University of Essex

School of Mathematics, Statistics
and Actuarial Science

MA981 DISSERTATION

ADVANCING CARDIAC CARE: MACHINE LEARNING APPLICATIONS IN HEART DISEASE PREDICTION

OBED MAWUKO KWADZO BANINI

Supervisor: **DR. NA YOU**

January 14, 2024
Colchester

Contents

1 Introduction and Abstract 7
1.1 Background 8
1.2 Objectives 9
1.3 Research Structure 10

2 Review of related literature 11
2.1 Definition of Machine Learning and It's Models 11
2.2 Overview of general approaches and related works 15

3 Methodology 19
3.1 Data loading and cleaning 20
3.2 Exploratory data analysis 20
3.3 Data pre-processing and transformation 21
3.4 Hypothesis testing 21
3.5 Splitting and scaling data 22
3.6 Feature Engineering 22
3.7 Building Initial models 22
3.8 Model Evaluation 23
3.9 Build Alternative Models and evaluate results 23
3.10 Best model selection and further assessment of model 24

4 Results and Discussion 25
4.1 Exploratory data analysis 25
4.2 Results of Research objectives 31
4.2.1 Hypothesis Testing 31
4.2.2 Racial Disparities 34

4.2.3 Predictive models 35

5 Conclusions 45

A A Long Proof 47

A.1 Numeric Distributions in Heart Data 47

A.2 Categorical Distributions in Heart Data 47

A.3 Logistic Regression Results of Demographic and lifestyle factors for Predicting Heart Disease 47

B Additional Figures 51

B.1 Skin Cancer Prevalence by Race 51

C Another Appendix 53

List of Figures

4.1 Distribution of Heart Disease Cases 25

4.2 Box Plot comparing Numeric Distributions of Heart Disease cases 27

4.3 Box Plot comparing Numeric Distributions of Heart Disease cases 30

A.1 Distributions of Physical, Mental health Days and Sleep Time. 48

A.2 Categorical Distributions in Heart Data. 49

B.1 Proportional Differences in Heart Disease Across Various Variables . . . 52

List of Tables

4.1 Summary Statistics of Numeric Categories . . . 26
4.2 Comprehensive Logistic Regression Results for Health-Related Factors Predicting Heart Disease . . . 33
4.3 Counts of Individuals by Race . . . 34
4.4 Proportion of Individuals with Heart Disease by Race . . . 34
4.5 Performance Metrics of Machine Learning Models for Heart Disease Prediction . . . 35
4.6 Classification Report of Logistic regression with smote and other opitmization techniques(Hyperparemeter tuning, threshold optimization, cross validation) . . . 37
4.7 Classification Report of xgb with class weight on scaled top 7 features . 39
4.8 Classification Report of xgb with smote on scaled top 7 features . . . 39
4.9 Classification report of xgb with class weight and all features unscaled . 39
4.10 Classification Report of xgb with class weight, best hyperparameters, 3 fold cross validation, threshold optimization on top 7 features train data 40
4.11 Classification report with all features using class weight and all other optimized features . . . 40
4.12 Classification Report of xgb trained with all features, smote and all other optimized methods already mentioned . . . 40
4.13 Predicted vs Actual Output . . . 41
4.14 Error Analysis . . . 42
4.15 Top 10 Features by Importance . . . 42

A.1 Logistic Regression Results of Demographic and lifestyle factors for Predicting Heart Disease . . . 50

B.1 Proportion of Skin Cancer Cases in Different Racial Categories 51

C.1 Classification Report on Downsampled Majority Class with Optimization techniques 53

C.2 Classification Report with SMOTEENN and Optimization techniques . . 54

CHAPTER 1

Introduction and Abstract

Abstract

Coronary Heart Diseases are one of the most common causes of deaths worldwide. Major risk factors include Black race, smoking, obesity and the like. Machine learning models are one of the many technological advancements that are being of great help in the field of medicine. Machine learning models are being used to predict various forms of heart diseases using data obtained from patients pertaining to risk factors. The goal is to ensure that heart diseases are detected early in order to prevent fatalities and associated complications. It will also provide a better understanding of predictors of coronary heart diseases and support the development of prevention strategies targeted for individual patients.Therefore, the main objective of this study is to leverage advanced data science techniques for an in-depth analysis modelling of heart disease risk factors and build a predictive framework for heart disease predictions.

Various machine learning models, including Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors, are evaluated. While these models exhibit high accuracy, their recall for heart disease cases is low. Further exploration involves the application of Extreme Gradient Boosting (XGBoost), with the model trained on class weight on all features proving the most effective, achieving a high ROC AUC score of 83.15% and 76% recall score.

Despite the success, the study has limitations, and the generalizability of the models is cautioned. Future studies should focus on refining models and considering a broader range of factors for improved applicability in diverse healthcare scenarios.

1.1 Background

Coronary heart disease (CHD) is the most common type of heart disease, and it is described as a disease of the coronary arteries [1].

Definition 1.1. It is usually referred to as coronary artery disease (CAD), which is the development of atherosclerotic plaques within the walls of the coronary arteries and as a result reducing the flow of blood to the heart muscle and causing a reduction in the strength of the muscle walls.[1].

In recent times, the worldwide prevalence of coronary heart diseases is 1,655 per 100,000 people [1]. Globally, Latin America and the Middle East have the highest numbers of coronary heart disease[2]. it was predicted that the number of deaths from CAD would rise in developing nations such as South America, Asia, Latin America, Sub-Saharan Africa, and the Far East from nine million in the year 1990 to nineteen million in the year 2010.[2].

Although coronary artery disease (CAD) affects over 20 million Americans, there are notable racial and ethnic differences in the prevalence of the condition [3]. Black Americans have a greater CAD death rate than White Americans, even though they may have a lower prevalence of CAD with more blacks dying from the disease at a reported rate of 21% greater than whites. [3].This may be due to the presence of CAD risk factors, socio-economic variables, delayed presentation and diagnosis of coronary heart diseases [3].

Intriguingly, Black populations tend to experience higher rates and earlier onset of heart disease compared to Whites, despite lower or similar levels of known risk factors like smoking. This discrepancy poses critical questions about the underlying factors contributing to heart disease, making it a compelling area for data-driven investigation, hence the motivation behind this study. Advanced computing technologies and the widespread use of electronic medical record databases have enabled the integration of artificial intelligence (AI) and machine learning in clinical research [4]. Machine learning methods show promise over traditional risk assessment tools, such as those from the American College of Cardiology or American Heart Association. These advancements extend to multiple metrics involved in predicting the risk, incidence, and outcomes of cardiovascular disease (CVD) [4]. Due to the number of deaths that occur as a result of heart disease,it is important that a technique which is reliable and efficient be developed to detect the heart disease early enough to reduce the number of deaths hence the importance of data science techniques and machine learning modelling [5] which is what part of this study aims to achieve. That is to develop a practical predictive model that accurately assesses the risk of coronary heart disease using real-world, imbalanced data sets. This model will be designed to effectively handle the diversity and irregularities often found in practical data scenarios. Special attention will be given to training and fine-tuning the model to ensure it remains reliable and precise in its predictions, particularly when dealing with the class imbalances that are typical in real-world datasets.

The next sections will outline the research objectives as well as the general structure of the study.

1.2 Objectives

The main objectives of this study is to leverage advanced data science techniques for an in-depth analysis and modelling of heart disease risk factors to build a machine learning algorithm that effective predicts heart disease. This will facilitate a better understanding of its predictors, help in early detection of the Coronary Heart Disease and support the development of targeted prevention strategies. This objective will further be broken down into the following;

1. Establish whether demographic, lifestyle, and health-related factors significantly affect the occurrence of heart disease? This will be done by testing the hypothesis.

2. To uncover whether racial disparities exist in the occurrence of heart disease, and if so, what are the variations in the proportion of positive cases within different racial groups.

3. Build an effective machine learning algorithm for predicting coronary heart disease given the presence of certain risk features.

1.3 Research Structure

The dissertation will apply advanced data science techniques to analyze heart disease predictors. It will offer insights into how demographic, lifestyle, and health-related factors, race, contribute to heart disease risk. The research aims to develop a predictive model for early identification of high-risk individuals, contributing to targeted prevention and intervention strategies. This will start by an introduction, which talks about motivation for the study, set tone for what this study seeks to achieve as well as the general structure of study. The next chapter is a review of related literature on works done on heart disease prediction models. The third chapter will document exactly how heart data was used to answer the research objectives and the results produced. The final chapter will look at the conclusions of the study.

CHAPTER 2

Review of related literature

This chapter will explain the methods used in the study as well as what others have done in trying to build models that can effectively predict Coronary Heart Disease.

2.1 Definition of Machine Learning and It's Models

This section will look to explain Machine Learning and key models used in this study.

Definition 2.1. Machine learning is an emerging area in the field of technology, which is designed to mimic the intelligence of man [6]. The aim of this field in computer science is to enable computers to acquire knowledge through learning, eliminating the necessity for explicit programming [14]. Machine learning is one of the many fields in advanced technology that is currently developing in our society to make life easier for everyone [14].

Machine learning can be applied in various fields and can be used for different purposes such as how to design autonomous mobile robots that learn to navigate from their own experience, how to data mine historical medical records to learn which future patients will respond best to which treatments, as well as making predictions for early detection of diseases and how to build search engines that automatically customize to their user's interests [7]. Common algorithms that makes machine learning able to make these predictions are logistic regression, decision tree, k-nearest neighbors, gradient booster, random forest, among others.

Logistic regression is an efficient and powerful way to analyze the effect of a group of independent variables on a binary outcome by quantifying each independent variableâs unique contribution [8]. Basic assumptions that must be met for logistic regression to work efficiently include independence of errors, linearity in the logit for continuous variables, absence of multicollinearity, and lack of strongly influential outliers.Additionally, it is crucial to have a sufficient number of occurrences for each factor to avoid creating a model that is too tailored (overfit). Experts typically recommend having a minimum of 10 to 20 events for each variable being studied [15]. Moreover, regression models operate under the assumption that the link between predictor variables and the outcome is consistent, whether positive or negative, linear or nonlinear, across the entire range of values [9].

Logistic regression helps us understand how two choices (like alive or dead, success or failure, yes or no) are related to different factors. These factors can be categories or numbers. The key to a good logistic regression model is picking the right factors. Even though it might seem smart to include many factors, doing that could make it harder to find real connections. It might give us results that are not very accurate or even show connections that aren't actually there. [9].

Random Forest is a type of advanced computer learning that's been developed recently. In this method, it uses a bunch of decision-makers, called tree classifiers, working together. Each decision-maker is created using a random set of information from the original data, and they all vote on the best category for a given situation. [10]. In the Random Forest approach, decision trees serve as the fundamental classifiers, and multiple trees are generated. The randomness is incorporated through two methods: firstly, by randomly picking data for creating bootstrap samples, following a bagging-like approach, and secondly, by randomly selecting input features when constructing each individual decision tree. [11].

The K Nearest Neighbor (kNN) classifier is employed to categorize unlabeled observations by associating them with the class of the most similar labeled examples [12]. Among the diverse machine learning algorithms available, the K Nearest Neighbor (KNN) algorithm stands out as one of the most commonly used, as emphasized by Uddin (2022) [13]. The different KNN variants include Classic one, Adaptive, Locally adaptive, k-means clustering, Fuzzy, Mutual, Ensemble, Hassanat and Generalised mean distance [13]. Generally, the K Nearest Neighbor (KNN) algorithm classifies datasets using a training model similar to the testing query, considering the k nearest training data points (neighbors) closest to the query being tested [13]. Subsequently, the algorithm employs a majority voting rule to finalize the classification [13]. Among machine learning algorithms, KNN is recognized for its simplicity and widespread use in classification tasks due to its adaptive and easy-to-understand design [13]. The algorithm is well-known for its effectiveness in solving regression and classification challenges across diverse data sizes, label numbers, noise levels, ranges, and contexts [13].

Gradient boosting machines, or GBMs, are sophisticated learning systems that continuously enhance their predictions by adding new models sequentially, each dedicated to correcting the mistakes of the preceding ones [14]. The fundamental concept revolves around constructing these new models to align closely with the areas where the collective learning system made errors. The process involves utilizing a loss function to measure the discrepancies between predictions and actual outcomes. While various loss functions are available, a common example is the squared-error loss, where the system systematically reduces errors by squared amounts during the learning process [14].

One notable feature of GBMs is their adaptability and its ability to handle of linear and non linear relationship between dependent and independent variables. They can be finely tuned to suit unique data scenarios, allowing for a broad spectrum of possibilities in customizing the models to specific requirements. However, finding the optimal configuration often requires a degree of trial and error, similar to adjusting a versatile tool to handle various tasks [14].

In addition to the models, There are various metrics used to assess the efficiency of a machine learning model [28]. These are precision, recall, f1 score, area under the curve (ROC AUC), accuracy and others.

Recall: The proportion of actual positive cases that are predicted correctly[29].

Precision: A measure of how often an ML model is correct when predicting the target class[29].

The F1 score is a blend of precision and recall, penalizes outliers in either metric. Its asymmetry is contingent on the chosen positive and negative classes, as it relies on their definition[[30]]. As a result, researchers will rather opt for F1 score frequently along with accuracy in the evaluation of their built model[31]. Receiver Operating Characteristic (ROC) curves visually demonstrate the effectiveness of a risk prediction model in differentiating between individuals with a specific condition and those without it. The curve presents a graphical depiction of the balance between sensitivity and specificity across different threshold values utilized for classification. Essentially, it illustrates the model's capability to accurately identify positive cases while minimizing false positives[32].

Also some common terms you will come across in this write up are explained below;

[Feature Engineering:] The process involves transforming the existing feature space using mathematical functions. The primary objective is to reduce modeling errors associated with a specific target[33].

Machine learning algorithms usually reframe problems as optimization challenges and use a variety of optimization techniques to solve them[34]. The optimization function comprises multiple **hyperparameters** that are set before the learning process. These hyperparameters play a key role in shaping how the machine learning algorithm adjusts the model to suit the data. It's important to note that hyperparameters are distinct from internal model parameters, such as the weights in a neural network, which are learned from the data during the model training phase[34].Before initiating the training phase, the objective is to identify a set of hyperparameter values that yield optimal performance on the data within a reasonable timeframe. This procedure, known as hyperparameter optimization or tuning, is crucial for enhancing the prediction accuracy of machine learning algorithms[34]. The goal is to fine-tune these hyperparameter values to achieve the best possible model performance during the subsequent training phase.

In the context of machine learning, class weights serve as a mechanism to address class imbalance within a dataset during the training phase[35]. Class imbalance arises when the number of instances in different classes varies significantly, typically occurring when one class has a substantially larger number of samples compared to others[35]. Utilizing class weights helps the algorithm appropriately account for and mitigate the impact of imbalances, ensuring fair consideration of all classes during the learning process.

Cross-validation is a widely employed technique in machine learning for comparing and selecting the most suitable model for a given problem[36]. This method involves dividing the dataset into subsets, such as k folds, train-test splits, and similar strategies. Each subset is used alternately as both a training and a testing set, allowing for a comprehensive evaluation of a model's performance across different data partitions. By assessing a model's consistency and effectiveness across various subsets, cross-validation aids in making informed decisions about model selection and ensures a more robust evaluation of its generalization capability..

2.2 Overview of general approaches and related works

Coronary artery disease (CAD) is a cardiovascular disease which has been found to be the leading cause of death in both developed and developing countries [15]. CAD commonly manifested as stable angina, unstable angina, myocardial infarction (MI), or sudden cardiac death [15]. Atherosclerosis, which is the main cause of cardiovascular diseases, does not have a single cause but a myriad of causes, which therefore makes it a multifactorial disease [16]. Risk factors of the disease include gender, age, heredity, hypercholesterolemia, smoking, hypertension, obesity and sedentary lifestyle, diabetes mellitus, metabolic syndrome, chronic renal failure and stress [16].

Research studies conducted in New Zealand and at Uppsala University in Sweden grouped the risk factors for CAD into two main categories [17]. Ethnicity, age, gender, and family history of CAD are non-modifiable risk factors while sedentary lifestyle, poor diet, smoking, obesity, diabetes, hypertension, hyperlipidemia, and stress are examples of modifiable risk factors [17]. In the context of age, studies conducted by Carnethon et al. suggest a notable increase in the risk of coronary artery disease (CAD) after the age of 35. Beyond the age of 40, the lifetime probability of acquiring CAD is reported to be 32% for women and 49% for men [17]. Additionally, men exhibit a considerably higher risk of developing CAD when compared to women [17].

Ethnic groups with a higher risk of CAD morbidity and mortality include Blacks, Hispanics, Latinos, and Southeast Asians [17]. An important additional risk factor is family history. A higher risk of CAD death exists for patients under 50 years old who have a family history of early heart illness [17]. Smoking is one of the major risk factors for coronary artery disease, with frequent smokers being at greater risk of cardiovascular events. Cessation of smoking is associated with a marked reduction in the risk of cardiovascular diseases [18].

In a study done by Zhang et al., an inverse association was shown for moderate alcohol consumption and the risk of CHD even though alcohol consumption has been consistently considered an important risk for some chronic diseases, including hypertension and diabetes [19]. Studies done by Wolk and colleagues also highlighted the fact that primary sleep abnormalities are associated with cardiovascular diseases [20]. In contrast, depressive disorders are recognized for causing disruptions in sleep patterns, diminished physical activity, and challenges in adhering to health advice. These factors are associated with a higher probability of developing cardiovascular conditions. [21].

Machine learning (ML) is used in various fields. In the data sets related to medicine, ML techniques aid in the prevention of cardiac conditions [22]. Finding such crucial information gives researchers important new perspectives on how to apply their diagnosis and treatment for a specific patient [22] Healthcare providers can also forecast diseases with the help of researchers that analyze vast volumes of complex healthcare data using a variety of Machine Learning techniques [22].

In a research conducted by Chala Beyene et al., it was recommended that data mining techniques should be used to predict and analyze the occurrence of heart diseases. Predicting the emergence of cardiac disease is the primary goal in order to quickly and automatically diagnose the condition early on to prevent complications and fatalities [23].

In another study by Muhammad et al., a dataset labeled by medical experts for coronary artery disease (CAD) was employed in various machine learning algorithms, including support vector machine, K nearest neighbor, random tree, NaÃ¯ve Bayes, gradient boosting, and logistic regression. These algorithms aimed to construct predictive models for CAD diagnosis. The random forest-based machine learning model demonstrated the highest accuracy, achieving 92.04%. In terms of specificity, the Naive Bayes-based machine learning model excelled with a specificity of 92.40%. For sensitivity, the support vector machine-based machine learning model showed the highest performance at 87.34%. Regarding the Receiver Operating Characteristic (ROC), the random forest-based machine learning model emerged as the most effective with a sensitivity of 92.20% [24].

In a study by Yadav and colleagues, an analysis of different algorithms using a designated dataset for predicting coronary heart diseases revealed the Decision Tree algorithm and Random Forest classifier as top performers, both achieving an accuracy of 97.08%. Logistic Regression followed with the second-highest accuracy at 80.52%, while the K-NN algorithm demonstrated the lowest accuracy, reaching 70.13%.[25].

In a research conducted by Ramesh et al., the findings indicated that Machine Learning methods exhibited superior performance compared to statistical techniques. This study substantiates the conclusions drawn by various researchers, suggesting that the utilization of Machine Learning models is the optimal approach for predicting and classifying heart disease [22].

Alotaibi et al. conducted a study where diverse machine learning techniques were utilized to analyze data and predict the probability of heart disease within a medical database [9]. Moreover, utilizing real patient data, medical information systems could be employed to predict heart diseases or other illnesses, as highlighted by the comparative study and its results. The model employed exhibited enhanced accuracy compared to previous models [25]

Linear Regression proves to be an apt approach for predicting the likelihood of heart disease, as evidenced by the Multiple Regression Model for Prediction of Heart Disease presented by K. Polaraju et al. The findings from this investigation showcase that the Regression method exhibits superior classification accuracy compared to alternative algorithms [26].

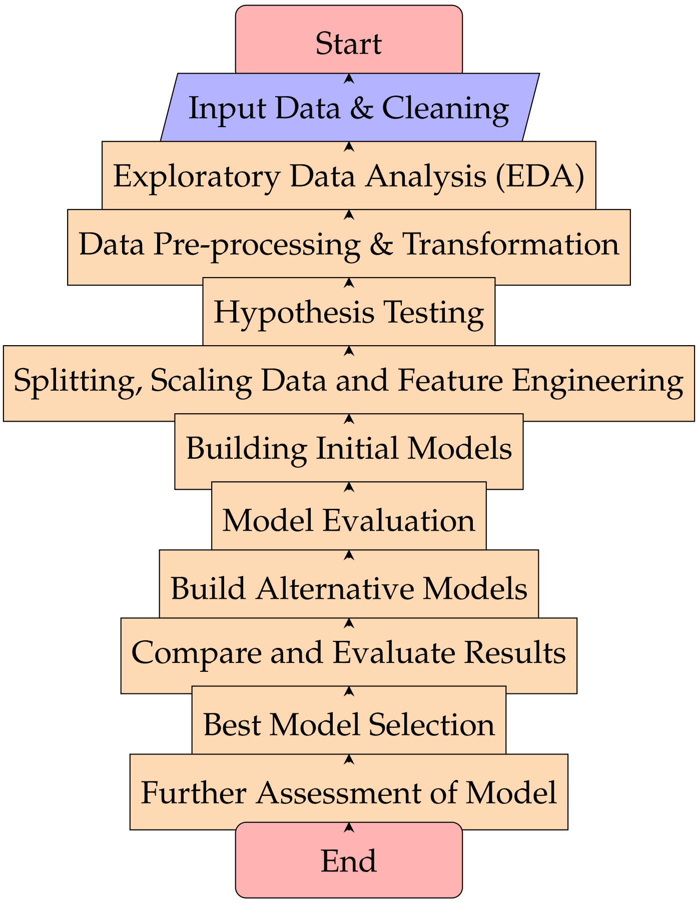
Bhatt and colleagues conducted a research study with the aim of creating a model capable of accurately predicting cardiovascular diseases to mitigate the fatality associated with such conditions [27]. The study introduced a k-modes clustering method with Huang initialization to enhance the accuracy of classification [27]. Various models, including random forest (RF), decision tree classifier (DT), multilayer perceptron (MP), and XGBoost (XGB), were utilized. The results indicated that, in terms of accuracy, the multilayer perceptron with cross-validation outperformed all other algorithms, achieving the highest accuracy of 87.28

The next chapter will address how the heart data from kaggle was used to answer the research objectives including testing hypothesis and building the prediction model using various data science and machine learning techniques.

CHAPTER 3

Methodology

This chapter will document every step used in deriving insights out of the dataset and the results achieved. Here is a flow chat for guidance.


Start
Input Data & Cleaning
Exploratory Data Analysis (EDA)
Data Pre-processing & Transformation
Hypothesis Testing
Splitting, Scaling Data and Feature Engineering
Building Initial Models
Model Evaluation
Build Alternative Models
Compare and Evaluate Results
Best Model Selection
Further Assessment of Model
End

3.1 Data loading and cleaning

The study adopts a multifaceted approach, integrating data science and machine learning techniques, to analyze heart disease data. The entire process was done in a python, a powerful tool for model development. It commences with the importation of essential Python libraries, facilitating data handling, visualization, and analysis. The dataset used in this study found at the following link: Kaggle Dataset: Heart Disease 2020 with 319,795 instances is loaded using Pandas. Preliminary data exploration involves assessing missing values and understanding the dataset's structure and characteristics.

3.2 Exploratory data analysis

In the exploratory data analysis (EDA) phase, the study focuses on generating summary statistics to understand the distribution of numerical variables. Visualization tools, including bar plots, histograms and boxplots, are employed to illustrate the distribution of heart disease cases and crucial variables like BMI, Physical Health, Mental Health, and Sleep Time. Although per the box plot outliers were seen, I still included in the analysis as they were not conclusive and figures are well within range and possible in real world. The analysis also extends to categorical variables, such as Smoking, Physical Activity, and Race, examined through count plots and bar charts to reveal underlying patterns and distributions. To uncover if racial disparities exist in the occurrence of heart disease, and if so, what are the variations in the proportion of positive cases within different racial groups, this method was applied;

Let R_i be the total count of individuals in a specific race category i, and let D_i be the count of individuals with heart disease in the same race category i. The proportion P_i of individuals with heart disease in race category i can be calculated as:

$$P_i = \left(\frac{D_i}{R_i}\right) \times 100 \tag{3.1}$$

This formula multiplies the ratio of individuals with heart disease by 100 to convert it into a percentage, representing the proportion of the population with heart disease in each racial category.

3.3 Data pre-processing and transformation

Data pre-processing is a key step, involving the encoding of categorical variables like 'Sex', 'Smoking', and 'Alcohol Drinking', and transforming 'HeartDisease' into a binary format. Dummy variables are created for categories like 'AgeCategory', 'Race', 'Diabetic', and 'GenHealth'. This phase also includes a thorough check for missing values to ensure data integrity.

A correlation matrix is then generated to examine the interrelationships among features. This analysis is instrumental in identifying multicollinearity and fixing it, which informs feature selection for hypothesis testing. Two distinct sets of features, focusing on demographic, lifestyle, and health-related factors, are delineated for testing hypotheses H1 and H2.

3.4 Hypothesis testing

Logistic regression models are applied to test these hypotheses, considering the binary nature of the dependent variable HeartDisease. For $H1$, demographic and lifestyle factors are analyzed, while $H2$ focuses on health-related factors. Missing values are imputed using a simple mean strategy before model fitting.

- **Hypothesis 1 (H1): Demographic and Lifestyle Factors**
  - **Null Hypothesis** H_{0_1}: *Demographic and lifestyle factors, namely Sex, Smoking, Alcohol Drinking, Physical Activity, Age Category, and Race, do not significantly affect the occurrence of heart disease.*
  - **Alternative Hypothesis** H_{A_1}: *Demographic and lifestyle factors, specifically Sex, Smoking, Alcohol Drinking, Physical Activity, Age Category, and Race, significantly affect the occurrence of heart disease.*
- **Hypothesis 2 (H2): Health-Related Factors**
  - **Null Hypothesis** H_{0_2}: *Health-related factors, including BMI, Physical Health, Mental Health, Sleep Time, Diabetic status (Yes/No), Stroke, Asthma, Kidney Disease, Skin Cancer, and General Health Perception (various categories), do not significantly predict heart disease.*

- **Alternative Hypothesis** H_{A_2}: *Health-related factors, such as BMI, Physical Health, Mental Health, Sleep Time, Diabetic status, Stroke, Asthma, Kidney Disease, Skin Cancer, and General Health Perception, significantly predict heart disease.*

3.5 Splitting and scaling data

The study progresses to an extensive model-building phase. The dataset is split into training and test sets 80% and 20% respectively, with a stratified approach based on the target variable to preserve class proportions. feature scaling is performed using a standard scaler.

3.6 Feature Engineering

Feature selection is conducted via the SelectKBest method, selecting the most relevant features based on the ANOVA F-test. This enhances model performance by concentrating on informative variables.

3.7 Building Initial models

I began the building the predictive model phase by exploring 4 different machine learning algorithms including Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and K-Nearest Neighbors and trained on the 7 top features first without SMOTE and secondly with SMOTE to address the dataset's class imbalance. This is how the logistic regression model good at predicting linear relations for example makes prediction;

The formula

$$P(Y=1) = \frac{1}{1+e^{-z}} \tag{3.2}$$

The formula above is the key component in logistic regression that makes predictions. Let me break down how this works:

Linear Combination (z): The linear combination z is calculated using the coefficients $(\beta_0, \beta_1, \ldots, \beta_n)$ and input features $(x_0 = 1, x_1, x_2, \ldots, x_n)$:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{3.3}$$

Logistic Function (Sigmoid): The logistic function (sigmoid function) is applied to the linear combination z:

$$P(Y = 1) = \frac{1}{1 + e^{-z}} \tag{3.4}$$

The sigmoid function maps the linear combination to a range between 0 and 1, representing the probability of the positive class ($Y = 1$).

Decision Rule: Typically, a threshold (e.g., 0.5) is chosen, and if the predicted probability $P(Y = 1)$ is greater than or equal to this threshold, the model predicts the positive class; otherwise, it predicts the negative class.

In a nutshell, the logistic regression model predicts the probability of the positive class using the logistic function applied to the linear combination of input features and coefficients.

3.8 Model Evaluation

The Models' performances is evaluated using metrics like accuracy, precision, recall, F1 score, and ROC-AUC score, supplemented by detailed classification reports. However, for the class imbalance of the heart data the focus will be on recall, f1 score and ROC-AUC to give an accurate performance of models.

In binary classification, the formulas for Recall, F1 Score, and ROC-AUC are:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}} \tag{3.5}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}} \tag{3.6}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.7}$$

$$\text{ROC-AUC} = \text{Area Under the ROC Curve} \tag{3.8}$$

3.9 Build Alternative Models and evaluate results

In pursuit of innovation to enhance the model's performance, I explored various model optimization techniques to empower the best-performing model in potentially predicting heart disease cases better, especially given their status as the minority class. Advanced resampling techniques, 3-fold cross-validation, class weight adjustment, threshold optimization, and hyperparameter tuning, were implemented in various ways and combinations across two models: Logistic Regression and XGBoost, each trained on different sets of data.

XGBoost was introduced due to its proficiency in handling complex data with non-linear relationships. One set of training data included all features from the dataset, while the other contained only the top 7 most important features in predicting heart disease, determined using the Anova F-test. The primary objective is to optimize model performance, with a particular focus on evaluation metrics such as roc-auc, recall scores, and f1 scores, as presented in the classification report.

3.10 Best model selection and further assessment of model

Finally, Based on the evaluation results, the best model in heart disease prediction was selected. The model underwent a comprehensive evaluation, leveraging its inherent feature importance functionality. This analysis aimed to discern the crucial features influencing early heart disease prediction. The identified features were meticulously sorted and examined based on their respective importances, providing valuable insights into each feature's contribution.

To enhance interpretability, a visualization was created. This visualization allowed a nuanced comparison between the model's predictions and the actual outcomes on the test data, facilitating a clearer understanding of the predictive performance.

In pursuit of a deeper understanding of the model's behavior, a thorough error analysis was conducted. This step involved visualizing instances where the model's predictions deviated from the actual outcomes. By pinpointing areas of prediction inaccuracies, this analysis provided crucial insights for potential enhancements in the model's performance.

CHAPTER 4

Results and Discussion

This chapter will further be broken into two parts. One part will highlight results from the EDA phase to have a deeper understanding of the Heart Disease dataset and the second part will focus on the results of the research objectives.

4.1 Exploratory data analysis

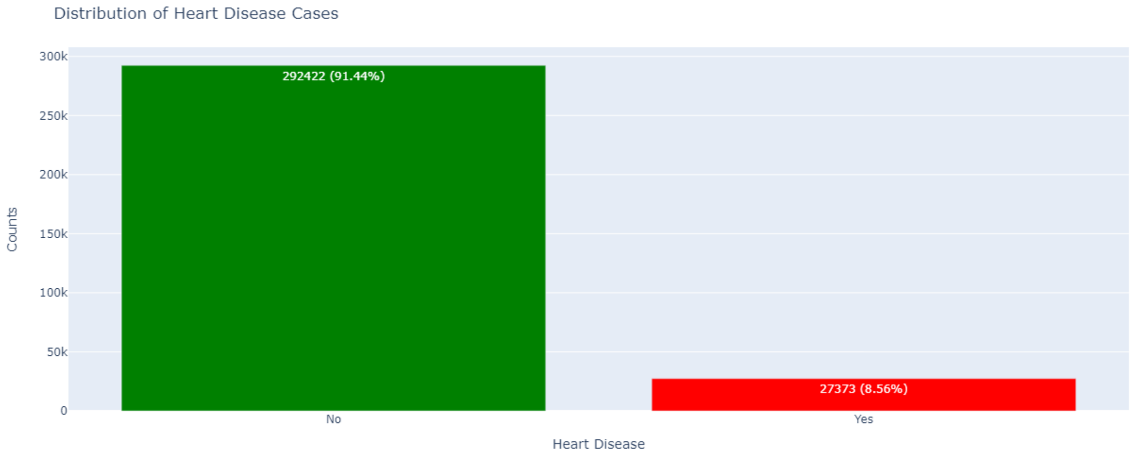


Figure 4.1: Distribution of Heart Disease Cases

The bar chart in Figure 4.1 illustrates the distribution of heart disease cases within the dataset. It shows a significant disparity between the number of cases reported as ’No’ (green bar) representing those without heart disease and ’Yes’ (red bar) representing those with heart disease. The green bar is substantially higher, indicating that the majority of the dataset's subjects do not have heart disease, quantified at 292,422 individuals (91.44%). In contrast, the red bar represents a smaller portion of the dataset, with 27,373 individuals (8.56%) having heart disease making 325,795 instances in total.

The summary statistics of the numerical features in the dataset are presented in Table 4.1.

Table 4.1: Summary Statistics of Numeric Categories

	BMI	**PhysicalHealth**	**MentalHealth**	**SleepTime**
Count	319,795.000	319,795.000	319,795.000	319,795.000
Mean	28.325	3.372	3.898	7.097
Std	6.356	7.951	7.955	1.436
Min	12.020	0.000	0.000	1.000
25%	24.030	0.000	0.000	6.000
50%	27.340	0.000	0.000	7.000
75%	31.420	2.000	3.000	8.000
Max	94.850	30.000	30.000	24.000

These statistics, including mean, standard deviation, minimum, and maximum values and the quartiles, are examined to gain insights into the central tendency, dispersion, and distribution of numeric data and to identify the presence of outliers. Physical health and mental health represent the number of days, within the past 30 days, where individuals experienced poor physical or mental health, with higher values indicating poorer health. Similarly, higher Body Mass Index (BMI) numbers are associated with poorer health.

The average number of days of being physically sick in the dataset is 3.37, while the average mental health is 3.90 days. The mean BMI is approximately 28.33, representing the average body mass index across all individuals. The standard deviation of 6.36 suggests a moderate amount of variability in BMI values, with higher values indicating greater variability. The minimum BMI observed is 12.02, representing the lowest body mass index. The first quartile (25%) falls below 24.03, and the median BMI is approximately 27.34. The third quartile (75%) falls below 31.42.

The maximum BMI observed is 94.85, which may indicate the presence of outliers due to its high value. Sleep time, representing the number of hours for sleeping, has an average of approximately 7.10 hours. The standard deviation of 1.44 indicates variability in sleep time, with a minimum sleep time of 1.0 hour. The first quartile (25%) of sleep times falls below 6.0 hours, the median sleep time is 7.0 hours, and the third quartile (75%) falls below 8.0 hours. The maximum sleep time observed is 24.0 hours.

For a visual representation of the distributions of physical health, mental health, and sleep time, refer to Figure A.1.

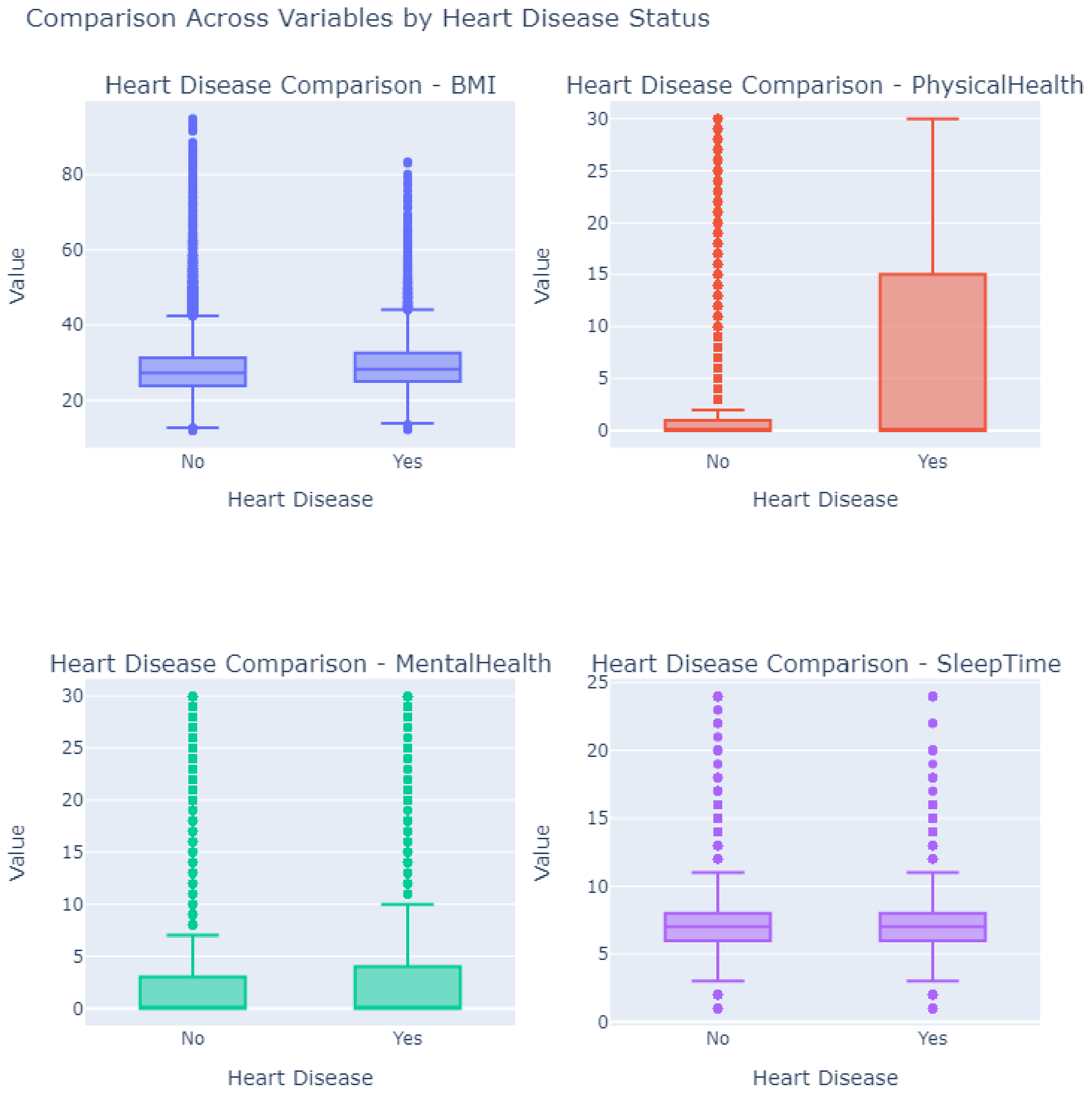


Figure 4.2: Box Plot comparing Numeric Distributions of Heart Disease cases

The box plot as shown above 4.2 was developed to compare the summary statistics values of instances with and without heart diseases and to further confirm the presence of outliers for both classes the in data.

As shown in the Appendix, Figure A.2 displays comprehensive view of the distribution of various categorical factors pertinent to the study of heart disease. The distribution of smoking is noteworthy, with a significant number of smokers (131,908) in the dataset out of 319,795 instances. Mobility challenges, as indicated by the 'DiffWalking' responses, are present in a smaller yet significant portion of the population (44,410 individuals), which could signify an increased cardiovascular risk that merits targeted interventions. The gender distribution, represented by 167,805 females and 151,990 males, where as racial representation is predominantly White, with lower counts for other racial groups.

In terms of diabetes, the majority of the sample does not have diabetes. However, a substantial number report having diabetes or being borderline diabetic. Physical activity levels are high within the dataset, suggesting a potential protective factor against heart disease within the studied cohort. However, this also raises questions about the representativeness of the sample in relation to the general population, where sedentary lifestyles are increasingly common. Self-assessed general health status shows that most individuals rate their health positively, with the categories 'Very good' and 'Good' being the most reported. Finally, asthma-Its presence in 42,872 individuals nevertheless underscores the importance of considering comorbid conditions in heart disease research.

Further analyzing the differences between individuals with and without heart disease across the variables as shown in Figure B.1 The analysis of the data reveals noteworthy disparities in the prevalence of heart disease across various health and lifestyle factors. Individuals facing mobility challenges, specifically those with difficulty walking, show a pronounced prevalence of heart disease at 22.58%, a figure significantly higher than their counterparts with no such difficulties. This highlights the profound impact of physical mobility on cardiovascular health.

In the realm of smoking, the data points to a heightened risk among smokers, where 12.16% have heart disease, suggesting a substantial increase in risk compared to non-smokers. This underlines the critical health implications of smoking on heart health.

A striking observation emerges in the case of stroke survivors, where an alarming

36.37% also suffer from heart disease, indicating a strong interconnection between stroke and heart disease. This rate is notably higher than in individuals without a history of stroke, emphasizing the severe cardiovascular consequences of stroke.

The diabetic population exhibits a significant 21.95% prevalence of heart disease, a rate that is considerably higher compared to non-diabetic individuals. This underscores the critical need for effective diabetes management as a key strategy in reducing heart disease risk.

In terms of general health perception, a gradient is observed. Those rating their health as ′poor′ show a high prevalence of heart disease at 34.10%, while those with ′fair′ health report a 20.43% prevalence, and the rate decreases further to 10.26% among those with ′good′ health. The lowest prevalence is seen in those perceiving their health as ′very good′ at 4.73%. These figures suggest a direct correlation between self-assessed health status and the likelihood of heart disease, highlighting the importance of overall health perception as an indicator of cardiovascular risk.

The analysis also reveals gender differences in heart disease prevalence, with males showing a slightly higher rate (10.62%) compared to females (6.69%). This points to gender-specific variations in heart disease risk, necessitating tailored approaches in prevention and treatment strategies.

Finally, EDA ended by taking a look at the correlation heatmap, a visual representation of the Pearson correlation coefficients between various health-related variables.The heatmap is a useful tool for quickly identifying potential relationships between variables that might warrant further analysis or consideration when building predictive models or conducting health-related research. Each cell in the heatmap shows the correlation coefficient between two selected variables, ranging from -1 to 1 and a spectrum of colours.This is shown in 4.3 below.

The spectrum with dark red or 1 signalling a perfect positive correlattion and dark blue or -1 signalling a perfect negative correlation with 0 indicating no correlation. It's also important to note that correlation does not imply causation; these relationships simply indicate a tendency to vary together in the dataset and further investigation would be needed to determine any causal relationships.

The correlation analysis reveals insights into the relationships between various factors. Here are the key correlations observed:

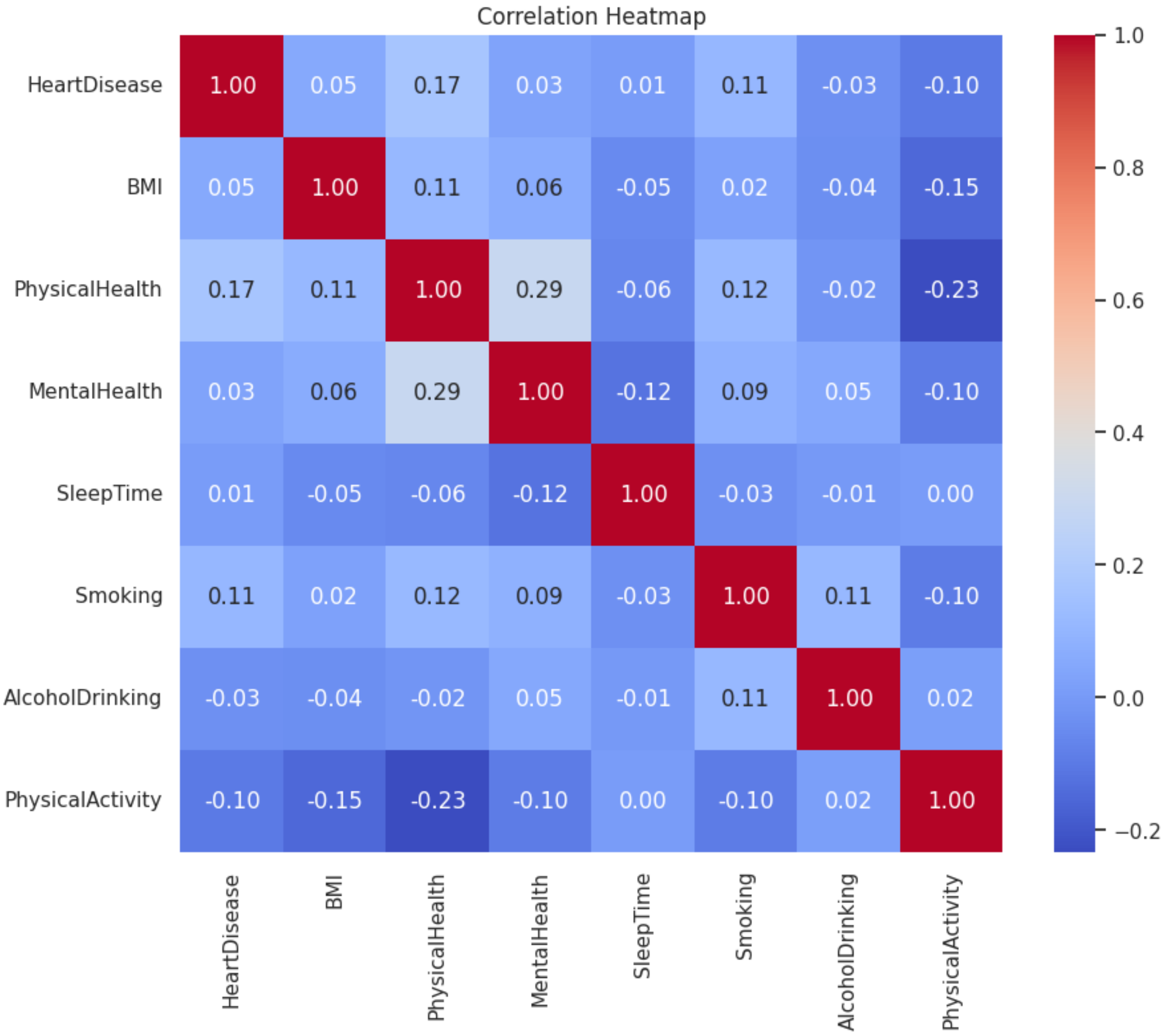


Figure 4.3: Box Plot comparing Numeric Distributions of Heart Disease cases

Correlations with Heart Disease Heart Disease and Physical Health (0.17): A positive correlation suggests that individuals with more physical health problems tend to have a higher prevalence of heart disease.

Heart Disease and Smoking (0.11): A slight positive correlation indicates that smoking is associated with a higher prevalence of heart disease, aligning with common medical knowledge.

Heart Disease and BMI (0.051803): A very weak positive correlation, indicating that higher BMI has a slight association with the presence of heart disease.

Correlations between Other Factors

Physical Health and Mental Health (0.29): A moderate positive correlation suggests that individuals who report more physical health problems also tend to report more mental health issues.

Sleep Time and Mental Health (-0.12): A slight negative correlation indicates that individuals who report more mental health issues tend to have less sleep, although the relationship is not strong.

Smoking and Alcohol Drinking (0.11): A slight positive correlation suggests that individuals who smoke may also be more likely to drink alcohol.

These correlations provide valuable insights into the interplay of different factors and their potential impact on heart disease.

This detailed quantitative assessment provides a foundational understanding of the sample's demographics and health characteristics, setting the stage for rigorous statistical analysis to uncover underlying patterns and associations within the data.

In summary, the data paints a complex picture of heart disease risk, with foundational understanding of the it's demographics and health characteristics play significant roles in influencing the likelihood of developing heart disease.The next subsection will seek to show the results in accordance to the research objectives of the study.

4.2 Results of Research objectives

4.2.1 Hypothesis Testing

Firstly with the hypothesis testing, the logistic regression analysis as shown in Table A.1 insightful details on how demographic and lifestyle factors influence the risk of heart disease. It uncovered that:

The positive coefficient for 'Sex' ($coef = 0.6497$) with a highly significant p-value ($p < 0.001$) indicates a higher likelihood of heart disease in males compared to females. The 'Smoking' variable, having a positive coefficient ($coef = 0.5381$) and significant p-value ($p < 0.001$), suggests an increased risk of heart disease among smokers. The negative coefficient for 'AlcoholDrinking' ($coef = -0.4074$) with significance ($p < 0.001$) points to a possible inverse relationship between moderate alcohol consumption and heart disease risK. A negative coefficient for 'PhysicalActivity' ($coef = -0.5140$), significant at $p < 0.001$, implies that engaging in physical activities reduces the risk of heart disease. All age categories beyond 30 years show positive and significant coefficients, highlighting an increased heart disease risk with advancing age.The negative coefficients for racial categories such as 'Asian', 'Black', 'Hispanic', 'Other', and 'White' compared to the baseline category suggest variations in heart disease risk across races.

These results lead us to **reject the null hypothesis H_{0_1} that demographic and lifestyle factors do not significantly influence heart disease occurrence**. Instead, we accept the alternative hypothesis H_{A_1} that these factors, including sex, smoking habits, alcohol consumption, physical activity, age, and race, substantially impact heart disease risk. This supports the critical role of these variables in assessing and managing heart disease.

For the Health related factors the table in 4.2, the logistic regression analysis, based on a substantial dataset of 319,795 individuals, utilized the Logit model to predict heart disease. The model's adequacy is indicated by a Pseudo R-squared value of 0.1469, suggesting a moderate level of explanation of heart disease variability. The statistical significance of the model is confirmed by an LLR p-value less than 0.000, indicating that the included health-related factors collectively contribute to predicting heart disease.

In this analysis, each health-related factor's coefficient reveals its unique impact on heart disease risk. The negative coefficient for BMI (-0.0053), despite its statistical significance (p < 0.001), is counter intuitive to usual medical expectations. This result might be influenced by complex interactions with other variables in the model or specific characteristics of the study population. The positive coefficients for PhysicalHealth (0.0099) and MentalHealth (-0.0153), both statistically significant, suggest a nuanced relationship with heart disease risk. The former indicates that deteriorating physical

<table>
<tr><th colspan="5">Logit Regression Results for Health-Related Factors</th></tr>
<tr><td colspan="5"><b>Dep. Variable:</b> HeartDisease <b>No. Observations:</b> 319795</td></tr>
<tr><td colspan="5"><b>Model:</b> Logit <b>Df Residuals:</b> 319781</td></tr>
<tr><td colspan="5"><b>Method:</b> MLE <b>Df Model:</b> 13</td></tr>
<tr><td colspan="5"><b>Date:</b> Wed, 10 Jan 2024 <b>Time:</b> 11:21:33</td></tr>
<tr><td colspan="5"><b>Pseudo R-squ.:</b> 0.1469 <b>Log-Likelihood:</b> -79727</td></tr>
<tr><td colspan="5"><b>converged:</b> True <b>LL-Null:</b> -93453</td></tr>
<tr><td colspan="5"><b>Covariance Type:</b> nonrobust <b>LLR p-value:</b> < 0.000</td></tr>
<tr><td><b>Variable</b></td><td><b>Coef.</b></td><td><b>Std. Err.</b></td><td><b>z</b></td><td><b>P>|z|</b></td></tr>
<tr><td>const</td><td>-3.9407</td><td>0.049</td><td>-81.109</td><td><0.001</td></tr>
<tr><td>BMI</td><td>-0.0053</td><td>0.001</td><td>-5.160</td><td><0.001</td></tr>
<tr><td>PhysicalHealth</td><td>0.0099</td><td>0.001</td><td>12.069</td><td><0.001</td></tr>
<tr><td>MentalHealth</td><td>-0.0153</td><td>0.001</td><td>-17.946</td><td><0.001</td></tr>
<tr><td>SleepTime</td><td>0.0251</td><td>0.004</td><td>6.075</td><td><0.001</td></tr>
<tr><td>Diabetic_Yes</td><td>0.7552</td><td>0.016</td><td>46.962</td><td><0.001</td></tr>
<tr><td>Stroke</td><td>1.3149</td><td>0.022</td><td>59.324</td><td><0.001</td></tr>
<tr><td>Asthma</td><td>0.0578</td><td>0.018</td><td>3.149</td><td>0.002</td></tr>
<tr><td>KidneyDisease</td><td>0.7262</td><td>0.024</td><td>30.193</td><td><0.001</td></tr>
<tr><td>SkinCancer</td><td>0.6879</td><td>0.019</td><td>36.999</td><td><0.001</td></tr>
<tr><td>GenHealth_Fair</td><td>1.9421</td><td>0.032</td><td>61.318</td><td><0.001</td></tr>
<tr><td>GenHealth_Good</td><td>1.3760</td><td>0.029</td><td>47.618</td><td><0.001</td></tr>
<tr><td>GenHealth_Poor</td><td>2.4087</td><td>0.039</td><td>61.289</td><td><0.001</td></tr>
<tr><td>GenHealth_Very good</td><td>0.6847</td><td>0.030</td><td>22.954</td><td><0.001</td></tr>
</table>

Table 4.2: Comprehensive Logistic Regression Results for Health-Related Factors Predicting Heart Disease

health increases heart disease likelihood, while the latter's negative value suggests a decrease in heart disease risk with poorer mental health, warranting a deeper investigation into the underlying dynamics.

The positive coefficient for SleepTime (0.0251) aligns with medical literature linking longer sleep duration with increased heart disease risk, potentially due to underlying health issues affecting both sleep and cardiovascular health.

Notably, Diabetic status ('DiabeticYes' coefficient at 0.7552), Stroke (1.3149), Asthma (0.0578), Kidney Disease (0.7262), and Skin Cancer (0.6879) exhibit positive coefficients, all significant at $p < 0.001$. These findings reaffirm the established links between these conditions and heightened heart disease risk.

The analysis also delves into self-perceived health status through the General Health Perception variable, with categories like Fair, Good, Poor, and Very Good. The positive and statistically significant coefficients for these categories underscore the correlation between poorer self-assessed health and increased likelihood of heart disease.

Given these detailed results, we reject the null hypothesis, which posited that health-related factors do not significantly predict heart disease, concluding that factors like BMI, Physical Health, Mental Health, Sleep Time, Diabetic status, Stroke, Asthma, Kidney Disease, Skin Cancer, and General Health Perception hold significant predictive power for heart disease..

4.2.2 Racial Disparities

Secondly, the data presented in the tables below provide a comprehensive breakdown of individuals by race and the corresponding proportions with heart disease. Table 4.3 lists the count of individuals from different racial backgrounds within the data, with the White race being the most represented. Table 4.4 further explores these groups by detailing the percentage of individuals with heart disease within each racial category. Notably, American Indian/Alaskan Native individuals have the highest reported proportion of heart disease at 10.30%, which is a significant contrast to the Asian population at 3.40%, the lowest among the listed races. This stark difference suggests that racial factors may play a significant role in the prevalence of heart disease, warranting further investigation into the socio-economic and genetic factors that could influence these disparities.

The data further revealed as shown in Table B.1 that the White population at 11.60% were dominant in skin cancer cases in the group compared to others.

Race	**Count**
White	204934
Black	21051
Hispanic	20133
Other	8725
Asian	6580
American Indian/Alaskan Native	4707

Table 4.3: Counts of Individuals by Race

Race	**Proportion (%)**
American Indian/Alaskan Native	10.30
Asian	3.40
Black	7.51
Hispanic	4.88
Other	8.03
White	9.25

Table 4.4: Proportion of Individuals with Heart Disease by Race

These results and the variations in proportions of heart disease prevalence per race in that dataset do confirm that racial disparities do exist in the occurrence of heart disease and other diseases like skin cancer according to the heart disease 2020 data.

4.2.3 Predictive models

In the third objective that seeks to build the most effective data science and machine learning model for predicting heart disease especially when the data in real world setting can be highly imbalanced, various techniques were used to achieve this aim. This piece will document each technique used and model performance or results obtained right through to the best performing model to predict heart disease.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	91.48%	51.49%	7.91%	13.71%	73.98%
Decision Tree	91.46%	51.68%	4.49%	8.27%	73.41%
Random Forest	91.45%	50.76%	4.89%	8.93%	73.78%
K-Nearest Neighbors	91.17%	41.58%	7.80%	13.13%	66.30%
Logistic Regression (SMOTE)	75.73%	20.31%	62.74%	30.68%	73.97%
Decision Tree (SMOTE)	73.60%	19.15%	64.69%	29.55%	72.92%
Random Forest (SMOTE)	73.56%	19.15%	64.84%	29.57%	73.07%
K-Nearest Neighbors (SMOTE)	90.91%	37.39%	9.26%	14.84%	66.83%

Table 4.5: Performance Metrics of Machine Learning Models for Heart Disease Prediction

The performance metrics for various machine learning models 4.5 in predicting heart disease reveal insightful contrasts, particularly when comparing the models in their standard forms versus those enhanced with SMOTE (Synthetic Minority Over-sampling Technique). These metrics—accuracy, precision, recall, f1_score, and roc_auc—are essential for evaluating the models' effectiveness in identifying heart disease cases amidst non-heart disease instances.

In their standard forms, models like Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and K-Nearest Neighbors demonstrate high accuracy, around 91.48%, 91.46%, 91.45%, and 90.17%, respectively. This high accuracy suggests that these models are generally proficient in classifying the instances correctly. However, a deeper look into precision and recall for heart disease cases (class 1) exposes some challenges. For instance, Logistic Regression, with a precision of 51.49% for class 1, implies that it correctly identifies heart disease cases slightly over half the time when it predicts them. The recall for class 1 is markedly low at 7.91%, indicating it misses a substantial number of actual heart disease cases. This pattern of low recall is consistent across all models, with Decision Tree (4.49%), Random Forest (4.89%), and K-Nearest Neighbors (7.80%), suggesting a general difficulty in accurately identifying heart disease cases.

The introduction of SMOTE significantly alters the performance landscape of these models. SMOTE, designed to address class imbalance by artificially oversampling the minority class, results in a trade-off between different metrics. For instance, the accuracy of Logistic Regression drops to 75.73% with SMOTE, indicating a reduction in overall correct classifications. However, this is counterbalanced by a substantial increase in recall for class 1, soaring to 62.74%, a marked improvement over the 7.91% without SMOTE. This pattern of decreased accuracy but increased recall is evident in all models with SMOTE. For example, the Decision Tree's recall jumps from 4.49% to 64.69%, and the Random Forest's from 4.89% to 64.84%.

This shift is critical in the context of heart disease prediction. While the standard models are more accurate overall, their low recall for heart disease cases means they are likely to miss many actual cases. The SMOTE-enhanced models, despite being less accurate overall and suffering a drop in precision for class 1 (indicating more false positives), are much more effective in identifying heart disease cases. The f1_scores for class 1 in these models, which are higher than in their non-SMOTE counterparts, reflect a more balanced trade-off between precision and recall, indicating improved performance in this critical aspect.

In all the Logistic Regression model enhanced with SMOTE is identified as the optimal choice for scenarios prioritizing high sensitivity in the detection of heart disease. This model's performance is characterized by its ability to achieve a high recall for class 1 and other combinations of F1 Score and ROC AUC, This high recall rate is crucial in medical contexts such as early screening or diagnostic settings, where the cost of missing a heart disease case can be particularly grave.

To potentially explore the possibility of improving the model to be able to recall the Heat Disease positive cases, the following optimization techiques were adopted;

hyperparameter tuning, threshold optimization, 5 fold cross validation and of course feature engineering where the top 7 features were used to train all models according to their importance in predicting the Coronary Heart Disease. These were used on the logistic regression with smote model. Here are the findings the table below:

	Precision	Recall	F1-Score	Support
0	0.95	0.88	0.91	58484
1	0.27	0.46	0.34	5475
Accuracy		0.85		
Macro Avg	0.61	0.67	0.63	63959
Weighted Avg	0.89	0.85	0.86	63959
ROC AUC Score		0.6683		

Table 4.6: Classification Report of Logistic regression with smote and other opitmization techniques(Hyperparemeter tuning, threshold optimization, cross validation)

With best parameters: {'logisticregression__C': 0.1, 'logisticregression__penalty': 'l2'}, The model's performance in table 4.6 can be dissected into two distinct categories based on the target classes. For Class 0, which represents Non-Heart Disease Cases, the model exhibits high precision at 95%. This high precision implies that when the model predicts an instance as non-heart disease, it is correct 95% of the time. The recall for this class stands at 88%, indicating the model's proficiency in identifying 88% of all actual non-heart disease cases. The F1-Score, being 92%, suggests a strong balance between precision and recall for non-heart disease predictions.

In contrast, for Class 1, the Heart Disease Cases, the model's precision drops significantly to 27%. This lower precision indicates a higher rate of false positives when predicting heart disease cases. The recall for this class is 46%, which, while better than random guessing, points to a scenario where over half of the true heart disease cases might be missed. The F1-Score for Class 1 is relatively low at 34%, highlighting a weaker balance between precision and recall for heart disease predictions. ROC AUC Score is 66.83%

Comparing The evaluation of logistic regression models reveals that the model with SMOTE, as the sole optimization technique, still surpasses the logistic regression with additional optimization techniques based on ROC AUC scores and recall for Class 1. The former achieved an ROC AUC score of 73.97%, outperforming the latter with a score of 66.83%, indicating superior overall performance. In specific cases, the former model excelled with a recall of 62.74% for Class 1 compared to the latter's 46%. Notably, the highly optimized model only demonstrated superiority in predicting Class 0 cases. Consequently, the logistic regression with SMOTE, focusing solely on the top 7 features, remains the most effective model for predicting heart disease cases.

I explored further by trying out different resampling of training data on the logistic regression with Smote together with other optimization techniques to see if the model will be improved. The first i tried was the option of balancing out the train data to have equal classes for the minority class(yes) and and the minority class (no) of heart disease by downsampling the majority class then use it to train the data in a bid to evaluate if model performs better overall and also effectively predict the yes cases. These are the findings as tabulated in the appendix C.1. Comparing the figures this model did not perform better than Logistic regression with SMOTE and it associated techniques.

Another resampling technique used with same optimization techniques as mentioned before was the use of SMOTEENN and this were the results C.2

It was still evident that based on the results using ROC AUC and Recall on the Logistic Regression with SMOTE without any other optimization technique was the best out of all the other ones.

In further exploration, I delved into a distinct machine learning algorithm renowned for its effectiveness in handling both non-linear and linear relationships within datasets. The chosen algorithm is Extreme Gradient Boosting (XGBoost). Employing resampling techniques, including class weights, alongside the optimization methods mentioned earlier, I conducted experiments with two sets of training data. The first set involved training on scaled data, specifically the top 7 features identified through an ANOVA F-test. The second set involved training on not scaled data, encompassing all available features. Here are the outcomes:

In this analysis 4.9, 4.7, 4.8, 4.12, 4.10 and 4.11we evaluate and compare several XGBoost models trained on different feature sets and using various optimization techniques. The key metrics considered for model selection are the ROC AUC score and the recall for Class 1 due to the imbalanced nature of the dataset and objective of this study

Table 4.7: Classification Report of xgb with class weight on scaled top 7 features

	Precision	Recall	F1-Score	Support
0	0.96	0.74	0.83	58484
1	0.19	0.67	0.30	5475
Accuracy			0.73	63959
Macro Avg	0.58	0.70	0.57	63959
Weighted Avg	0.89	0.73	0.79	63959
ROC AUC Score		0.7388079993154276		

Table 4.8: Classification Report of xgb with smote on scaled top 7 features

	Precision	Recall	F1-Score	Support
0	0.96	0.74	0.84	58484
1	0.19	0.66	0.30	5475
Accuracy			0.73	63959
Macro Avg	0.58	0.70	0.57	63959
Weighted Avg	0.89	0.73	0.79	63959
ROC AUC Score		0.7341492751871564		

Table 4.9: Classification report of xgb with class weight and all features unscaled

	Precision	Recall	F1-Score	Support
0	0.97	0.75	0.85	58484
1	0.22	0.76	0.34	5475
Accuracy			0.75	63959
Macro Avg	0.60	0.76	0.59	63959
Weighted Avg	0.91	0.75	0.80	63959
ROC AUC Score		0.8315351581933661		

Table 4.10: Classification Report of xgb with class weight, best hyperparameters, 3 fold cross validation, threshold optimization on top 7 features train data

	Precision	**Recall**	**F1-Score**	**Support**
0	0.94	0.90	0.92	58484
1	0.28	0.44	0.34	5475
Accuracy			0.86	63959
Macro Avg	0.61	0.67	0.63	63959
Weighted Avg	0.89	0.86	0.87	63959
ROC AUC Score		0.7384544217534109		

Table 4.11: Classification report with all features using class weight and all other optimized features

	Precision	**Recall**	**F1-Score**	**Support**
0	0.91	1.00	0.96	58484
1	0.00	0.00	0.00	5475
Accuracy			0.91	63959
Macro Avg	0.46	0.50	0.48	63959
Weighted Avg	0.84	0.91	0.87	63959
ROC AUC Score		0.8371127567497679		

Table 4.12: Classification Report of xgb trained with all features, smote and all other optimized methods already mentioned

	Precision	**Recall**	**F1-Score**	**Support**
0	0.95	0.84	0.89	58484
1	0.23	0.49	0.31	5475
Accuracy			0.81	63959
Macro Avg	0.59	0.66	0.60	63959
Weighted Avg	0.88	0.81	0.84	63959
ROC AUC Score		0.7651203139039082		

which is to predict heart disease.

The best-performing model among the evaluated XGBoost variants is the one trained with class weight on all features of the train dataset as shown in table 4.9.Not only is this model best amongst the other xgboost variants, it improves upon the Logistic regression model with SMOTE trained on the top 7 features of the dataset.

This model achieved a high ROC AUC score of 83.15%, indicating its effectiveness in distinguishing between positive and negative cases. Additionally, the recall for Class 1 is 76%, signifying the model's ability to effectively predict positive cases.

It is interesting to note that the use of SMOTE and class weight had a similar impact on the model's performance, with comparable ROC AUC scores.

Surprisingly, all the optimization techniques applied did not lead to improvements in the models; in fact, they seemed to make the model perform less effectively, particularly in the case of XGBoost and Logistic Regression. This may be because those optimization techniques were to complex for the trained data. A further analysis needs to done to review and investigate this claim.

These findings suggest that for this specific dataset and model architecture, the inclusion of certain optimization techniques may not necessarily enhance predictive performance and might even lead to a decrease in model effectiveness.

To comprehensively evaluate the selected model, we delve into its output, examining the actual and predicted values. We conduct a detailed error analysis, particularly focusing on false positives and false negatives, to gain insights that can inform further model refinement. Additionally, we conclude the assessment by scrutinizing the top features and scores generated by the model, providing a deeper understanding of its inner workings.

Table 4.13: Predicted vs Actual Output

Actual	**Predicted**
0	1
0	0
0	0
0	1
0	0

Table 4.14: Error Analysis

Actual	Predicted
0	1
0	1
0	1
0	1
0	1

Table 4.15: Top 10 Features by Importance

Feature	Importance
DiffWalking	572.90
AgeCategory_80 or older	417.31
Diabetic_Yes	344.69
AgeCategory_70-74	342.96
AgeCategory_75-79	329.97
GenHealth_Fair	273.57
GenHealth_Poor	223.49
AgeCategory_65-69	220.08
Stroke	216.07
AgeCategory_25-29	200.25

Concluding this chapter, Based on the classification report, ROC AUC score, top 10 features of importance by the model and error analysis as shown in 4.9 4.15, 4.14 and 4.13, gives us a clearer picture about the model, let's analyze the performance of the model:

The model's performance for predicting the negative class (0) is quite good. It achieves a high precision of 97%, meaning that when it predicts the negative class, it is correct 97% of the time. The recall for the negative class is 75%, indicating that the model correctly identifies 75% of the actual negative cases. The F1-Score, which balances precision and recall, is 85% for the negative class. The model's overall accuracy is 75%, suggesting that 75% of the predictions across both classes are correct.

For the positive class (1), the model's precision is 22%, meaning that when it predicts the positive class, it is correct 22% of the time. However, the recall for the positive class is 76%, indicating that the model correctly identifies 76% of the actual positive cases. The F1-Score for the positive class is 34%, representing a balance between precision and recall.

The ROC AUC score is 83.15%, indicating a good overall performance of the model in distinguishing between the two classes. The predicted vs. actual output and error analysis tables provide specific instances where the model made correct and incorrect predictions.

The XGBoost model, trained with class weight on all features, highlighted key factors influencing predictions. The top 10 features by importance include age categories ('80 or older,' '70-74,' '75-79,' '65-69,' '25-29'), health indicators ('Fair' and 'Poor' general health), specific health conditions (diabetes and stroke), and a mobility-related feature ('DiffWalking'). These features collectively contribute to the model's ability to make nuanced predictions. The insights suggest that demographic information, health status, and medical history are crucial in assessing the risk of the target condition.

Predicted vs Actual Output shows a few instances where the model made predictions (Predicted) compared to the actual class labels (Actual).

Error Analysis It provides examples of false positive cases where the model predicted 1 while the actual label was 0.

> Since Class 0 (Negative Class)Has a recall of 0.75. Recall is the ratio of correctly predicted positive observations to all observations in the actual class. Since there are 58,484 actual negatives (class 0), and the recall is 0.75, it means 75% of the actual negatives are correctly identified (True Negatives), and 25% are incorrectly identified (False Positives).

$$\text{False Positives} = 58,484 \times (1 - 0.75)$$

$$\text{False Positives} = 58,484 \times (1 - 0.75)$$

$$\text{False Positives} = 58,484 \times 0.25$$

This calculation will give the total number of false positives as predicted by the model. Remember, this is an estimation based on the recall value for the negative class. The actual count might vary slightly, but this provides a good approximation.

The next chapter will talk about the conclusions based on this chapter′s results.

CHAPTER 5

Conclusions

In summary, this study confirms the existence of racial disparities in heart disease prevalence, with the hypothesis testing revealing the significant impact of demographic, lifestyle, and health factors on heart disease occurrence. The imbalance in the dataset posed challenges for traditional machine learning models in accurately identifying heart disease cases, leading to the incorporation of Synthetic Minority Over-sampling Technique (SMOTE) for improved model effectiveness.

Initially, models like Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors exhibited high overall accuracy, masking their limited recall for heart disease cases. The introduction of SMOTE addressed the class imbalance issue, resulting in enhanced recall rates for heart disease cases despite a slight decrease in overall accuracy. Further optimization techniques, including hyperparameter tuning, threshold optimization, and 5-fold cross-validation, were implemented.

Surprisingly, the Logistic Regression model enhanced with SMOTE proved to be optimal for high-sensitivity scenarios in heart disease detection, crucial for early screening or diagnostics, until the exploration of the Extreme Gradient Boost (XGBoost) model. The XGBoost model, trained with class weight on all features, outperformed other variants, achieving a remarkable ROC AUC score of 83.15%, with notable recall rates and precision for heart disease cases.

Adding specific figures to the analysis provides a quantitative perspective on model performance, highlighting trade-offs between accuracy and sensitivity. The XGBoost model, with its emphasis on class weight and all features, demonstrated strong overall performance, particularly in distinguishing between positive and negative cases.

Logistic Regression with SMOTE:

- Accuracy: 75.73%
- Precision (Class 1): 20.31%
- Recall (Class 1): 62.74%
- F1 Score (Class 1): 30.68%
- ROC AUC: 73.97%

Best XGBoost Model (all features and class weight):

- ROC AUC Score: 83.15%
- Recall (Class 1): 76%
- Precision (Class 1): 22%
- F1 Score (Class 1): 34%

The performance of Logistic Regression and XGBoost models is influenced by their ability to handle linear and non-linear relationships. Logistic Regression excels in linear scenarios, while XGBoost's strength lies in capturing complex, non-linear relationships and also linear scenarios, reflected in the evaluation metrics.

However, the chosen XGBoost model is not without limitations. Significant class imbalance and the potential for overfitting raise concerns about generalizability and interpretability. Despite these limitations, the model's benefits, such as handling non-linear relationships and linear relationships, addressing imbalanced data, providing insights into feature importance, and robustness to overfitting, position it as a valuable tool in real-world cardiac care.

Translating these findings into practice emphasizes the importance of balancing overall accuracy with the correct identification of heart disease cases. The study suggests that the chosen XGBoost model, particularly in scenarios prioritizing high sensitivity, holds promise for early screening and diagnostics in cardiac care. However, acknowledging the study's limitations, future research should explore additional optimization techniques and consider a broader array of factors to enhance the applicability of machine learning models in real-world cardiac care.

APPENDIX A

A Long Proof

A.1 Numeric Distributions in Heart Data

A.2 Categorical Distributions in Heart Data

A.3 Logistic Regression Results of Demographic and lifestyle factors for Predicting Heart Disease

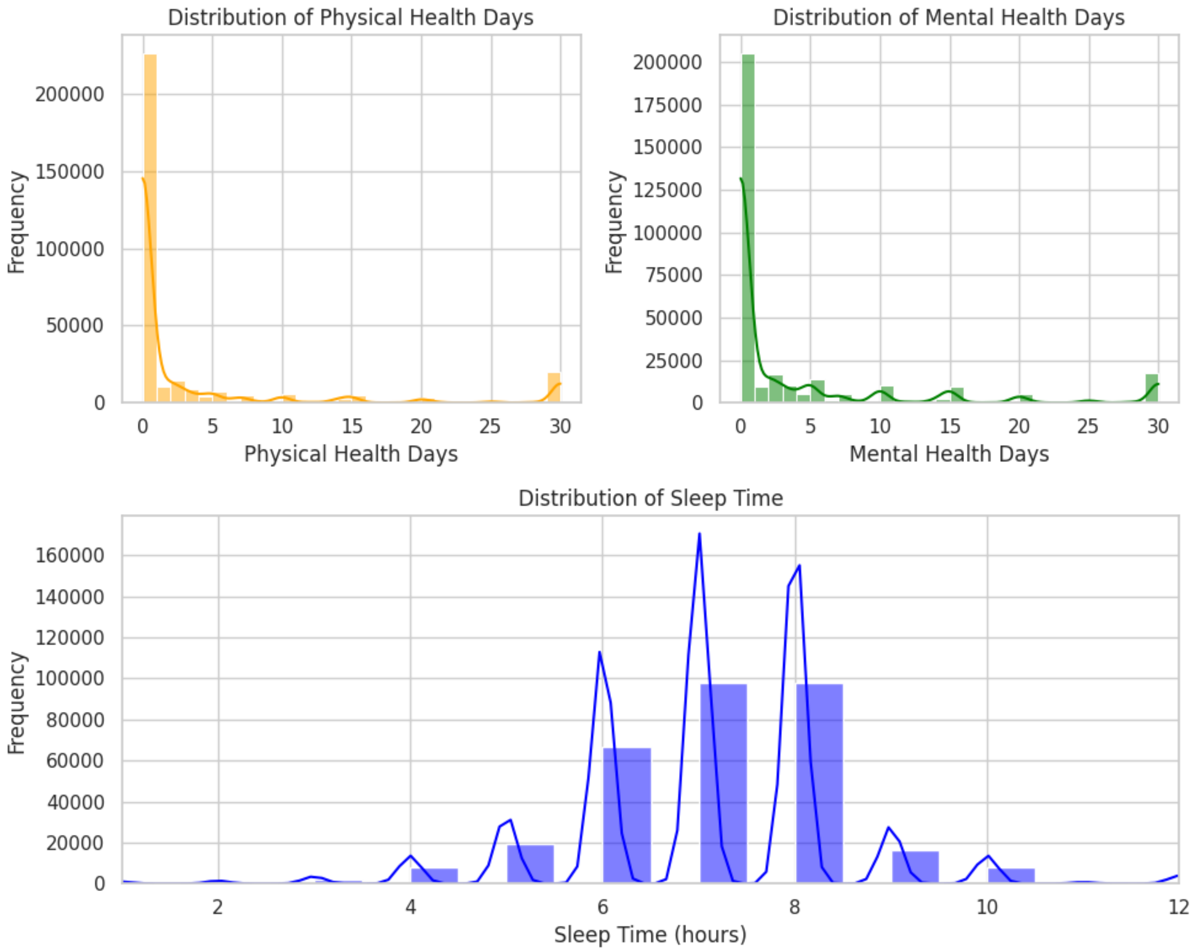
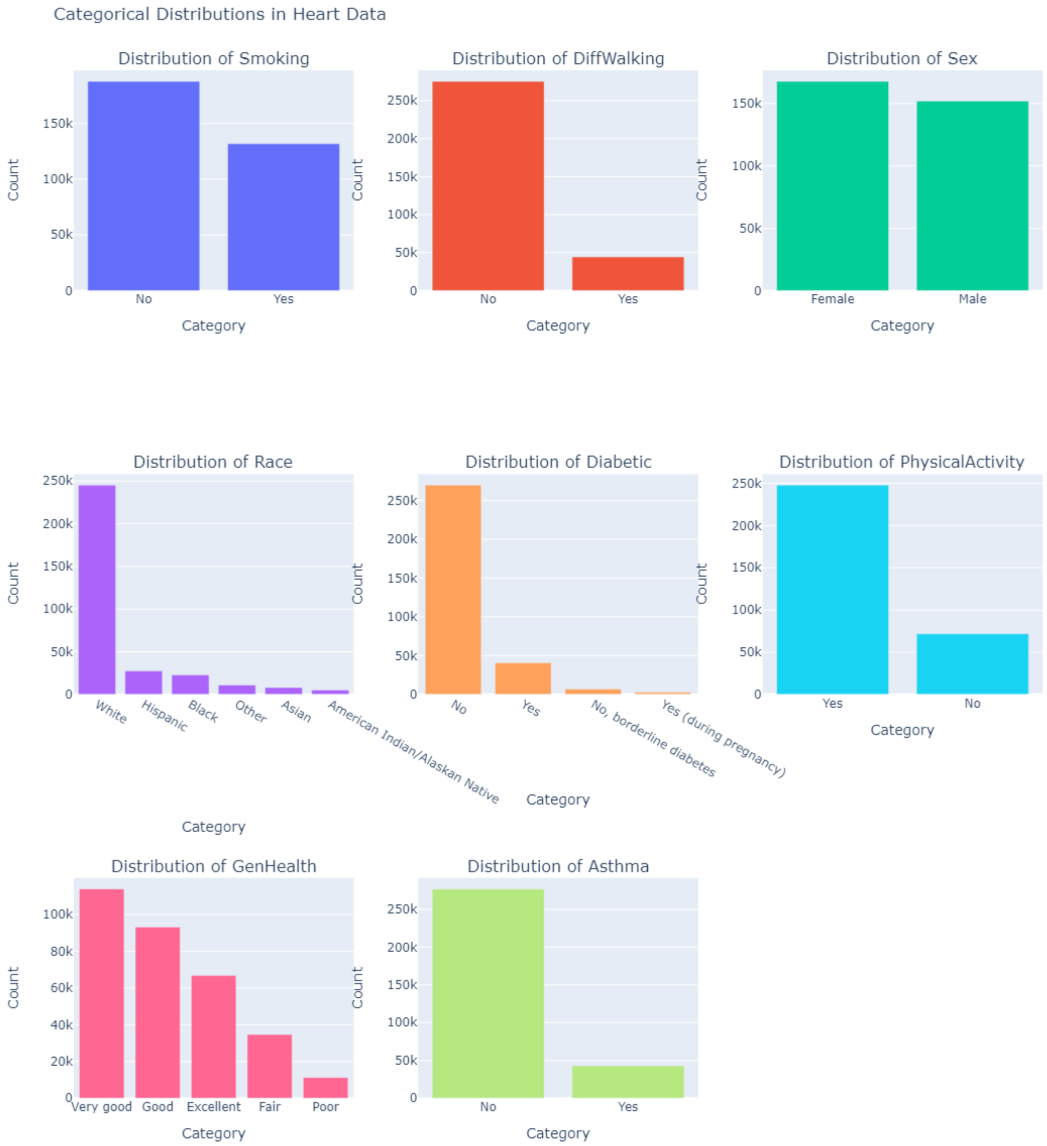


Figure A.1: Distributions of Physical, Mental health Days and Sleep Time.


Categorical Distributions in Heart Data
Distribution of Smoking
Count
0
50k
100k
150k
No
Yes
Category
Distribution of DiffWalking
Count
0
50k
100k
150k
200k
250k
No
Yes
Category
Distribution of Sex
Count
0
50k
100k
150k
Female
Male
Category
Distribution of Race
Count
0
50k
100k
150k
200k
250k
White
Hispanic
Black
Other
Asian
American Indian/Alaskan Native
Category
Distribution of Diabetic
Count
0
50k
100k
150k
200k
250k
No
Yes
No, borderline diabetes
Yes (during pregnancy)
Category
Distribution of PhysicalActivity
Count
0
50k
100k
150k
200k
250k
Yes
No
Category
Distribution of GenHealth
Count
0
20k
40k
60k
80k
100k
Very good
Good
Excellent
Fair
Poor
Category
Distribution of Asthma
Count
0
50k
100k
150k
200k
250k
No
Yes
Category

Logit Regression Results

Dep. Variable: HeartDisease	**No. Observations:** 319795
Model: Logit	**Df Residuals:** 319773
Method: MLE	**Df Model:** 21
Date: Wed, 10 Jan 2024	**Time:** 10:30:17
Pseudo R-squ.: 0.1416	**Log-Likelihood:** -80224
converged: True	**LL-Null:** -93453
Covariance Type: nonrobust	**LLR p-value:** < 0.000

Variable	**Coef.**	**Std. Err.**	**z**	**P>|z|**
const	-4.7605	0.101	-47.138	<0.001
Sex	0.6388	0.014	47.116	<0.001
Smoking	0.5343	0.014	39.373	<0.001
AlcoholDrinking	-0.4329	0.032	-13.391	<0.001
PhysicalActivity	-0.5233	0.014	-36.871	<0.001
AgeCategory25-29	0.1615	0.124	1.303	0.193
AgeCategory30-34	0.5529	0.111	4.994	<0.001
AgeCategory35-39	0.7195	0.106	6.795	<0.001
AgeCategory40-44	1.1941	0.099	12.005	<0.001
AgeCategory45-49	1.6022	0.096	16.723	<0.001
AgeCategory50-54	2.0900	0.092	22.600	<0.001
AgeCategory55-59	2.3997	0.091	26.368	<0.001
AgeCategory60-64	2.6943	0.090	29.876	<0.001
AgeCategory65-69	2.9253	0.090	32.546	<0.001
AgeCategory70-74	3.2265	0.090	35.951	<0.001
AgeCategory75-79	3.4431	0.090	38.197	<0.001
AgeCategory80 or older	3.6961	0.090	41.169	<0.001
RaceAsian	-0.8544	0.080	-10.627	<0.001
RaceBlack	-0.3563	0.055	-6.508	<0.001
RaceHispanic	-0.3617	0.056	-6.476	<0.001
RaceOther	-0.1633	0.061	-2.690	0.007
RaceWhite	-0.3787	0.049	-7.770	<0.001

Table A.1: Logistic Regression Results of Demographic and lifestyle factors for Predicting Heart Disease

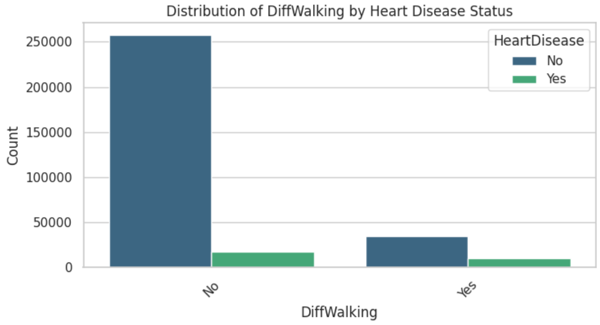
APPENDIX B

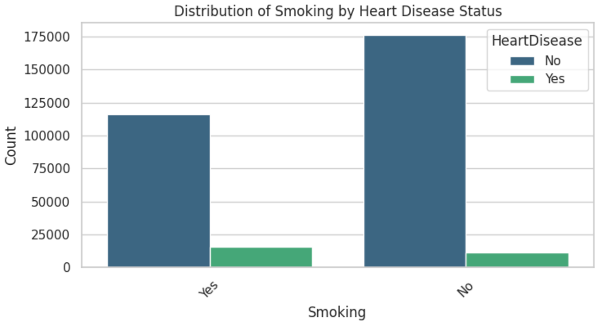
Additional Figures

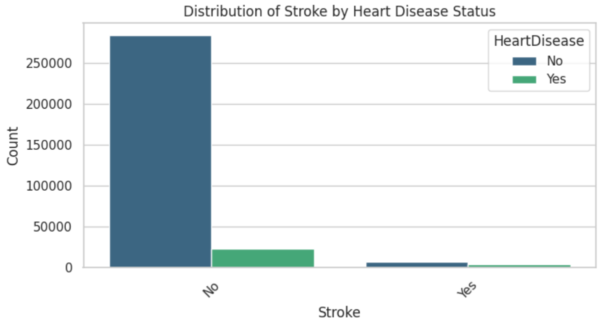
B.1 Skin Cancer Prevalence by Race

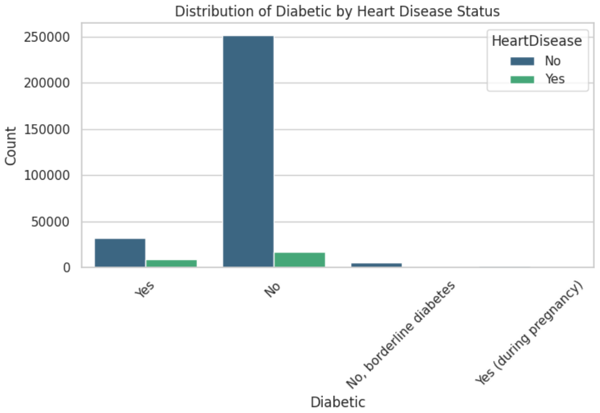
Race	Proportion (%)
American Indian/Alaskan Native	3.23
Asian	0.82
Black	0.62
Hispanic	1.60
Other	4.78
White	11.60

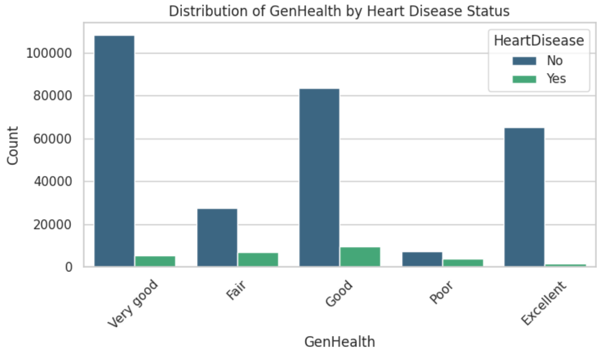
Table B.1: Proportion of Skin Cancer Cases in Different Racial Categories

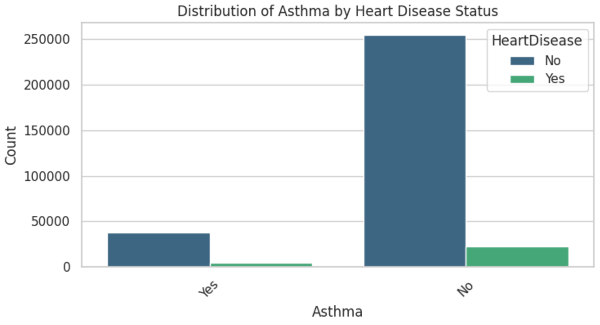

Distribution of DiffWalking by Heart Disease Status
HeartDisease
No
Yes
Count
250000
200000
150000
100000
50000
0
No
Yes
DiffWalking

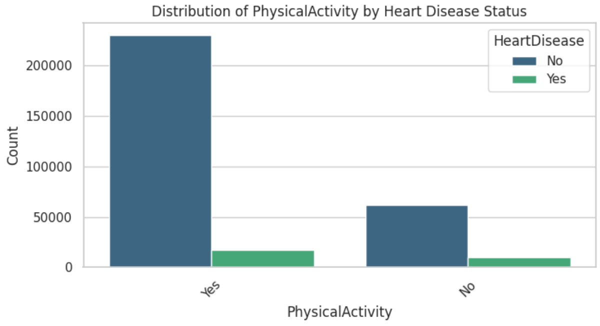

Distribution of Smoking by Heart Disease Status
HeartDisease
No
Yes
Count
175000
150000
125000
100000
75000
50000
25000
0
Yes
No
Smoking

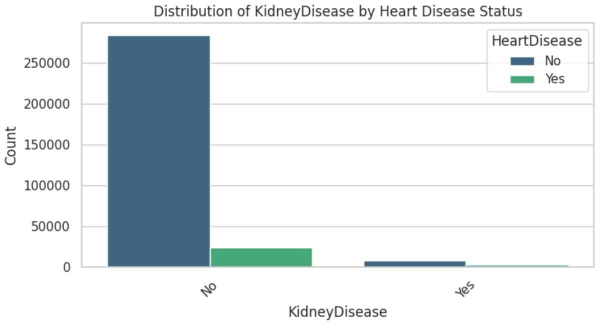

Distribution of Stroke by Heart Disease Status
HeartDisease
No
Yes
Count
250000
200000
150000
100000
50000
0
No
Yes
Stroke

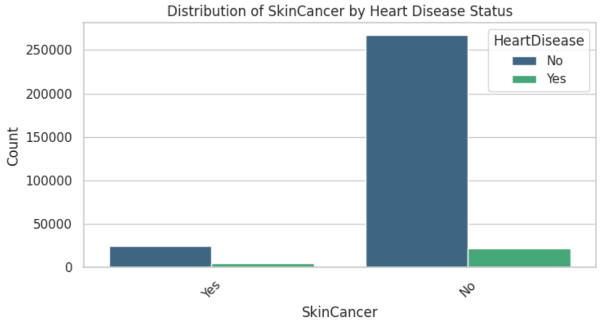

Distribution of Diabetic by Heart Disease Status
HeartDisease
No
Yes
Count
250000
200000
150000
100000
50000
0
Yes
No
No, borderline diabetes
Yes (during pregnancy)
Diabetic

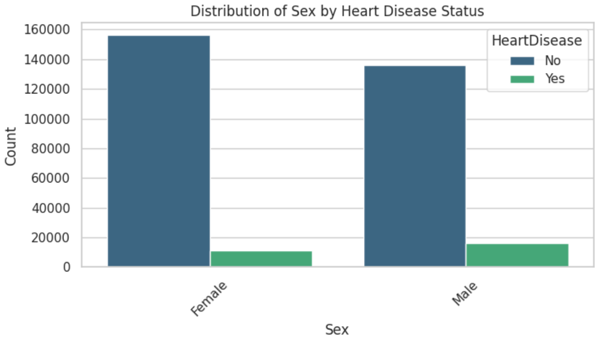

Distribution of GenHealth by Heart Disease Status
HeartDisease
No
Yes
Count
100000
80000
60000
40000
20000
0
Very good
Fair
Good
Poor
Excellent
GenHealth


Distribution of Asthma by Heart Disease Status
HeartDisease
No
Yes
Count
250000
200000
150000
100000
50000
0
Yes
No
Asthma


Distribution of PhysicalActivity by Heart Disease Status
HeartDisease
No
Yes
Count
200000
150000
100000
50000
0
Yes
No
PhysicalActivity


Distribution of KidneyDisease by Heart Disease Status
HeartDisease
No
Yes
Count
250000
200000
150000
100000
50000
0
No
Yes
KidneyDisease


Distribution of SkinCancer by Heart Disease Status
HeartDisease
No
Yes
Count
250000
200000
150000
100000
50000
0
Yes
No
SkinCancer


Distribution of Sex by Heart Disease Status
HeartDisease
No
Yes
Count
160000
140000
120000
100000
80000
60000
40000
20000
0
Female
Male
Sex

Appendix C

Another Appendix

	Precision	Recall	F1-Score	Support
0	0.94	0.84	0.89	58484
1	0.19	0.38	0.25	5475
Accuracy			0.80	63959
Macro Avg	0.56	0.61	0.57	63959
Weighted Avg	0.87	0.80	0.83	63959
ROC AUC Score			0.6682972465013263	

Table C.1: Classification Report on Downsampled Majority Class with Optimization techniques

	Precision	**Recall**	**F1-Score**	**Support**
0	0.94	0.92	0.93	58484
1	0.30	0.38	0.34	5475
Accuracy			0.87	63959
Macro Avg	0.62	0.65	0.63	63959
Weighted Avg	0.89	0.87	0.88	63959
ROC AUC Score			0.6682972465013263	

Table C.2: Classification Report with SMOTEENN and Optimization techniques

Bibliography

[1] A. Khoja, P.H. Andraweera, Z.S. Lassi, A. Ali, M. Zheng, M.M. Pathirana, E. Aldridge, M.R. Wittwer, D.D. Chaudhuri, R. Tavella, M.A. Arstall, Risk Factors for Early-Onset Versus Late-Onset Coronary Heart Disease (CHD): Systematic Review and Meta-Analysis, Heart Lung Circ. 32, 1277â1311 (2023). https://doi.org/10.1016/J.HLC.2023.07.010

[2] U. Ralapanawa, R. Sivakanesan, Epidemiology and the Magnitude of Coronary Artery Disease and Acute Coronary Syndrome: A Narrative Review, J Epidemiol Glob Health. 11, 169â177 (2021). https://doi.org/10.2991/jegh.k.201217.001

[3] C.L. Grines, A.J. Klein, H. Bauser-Heaton, M. Alkhouli, N. Katukuri, V. Aggarwal, S.E. Altin, W.B. Batchelor, J.C. Blankenship, F. Fakorede, B. Hawkins, G.A. Hernandez, N. Ijioma, B. Keeshan, J. Li, R.A. Ligon, A. Pineda, Y. Sandoval, M.N. Young, Racial and ethnic disparities in coronary, vascular, structural, and congenital heart disease, Catheterization and Cardiovascular Interventions. 98, 277â294 (2021). https://doi.org/10.1002/CCD.29745

[4] Y. Zhao, E.P. Wood, N. Mirin, R. Vedanthan, S.H. Cook, R. Chunara, Machine Learning for Integrating Social Determinants in Cardiovascular Disease Prediction Models: A Systematic Review, medRxiv. 2020.09.11.20192989 (2020). https://doi.org/10.1101/2020.09.11.20192989

[5] S. Begum, A. Siddique, R. Tiwari, A Study for Predicting Heart Disease using Machine Learning, Turkish Journal of Computer and Mathematics Education (TURCOMAT). 12, 4584â4592 (2021).

[6] I. El Naqa, M.J. Murphy, What Is Machine Learning? Machine Learning in Radiation Oncology, 3â11 (2015). https://doi.org/10.1007/978-3-319-18305-3_1

[7] T.M. Mitchell, The Discipline of Machine Learning, (2006).

[8] J.C. Stoltzfus, Logistic Regression: A Brief Primer, Academic Emergency Medicine. 18, 1099â1104 (2011). https://doi.org/10.1111/J.1553-2712.2011.01185.X

[9] P. Ranganathan, C. Pramesh, R. Aggarwal, Common pitfalls in statistical analysis: Logistic regression, Perspect Clin Res. 8, 148 (2017). https://doi.org/10.4103/PICR.PICR_87_17

[10] E. Szczerbicki, Management of Complexity and Information Flow, Agile Manufacturing: The 21st Century Competitive Strategy. 247â263 (2001). https://doi.org/10.1016/B978-008043567-1/50013-9

[11] Decision Tree Algorithm in Machine Learning - Javatpoint, https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[12] Z. Zhang, Introduction to machine learning: k-nearest neighbors, Ann Transl Med. 4, (2016). https://doi.org/10.21037/ATM.2016.03.37

[13] S. Uddin, I. Haque, H. Lu, M.A. Moni, E. Gide, Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction, Scientific Reports 2022 12:1. 12, 1â11 (2022). https://doi.org/10.1038/s41598-022-10358-x

[14] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Front Neurorobot. 7, (2013). https://doi.org/10.3389/FNBOT.2013.00021

[15] A.K. Malakar, D. Choudhury, B. Halder, P. Paul, A. Uddin, S. Chakraborty, A review on coronary artery disease, its risk factors, and therapeutics, J Cell Physiol. 234, 16812â16823 (2019). https://doi.org/10.1002/JCP.28350

[16] M. Trigka, E. Dritsas, Long-Term Coronary Artery Disease Risk Prediction with Machine Learning Models, Sensors 2023, Vol. 23, Page 1193. 23, 1193 (2023). https://doi.org/10.3390/S23031193

[17] JC, B., TE, G., E, K., Risk Factors For Coronary Artery Disease, Risk Factors in Coronary Artery Disease. 1â219 (2020). https://doi.org/10.3109/9781420014570

[18] G.F. Gensini, M. Comeglio, A. Colella, Classical risk factors and emerging elements in the risk profile for coronary artery disease, Eur Heart J. 19 Suppl A, A53-61 (1998).

[19] X.Y. Zhang, L. Shu, C.J. Si, X.L. Yu, D. Liao, W. Gao, L. Zhang, P.F. Zheng, Dietary Patterns, Alcohol Consumption and Risk of Coronary Heart Disease in Adults: A Meta-Analysis, Nutrients 2015, Vol. 7, Pages 6582-6605. 7, 6582â6605 (2015). https://doi.org/10.3390/NU7085300

[20] R. WOLK, A. GAMI, A. GARCIATOUCHARD, V. SOMERS, Sleep and Cardiovascular Disease, Curr Probl Cardiol. 30, 625â662 (2005). https://doi.org/10.1016/J.CPCARDIOL.2005.07.002

[21] S. Amarasekera, P. Jha, Understanding the links between cardiovascular and psychiatric conditions, Elife. 11, (2022). https://doi.org/10.7554/ELIFE.84524

[22] T.R. Ramesh, U.K. Lilhore, M. Poongodi, S. Simaiya, A. Kaur, M. Hamdi, PREDICTIVE ANALYSIS OF HEART DISEASES WITH MACHINE LEARNING APPROACHES, Malaysian Journal of Computer Science. 2022, 132â148 (2022). https://doi.org/10.22452/MJCS.SP2022NO1.10

[23] L.J. Muhammad, I. Al-Shourbaji, Â· Ahmed, A. Haruna, I.A. Mohammed, A. Ahmad, Â· Muhammed, B. Jibrin, Machine Learning Predictive Models for Coronary Artery Disease, SN Comput Sci. 2, 350 (2021). https://doi.org/10.1007/s42979-021-00731-4

[24] A.L. Yadav, K. Soni, S. Khare, Heart Diseases Prediction using Machine Learning, 2023 14th International Conference on Computing Communication and Net-

working Technologies (ICCCNT). 1â7 (2023). https://doi.org/10.1109/ICCCNT56998.2023.10306469

[25] F.S. Alotaibi, Implementation of Machine Learning Model to Predict Heart Failure Disease, International Journal of Advanced Computer Science and Applications. 10, 261â268 (2019). https://doi.org/10.14569/IJACSA.2019.0100637

[26] M. Muthuvel, M. Marimuthu, M. Abinaya, K.S. Hariesh, K. Madhankumar, A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach, Article in International Journal of Computer Applications. 181, 975â8887 (2018). https://doi.org/10.5120/ijca2018917863

[27] C.M. Bhatt, P. Patel, T. Ghetia, P.L. Mazzeo, Effective Heart Disease Prediction Using Machine Learning Techniques, Algorithms 2023, Vol. 16, Page 88. 16, 88 (2023). https://doi.org/10.3390/A16020088

[28] E.W. Steyerberg, A.J. Vickers, N.R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M.J. Pencina, M.W. Kattan, *Assessing the performance of prediction models: a framework for some traditional and novel measures*, *Epidemiology*, **21**, 128, (2010), https://doi.org/10.1097/EDE.0B013E3181C30FB2

[29] *Precision and Recall: How They Affect Predictive Model Evaluation*, (Year), https://www.linkedin.com/advice/0/how-do-precision-recall-affect-predictive-model.

[30] S.A. Hicks, I. StrÃ¼mke, V. Thambawita, M. Hammou, M.A. Riegler, P. Halvorsen, S. Parasa, *On evaluation metrics for medical applications of artificial intelligence*, *Sci Rep*, **12**, (2022), https://doi.org/10.1038/S41598-022-09954-8

[31] *F1 Score in Machine Learning: Intro & Calculation*, (Year), https://www.v7labs.com/blog/f1-score-guide.

[32] J.Y. Verbakel, E.W. Steyerberg, H. Uno, B. De Cock, L. Wynants, G.S. Collins, B. Van Calster, *ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models*, *J Clin Epidemiol*, **126**, 207â216, (2020), https://doi.org/10.1016/J.JCLINEPI.2020.01.028.

[33] U. Khurana, H. Samulowitz, D. Turaga, *Feature Engineering for Predictive Modeling Using Reinforcement Learning*, *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**, 3407â3414, (2018), https://doi.org/10.1609/AAAI.V32I1.11678.

[34] J. Wu, X.Y. Chen, H. Zhang, L.D. Xiong, H. Lei, S.H. Deng, *Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization*, *Journal of Electronic Science and Technology*, **17**, 26â40, (2019), https://doi.org/10.11989/JEST.1674-862X.80904120.

[35] Tracyrenee, *Define class weights to make predictions on a class imbalanced dataset*, *Medium*, (Year), https://medium.com/mlearning-ai/define-class-weights-to-make-predictions-on-a-class-imbalanced-datas

[36] Author's Name, *The Importance of Cross Validation - Data Science Courses | DataScientest*, (Year), https://datascientest.com/en/the-importance-of-cross-validation.