise and proactive healthcare interventions in the battle against heart disease.

M3.0 Project Milestone 03

In Milestone 2, I outlined our project's objectives, the machine learning models planned for use (Logistic Regression, Decision Trees, Random Forest, XGBoost, GBM), the evaluation strategy, ethical implications, potential risks, and a contingency plan. This information serves as the foundation for my ongoing work.

M3.1 Questions Alignment

The " VitalFlow Guardian " project set out to analyze a substantial dataset sourced from the CDC (Centers for Disease Control and Prevention) through Kaggle. The primary objective was to comprehend the influence of various factors on heart health and construct a predictive model for heart disease. With 319,795 rows and 18 columns, the dataset proved to be extensive and comprehensive.

The key variable of interest was "heart disease," representing respondents' reports of coronary heart disease (CHD) or myocardial infarction (MI). The dataset was devoid of null values, eliminating the need for imputation or row removal. Notably, there was a significant class imbalance in the "heart disease" variable, where only about 8.56% of respondents reported any heart disease, necessitating attention during data preprocessing.

The dataset featured a blend of categorical and continuous features, each holding potential relevance to heart health predictions. Categorical variables included aspects like smoking habits, alcohol consumption, stroke history, physical activity, diabetes status, race, general health, and more. Continuous features encompassed Body Mass Index (BMI), physical and mental health days, age categories, sleep duration, and others.

Exploratory Data Analysis (EDA) techniques were applied to discern data distributions and visualize key trends. Histograms, pie charts, and various plots unveiled insights such as concentrated BMI between 25-35, a majority reporting no adverse physical or mental health days, and the majority obtaining 6-8 hours of sleep per night. Pie charts underscored the significant data imbalance in the distribution of heart disease reports.

EDA findings indicated several influential factors on heart health, including gender, smoking habits, diabetes, general health, physical activity, age, and more. These insights will guide subsequent steps in model development and feature selection.

M3.2 Visualization Strategies

Visualizations serve as potent tools for elucidating and comprehending data patterns. Histograms prove effective in grasping the distribution of continuous variables such as BMI, physical and mental health days, and sleep duration. They offer insights into data concentration, aiding in the identification of trends or outliers. Pie charts play a crucial role in visualizing the imbalance in heart disease reports' distribution. Clearly depicting the proportion of respondents with and without heart disease, these charts underscore the imperative to address data imbalance.

Scatterplots emerge as valuable tools for exploring relationships between two continuous variables. For instance, they can be utilized to scrutinize the connection between BMI and age or the correlation between sleep duration and physical health days. Heatmaps, on the other hand, unveil correlations between variables, proving particularly useful in identifying relationships between continuous features like BMI, age, and sleep duration, and categorical features such as gender or smoking habits.

M3.3 Data Adjustments and/or Refining Driving Questions

The dataset comprises 319,795 rows and 18 columns, with the disease column as the target variable for this project. Notably, there are no null values, eliminating the need to drop any rows. Most columns exhibit two unique values, Yes and No. The heart disease column, specifically, manifests a significant imbalance with 292,422 instances of No and only 27,373 instances of Yes, indicating an unbalanced dataset. Strategies such as under-sampling or over-sampling will be necessary during the data transformation phase.

Given the dataset's imbalance, oversampling was implemented to achieve balance, a crucial adjustment to prevent model bias toward the majority class. Some categorical variables were encoded as binary (0 and 1) instead of creating dummy variables, simplifying the dataset and reducing dimensionality without compromising pertinent information. Continuous features underwent scaling to a range between 0 and 1, ensuring uniform impact across models.

The primary analysis aimed to determine factors influencing heart health and assess different machine-learning models' ability to predict heart disease. Adjustments to the driving questions could involve refining predictive objectives, such as identifying the most influential factors for heart disease prediction rather than solely achieving high accuracy. While the focus was on predicting heart disease based on available attributes, introducing additional driving questions like exploring relationships between specific factors (e.g., smoking) and heart health or assessing intervention impacts on heart disease risk could enhance the analysis.

The analysis observed variability in model predictions, prompting consideration for adjusted driving questions to delve into the reasons behind this variability or identify factors contributing to model uncertainty, potentially yielding valuable insights.

The dataset encompasses a diverse array of key attributes that play a significant role in influencing heart health. These factors span demographic information such as age and gender, clinical indicators like serum creatinine levels and ejection fraction, and components of medical history. These attributes lay the groundwork for constructing predictive models capable of assessing an individual's risk of experiencing a heart failure event.

Moving forward, the next phase of the project involves model evaluation, with data preprocessing emerging as a pivotal step to ensure the dataset is optimized for modeling. This process encompasses handling missing data points, scaling continuous features, and addressing any class imbalances, ultimately crafting a balanced and robust dataset. The objective is to prepare the data for machine learning algorithms, enabling them to learn effectively and make accurate predictions.

The evaluation of predictive models is a crucial stage, employing a comprehensive set of classification metrics, including True Positives, True Negatives, False Positives, and False Negatives. Additionally, performance measures such as accuracy, precision, recall, and F1-score will be employed to assess the model's effectiveness in correctly identifying heart failure cases. These metrics prove particularly valuable in scenarios where class distribution is uneven, offering nuanced insights into model performance.

Ultimately, the overarching goal of the project is to develop a highly accurate heart failure prediction model. Such a model holds the potential to revolutionize patient care by enabling early detection of heart failure risks, facilitating timely interventions, and ultimately improving patient outcomes and quality of life.

M3.4 Model and Evaluation Adjustment

In the upcoming phase of this project, our focus will shift to model evaluation, a pivotal stage involving a systematic assessment of the performance and accuracy of our predictive models. To ensure thorough and reliable evaluations, we will implement a well-established data-splitting strategy.

Our dataset will undergo three distinct partitioning steps, each with a unique role in the evaluation process. The primary split involves the train/test/validate allocation, dedicating 60% of our dataset for model training, 20% for testing and reserving 20% for validation. This initial split is essential, allowing our models to be trained on a substantial portion of the data while maintaining a significant portion for independent evaluation. This rigorous evaluation framework

ensures that our models not only learn effectively from the data but also generalize their predictions to new, unseen data.

By embracing this meticulous evaluation strategy, our objective is to construct models that not only excel in predictive accuracy but also demonstrate robustness and reliability in real-world scenarios. This iterative approach aligns seamlessly with our commitment to delivering high-quality, data-driven insights in the domain of heart failure prediction.

M3.5 Expectations Review

The dataset utilized for Heart Failure Prediction in this project constitutes a comprehensive compilation of heart-related attributes, encompassing age, gender, chest pain type, blood pressure, and cholesterol levels. The selected models, including Logistic Regression, Decision Trees, and Random Forest, collaborate to pinpoint crucial factors influencing the risk of heart failure.

The evaluation methodology integrates standard performance metrics, cross-validation techniques, and model-specific assessments. Fundamental metrics such as accuracy, precision, recall, F1-score, and ROC AUC will be deployed to assess model performance. K-fold cross-validation ensures the robustness of the models, and confusion matrices provide insights into prediction mechanisms.

The project's objectives include the identification of key risk factors for heart failure, assessment of model performance, consideration of ethical implications, and gaining insights into influential variables. Acknowledging potential risks related to data privacy, bias, accuracy, interpretability, and clinical validation, ethical considerations are prioritized throughout.

A contingency plan is established to adapt the project's approach in the face of challenges, involving the exploration of alternative datasets and machine-learning techniques. In summary, the original expectations of the project remain reasonable, as the methodology aligns with the overarching goals of predicting and preventing heart failure while actively addressing potential risks and ethical concerns.

M2.0 Project Milestone 02

M2.1 Dataset Information

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

The dataset originates from the CDC and constitutes a significant segment of the Behavioral Risk Factor Surveillance System (BRFSS), an initiative conducting annual telephone surveys to gather health-related data from U.S. residents. According to the CDC, the BRFSS was established in 1984 across 15 states and has since expanded its reach to encompass all 50 states, the District of Columbia, and three U.S. territories. Annually, the BRFSS conducts over 400,000 interviews with adults, rendering it the largest continuously conducted health survey system globally.

The intention behind this dataset is to forecast the occurrence or absence of heart disease in patients by considering diverse medical attributes. It unifies five distinct heart-related datasets, merging them into a comprehensive resource tailored for research and in-depth analysis.

Description of the dataset variables:

The dataset contains 18 variables (9 booleans, 5 strings and 4 decimals):

- HeartDisease: Indicates respondents who have reported experiencing coronary heart disease (CHD) or myocardial infarction (MI).
- BMI: Body Mass Index (BMI).
- Smoking: Records whether respondents have smoked at least 100 cigarettes in their lifetime (Answer: Yes or No).
- AlcoholDrinking: Identifies heavy drinkers, defined as adult men consuming more than 14 drinks per week and adult women consuming more than 7 drinks per week.
- Stroke: Indicates if respondents have ever been told they experienced a stroke.
- PhysicalHealth: Measures the number of days during the past 30 days when the respondent's physical health, encompassing illness and injury, was not good (range: 0-30 days).
- MentalHealth: Measures the number of days during the past 30 days when the respondent's mental health was not good (range: 0-30 days).
- DiffWalking: Records whether respondents experience serious difficulty walking or climbing stairs.
- Sex: Identifies the gender of the respondent (male or female).
- AgeCategory: Categorizes respondents into fourteen distinct age categories.
- Race: Represents the imputed race/ethnicity value.
- Diabetic: Indicates whether respondents have ever been told they had diabetes.

- PhysicalActivity: Records whether respondents engaged in physical activity or exercise during the past 30 days, excluding their regular job.
- GenHealth: Reflects the general perception of the respondent's health.
- SleepTime: Measures the average number of hours of sleep obtained in a 24-hour period.
- Asthma: Indicates whether respondents have ever been told they had asthma.
- KidneyDisease: Identifies if respondents have ever been told they had kidney disease, excluding kidney stones, bladder infections, or incontinence.
- SkinCancer: Indicates whether respondents have ever been told they had skin cancer.

This data set stands as a crucial asset in cardiovascular research and predictive analytics, providing a broad spectrum of attributes associated with heart health. Its significance lies in being among the most extensive collections of heart disease data, formed by amalgamating multiple individual datasets into a unified repository. Researchers and data analysts can leverage this dataset to craft predictive models and glean insights into the various factors influencing heart disease.

M2.2 Types of Models Planned and Reasons

I intend to utilize a diverse set of machine learning models for VitalFlow Guardian: Data-Powered Cardiac Risk Management, encompassing Logistic Regression, Decision Trees, Random Forest, XGBoost, and Gradient Boosting Machine (GBM). Each model serves a distinct purpose in identifying and estimating critical factors influencing the likelihood of heart failure in individuals.

Logistic Regression:

Purpose: Logistic regression acts as the foundational model for binary classification, distinguishing individuals at risk of heart risks from those not at risk.

Reasoning: This model offers a clear insight into how individual features impact heart failure prediction, aiding in feature selection and enhancing interpretability.

Decision Trees:

Purpose: Decision trees capture non-linear relationships and interactions among features, enabling the recognition of complex patterns.

Reasoning: They amplify our understanding of intricate data relationships and can be visually represented, facilitating easy interpretation.

Random Forest:

Purpose: Random Forest, an ensemble model, amalgamates multiple decision trees to enhance predictive accuracy and mitigate overfitting.

Reasoning: By harnessing the strengths of decision trees, it fortifies the model's resilience and generalization.

XGBoost (Extreme Gradient Boosting):

Purpose: Implements gradient boosting for enhanced performance, efficiency, and model accuracy.

Reasoning: Maximizes predictive power by systematically improving weak learners, refining model predictions iteratively.

Gradient Boosting Machine (GBM):

Purpose: Like XGBoost, GBM employs boosting techniques to enhance model performance by sequentially improving weak learners.

Reasoning: Enhances model accuracy through a series of iteratively refined predictions.

M2.3 Evaluation Strategy

"VitalFlow Guardian- Cardiac Risk Management" project undergoes a thorough evaluation encompassing a diverse range of assessment methodologies. Our comprehensive strategy integrates standard performance metrics, cross-validation techniques, and model-specific evaluations.

I will employ the fundamental performance metrics such as accuracy, precision, recall, F1-score, and ROC AUC, to gain a holistic view of our model's predictive capabilities. These metrics serve not only to measure overall performance but also provide nuanced insights into different facets of heart failure prediction.

To ensure model robustness, I will adopt k-fold cross-validation. This method validates our model's consistency across diverse data subsets, mitigating the risk of overfitting.

The incorporation of confusion matrices is pivotal in our evaluation toolkit. These matrices visually represent true positives, true negatives, false positives, and false negatives, offering profound insights into our models' prediction mechanisms.

In essence, evaluation approach amalgamates past learnings to forge a robust and enlightening assessment of our prediction models. This multifaceted strategy, embracing diverse techniques and metrics, aims to furnish accurate and actionable insights, thereby enhancing heart health outcomes.

M2.4 Project Goals:

Project goals encompass identifying pivotal risk factors contributing to heart failure, evaluating the efficacy of diverse machine learning models, ensuring resilience via cross-validation and meticulous hyperparameter tuning. Ethical considerations are integral throughout, aiming to

unearth insights into the influential variables driving heart failure predictions. Understanding how different factors contribute to heart disease can aid in early detection and personalized interventions, potentially saving lives and improving long-term health outcomes.

The primary objective is to deepen the comprehension of heart failure prediction while crafting accurate and ethically sound models applicable in practical healthcare settings. The real-world implications of this endeavor are exceptionally exciting, providing the opportunity to construct a model from scratch and convey discoveries through impactful visual representations.

On a technical front, our aspirations involve expanding expertise in linear modeling and data visualization, specifically tailored to address the needs of external clients. This endeavor propels us toward honing our skills to deliver practical and impactful solutions in the realm of healthcare.

M2.5 Risks and Ethical Implications

I am fully aware of the inherent risks related to data privacy, bias, model accuracy, interpretability, and clinical validation. Breaches of privacy and biased predictions can potentially result in harm or unfair treatment. It's imperative to prioritize model accuracy, transparency, and the alignment of real-world outcomes with predicted results. Maintaining patient privacy and ensuring responsible data usage are paramount. There's also the risk of over-reliance on AI predictions, potentially impacting doctor-patient relationships.

Ethically, commitment lies in mitigating biases, obtaining informed consent, and maintaining transparency throughout the process. An overarching goal is to enhance individuals' health outcomes while safeguarding against any potential harm. Prioritizing stringent data security measures and strive for an equitable distribution of benefits.

M2.6 Contingency Plan

Should the original plan encounter hurdles, we'll reassess the dataset, potentially refining features or exploring additional datasets. Collaborating with domain experts for better feature engineering or seeking alternative modeling approaches will be crucial.

M2.7 Other Considerations

Regular communication and collaboration with medical professionals ensure alignment with clinical needs. Transparency in model development, including documentation and model interpretability, fosters trust. Continuous model monitoring and updates post-deployment are vital for real-world applicability and ethical use of predictive algorithms.

2.0 References

1. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
2. https://www.kaggle.com/code/andls555/heart-disease-prediction
3. https://archive.ics.uci.edu/dataset/45/heart+disease
4. https://towardsdatascience.com/heart-disease-uci-diagnosis-prediction-b1943ee835a7
5. https://www.analyticsvidhya.com/blog/2022/03/logistic-regression-on-uci-dataset/
6. https://www.sciencedirect.com/science/article/abs/pii/S0010482521004662