Machine Learning Based Approach for Accurate Heart Disease Prediction

A Research Project Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science (Engg.) in Computer Science and Engineering

Submitted By

Muhammad Abdul Mukit (CE18040)

Minhajul Islam Tapadar (CE18042)

Supervised By

Mohd. Sultan Ahammad

Assistant Professor

Department of Computer Science and Engineering

Department of Computer Science and Engineering

Mawlana Bhashani Science and Technology University

Santosh, Tangail-1902, Bangladesh

Machine Learning Based Approach for Accurate Heart Disease Prediction

A Research Project Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science (Engg.) in Computer Science and Engineering

Submitted By

Muhammad Abdul Mukit (CE18040)

Minhajul Islam Tapadar (CE18042)

Supervised By

Mohd. Sultan Ahammad

Assistant Professor

Department of Computer Science and Engineering

Board of Examiners:

…… (Supervisor)

…… (Examiner)

…… (Examiner)

Thesis Approval

Muhammad Abdul Mukit (CE18040)
Minhajul Islam Tapadar (CE18042)

Machine Learning Based Approach for Accurate Heart Disease Prediction

We the undersigned, recommend that the thesis completed by the students listed above, in partial fulfilment of B.Sc. (Engg.) in Computer Science and Engineering degree requirements, be accepted by the Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University for deposit

Supervisor Approval

……
Mohd. Sultan Ahammad
Assistant Professor
Department of Computer Science and Engineering

Departmental Approval

……
Dr. Md. Sazzad Hossain
Professor
Department of Computer Science and Engineering

Mawlana Bhashani Science and Technology University
Santosh, Tangail-1902, Bangladesh

Declaration

We hereby declare that the research presented in this project is the result of our own investigation conducted under the guidance of Assistant Professor Mohd. Sultan Ahammad from the Department of Computer Science and Engineering (CSE) at Mawlana Bhashani Science and Technology University (MBSTU), Bangladesh. We affirm that this thesis, in its entirety, has not been submitted elsewhere for any degree or diploma. Any information obtained from the published work of others has been duly acknowledged within the text, and a comprehensive list of references is provided.

Muhammad Abdul Mukit
Student ID: CE18040

.....
Signature of Candidate

Minhajul Islam Tapadar
Student ID: CE18042

.....
Signature of Candidate

Acknowledgements

All praise goes to the Almighty Allah, who enabled us to complete and submit the thesis "Machine Learning Based Approach for Accurate Heart Disease Prediction" successfully for the completion of the degree of Bachelor of Science (Engg.) in this study. Besides my supervisor, we would like to thank the rest of our supervisory committee members for their insightful comments and encouragement. We extend our heartfelt appreciation and deepest gratitude to our esteemed supervisor, Assistant Professor Mohd. Sultan Ahammad, from the Department of Computer Science and Engineering (CSE) at Mawlana Bhashani Science and Technology University. We express our sincerest thanks for his unwavering supervision, invaluable intellectual guidance, and unwavering encouragement throughout the entirety of this thesis. His comments and suggestions were very stimulating and developed our ideas for accomplishing them.

Finally, we would like to thank our parents for giving us mental support at different stages during the completion of my research work.

Muhammad Abdul Mukit
Minhajul Islam Tapadar
June,2023

Abstract

Accurate prediction of heart disease is essential for effective diagnosis and timely intervention. This thesis proposes various machine learning approach that was used to find the best accuracy of heart disease prediction from this dataset. The research methodology involves preprocessing the heart disease dataset, including data cleaning, feature engineering, and normalization. Filter Feature selection techniques are applied to identify the most relevant features for heart disease prediction. We used smoothing to generate synthetic data and merge it with the original data in order to improve performance. To determine which algorithm is best for achieving the goal, we tested different classification algorithms, including Logistic Regression, Decision Tree, KNN, Naïve Bayes, Random Forest Classifier. Then, to train our model, we used ensemble techniques Gradient Boosting. In our case, Naïve Bayes works well with an accuracy of 69.803%, Regression works a little better with an accuracy of 73.445%. KNN, Random Forest, and Decision Tree work with much better accuracy of 82.910%, 86.323% and 86.829% respectively. Our best finding ensemble techniques model (gradient boosting) works well with an accuracy of 91.596%, precision values of 91.404%, recall values of 91.837% and f1 score values of 91.620%. Ultimately, the experimental findings demonstrated noteworthy enhancements in the accuracy of the prediction classifiers when utilizing the gradient boosting model.

Keywords: classification, heart disease, ensemble, machine learning, random forest, KNN, decision tree, gradient boosting

Contents

List of Figures

List of Tables

List of Abbreviations

ML	Machine Learning
DL	Deep Learning
SOM	Self – Organizing Map
CNN	Convolutional Neural Network
DNN	Deep Neural Network
WHO	World Health Organization
AUC	Area under the ROC Curve
ROC	Receiver Operating Characteristic Curve
DT	Decision Tree
KNN	K-Nearest Neighbors
SVM	Support Vector Machines
NB	Naïve Bayes
GB	Gradient Boosting
LR	Logistic Regression
RF	Random Forest
RNN	Recurrent Neural Networks

Chapter 1

Introduction

1.1 Overview

Heart disease is a leading cause of mortality worldwide. According to estimates, 17.9 million people worldwide die from heart disease annually [1]. Effective treatment and management of heart disease depend on early identification and precise diagnosis. Since current approaches for diagnosing heart illnesses are inefficient for early detection for a variety of reasons, including accuracy and computational time [2], researchers are working to create an efficient strategy for the timely identification of heart disorders. It is extremely difficult to diagnose and manage cardiac disease when cutting-edge equipment and qualified medical professionals are not available [3]. A proper diagnosis and course of action can save the lives of several patients [4]. Heart illnesses are diagnosed by a doctor using the patient's medical history, the results of the physical exam, and an analysis of any alarming symptoms. However, the findings of this method of diagnosis are insufficient to detect heart disease patients. Furthermore, it is both costly and computationally challenging to examine [5]. Given the complexity of the data, data mining and machine learning (ML) approaches [6] are gaining popularity as tools for data analysis. Accurate health tools may be created using machine learning and data-driven strategies. The objective of the current study is to recognize and evaluate the organizational barriers that prevent medical institutions from implementing a successful strategy and provide managers with tactical answers to these challenges [7]. Machine learning can aid in accurate heart disease detection for effective treatment. Accurate prediction of heart disease plays a crucial role in facilitating early diagnosis, timely intervention, and effective patient management. Traditional approaches to heart disease prediction have relied on single algorithms, but recent advancements in machine learning techniques have opened up new possibilities for improving the accuracy of predictions. Our aim is to provide a better hybrid machine learning model for classifying heart diseases.

1.2 Motivation

The world population is growing very fast, and in this fast-growing world, the rate of heart disease patients is increasing because of unhealthy lifestyles [8], making it a global concern for researchers. Traditional risk assessment models often rely on limited variables and may not capture the complexity of the disease accurately, and there are fewer experts to detect heart diseases easily. In recent years, machine learning techniques have shown promising results in various medical domains, including heart disease prediction. Over the past few decades, there has been a notable increase in interest regarding the application of machine learning techniques in the field of heart disease.

Heart diseases can be predicted using ML. By analyzing large datasets of patient data, machine learning algorithms can identify patterns that are associated with heart disease. This information can then be used to develop models that can predict who is at risk of developing heart disease. However, the development of such models includes numerous challenging tasks, including accuracy. The early and accurate prediction of heart disease plays a vital role in its diagnosis. This crucial step has a significant impact on patient outcomes as it enables timely interventions and the implementation of personalized treatment plans. By leveraging advanced prediction models, healthcare professionals can identify individuals at risk of heart disease at an early stage, allowing for proactive measures to be taken. This not only improves patient outcomes but also has the potential to save countless lives.

1.3 Aims and Objectives

Our Primary Objectives are:

- To Prevent Heart diseases by predicting the possibility of heart diseases using ML model.
- To see which features are affecting our model mostly.
- To see if our model perfumes better with more data generated by artificial smoothing.
- To build a model where we can minimize the losses of the model as much as possible and to provide the most accurate outcome.

1.4 Challenges

One of the biggest challenges we face in developing this machine learning models for heart disease prediction is the availability of high-quality data. Heart disease is a complex disease with many contributing factors, and it can be difficult to collect data on all of these factors. Additionally, the data that is available may be incomplete or inaccurate, which can lead to inaccurate models.

Machine learning models can be complex, and it can be difficult to understand how they make predictions. This can make it difficult to interpret the results of the model and to use the model to make clinical decisions.

Machine learning models can be biased, which means that they may not be accurate for all populations. This can be a challenge for heart disease prediction, as the disease affects different populations in different ways.

By addressing these challenges, hybrid machine learning approaches have the potential to improve the accuracy of heart disease prediction and make a significant impact on the prevention and treatment of the disease.

1.5 Thesis Organization

This research is organized as follows:

Chapter 1 Introduction
Provides an introduction to Machine Learning Based Approach for Accurate Heart Disease Prediction. Then motivation of this thesis is discussed briefly. Later research questions are discussed. And finally, we have been identified the aims and objective with the challenge we faced.

Chapter 2 Literature Review
In this chapter we highlight the previous research works on heart disease that based on Machine Learning as well as limitations.

Chapter 3 Methodology
Illustrates in depth the model and operating principle of the planned ML-based research Endeavor.

Chapter 4 Result Analysis and Discussion
Result analysis and discussion presents simulation results and taken decisions.

Chapter 5 Limitation and Future Work

Limitation and future work focus on future study directions of the entire research project.

Chapter 6 Conclusion

We concentrate on how to wrap up the complete study endeavor in this chapter.

Chapter 2

Literature Review

In this section, we have mentioned some literature reviews of recent works which are given below:

In a research paper by Mohammad Shabaz et al. [9], a new approach called "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning" is proposed. The authors utilized the Public Health Dataset, which consists of 76 attributes. However, previous studies only utilized a subset of 14 attributes. Feature selection was conducted using the Lasso method. Results from the first strategy, which did not include feature selection and outlier detection, showed that the random forest achieved an accuracy of 76.7%, logistic regression achieved 83.64%, KNN achieved 82.27%, support vector machine achieved 84.09%, decision tree achieved 75.0%, and XGBoost achieved 70.0%. Results from the second strategy, which included feature selection but not outlier detection, revealed that the random forest achieved an accuracy of 88%, logistic regression achieved 85.9%, KNN achieved 79.69%, support vector machine achieved 84.26%, decision tree achieved 76.35%, and XGBoost achieved 71.1%. Lastly, by employing the 4ird Approach, which involved both feature selection and outlier detection, the Random Forest achieved an accuracy of 80.3%, Logistic Regression achieved 83.31%, KNN achieved 84.86%, Support Vector Machine achieved 83.29%, Decision Tree achieved 82.33%, and XGBoost achieved 71.4%. KNN emerged as the winner with an accuracy of 77.7% and a specificity of 80%.

In their article titled "Heart Disease Prediction Using Machine Learning Techniques," authors Devansh Shah et al. [10] introduced a distinct approach. This study explored numerous attributes related to heart disease and utilized supervised learning methods such as K-nearest neighbor, Naive Bayes, decision trees, and the random forest algorithm to develop a predictive model. Out of the 76 attributes available, only 14 were selected for testing. Interestingly, K-nearest neighbor exhibited the highest accuracy score among the employed methods.

Due to the complexity of the heart, it is crucial to handle it with care to avoid any potential fatalities. Various approaches, including K-nearest neighbor (KNN), decision trees, genetic algorithms, and naive Bayes, are employed to categorize the severity of heart disease [11].

According to Mohan et al. [12], a hybrid strategy combining two different methods yields the highest accuracy among all others, achieving an impressive 88.4%.

Academic researchers have utilized data mining techniques to make predictions regarding cardiac disorders. In their study, Kaur et al. [13] delve into the extraction of intriguing patterns and information from a vast dataset. The researchers compared the accuracy of various machine learning and data mining methods to identify the most effective approach, and the results strongly favor Support Vector Machine (SVM).

Numerous researchers, such as Kohali et al. [14], are actively engaged in developing machine learning algorithms for predicting various illnesses. Their studies encompass the prediction of heart disease using logistic regression, diabetes using support vector machines, and breast cancer using the Adaboot classifier. The findings indicate that logistic regression achieved an accuracy of 87.1%, support vector machines achieved 85.71% accuracy, and the Adaboot classifier exhibited an accuracy of up to 89.57%. These results demonstrate the effectiveness of these algorithms in making accurate predictions for different illnesses.

In a study by Shantakumar B. Patil et al. [15], an efficient approach for heart attack prediction was presented, focusing on extracting significant patterns from the dataset. The K-means clustering algorithm was employed, and the weight of each item was determined using the MAFIA algorithm. Patterns with a weightage higher than the threshold were considered for prediction, enhancing the accuracy of the prediction model.

Jyoti Soni et al. [16] proposed a heart disease prediction model utilizing a 15-attribute dataset and employing data mining techniques such as time series analysis, artificial neural networks (ANN) clustering rules, and association rules. The paper also introduced a technique to increase accuracy by reducing the dataset size through the application of a genetic algorithm. This approach aimed to optimize the prediction process by selecting the most informative attributes and reducing unnecessary data.

Headey Takci et al. [17] introduced a method to enhance heart attack prediction through the use of feature selection. The study employed twelve classification methods and four feature selection algorithms to predict heart attacks. Processing time, model accuracy, and ROC analysis were utilized for comparing the different approaches.

Fizar Ahmed et al. [18] proposed an Internet of Things (IoT)-based application for heart attack prediction. The system leveraged IoT technologies to develop a predictive model.

Poornima Singh et al. [19] presented an effective heart disease prediction system. The proposed system utilized the Multilayer Perceptron Neural Network (MLPNN) with the backpropagation (BP) algorithm to predict heart disease.

Numerous methods, including deep learning, have been developed for the classification of heart diseases. However, there is still scope for improvement in achieving satisfactory accuracy in classification experiments. Further advancements in medical treatment development in this area can bring a new dimension to the field of medical computation.

Chapter 3

Methodology

This chapter has been organized by the proposed methodology with its architecture. After that, each of the parts of our procedures is discussed in different sections.

3.1 Proposed Methodology

Building a machine learning based model for accurate heart disease prediction is the foundation of our suggested methodology.

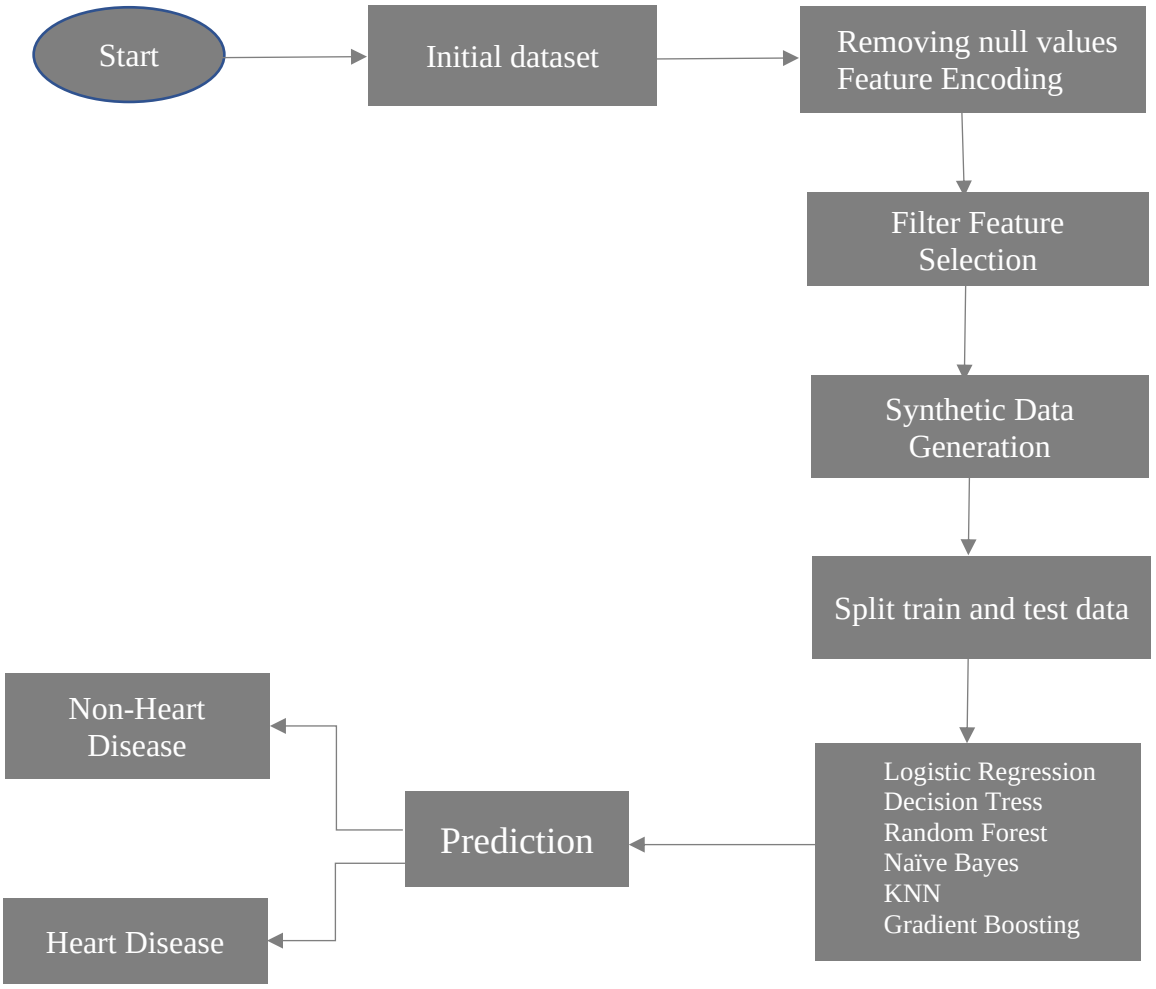

Figure 3.1: Proposed Methodology

3.2 Data Description

Sourced initially from the CDC, the dataset forms a significant portion of the Behavioral Risk Factor Surveillance System (BRFSS) and was acquired from the Kaggle platform [20] were used in our study. The majority of columns in the dataset pertain to inquiries regarding the health status of respondents, such as questions like "Do you have serious difficulty walking or climbing stairs?" or "Have you ever smoked at least 100 cigarettes in your entire life?" The dataset encompasses information from 319,795 individuals. The survey comprises multiple sections. Within the dataset, there are 18 feature variables, including 9 Boolean values, 5 strings, and 4 decimal numbers. In machine learning endeavors, the variable "HeartDisease" can be utilized as the explanatory variable; however, it is important to note that the classes within this variable exhibit significant imbalances.

These data were provided as a csv file. Then we performed some pre-processing, such as deleting unnecessary features and rows with just null values.

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity
0	No	16.60	Yes	No	No	3	30	No	Female	55-59	White	Yes	Yes
1	No	20.34	No	No	Yes	0	0	No	Female	80 or older	White	No	Yes
2	No	26.58	Yes	No	No	20	30	No	Male	65-69	White	Yes	Yes
3	No	24.21	No	No	No	0	0	No	Female	75-79	White	No	No
4	No	23.71	No	No	No	28	0	Yes	Female	40-44	White	No	Yes
5	Yes	28.87	Yes	No	No	6	0	Yes	Female	75-79	Black	No	No
6	No	21.63	No	No	No	15	0	No	Female	70-74	White	No	Yes
7	No	31.64	Yes	No	No	5	0	Yes	Female	80 or older	White	Yes	No
8	No	26.45	No	No	No	0	0	No	Female	80 or older	White	No, borderline diabetes	No
9	No	40.69	No	No	No	0	0	Yes	Male	65-69	White	No	Yes
10	Yes	34.30	Yes	No	No	30	0	Yes	Male	60-64	White	Yes	No
11	No	28.71	Yes	No	No	0	0	No	Female	55-59	White	No	Yes
12	No	28.37	Yes	No	No	0	0	Yes	Male	75-79	White	Yes	Yes
13	No	28.15	No	No	No	7	0	Yes	Female	80 or older	White	No	No
14	No	29.29	Yes	No	No	0	30	Yes	Female	60-64	White	No	No
15	No	29.18	No	No	No	1	0	No	Female	50-54	White	No	Yes
16	No	26.26	No	No	No	5	2	No	Female	70-74	White	No	No
17	No	22.59	Yes	No	No	0	30	Yes	Male	70-74	White	No, borderline diabetes	Yes
18	No	29.86	Yes	No	No	0	0	Yes	Female	75-79	Black	Yes	No
19	No	18.13	No	No	No	0	0	No	Male	80 or older	White	No	Yes
20	No	21.16	No	No	No	0	0	No	Female	80 or older	Black	No, borderline diabetes	No
21	No	28.90	No	No	No	2	5	No	Female	70-74	White	Yes	No
22	No	26.17	Yes	No	No	0	15	No	Female	45-49	White	No	Yes
23	No	25.82	Yes	No	No	0	30	No	Male	80 or older	White	Yes	Yes
24	No	25.75	No	No	No	0	0	No	Female	80 or older	White	No	Yes

Figure 3.2: Dataset after cleaning

3.2.1 Data frame Count Plot

Count plots are utilized to visually represent the frequencies of observations within categorical bins using bars. The count plot function typically accepts two parameters, namely x and y, which can be the names of variables in the dataset or vector data. These parameters serve as inputs for plotting long-form data. In this context, the Seaborn Library was employed to create the count plots.

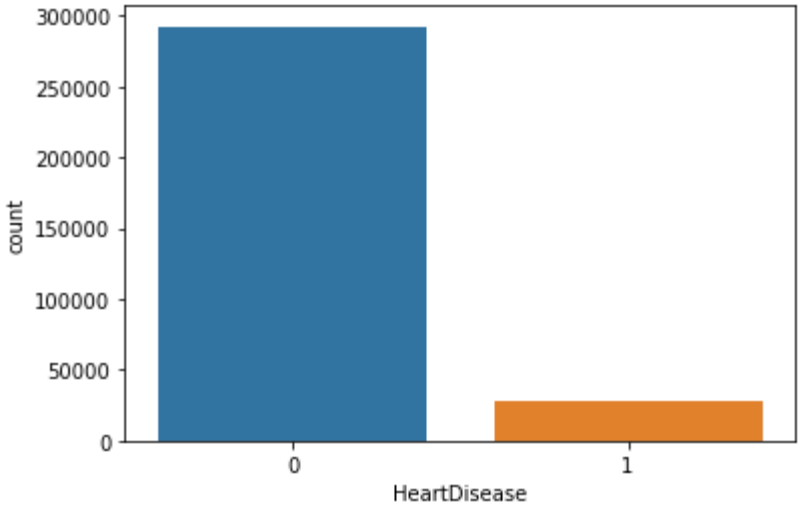

Figure 3.3: Data frame counter plot

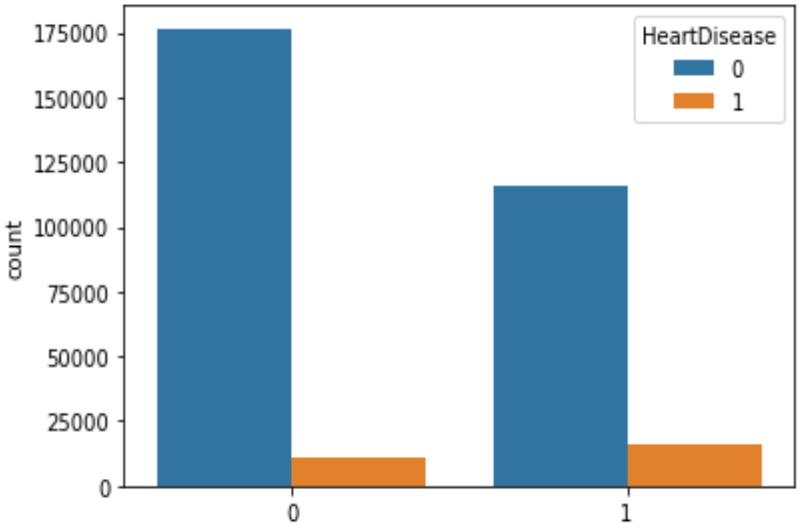

Figure 3.4: Counter plot with respect to is he/she smoking

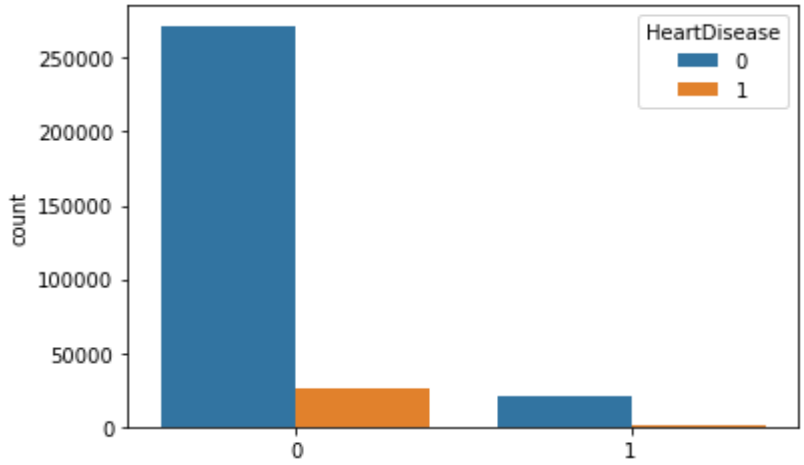

Figure 3.5: Counter plot with respect to if he/she drink alcohol

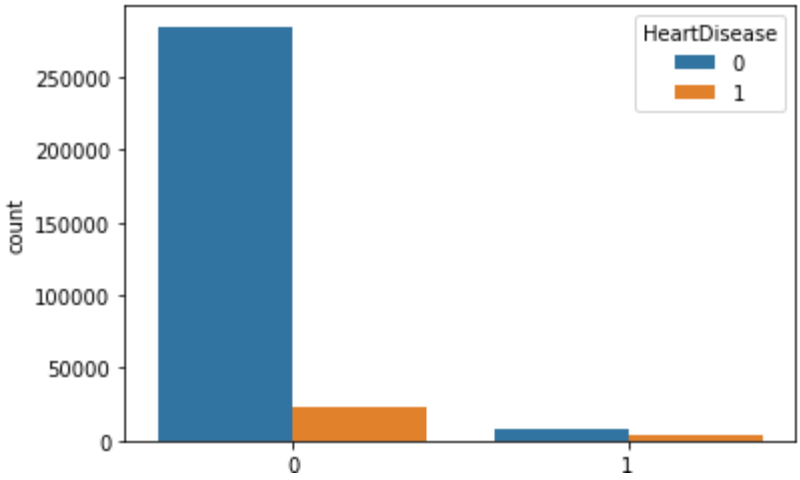

Figure 3.6: Counter plot with respect to if he/she had stroke

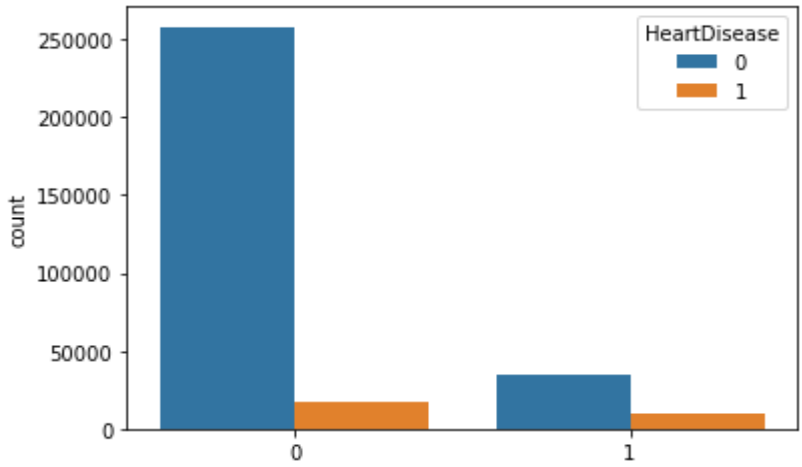

Figure 3.7: Counter plot with respect to if he/she has difficulty in walking

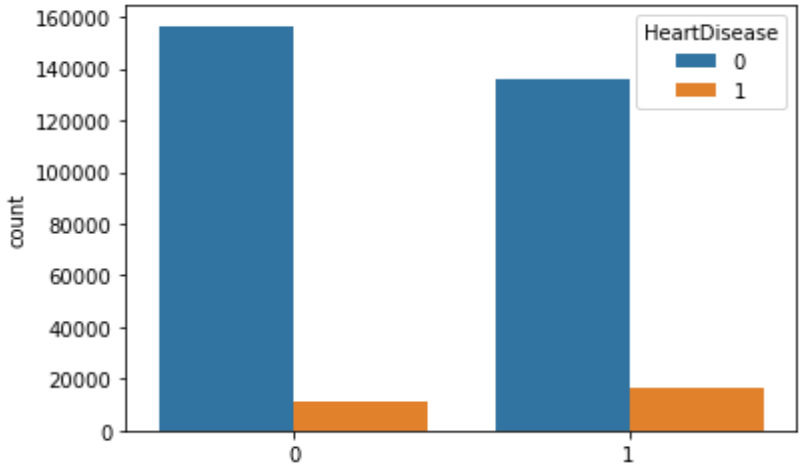

Figure 3.8: Counter plot with respect to gender

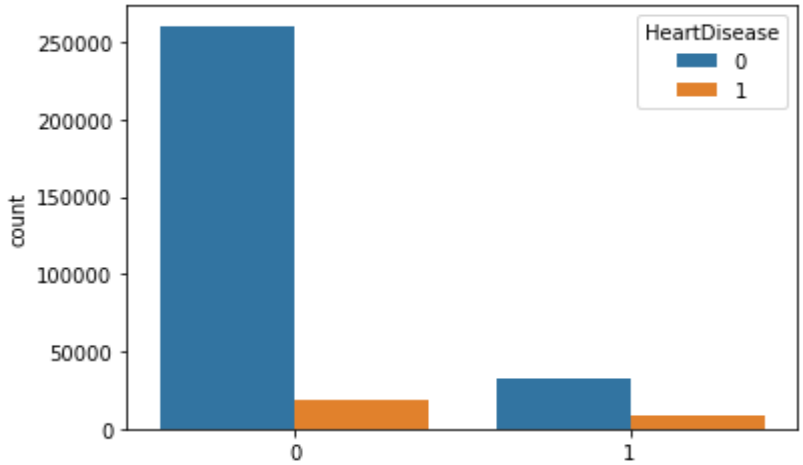

Figure 3.9: Counter plot with respect to if he/she is a Diabetic Patient

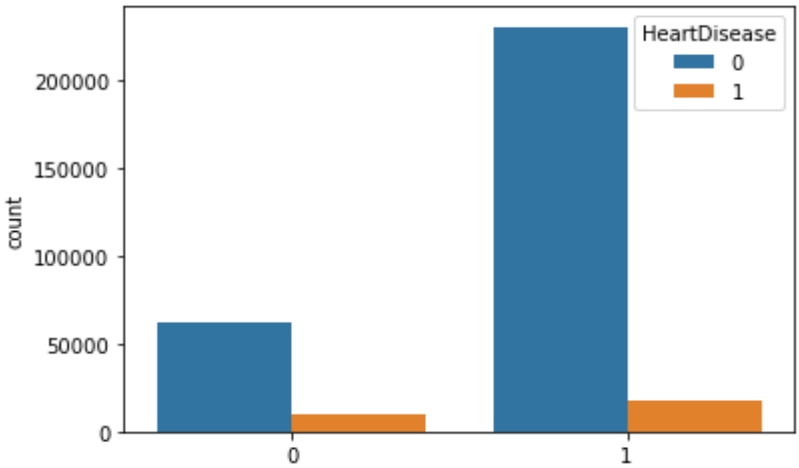

Figure 3.10: Counter plot with respect to Physical Activity

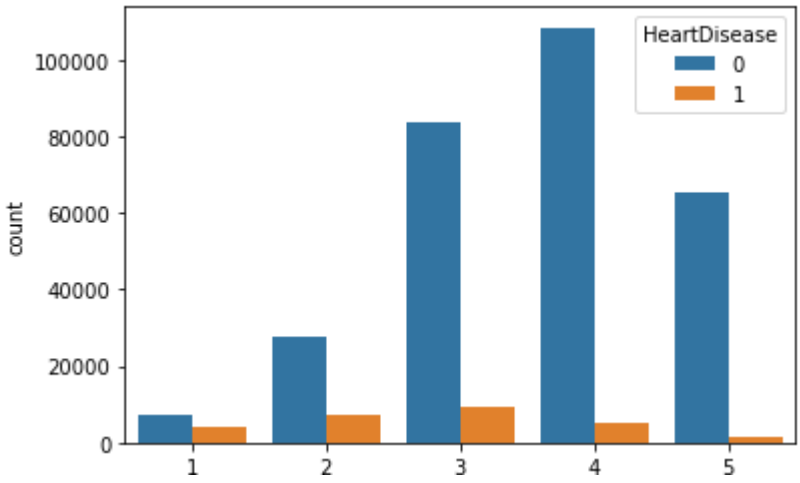

Figure 3.11: Counter plot with respect to general health

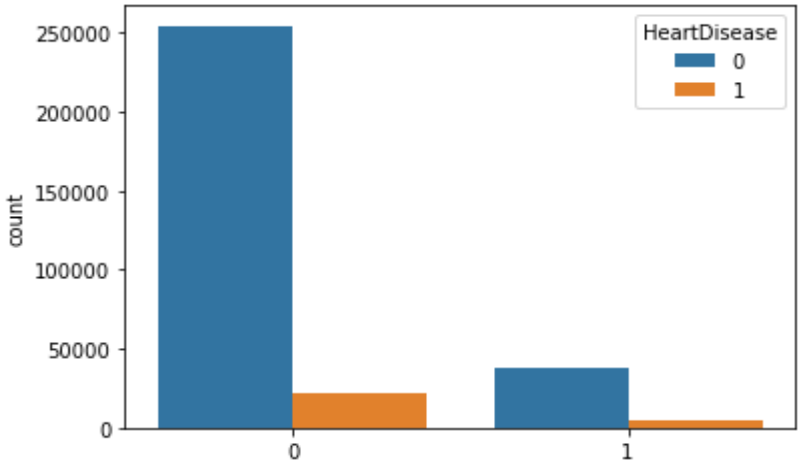

Figure 3.12: Counter plot with respect to if he/she has Asthma

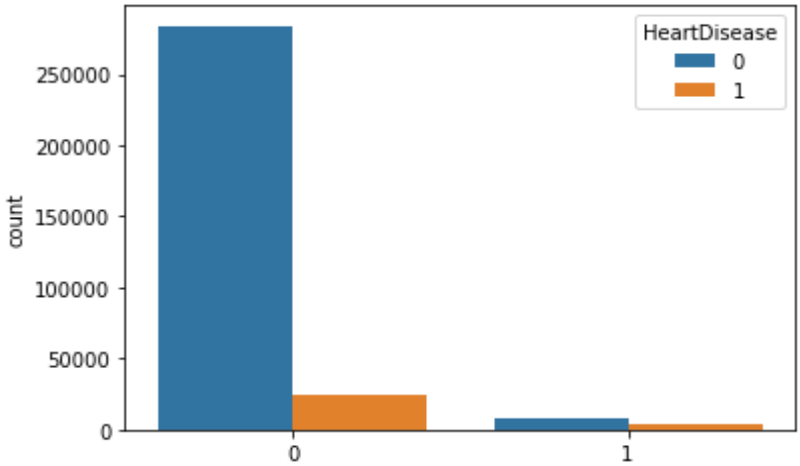

Figure 3.13: Counter plot with respect to if he/she has kidney disease

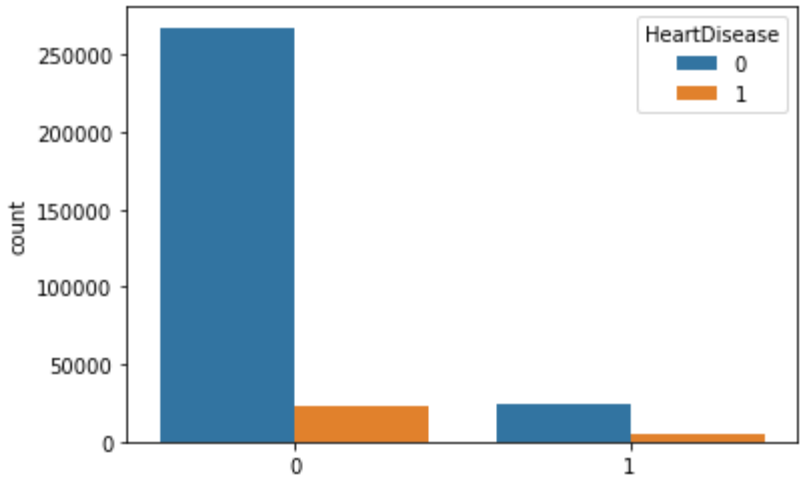

Figure 3.14: Counter plot with respect to if he/she has Skin Cancer

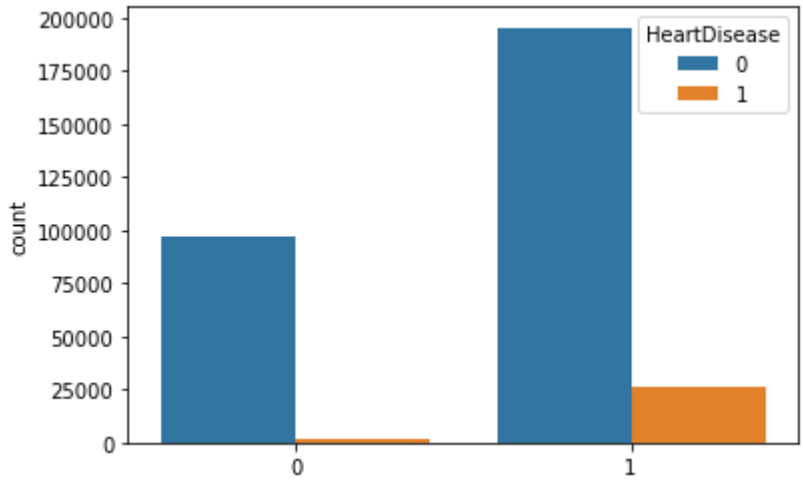

Figure 3.15: Counter plot with respect to Age Category

3.3 Data Preprocessing

Data preprocessing serves as the initial stage in machine learning model development. Data preprocessing includes imputation of missing data, elimination or modification of outlier observations, and data transformation (often normalization and standardization).

3.3.1 Missing Data Imputation

There were instances in our dataset where all values were null; we eliminated these instances by removing them entirely. Imputation is applied to other null values. Imputation is a technique for substituting viable alternatives for missing data values in a dataset. We replaced all NaN values

for column BMI, Physical Health, Mental Health, Age Category, and Sleep Time with the mean of the variable, for the rest of column, where we replaced with the most frequent values.

3.3.2 Feature Encoding

Categorical variables can be easily converted into numerical values by utilizing the method of encoding, which enables a machine learning model to be quickly customized to the collected data. Machine learning models are limited to using numeric values. This is why it's crucial to turn the relevant attributes' categorical values into numeric ones.

We used self-defined numbers for encoding to the column of Smoking, HeartDisease, Stroke, AlcoholDrinking, Sex, Diabetic, PhysicalActivity, AgeCategory, GenHealth, DiffWalking, Asthma, KidneyDisease, SkinCancer.

3.4 Machine Learning Models

To achieve accurate heart disease prediction, we employed a range of machine learning models on our dataset. By leveraging these models, we aimed to improve the accuracy and reliability of our predictions. The specific models utilized include algorithms such as Logistic Regression, Decision Tree, Naive Bayesian, Random Forest Classifier, K-Nearest Neighbors, Gradient Boosting. By employing multiple models, we sought to explore different approaches and identify the most accurate and robust model for heart disease prediction.

3.4.1 Logistic Regression

Logistic regression is a well-known supervised learning algorithm renowned for its ability to predict categorical outcomes based on independent variable values. By analyzing the relationship between the independent variables and the categorical target variable, logistic regression estimates the probability of belonging to a specific category. This algorithm is widely used in various domains and offers valuable insights into classification problems where the outcome is discrete or binary in nature. [21][22].

Logistic regression is a predictive modeling technique that focuses on categorical dependent variables. It is specifically designed to handle variables that are categorical or discrete in nature.

The predicted outcomes from logistic regression can be in the form of "Yes" or "No," "0" or "1," "true" or "false," or other similar categories. However, what sets logistic regression apart is that it provides probability values ranging between 0 and 1, offering insights into the likelihood of belonging to a specific category

Linear and logistic regression are comparable but used differently. Linear regression solves regression issues, whereas logistic regression classifies.

Using logistic regression, which predicts two maximum values (0 and 1), the best "S"-shaped logistic function is sought for rather than the best straight line. The logistic function curve may predict outcomes like cell cancer, mouse obesity, and more. Logistic Regression can categorize observations from a broad range of data to rapidly determine which variables are best for categorization.

The flow chart of Logistic Regression is given below:

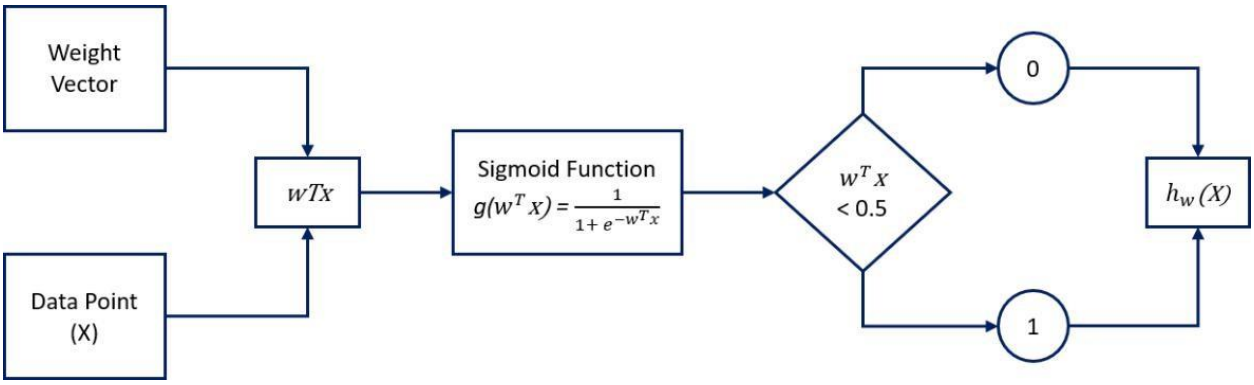

Figure 3.16: Flow-chart of logistic regression

You'll find the formula for calculating the Logistic Regression cost function down below.

Here, h_w (x) is the value of hypothesis which we are calculate using sigmoid function.

$J(w_0, w_1, w_2, \ldots, w_n) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}\{hw(xi), yi\}$

where,

Cost $\{h_w(x^i), y^i\} = -y^i \log(h_w(x)) - (1 - y^i)\log(1 - h_w(x))$

The logistic function is depicted in the image below.

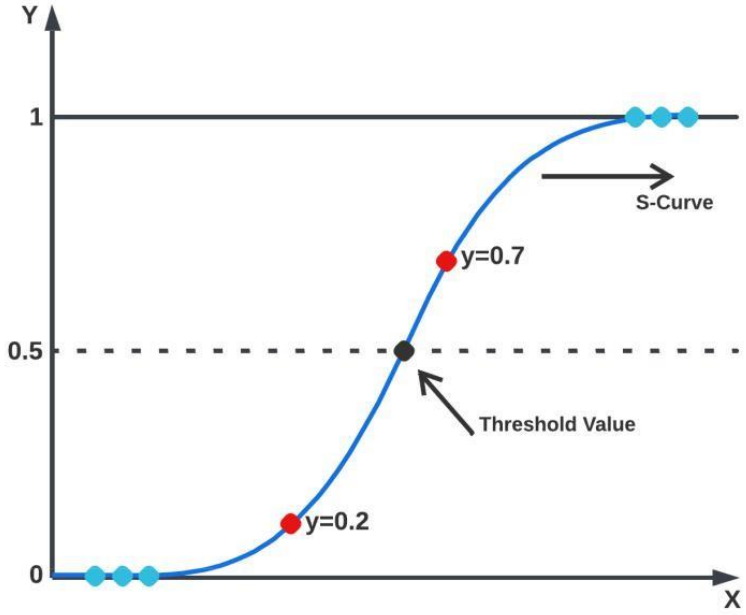

Figure 3.17: Curve of sigmoid function

3.4.2 Decision Tree

The decision tree algorithm excels at prediction and classification tasks. According to Patel & Prajapati, a decision tree can be described as a tree-like structure resembling a flowchart. Each internal node in the tree represents a test on a specific attribute, while each branch represents the potential outcomes of the test. The leaf nodes, also known as terminal nodes, contain the class labels associated with the final predictions made by the decision tree.[24].

Divide the source data into subgroups and evaluate their attribute values to "learn" a tree. Recursive partitioning repeats this procedure for each new subgroup. When all subsets at a node have the same target variable value or splitting does not enhance predictions, the recursion terminates. Decision tree classifiers are ideal for exploratory knowledge discovery since they don't need domain knowledge or parameters. Decision trees analyze high-dimensional data. Decision trees classify well.

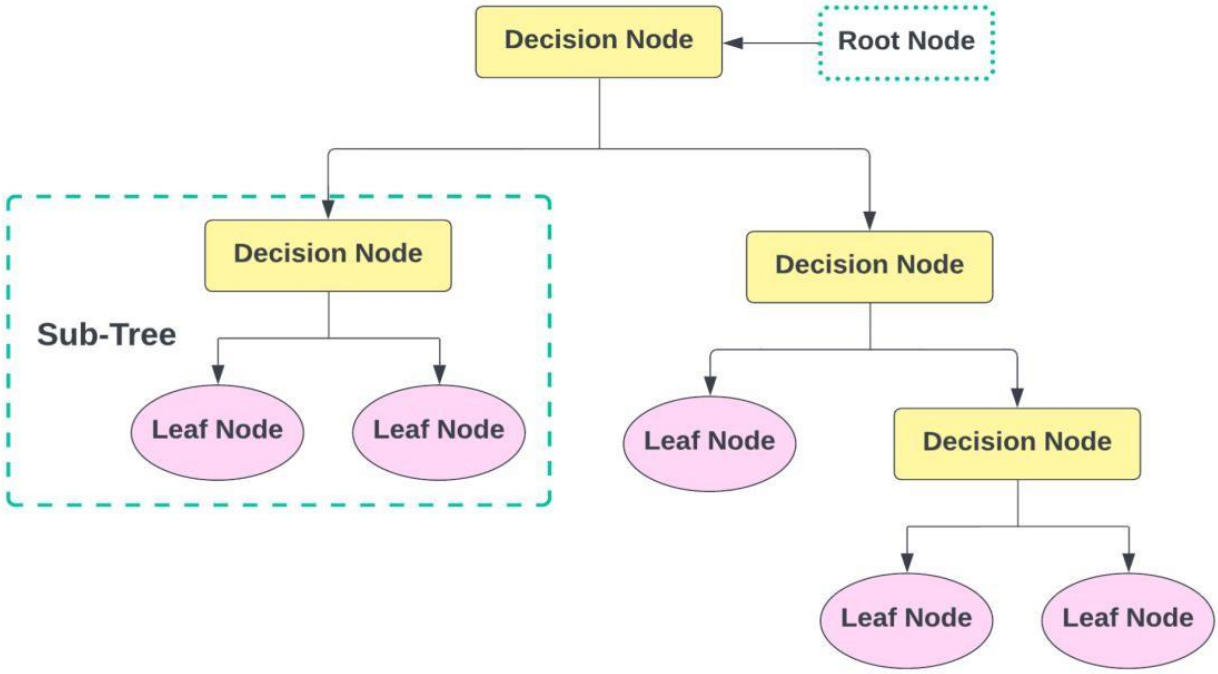

Figure 3.18: Example of decision tree

The most difficult aspect of using a Decision Tree is determining the attribute that should be assigned to the root node of each level. The procedure is called attribute selection. There are two widely used metrics for attribute selection:

- Information Gain
- Gini Index

3.4.2.1 Information Gain

The entropy changes as the training examples are split up into smaller groups using a decision tree node. Entropy change is quantified by information gain.

If we define S as a set of instances, A as an attribute, Sv as the subset of S where A = v, and Values (A) as the set of all possible values of A, we can formulate the following equation.

$$Gain\ (S, A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|Sv|}{|S|} Entropy\ (Sv)$$

$$Entropy(S) = -\sum_{i}^{m} Pi\ log\ Pi$$

Where, P_i probability of the value of any target variable.

3.4.2.2 Gini Index

Using a metrices called the Gini Index, we can find out how often an arbitrarily chosen component is misidentified. This implies that a preference for a characteristic with a lower Gini index is warranted. If the Gini Index is specified, Sklearn will use the "gini" value as its default.

The Gini Index calculation formula is provided below:

Gini Index = $1 - \sum_J x^2$

3.4.3 Random Forest Classifier

By integrating several decision trees with the Bootstrap and Aggregation (or bagging) methodology, Random Forest is an ensemble technique that can carry out both classification and regression. The goal is to combine the results of several different decision trees rather than just using one to arrive at a single conclusion.

The learning models of Random Forest are based on several decision trees. To test the efficacy of each model, we generate sample datasets by selecting rows and attributes at random from the entire dataset. Bootstrap refers to the content that follows. Here are the measures that make up the random forest algorithm:

Step 1: The first thing that Random Forest does is pick n random records out of a dataset of k records.
Step 2: In the subsequent stage, distinct decision trees are constructed for each sample.
Step 3: Each tree in the decision-making process will result in a certain result.
Step 4: For both classification and regression, the final product is evaluated using majority voting and averaging, respectively.

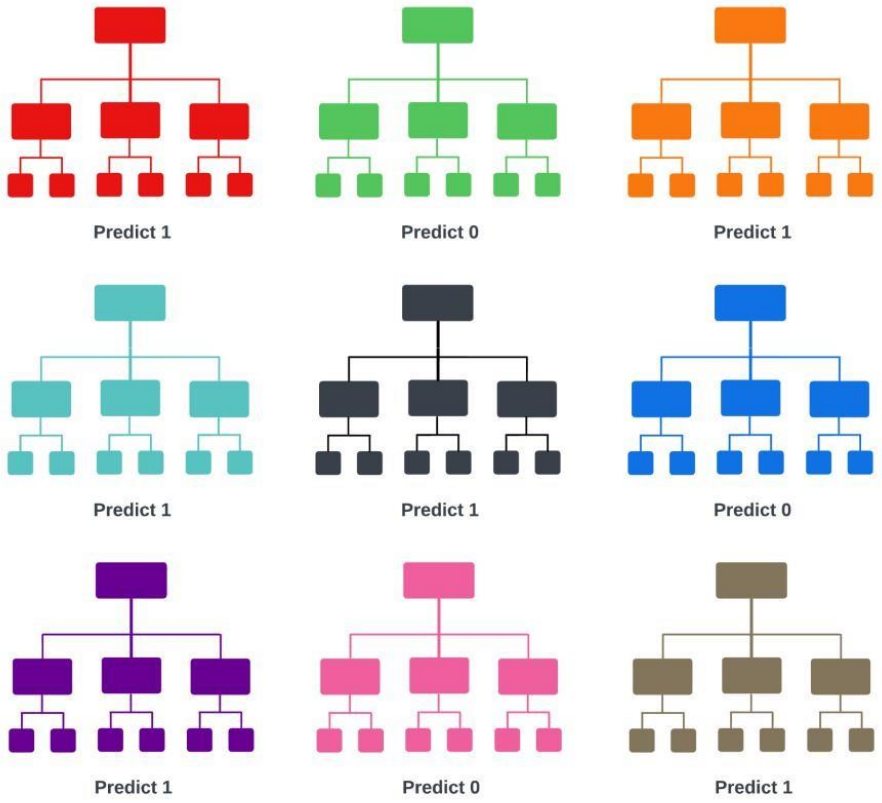

Figure 3.19: Example of random forest classifier

3.4.4 Naive Bayes Classifier

The principles and practical applications of Naive Bayes classifiers are discussed in this article.

Naive Bayes classifiers are a family of methods for categorizing data that are inspired by Bayes' Theorem. In this family of algorithms, each pair of features being categorized is treated as separate from the others.

The following illustration will shed light on the operation of the Naive Bayes' Classifier: Let's pretend we have a dataset of weather observations and a variable we're interested in measuring called "Play."

In light of the aforementioned data set, we must make a call on whether or not to play on a certain day based on the weather forecast. Consequently, the following procedures are required for the resolution of this issue:

- Formulate a set of frequency tables from the provided data.
- Produce a probability table by calculating the probabilities associated with a set of features.

$$P(A\backslash B) = \frac{P(B\backslash A) * P(A)}{P(B)}$$

where,

P(A\B) = Probability of the Hypothesis after it has been determined that the Evidence is True.

P(B\A) = Probability that the Hypothesis is Correct, given the Available Evidence.

P(A) = Probability of the Hypothesis Before It Is Tested.

P(B) = Possibility or likelihood that the evidence is correct.

3.4.5 K-nearest neighbors

K-nearest neighbors (KNN) is a straightforward but efficient technique used in machine learning for both classification and regression problems. It is a non-parametric technique that bases predictions on how closely data points are related.

The KNN algorithm works as follows:

Training Phase: In the training phase, the algorithm straightforwardly retains the feature vectors along with their corresponding labels from the training dataset. This process involves storing the input data and the associated target labels, which serve as the basis for training the machine learning model. No explicit training or model building is performed.

Prediction Phase: KNN takes into account the K nearest neighbors to a new data point based on a distance metric (such Euclidean distance) while formulating predictions for that point. The number of neighbors to take into account is determined by the value of K, a user-defined parameter.

a. For classification tasks, the class label of a new data point is determined using a majority voting mechanism among its K nearest neighbors. The algorithm identifies the K data points that are closest to the new data point based on a chosen distance metric. It then examines the class labels

of these neighbors and assigns the predicted class label to the new data point based on the most frequently occurring class among its K nearest neighbors. This approach leverages the collective decision of the nearby data points to make accurate predictions for the new data point.

b. In regression tasks, the predicted value for a new data point is computed as the average or weighted average of the target values belonging to its K nearest neighbors. The algorithm identifies the K data points that are closest to the new data point based on a specified distance metric. It then calculates the predicted value by taking the average (or weighted average) of the target values associated with these neighboring data points. This approach allows the algorithm to estimate the value for the new data point based on the collective behavior of its closest neighbors.

KNN is a versatile algorithm with various applications, including recommendation systems, image recognition, anomaly detection, and more. Its simplicity, interpretability, and effectiveness make it a popular choice in many machine learning tasks.

3.5 Ensemble Techniques

By integrating numerous models into a single, more accurate one, ensemble methods strive to improve the precision of model findings. By combining many models, we can get much higher precision in our findings. Because of this, ensemble approaches have become increasingly prominent in machine learning.

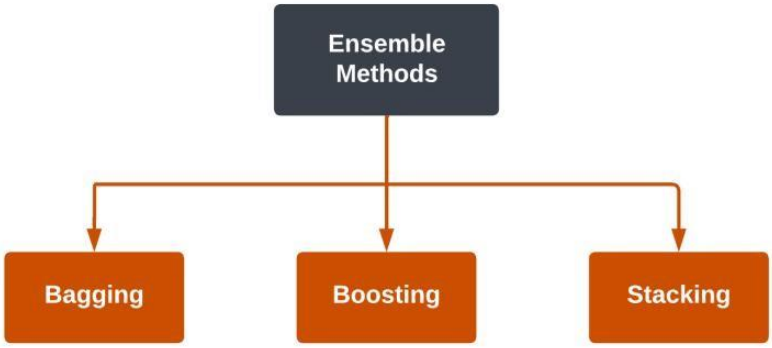

Figure 3.20: Ensemble techniques

There are two main categories for ensemble methods, and they are called sequential and parallel, respectively. In a sequential ensemble method, such as Adaptive Boosting, the base learners are generated first, and then the ensemble members are added (AdaBoost). Consistent performance among the basis learners is helped along by the successive production of new base learners. To further boost the model's performance, we increase the importance we place on the learners who were under-represented in the beginning.

Base learners, like random forest, are created simultaneously in parallel ensemble methods. To foster autonomy among the basis learners, parallel techniques take use of the parallel production of base learners. When using averages, the inaccuracy introduced by the independence of base learners is considerably reduced.

Most ensemble methods only employ a single algorithm in base learning, making all base learners seem the same. Learners who have the same traits across the board are said to have a homogenous foundation. In contrast, some approaches employ diverse base learners to build equally diverse ensembles. Learners with a heterogeneous basis come from a wide variety of backgrounds.

There different types of ensemble methods: Bagging, Boosting, Stacking.

3.5.1 Bagging

Bagging is commonly employed in these statistical procedures. Utilizing decision trees, it dramatically minimizes variation to boost model accuracy. Overfitting is a problem with many predictive models, and fixing it improves accuracy by lowering variability.

Bagging can be either bootstrapping or aggregating. Bootstrapping is a method for drawing representative samples from a larger pool of people, and it relies on a practice known as "replacement" sampling (set). Incorporating replacement samples into a sampling design helps ensure a random selection process. The method is complete after the samples have been run through the foundational learning algorithm.

Bagging makes use of aggregation to combine all forecast outcomes and randomly select one. Without aggregation, predictions will be off since no result will be taken into account. As a result, we utilize either probability bootstrapping methods or the aggregated results from all prediction models to arrive at our totals.

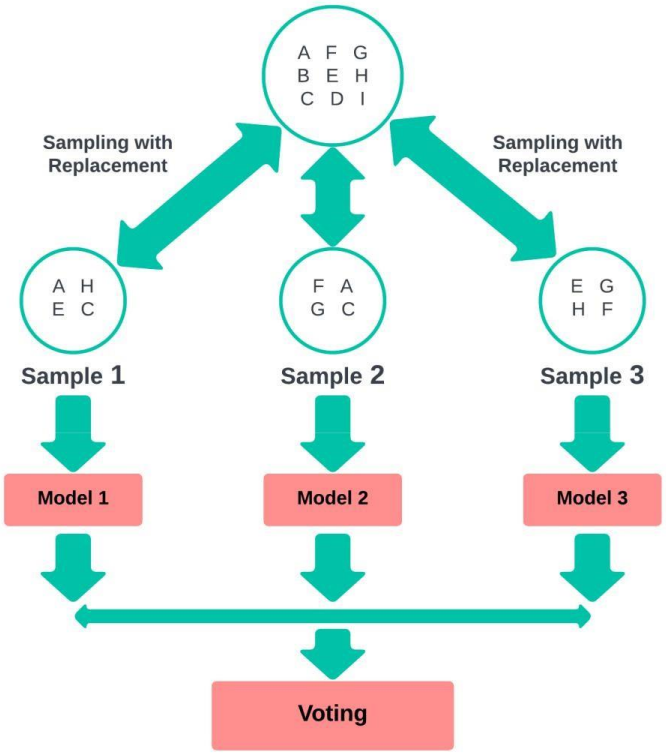

Figure 3.21: Example of bagging

Bagging is helpful because it combines weak base learners into a single strong learner that is more stable than single learners. It also gets rid of any differences, which stops the model from fitting too well. The problem with bagging is that it takes a lot of processing power. When the right way to bag things isn't done, it can make models more biased.

3.5.2 Boosting

Boosting is an ensemble technique employed to enhance predictions by utilizing previous predictor mistakes to improve future predictions. It transforms a collection of weak base learners into a single powerful learner, thereby facilitating easier model prediction. Boosting operates by aligning weak learners in a sequential manner, allowing them to learn from the subsequent learners in the sequence. This iterative process gradually improves the predictive models, leveraging the collective knowledge of the ensemble. By continuously refining the model based on past mistakes, boosting enables the creation of more accurate and robust predictions.

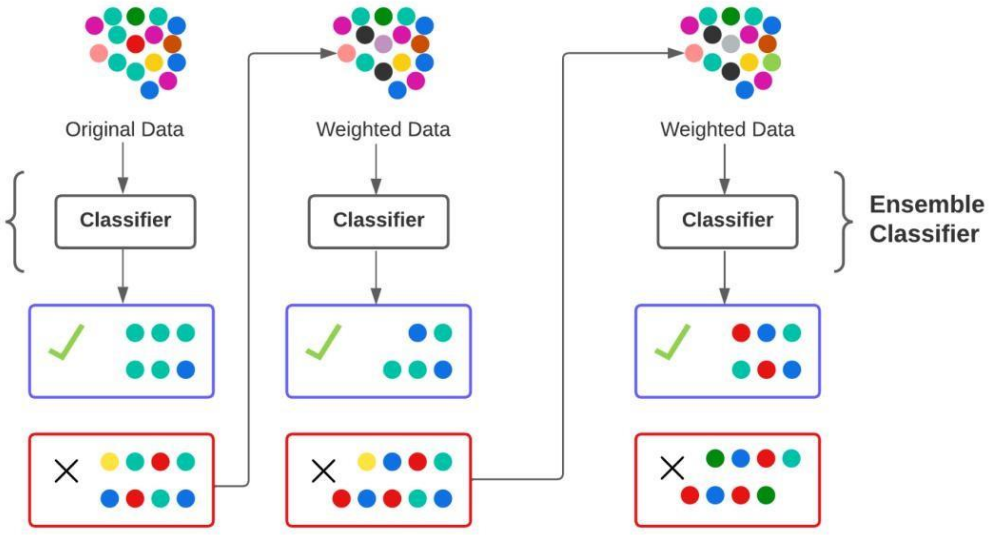

Figure 3.22: Example of boosting

Boosting comprises gradient boosting, AdaBoost, and XGBoost (Extreme Gradient Boosting). Because decision trees usually contain one branch, AdaBoost employs weak learners as decision stumps. AdaBoost's core decision mechanism uses equal-weight data points.

Gradient boosting staggers the addition of new predictors with the hope that the older forecasters may correct the newer ones, enhancing the model's predictive ability. We can repair incorrect predictors by re-fitting. The gradient booster corrects learner predictions by ascending the gradient.

XGBoost employs amplified gradient decision trees. It depends heavily on the model's processing power and speed. Gradient boosted devices need orderly model training, making installation time-consuming.

3.5.3 Stacking

This approach lets a training algorithm use the findings of numerous comparable learning algorithms to make more accurate predictions. Stacking helps in classification, density estimation, distance learning, and regression. It can also calculate bagging errors.

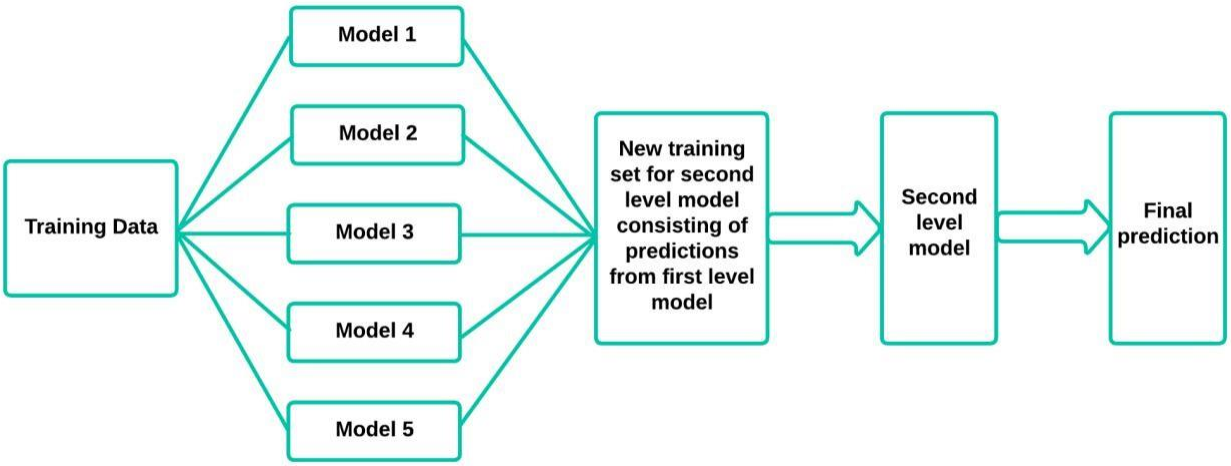

Figure 3.23: Example of stacking

3.6 Generative Adversarial Network

GANs use deep learning approaches like convolutional neural networks for generative modeling. Generative modeling is an unsupervised machine learning activity that automatically discovers and learns patterns in incoming data to produce or output new instances from the dataset [25].

GANs (Generative Adversarial Networks) employ an ingenious approach to train generative models by formulating the task as a supervised learning problem involving two sub-models: the generator model and the discriminator model. The generator model generates novel instances, while the discriminator model classifies samples as either real (belonging to the target domain) or fake (generated by the generator). By pitting these two models against each other in a competitive manner, GANs achieve a dynamic equilibrium where the generator continually improves its ability to generate realistic data, while the discriminator enhances its discrimination skills. This interplay drives the training process, resulting in a generative model that produces high-quality and authentic outputs.

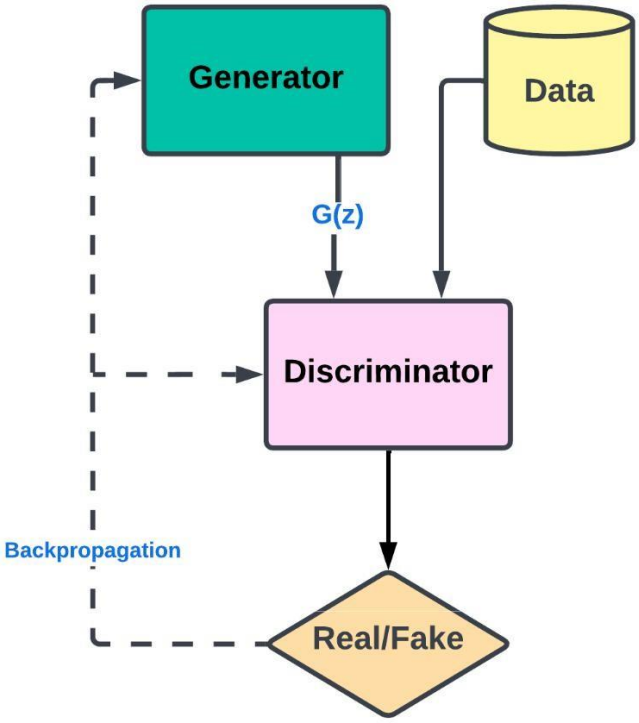

Figure 3.24: Flow-chart of GAN

Generative Adversarial Networks (GANs) [25] are a promising field that produces realistic examples in a variety of problem domains. This is especially true in image-to-image translation tasks like transforming images from summer to winter or day to night, and in Creating photorealistic image of object, scenes, and data that even human can't tell are lie.

Chapter 4

Results and Discussions

We have diligently followed the steps outlined in the approach presented in this section. Subsequent sections will now showcase a comprehensive demonstration of the outcomes achieved by our findings. Within this chapter, we will conduct a thorough evaluation of our model's performance in comparison to alternative models, encompassing the assessment of essential metrics including the confusion matrix, performance matrix, and overall model performance. By adhering to these rigorous evaluations, we aim to provide a robust understanding of the effectiveness and superiority of our findings.

4.1 Confusion Matrix

We calculated the Accuracy, Precision, Recall, and F1 score for each model to assess its performance. To calculate the performance parameters, we have used the confusion matrix. Although the confusion matrix is not a performance matrix, it is necessary to construct the other performance matrix. In table 4.1, we have shown the structure of a confusion matrix.

True Label	Predicted Label		
		0	1
	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

Table 4.1: Confusion matrix

Here, the true positive and true negative are the correctly predicted results, and the false positive and false negative are wrongly predicted.
There are four components in a confusion matrix. Below, we will explain -

True Positive (TP): This refers to the situation where a heart disease is correctly predicted to be present. In other words, the model accurately identifies a person as having a heart disease

True Negative (TN): This refers to the situation where a heart disease is correctly predicted to not be present. In other words, the model correctly identifies a person as being healthy or not having a heart disease.

False positive (FP): This occurs when the model incorrectly predicts the presence of a heart disease when the person is actually healthy. It is a type of error where the model produces a positive prediction (heart disease) when the actual condition is negative (no heart disease).

False Negative (FN): This happens when the model incorrectly predicts the absence of a heart disease when the person actually has the condition. It is a type of error where the model produces a negative prediction (no heart disease) when the actual condition is positive (heart disease).

To summarize:

- True Negative (TN): The model correctly predicts a healthy individual as healthy.
- False Positive (FP): The model predicts a person has a heart disease when they are actually healthy.
- False Negative (FN): The model predicts a person is healthy when they actually have a heart disease.
- True Positive (TP): The model correctly predicts a person with a heart disease as having a heart disease.

The following equations are used to calculate the performance metrics:

Accuracy: This term refers to how frequently the model predicts the correct outcomes. The ratio between the number of accurate predictions and the total number of predictions is known as accuracy. Accuracy measures the probability of a correctly classified into correct class. In a

situation with equal false positives and false negatives, accuracy works well. Equation (i) shows the formula to calculate accuracy.

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN) \quad \text{..... (i)}$$

Precision: The ratio between accurate predictions and overall accurate predictions. It measures the number of positive predictions (true positives) that are correctly made. Precision shows the probability that a ME (true) classified data is how probable its actual label is ME (true). To classify something as positive, the classifier must adhere to stringent criteria, so a high precision value indicates that there were few false positives. In our case high precision should not be the priority because number of FP is not our concern. Equation (ii) shows the formula to calculate the precision.

$$\text{Precision} = TP / (TP + FP) \quad \text{..... (ii)}$$

Recall: Recall is the percentage of total positive predictions that our model accurately predicted. Out of all positive cases in the data, it calculates the number of positive cases the classifier correctly predicted. Equation (iii) shows the formula to calculate recall.

$$\text{Recall} = TP / (TP + FN) \quad \text{.....(iii)}$$

F1- score: The weighted probability of sensitivity and specificity is known as the F1-score. By using the F1-score, we may simultaneously assess recall and precision. When recall and precision are equal, the f1-score is high. F1 is considered more useful than accuracy in case of uneven or imbalanced classes. Equation (v) shows the formula to calculate the f1-score.

$$\text{F1-score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad \text{..... (iv)}$$

In our particular context, it is crucial to minimize the occurrence of false negatives (FN). While the occurrence of false positives (FP) is of lesser concern, for instance, if the model predicts a person has a heart disease when they are actually healthy, that will not have as much of an impact as if the model predicts a person is healthy when they actually have a heart disease.

4.2 Dataset Distribution

The model is trained, validated and tested using the publicly available dataset. The dataset is from the Kaggle site, which is originally, came from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS) were used in our study. The dataset contains information of 319795 individuals. There are several segments in the survey. The dataset contains 18 feature variables (9 Booleans, 5 strings and 4 decimals). In machine learning projects, "HeartDisease" can be used as the explanatory variable, but note that the classes are heavily unbalanced.

Total Number of Instances	319795
Healthy	292422
Heart Disease	27373

Table 4.2: Information of Real Data

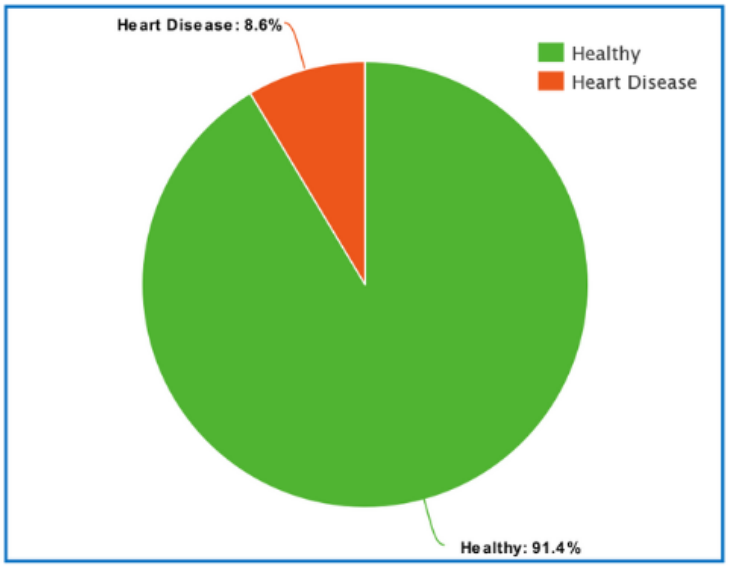

Figure 4.1: The Information of Real Data

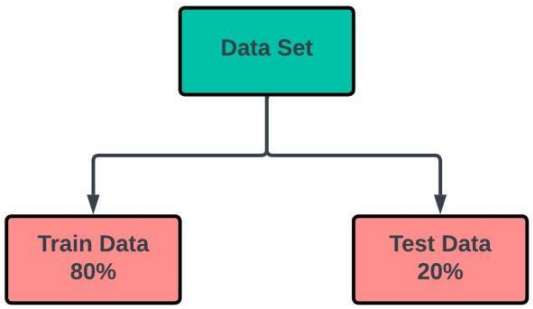

Figure 4.2: Dataset splitting paradigms

Training and testing datasets are created from the original dataset. The training dataset is then split into the training dataset and the validation dataset. The training and validation datasets are divided in a ratio of 80% to 20%. After Creating the Artificial Synthetic data, we have 584844 data. That means for training we have 467875 data and for testing we have 116969 data.

Total Number of Instances	584844
Training Data	467875
Testing Data	116969

Table 4.3: Information of Splitting Data

4.3 Modeling

We used five different classification algorithms to compare the result with Ensemble model (Gradient Boosting). The algorithms are Logistic Regression, Naive Bayes Classifier, k-nearest neighbors, Decision Tree, and Random Forest

4.3.1 Confusion Matrix

We calculate the confusion matrix of each model

	Logistic Regression	Decision Tree	Naïve Beys	Random Forest	KNN	Gradient Boosting
TP	43699	51611	38772	51350	53685	53732
FP	16256	8553	15583	8866	14776	5053
FN	14809	6897	19736	7158	5214	4776
TN	42205	49908	42878	49595	43685	53408

Table 4.4: Confusion matrix for each model

From the bellow Plot representation enables us to see that our Gradient Boosting has the lowest FN, which is what we need.

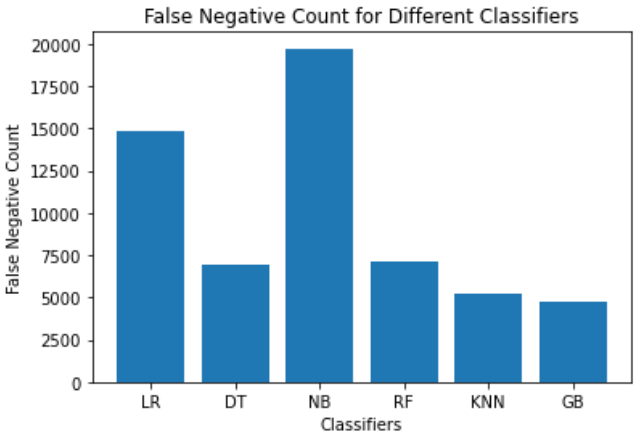

Figure 4.3: Comparison of False Negative data

4.3.2 Performance Analysis of Each Model

The performance of each model has been illustrated in this section. The machine learning models we have used are Logistic Regression, Naive Bayes Classifier, k-nearest neighbors, Decision Tree, Random Forest and Gradient Boosting model.

	Logistic Regression	Decision Tree	Naïve Beys	Random Forest	KNN	Gradient Boosting
Accuracy	73.445%	86.829%	69.803%	86.323%	82.910%	91.596%
Precision	72.887%	85.765%	71.341%	85.265%	78.292%	91.042%
Recall	74.697%	88.329%	66.240%	87.837%	91.088%	91.837%
F1-score	73.781%	87.029%	68.696%	86.532%	84.207%	91.620%

Table 4.5: Performance metrics of each model for data

In our pursuit of discovering the most effective model for heart disease detection, we meticulously explored numerous combinations of machine learning algorithms. In Section 4.3, we conducted a thorough evaluation of each combination, assessing their performance using precision, recall, and f1-score metrics.

Remarkably, in our pursuit of finding the optimal algorithm, the gradient boosting algorithm emerged as the top performer for heart disease detection. It outperformed the other five algorithms in terms of heart disease detection. It exhibited an impressive accuracy of 91.596%, a precision of 91.042%, a recall of 91.837%, and an f1-score of 91.620%.

The achieved accuracy of 91.596% underscores the outstanding capability of the gradient boosting algorithm in accurately predicting instances of heart disease, demonstrating its efficacy in distinguishing between those with and without the condition.

The precision metric, which gauges the proportion of correctly predicted heart disease cases out of all the predicted cases, achieved a remarkable value of 91.042%. This signifies that when the model predicted an instance of heart disease, it was correct 91.042% of the time. Such high precision implies a low false-positive rate, minimizing the chances of misclassifying healthy individuals as having heart disease.

The recall value of 91.837% showcases the algorithm's ability to correctly identify heart disease cases from the entire set of actual cases within the dataset. This remarkable recall value indicates a low false-negative rate, suggesting that the algorithm successfully captured a substantial proportion of the instances representing heart disease.

The f1-score, serving as the harmonic mean of precision and recall, provides a balanced measure of the algorithm's overall performance. With an impressive f1-score of 91.620%, the gradient boosting algorithm demonstrates a harmonious equilibrium between precision and recall, underscoring its efficacy in heart disease detection.

The results obtained from Section 4.3 serve as substantial validation of the algorithm's capability to accurately identify cases of heart disease while simultaneously minimizing the likelihood of misclassifying healthy individuals. The exceptional precision and recall values further reinforce the algorithm's ability to strike an optimal balance, avoiding false positives while capturing a significant portion of the heart disease instances.

Chapter 5

Limitation and Future Work

This chapter will examine the shortcomings of our current strategy and prospective directions for further investigation and development. By analyzing the obstacles and suggesting potential solutions.

Although our currently best finding strategy yields an accuracy of 91.596%, there is still room for improvement. Future research could focus on developing methods to enhance the accuracy and precision of our model, potentially through the integration of more advanced machine learning techniques or the exploration of statistical methodologies.

It is important to acknowledge that our study primarily focuses on classification algorithms, and there is potential for expanding the research to include regression methods. Incorporating regression models can provide a deeper understanding of the relationship between variables and allow for more nuanced predictions and analyses.

While our model demonstrates the ability to predict the presence of heart disease, it currently lacks the capability to predict multimodal heart diseases. Future work could concentrate on developing novel approaches or incorporating multimodal data sources to enable accurate prediction and classification of such complex heart diseases.

It is essential to recognize that our study's data collection process may have limitations due to the limited variables considered. To address this, future surveys should encompass a broader range of variables, capturing a more comprehensive set of features. This expanded dataset can contribute to a more robust and holistic understanding of the problem domain.

Furthermore, the current study primarily focuses on machine learning strategies, and there is an opportunity to leverage additional data science and statistical methodologies. Exploring alternative approaches, such as ensemble methods or deep learning architectures, may offer valuable insights and potentially enhance the overall performance of the model.

By acknowledging these limitations and outlining potential avenues for future research, we aim to encourage further exploration and advancements in the field, ultimately leading to improved accuracy, expanded capabilities, and a deeper understanding of the subject matter.

Chapter 6

Conclusion

This chapter discusses our research's findings. Furthermore, it conveys the completion of our findings and summarizing the key findings and insights gathered throughout the study. It will provide a comprehensive overview of our thesis project.

In conclusion, the early diagnosis of heart disease presents significant challenges. Leveraging the power of machine learning, we embarked on a comprehensive exploration of this problem using a vast survey dataset. We pre-processed the data obtained from the BRFSS survey. Our approach involved implementing various machine learning algorithms, including Naive Bayes, Random Forest Classifier, Logistic Regression, Decision Tree, K-NN Classifier, and Gradient Boosting. Among these methods, the gradient boosting algorithm exhibited the most promising outcomes.

However, despite our concerted efforts, we encountered a major hurdle—the limited correlation between the features in our dataset and the target variable. This intrinsic characteristic hindered the performance of our models, resulting in outcomes that fell short of our initial expectations. To address this limitation, we embarked on an innovative approach. We decided to augment our dataset by generating synthetic tabular data. By creating an additional 265,049 data instances for the minority class and merging them with the original dataset, we expanded the overall dataset size to 584,844 cases. With this enriched dataset, all five fundamental classification methods and gradient boosting demonstrated noticeable improvements in performance.

Furthermore, by leveraging the additional data, we achieved enhanced accuracy through the utilization of ensemble models. In addition to the insights gained from our research, we also identified specific attributes that hold the potential to significantly enhance machine learning model training and overall performance. We recommend including these attributes in future surveys, as gathering this additional information from individuals can lead to more robust and accurate predictions.

While our study encountered challenges in early heart disease diagnosis, our utilization of machine learning techniques, synthetic data augmentation, and ensemble models showcased the potential for improvement. By incorporating the suggested attributes and further exploring the complex interplay between features and the target variable, future research has the opportunity to make significant strides in advancing the accuracy and effectiveness of machine learning models for heart disease diagnosis.

References

[1] cardiovascular diseases (Cvds), “World health organization,”
https://www.who.int/health-topics/cardiovascular-diseases

[2] L. A. Allen, L. W. Stevenson, K. L. Grady et al., “Decision making in advanced heart failure: a scientific statement from the American heart association,” Circulation, vol. 125, no. 15, pp. 1928–1952, 2012.

[3] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, “Innovative artificial neural networks-based decision support system for heart diseases diagnosis,” Journal of Intelligent Learning Systems and Applications, vol. 5, no. 3, Article ID 35396, 2013.

[4] Q. K. Al-Shayea, “Artificial neural networks in medical diagnosis,”
Int. J. Comput. Sci. Issues, vol. 8, no. 2, pp. 150–154, 2011.

[5] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig,
“Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson’s disease symptom severity,” Journal of 'e Royal
Society Interface, vol. 8, no. 59, pp. 842–855, 2011.

[6]. Shamshirband, S.; Fathi, M.; Dehzangi, A.; Chronopoulos, A.T.; Alinejad-Rokny, H. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. J. Biomed. Inform. 2021, 113, 103627. [CrossRef] [PubMed]

[7]. Chen, P.-T.; Lin, C.-L.;Wu,W.-N. Big data management in healthcare: Adoption challenges and implications. Int. J. Inf. Manag. 2020, 53, 102078. [CrossRef]

[8] T. J. van Trier, N. Mohammadnia, M. Snaterse, R. J. G. Peters, H. T. Jørstad “Lifestyle management to prevent atherosclerotic cardiovascular disease: evidence and challenges”
https://link.springer.com/article/10.1007/s12471-021-01642-y

[9] Rohit Bharti, Aditya Khamparia, Mohammad Shabaz et al “Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning”
Computational Intelligence and Neuroscience Volume 2021, Article ID 8387680, 11 pages
https://doi.org/10.1155/2021/8387680

[10] Devansh Shah, Samir Patel, Santosh Kumar Bharti et al. “Heart Disease Prediction using Machine Learning Techniques” SN Computer Science (2020) 1:345
https://doi.org/10.1007/s42979-020-00365-y

[11] Aditi Gavhane, Gouthami Kokkula, Isha Panday, Prof. Kailash Devadkar, “Prediction of Heart Disease using Machine Learning”,Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology(ICECA), 2018.

[12] Senthil kumar mohan, chandrasegar thirumalai and Gautam Srivastva,“Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” IEEE Access 2019.

[13] Amandeep Kaur and Jyoti Arora,“Heart Diseases Prediction using Data Mining Techniques: A survey” International Journal of Advanced Research in Computer Science , IJARCS 2019.

[14] Pahulpreet Singh Kohli and Shriya Arora, “Application of Machine Learning in Diseases Prediction”, 4th International Conference on Computing Communication and Automation (ICCCA), 2018
.
[15] Patil, Shantakumar B., and Y. S. Kumaraswamy. "Extraction of significant patterns from heart disease warehouses for heart attack prediction." IJCSNS 9.2: 228-235.

[16] Soni, Jyoti, et al. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17.8: 43-48

[17] Takci, Hidayet. "Improvement of heart attack prediction by the feature selection methods." Turkish Journal of Electrical Engineering & Computer Sciences 26.1 (2018): 1-10.

[18] Fizar Ahmed,”An Internet of Things (IoT) Application for Predicting the Quantity of Future Heart Attack Patients “ International Journal of Computer Applications (0975 8887) Volume 164 No 6, April 2017.

[19] Singh, Poornima, Sanjay Singh, and Gayatri S. Pandi-Jain. "Effective heart disease prediction system using data mining techniques." International journal of nanomedicine 13.T-NANO 2014 Abstracts (2018): 121.

[20] https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-diseases

[21] Narkhede, Sarang. n.d. “Understanding AUC - ROC Curve | by Sarang Narkhede.”Towards Data Science. Accessed December 17, 2022. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

[22] “Numeracy, Maths and Statistics - Academic Skills Kit.” n.d. Numeracy, Maths and Statistics - Academic Skills Kit. Accessed December 17, 2022. https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regr ession-and-correlation/types-of-correlation.html.

[23] Peng, Joanne, Kuk L. Lee, and Gary M. Ingersoll. 2002. “An Introduction to Logistic Regression Analysis and Reporting.” The Journal of Educational Research 96 (1): 3-14.

[24] Patel, Harsh H., and Purvi Prajapati. 2018. “study and Analysis of Decision Tree Based Classification Algorithms.” International Journal of Computer Sciences and Engineering 6, no. 10 (October): 74-78.

[25] “tgan · PyPI.” 2019. PyPI. https://pypi.org/project/tgan/