PERSONAL KEY INDICATORS OF HEART DISEASE USING MACHINE LEARNING TECHNIQUES

BY

Shah Newaj Khan
ID: 181-15-1985

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md Assaduzzaman
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Mohammad Jahangir Alam
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

FEBRUARY 2023

APPROVAL

This Project titled **“Personal Key Indicators of Heart Disease Using Machine Learning Techniques”**, submitted by **Shah Newaj Khan**, ID No: **181-15-1985** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **5th February 2023**.

BOARD OF EXAMINERS

7/2/23

Chairman

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Nazmun Nessa Moon
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Raja Tariqul Hasan Tusher
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Ahmed Wasif Reza
Professor
Department of Computer Science and Engineering
East West University

DECLARATION

I hereby declare that, this project has been done by us under the supervision of **Md Assaduzzaman, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

04.02.23

Md Assaduzzaman
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

04.02.23

Mohammad Jahangir Alam
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:

Shahnewaj
04.02.23

Shah Newaj Khan
ID: -181-15-1985
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project successfully.

I am really grateful and wish my profound our indebtedness to **Md Assaduzzaman, Lecturer,** Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of “*Machine Learning*” to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to my Parents, my Family, and the Head of**,** the CSE Department “**Professor Dr. Touhid Bhuiyan”**, for his kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

The number one killer in the world is heart disease, which claims millions of lives each year. Coronary artery disease, often known as CAD, is one of the numerous conditions that may have an adverse impact on the heart. It is one of the leading causes of death among cardiovascular diseases. The prediction of cardiovascular illness is a serious challenge for the industry of clinical data analysis. Studies have shown that machine learning (ML) may assist with decision making and prediction based on data generated by industries such as healthcare, and these industries create a substantial amount of data. In the present body of research, the use of ML techniques to predict heart illness receives only a limited amount of attention. The goal of this study was to investigate which kinds of machine learning classifiers are the most reliable when it comes to achieving such high levels of diagnostic precision. Evaluation and comparison of the effectiveness and efficiency of a number of supervised machine learning algorithms for the prediction of cardiovascular disease were carried out. After utilizing a number of distinct machine learning approaches, such as the Decision tree (DT), Stochastic Gradient Descent (SGD), Random Forest (RF), Adaboost (Ada), and XGB, as well as logistic regression, I discovered that the accuracy of the Stochastic Gradient Descent (SGD) algorithm is 94.66%, making it the most accurate of all the algorithms. And AdaBoost (Ada) is the model with the lowest accuracy, with 93.07%.

TABLE OF CONTENTS

CONTENTS **PAGE**

Board of examiners i

Declaration ii

Acknowledgements iii

Abstract iv

CHAPTER

CHAPTER 1: Introduction **1-3**

1.1 Introduction 1

1.2 Motivation 2

1.3 Rationale of the Study 2

1.4 Research Questions 3

1.5 Expected Output 3

CHAPTER 2: Background **4-9**

2.1 Related Works 4

2.2 Comparative Analysis and Summary 8

2.3 Scope of the Problem 9

2.4 Challenges 9

CHAPTER 3: Research Methodology **10-16**

3.1 Research Subject and Instrumentation 10

3.2 Data Collection Procedure 10

3.3 Applied Methodology 10

3.4 Implementation Requirements 16

CHAPTER 4: Experimental Results and Discussion **17-21**

4.1 Experimental Setup 17

4.2 Experimental Results & Analysis 17

4.3 Discussion 21

CHAPTER 5: Impact on Society and Sustainability **22-23**

5.1 Impact on Society 22

5.2 Impact on Environment 22

5.3 Ethical Aspects 23

5.4 Sustainability Plan 23

CHAPTER 6: Summary, Conclusion and Implication for Future Research **24**

6.1 Summary of the Study 24

6.2 Conclusions 24

6.3 Implication for Further Study 24

REFERENCES **25-27**

LIST OF FIGURES

FIGURES	PAGE NO
FIGURE 3.1: Proposed Methodology	11
FIGURE 3.2: Checking Outliers	11
FIGURE 3.3: Data Imbalance Problem	12
FIGURE 3.4: Oversampling Using SMOTE	12
FIGURE 4.1: Confusion Matrix Generated by Decision Tree (DT)	18
FIGURE 4.2: Confusion Matrix Generated by Stochastic Gradient Descent (SGD)	18
FIGURE 4.3: Confusion Matrix Generated by Random Forest (RF)	19
FIGURE 4.4: Confusion Matrix Generated by AdaBoost (Ada)	19
FIGURE 4.5: Confusion Matrix Generated by XGB	20
FIGURE 4.6: Confusion Matrix Generated by Logistic	20

LIST OF TABLES

TABLES	PAGE NO
Table 3.5: The requirement of Hardware for this project	16
Table 3.6: The Requirement of software for This Project	16
Table 4.7: Performance Evaluation Metrics for All Applied Models	21

CHAPTER 1

Introduction

1.1 Introduction

Heart disease research is a relatively recent area of medical study. Before the early 1900s, heart illness actually occurred rarely. Infectious diseases have historically plagued humanity more than non-infectious ailments like heart disease [1]. Globally, cardiovascular diseases (CVD) account for an estimated 17.9 million deaths annually (according to the World Health Organization's projections for 2021) [2]. It is generally agreed that the major risk factors for cardiovascular disease are things like age, excessive alcohol consumption, an unhealthy diet, smoking, and inactivity [3] and that prolonged exposure to these things increases the likelihood of developing high blood pressure, diabetes, dyslipidemia, obesity, and stroke [4]. Several different risk factors make diagnosing heart disease challenging [5]. Human heart disease severity has been studied using a variety of data mining and machine learning techniques. Several different algorithms, such as K-Nearest Neighbor (KNN), Decision Trees (DT), Genetic Algorithm (GA), and Naive Bayes (NB), are used to categorize illness severity [6]. Using data mining and machine learning algorithms is the process known as "data mining.", statistics, and database systems to uncover previously unknown patterns (knowledge) in massive, existing data sets. The newly uncovered information can be utilized to construct intelligent predictive decision systems in a variety of industries, including healthcare, to facilitate timely, cost-effective diagnosis and treatment. Machine learning allows computers to learn without direct human involvement, using previously collected data to enhance performance based on experience [7].Heart disease diagnosis takes into account a wide range of potential causes. Hospitals are not the best places for people to keep track of their lives or get medical diagnoses. However, with the help of an automated service, everyone may monitor their health and seek professional treatment immediately if they see anything out of the ordinary. Everyone's everyday routines, from patients to doctors, will improve with this. In addition, we can better understand the root causes of cardiovascular illness and give more precise instructions on how to maintain a healthy heart by employing ML algorithms.

1.2 Motivation

Due to its internal nature, heart disease is often overlooked despite the fact that it is among world's leading causes of death. If a person's heart health has not been monitored, they could experience heart failure at any time. Humans can approach their death cycle limit in a relatively short amount of time. According to estimates, over 21 percent of all deaths in Bangladesh had been caused by heart attacks in 2020, totaling about 1,804,08. I've seen that many people close to us suffer from heart-related issues, but they can't seem to pinpoint the root cause of those issues. As an added downside, national incomes have been declining. Due to the prohibitive expense of diagnosis, people in this situation cannot obtain help for heart-related illnesses. Cardiovascular disease diagnosis is the medical industry's most challenging duty, and as a result, it has the highest expense. There is a vast region that can be used to make early predictions about heart disease because there are so many potential causes. Informed citizens can take preventative measures sooner if they receive early warnings, which has the potential to prevent many deaths. This is why I plan to investigate the database of cardiovascular disease-related syndrome information for the purpose of make a precise prediction of cardiovascular disease. The goal is to make medical care accessible and convenient for everyone who needs it.

1.3 Rationale of the Study

- Build a repository of cardiac-related information by collecting and organizing a massive amount of raw data.
- Examine what causes heart difficulties.
- Making early detection of heart disease more affordable.
- Helps evaluate cardiac health in underserved areas.
- Allow the patient to understand their disease so they can receive the appropriate treatment.

1.4 Research Questions

- When it comes to heart disease, what are the most important factors to consider?
- Does a more precise supervised algorithm exist?
- Can the current methods of cardiac detection be improved upon?
- When it comes to predicting cardiac issues, what method do you think will work best?

1.5 Expected Output

- Boost the accuracy of early the identification of heart illness.
- Properly identifying the severity of heart disease.
- Reduce the premature deaths caused by heart disease.
- Ordinary people are able to time their actions effectively.
- Raise public awareness of heart disease following prediction.

CHAPTER 2

Background

2.1 Related Works

Several researchers, all hoping to employ machine learning to forecast cardiac issues, had presented different decision support systems. Different researchers have developed various decision support systems for cardiac disease prediction, and these are discussed here. The authors, M. M. Ali et al. [8] stated that the purpose of this research was to determine the most accurate machine learning classifiers for cardiovascular disease (CVD). Kaggle was mined for its dataset. There are a total of 1025 patient records in the dataset, split roughly in half between males and females, with 499 classified as normal and 526 as having cardiac disease. By applying six different classification algorithms to the dataset and evaluating their accuracy and other statistical characteristics with 10-fold cross validation, they were able to determine the top performing approach. Multilayer perceptron (MP), K-nearest neighbor (KNN), random forest (RF), decision tree (DT), logistic regression (LR), and AdaboostM1 (ABM1) were some of the techniques they used. All of the predictions made by KNN, RF, and DT were right 100% of the time. This shows that they are the best at predicting cardiac diseases.

To better predict cardiovascular illness, S. Mohan et al. [9] suggested a unique strategy that employs machine learning techniques to identify meaningful information. A variety of feature combinations and well-established categorization methods are added as part of the prediction model. After compiling a wide range of sources, data on cardiovascular disease undergoes preliminary processing. Preprocessing findings for 297 patient records show that 137 of those records have a value of 1, indicating the presence of heart disease, while the remaining 160 reflect a value of 0. Different criteria are used to categorize the severity of the condition. In comparison to other methods, the accuracy of the hybrid random forest with a linear model classification approach (HRFLM) is the highest. The suggested hybrid HRFLM method combines the best features of the Random Forest (RF) and the Linear Method (LM). Using HRFLM to foresee cardiac issues has proven to be extremely reliable.

P. Rani et al. [10] authors of this research, suggest a hybrid decision-support system for the early diagnosis of cardiac illness using patient-specific clinical characteristics. Experiments were run using the Cleveland Heart Disease dataset retrieved from the UCI repository. Eight of the 14 features in this dataset are categorized, while the remaining six are numerical. They suggested a hybrid decision-support system for cardiac illness forecasting. The three phases that make up this hybrid system are data gathering, data pre-processing, and model building. The pre-processing phase involves a number of operations, including missing-value imputation, feature selection, feature scaling, and class balance. The MICE (multivariate imputation through chained equations) approach was utilized to fill in the blanks for the missing data. Then, a combination of a genetic algorithm and a recursive feature elimination method is used to pick features. SMOTE is used for class balancing. The suggested hybrid system for predicting the risk of heart disease has reached a level of accuracy that is better than some of the other systems that are already out there.

According to U. Nagavelli and coworkers. This publication's [11] principal purpose is to serve as a reference for medical professionals to use in the detection of heart problems at an early stage. As a first step, we use Naive Bayes with a weighted technique to calculate the likelihood that a given individual would develop cardiovascular disease. Second, we have an automated analysis of diagnosing ischemic heart disease that considers frequency domain, time domain, and information theoretic factors. By using this method, we may choose the two best-performing classifiers, such as a support vector machine (SVM) with XGBoost, to do the classification. The third automated method for diagnosing heart failure is a strengthened support vector machine (SVM) that employs a dual optimization technique. A clinical decision support system (CDSS) employs a sophisticated heart disease prediction model (HDPM) that combines DBSCAN for outlier identification and removal, SMOTE-ENN for balancing training data distribution, and XGBoost for heart disease prediction. To identify the most important non-invasive risk factors for cardiovascular disease.

S. I. Ansarullah [12] et al. use a number of feature selection methodologies and classification algorithms. Information about heart disease covers both qualitative and quantitative measurements of vulnerability. When filling in missing numbers, the mean imputation method is employed, whereas when filling in blank categories, the mode imputation method is used. The heart disease dataset is mined using random forest, decision tree, support vector machine, K nearest neighbor, and Naive Bayes, all tested with 10-fold cross-validation. The best outcomes are achieved by calculating a wide range of performance metrics from the medical and modeling fields, including sensitivity, specificity, accuracy, precision, AUROC score, misclassification rates, computational complexity, and interpretability. The random forest model performs best for predicting cardiac disease.

Three different categorization methods were used by H.M. Le et al. (2018) on the 58 attributes taken from the UCI Machine Learning Repository dataset [13]. They demonstrated that a support vector machine (SVM) with a linear kernel performed exceptionally well, with an accuracy of 89.93 percent.

C. A. Cheng and H. W. Chiu discovered [14] patterns with NN, DT, Support Vector Machines (SVM), and Naive Bayes on data from UCI's laboratory patients with heart illness. The efficiency and precision of these algorithms are evaluated in comparison to the results. When compared to previous approaches, the suggested hybrid approach achieves results of 86% for the F-measure.

In order to predict cardiac failure, Ali et al. [15] created a method based on two SVM (support vector machine) models. Both feature selection and prediction were accomplished by means of separate models. They used 70% of the data for training and 30% for testing. They used an L1 regularized linear SVM model to choose the features of interest. The prediction model was an SVM with an L2-regularized RBF kernel. The SVM's hyperparameters were tweaked to get optimal performance for both models.

Patients with hypertension are at increased risk for cardiac events, and Paragliola and Coronato [16] proposed a model to assess this threat. As input for their hybrid model, the authors offered ECG signals processed by a long short-term memory network and a convolutional neural network. Time-series signals were used by the system to anticipate a rise in hypertension.

F. J. Abdeldjouad et al. [17] observed that data mining techniques could aid doctors in making CVD diagnoses. A new hybrid methodology for cardiac disease prediction has been presented that integrates all existing methods into a single algorithm. Using WEKA and KEEL, they evaluate the performance of several classification algorithms according to sensitivity, accuracy, specificity, and error rate.

To make accurate predictions of cardiovascular disease, Martins et al. [18] used a combination of the Bayesian optimization XG boost classifier and the one-hot encoding method. The Cleveland heart disease dataset is used to gauge the model's efficacy, with comparisons made to other models in the field.

The use of ensemble classification algorithms for better cardiovascular risk prediction. [19] was conducted by C. Beulah Christalin Latha and S. Carolin Jeeva, who accessed their dataset through the UCI repository. They have employed ensemble approaches to improve classifier accuracy after first working with base classifiers like C4.5, RF, and Naive Bayes. On top of that, the outcomes varied depending on the features that were prioritized. Using feature selection and boosting, they were able to get an accuracy of about 84.85%.

When enormous volumes of data from various sources are analyzed, a process known as "data mining" is triggered. Heart disease can be diagnosed with this procedure. Risk factors allow for a rapid diagnosis of heart disease. Classification methods for cardiac diagnosis are the primary focus of this study [20].

According to the research published by Motarwar et al. [21] data visualization was utilized to model correlation or dependence in their dataset. Features with the most potential were selected from their dataset utilizing feature selection. This technique offers the most reliable information on which to run classification algorithms. The simple accuracy of each algorithm is improved by the application of additional augmentation methods. The dataset was trained with 242 samples, accounting for 80% of the total. In the remaining 20% of cases, we should see a total of 61 occurrences. ratio of how much each method enhances performance. Their precision is as shown below: SVM (90.16%), RF (95.08%), HT (81.24%), and Logistic Model Tree (80.69%) all outperformed the Gaussian NB (93.44%) and the other two trees (90.16%).

In their thesis, Rahma Atallah and Amjed Al-Mousa [22] report that their proposed model achieves an accuracy rate of almost 90%, making it more successful than any of the other algorithms they tested. In the first experiment, the classifier's default settings had an 80% success rate. After performing a GridsearchCV, they discovered the cross-validation-based parameters that increased accuracy to 88%. Also, the most current model is the logistic regression classifier. GridsearchCV's precision did not shift since its default settings were identical to the initial settings.

2.2 Comparative Analysis and Summary

70% of the population of Bangladesh has some form of cardiac disease. And they care even less about their physical wellbeing than that. As a result of their lack of awareness regarding their health, people are experiencing a wide range of symptoms. Because of their chronically poor diets, most people have elevated blood sugar levels, high cholesterol, and other health issues that put them at a higher risk for developing cardiovascular disease. This research, prediction, and analysis can help people avoid becoming heart disease patients by making them more aware of the risks posed by their food choices and the importance of regular health checks. There is nothing I can do to prevent heart disease through exercise or medication, but we can help raise awareness and encourage individuals to make positive lifestyle changes. There have been many studies conducted on the subject of heart disease prediction; however, these studies have failed to find a statistically significant correlation between risk factors for cardiovascular disease and mortality. In contrast, my study uses data on smoking and drinking habits, glucose tolerance, and other factors to establish a gender-specific risk index for cardiovascular disease. I have improved our ability to estimate the risk of heart disease by analyzing data from a number of different categories.

2.3 Scope of the Problem

In the process of completing this task, there were really few roadblocks that I ran through.

- Select the appropriate data set.
- Identifying the most effective library for cleaning data.
- Choose the most effective classifier.
- Correctly train my model.

2.4 Challenges

- To gather information.
- Reliability of model.
- My massive data set required complex preprocessing.
- Training takes a long time because of the massive amount of data.
- Superiority of my model.

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

So far as this instance is concerned, I've been applying machine learning to the problem at hand. There are various distinct approaches to machine learning, including DT, SGD, RF, Ada, XGB, and Logistic, have been compared in this presentation. In order to foretell the occurrence of heart disease, I employed a dataset and these algorithms. For more precise results, I ran all heart disease data through Google Colaboratory, using the Python programming language.

3.2 Data Collection Procedure

My anticipated model was developed in this research using data on cardiovascular disease. Kaggle was used to collect the data set. This dataset contains information about 18 different categories. There are about 400k persons of varying ages represented in the dataset, all of whom have health-related information included.

3.3 Applied Methodology

Selecting the best suitable algorithm for implementation might be a time-consuming process. When I applied the method on the dataset without cleaning it first, I obtained disappointing results. Consequently, before applying machine learning techniques, the dataset must be preprocessed. However, after trying out a variety of techniques and models, I looked to machine learning algorithms for a conclusive evaluation. Here I go over the whole process of data analysis utilizing a classification strategy and demonstrate how the selecting algorithms are put into practice. The picture below depicts the steps of the actual methodology used.

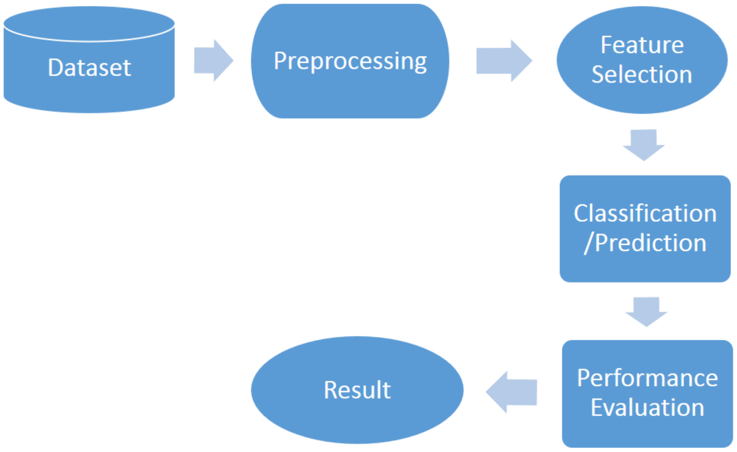

FIGURE 3.1: Proposed Methodology

3.3.1 Data Preprocessing

Because the outcome of a machine learning approach is contingent on the quality of the preparation and structure of the dataset, preprocessing is a crucial stage in any approach involving ML or data mining. I looked to see if there were any blanks or missing data points. Since almost all features are of the categorical kind, it stands to reason that some records will have the same values for all characteristics. So I looked to see if any values were being repeated. Then I looked for any unusual data points. Only the Body Mass Index (BMI) column might possibly include outliers.

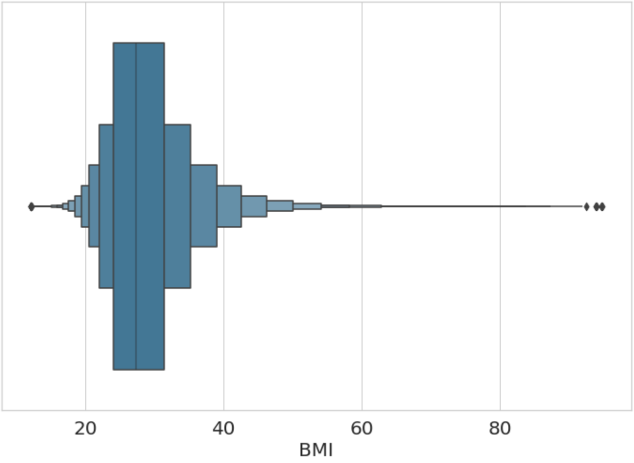

FIGURE 3.2: Checking Outliers

However, the range of values for this variable is so large (12-94.8) that we can safely ignore any values outside this range. There is a 91:9 imbalance in the dataset.

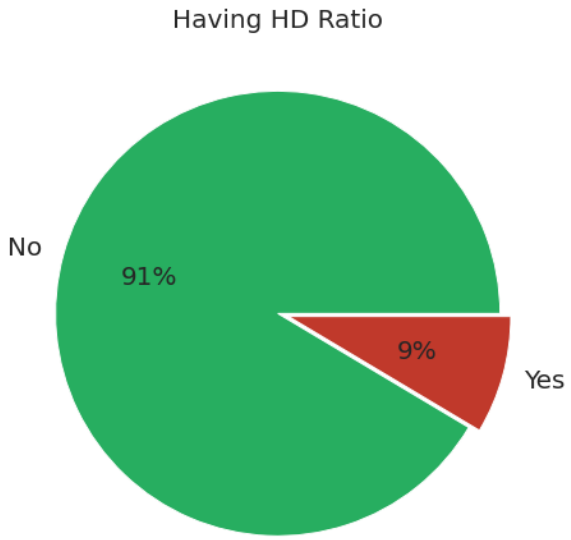

FIGURE 3.3: Data Imbalance Problem

Through the use of SMOTE, I oversampled the data.

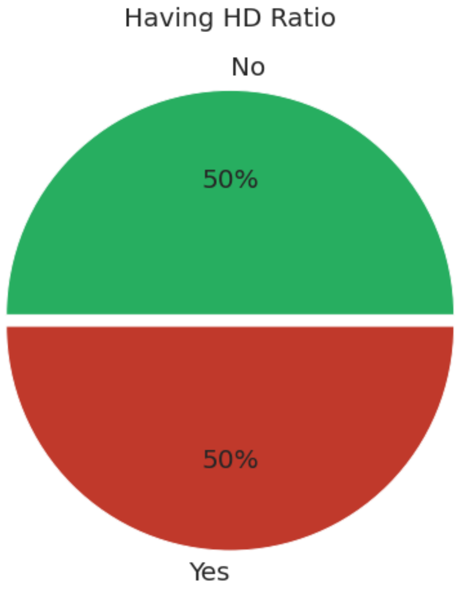

FIGURE 3.4: Oversampling Using SMOTE

3.3.2 Decision tree (DT)

As a machine learning algorithm, DT has been around for quite some time and is still widely used today. Classifying data into a hierarchy, such as a tree, is the goal of a DT, which is why its decision-making logic is designed to evaluate and match such outcomes. Nodes in a DT typically branch off from one another, the node at the very top of the tree is called the root node, and the nodes that branch off from it are called child nodes. Every internal node that has at minimum one child node reflects, in some way, the assessment of the variables or characteristics that were entered. Classification approaches begin with an analysis of each node, after which they proceed to select the offspring node that best fits the criteria at hand, continuing this pattern till they get the leaf node. The consequences of the choice are symbolized by the nodes at the leaves of the tree, which are also referred to as terminal nodes. Because of the intuitive character of DT, it has quickly become an essential component in a wide variety of medical diagnostic processes [23].

3.3.3 Stochastic Gradient Descent (SGD)

The one-versus-all classification strategy is utilized by stochastic gradient descent, which mixes many binary classifiers. Due to the fact that it uses all of the samples during each iteration, SGD has seen widespread application for processing big datasets. Because the method of operation is relatively comparable to that of the regression approach, it is not difficult to comprehend or put into practice. In order to acquire accurate results from SGD, the hyper parameters need to have their values adjusted appropriately. The sensitivity of SGD is great with regard to the scaling of features [24].

3.3.4 Random Forest (RF)

In most cases, even without hyper-parameter tuning, the results from using the machine learning technique known as random forest are excellent. Because of its flexibility and ease of implementation, it is also one of the most popular algorithms. Data classification using ensemble learning (RF) is an approach that builds on DT. During the learning phase, it produces a large number of trees as well as a thicket of decision trees. During the assessment phase of the process, every trees in the forest will make a prediction regarding the category that a particular occurrence will be placed in. Each tree makes a prediction for a class label, and then the results of those predictions are tallied to determine the ultimate verdict for each set of test data. As a rule of thumb, the most popular classification label is the one most likely to accurately describe the test data. This loop is repeated for each piece of information in the repository [25].

3.3.5 AdaBoost (Ada)

The name "Adaboost" stands for "adaptive boosting algorithm." The idea of boosting is employed here; this is an ensemble method for improving the efficiency of underperforming students. At the outset, this technique uses the primary dataset to train a classifier. After that, many instances of the classifier are created and trained, with each instance working to fix the mistakes made by its predecessors. The classification algorithm is replicated, and each instance is trained on a unique data set. The dataset is partitioned into many subsets via the application of weights to the data elements. As a result of being given a greater weight, a misclassified instance has a better probability of being chosen for the subsequent subgroup. This allows for the successive training of a large number of models. After that, a cost function is used to integrate these mediocre classifiers into a single robust one. Increased weight is given to classifiers that perform better overall. As a parameter, Adaboost allows you to specify a weak classifier on which you want to apply boosting. Adaboost uses the decision tree classifier by default while boosting [26].

3.3.6 XGB

To put it simply, XGB is a decision tree-based gradient boosting technique (GBDT). The Gradient-Based Decision Tree (GBDT) algorithm is a gradient-boosting technique. As an ensemble learning approach, gradient boosting focuses on training numerous weak classifiers and combining them into a single robust one. Through the use of negative gradients, the loss function is minimized while the misclassified classes are given more weight in the subsequent training cycle. When compared to GBDT, XGB incorporates a regularization technique to simplify the model, minimize the complexity of the loss function, and prevent overfitting. To further improve efficiency and scalability, an approximation approach is employed to determine the best option for splits, fine-tune the gradient boosting, and more. The algorithm's performance can be further enhanced by designating a separate branch for dealing with missing or sparse inputs. Last but not least, XGB allows for a parallel operation and an early halt to speed up the model's operation. To hasten the training process, the tree might be prematurely terminated when the prediction result is optimal. In addition, XGB can enhance the model's classification precision [27].

3.3.7 Logistic

When modeling with only two possible outcomes, logistic regression is the method of choice. The model is a generalized linear model, which describes the connection between independent and dependent variables X and Y. In the binomial distribution, the chance of a binary answer with the predictors is represented by the symbol and is written as p = P(Y = 1 | X), where X is any predictor.

$$\text{logit } (p) = \log \left(\frac{p}{1-p}\right) = \vec{\beta} \cdot \vec{X} = \beta_0 + \sum_{i=1}^{N} \beta_i X_i$$

The methodology described above is predicated on the hypothesis that the logarithm of the likelihood of the result is proportional to the predictors, as will be demonstrated in the following paragraphs.

$$\mathrm{E[Y \mid X]} = \mathrm{P}(\vec{\mathrm{X}}) = \frac{1}{1 + \mathrm{e}^{-\vec{\beta} \cdot \vec{\mathrm{x}}}}$$

By reducing the magnitude of parameter estimates, LASSO and Regression models approaches punish model complexity, making it more generally applicable to unseen data by making large magnitude features less suited to model parameters. This is handy with several predictions. Similar to how previous distributions with ongoing support pushed posterior distributions away from parameter extrema, penalization discourages parameters from attaining boundary values [28].

3.4 Implementation Requirements

The first and most crucial step is picking out the right program and machine. It is important to consider the goals of and the nature of this model.

3.4.1 Hardware Requirements

TABLE 3.5: THE REQUIREMENT OF HARDWORE FOR THIS PROJECT

Proxemics Area	Time of Target
Processor	Any Computer of modern era can run this model
Motherboard	Any modern eras motherboard
Ram	Minimum 2 Gigabyte
Internet Card	Internet Cards of any kind
Graphics Card	Video cards of any kind
Hard Disk	Minimum 50 Gigabyte
Casing	Any Type
Monitor	Colored Monitor
Keyboard	Any type
Mouse	Any type

3.4.2 Software Requirements

TABLE 3.6: THE REQUIREMENT OF SOFTWARE FOR THIS PROJECT

Software	Usage
Any windows and Mac operating system	To manage all the tools, programs, and equipment while running the computer.
Google Colaboratory	To run and Train and test the model.

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Setup

The ability to predict prediction levels is facilitated by the use of a wide variety of machine learning and data mining techniques. I implemented this data after collecting the necessary data and preprocessing it thoroughly. So far, I have collected data on 18 different features. The accuracy of the forecast was then estimated with a separate MLA. Decision Tree (DT), Stochastic Gradient Descent (SGD), Random Forest (RF), AdaBoost (Ada), XGB, and Logistic are some of the six methods I've utilized. I was able to attain varying degrees of accuracy after applying these machine learning classification techniques.

4.2 Experimental Results & Analysis

The Confusion Matrix is a useful tool for evaluating the accuracy of various credit scoring models [29]. In order to evaluate the effectiveness of a classifier using a confusion matrix, the system needs to be provided with labels containing values that are already known. These provide an indication of the classifier's overall performance. The more accurate the classifier was, the lower the values for false positives and negatives were [30]. A classifier's potential can be precisely evaluated using a confusion matrix. Once the classification procedure is complete, actual values and anticipated values are generated in a confusion matrix. The following matrix values are used to evaluate the system's efficiency. Based on the matrix values, the system's effectiveness is determined [31]. My confusion matrix for all known machine learning methods is shown here.

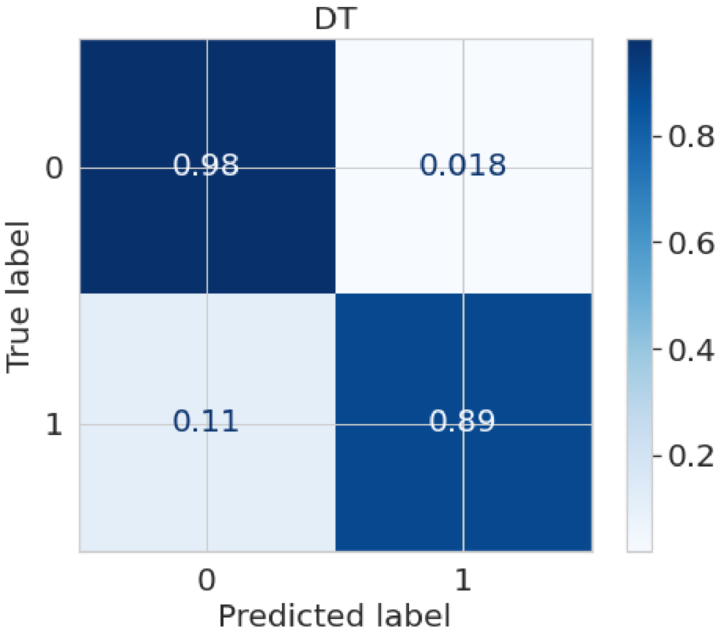

FIGURE 4.1: Confusion Matrix Generated by Decision Tree (DT)

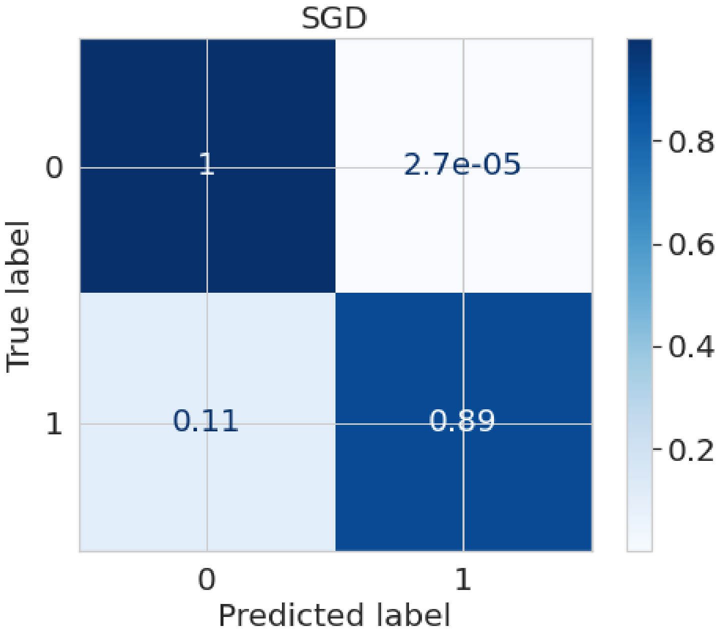

FIGURE 4.2: Confusion Matrix Generated by Stochastic Gradient Descent (SGD)

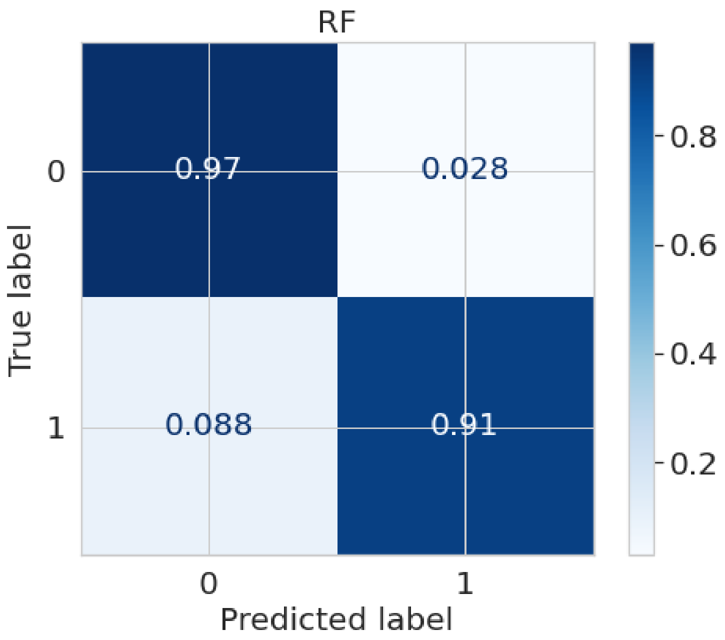

FIGURE 4.3: Confusion Matrix Generated by Random Forest (RF)

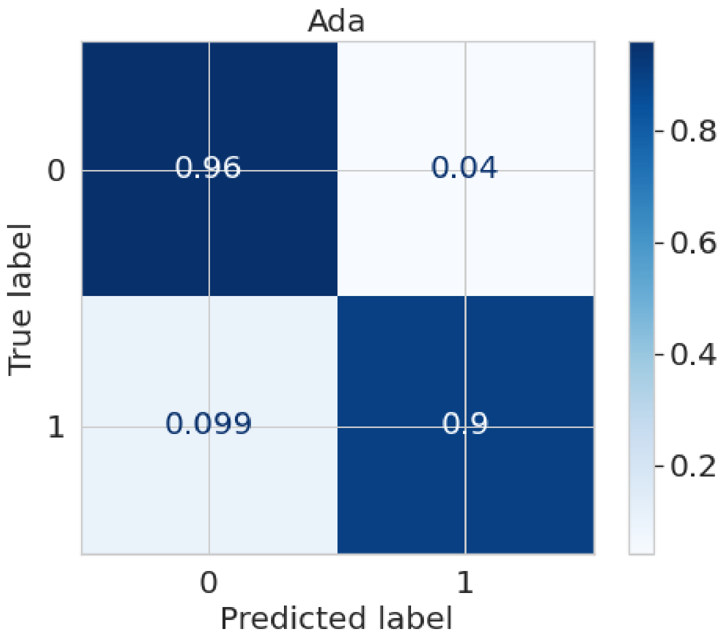

FIGURE 4.4: Confusion Matrix Generated by AdaBoost (Ada)

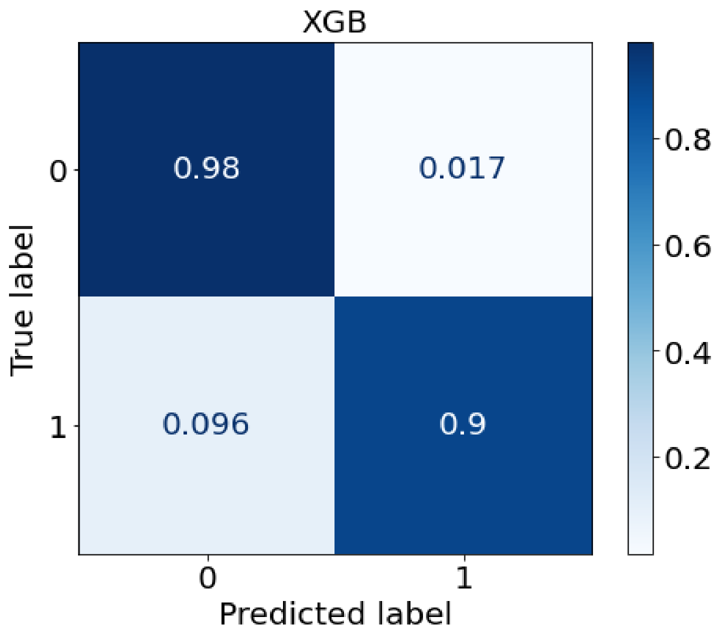

FIGURE 4.5: Confusion Matrix Generated by XGB

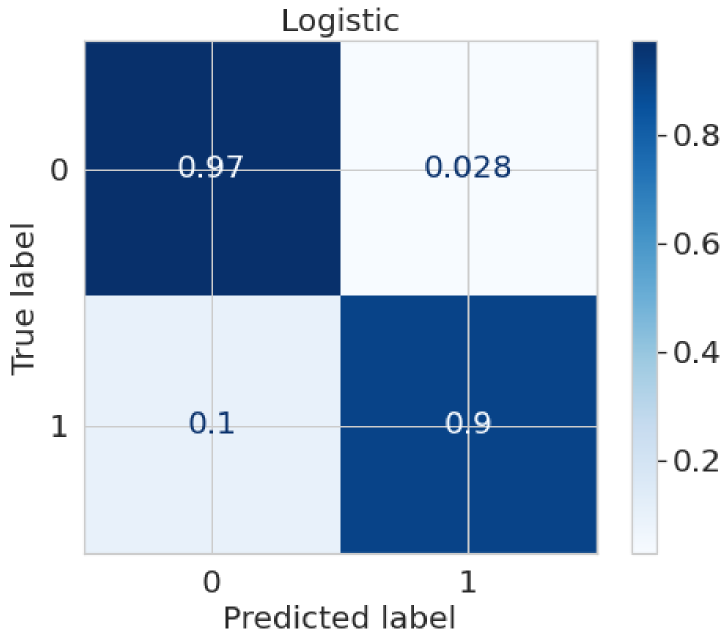

FIGURE 4.6: Confusion Matrix Generated by Logistic

TABLE 4.7: PERFORMANCE EVALUATION METRICS FOR ALL APPLIED MODELS

Models	Train Score	Test Score	Recall	Precision	f1-score	Accuracy
DT	0.944	0.936	88.92	98.06	93.27	93.58%
Logistic	0.935	0.936	89.94	96.96	93.32	93.56%
SGD	0.946	0.947	89.33	100.00	94.36	94.66%
RF	0.948	0.942	91.16	96.98	93.98	94.16%
Ada	0.930	0.931	90.12	95.77	92.86	93.07%
XGB	0.942	0.944	90.39	98.20	94.14	94.37%

I employed six distinct approaches to argue for the superiority of one over another in this context. The effectiveness of the ML technique is measured by five different metrics: Train Score, Test Score, Recall, Precision, and f1-score. Stochastic Gradient Descent (SGD) was found to be the most effective method once all was said and done, with an accuracy of 94.66%. AdaBoost (Ada) is the model with the lowest accuracy, with 93.07%.

4.3 Discussion

As a result of these results, we are able to deduce that the algorithms that we have been using are quite effective when applied to the task of making predictions. The levels of accuracy that could be achieved via the use of these algorithms were very close to one another. I was able to achieve higher levels of precision and cost savings because of the algorithms that we utilized.

Chapter 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

Most individuals in Bangladesh, as befits a nation of relatively low wealth, lack the resources necessary to pay for medical care when they become ill. Among the many leading causes of untimely mortality, cardiovascular illness ranks high. Many people in our community have cardiac problems but can't get diagnosed due to the high cost of medical care. Our heart disease prediction technology is cheap to run and produces reliable results. Those who previously went undiagnosed due to the high cost of medical care are now able to do so with ease. With the support of these systems, people are able to receive high-quality care at reasonable prices. Because our method was developed using relevant computer-based data and decision support techniques, low-cost clinical testing is a breeze. The lives of those impacted in our society can be saved as a result, and the death rate will gradually decline.

5.2 Impact on Environment

The data that we have used in this study was healthcare-related. A patient's information was part of it. The system can examine data and make healthcare judgments using data mining methods. These clinical decisions that assist with computer-based patient data could improve the medical system by reducing medical errors, enhancing patient safety, and lowering unnecessary practice variance. Thanks to data mining, we now live in a world brimming with information and insight. It can boost the quality of clinical judgments. The current method of making clinical decisions based on doctors' instincts and experience will be replaced by this computerized setting. Besides, consumers can discover the risk factors behind heart disease with this method, such as arsenic poisoning in water, food handling, and air pollution. So, we can enhance our consciousness about the aspects of the environment that are accountable for heart related disorders.

5.3 Ethical Aspects

- There will be no harm done to any patient.
- Protection of Individual Privacy.
- To avoid appearing genetically biased.
- Consider myself responsible for the outcomes of the study.

5.4 Sustainability Plan

The longevity improves daily as more people put it to professional usage. Over the following few days, I plan to add more functionality and fine-tune the accuracy, if necessary. I plan to include recent clinical data regarding heart disease into my prediction model.

Chapter 6

Summary, Conclusion and Implication for Future Research

6.1 Summary of the Study

In my dissertation, I developed a machine learning method for predicting cardiac disease. First, I gathered a dataset from the internet to use for this purpose. I did some thorough analysis of the dataset after I had collected it. I have created a data visualization that provides a graphical representation of the data for easier analysis. The data was then cleaned up so that I could use different classifier algorithms on it. In total, six classification algorithms have been used on the dataset. DT, SGD, RF, Ada, XGB, & Logistic were the techniques used. I've used these algorithms and gotten a variety of accuracy scores. After that, I did some research based on our findings.

6.2 Conclusions

Life-threatening cardiac disease may result in deadly consequences, among them cardiac arrest. DM and ML approaches have the ability to forecast the incidence of diseases with a high degree of accuracy. Here, I put ML methods for heart disease prediction to the test on a real-world dataset and found that six different classification algorithms—DT, SGD, RF, Ada, XGB, and Logistic—performed very well. The research attempted to identify the most effective ML methods by comparing many widely used and straightforward algorithms and found that they all delivered satisfactory results, at least with respect to the test dataset. This is still in the experimental stages of employing ML techniques, but it shows promise as a useful supplement to medical treatment.

6.3 Implication for Further Study

My dataset played a crucial role in the prediction of cardiac illnesses thanks to efficient data mining. After using six different classification algorithms, I finally found one that worked well enough for my purposes. Later on, I want to analyze the performance of these algorithms and replace them with others that provide more precision. I'll do everything I can to get information from hospitals in Bangladesh for usage as well.

References:

[1] L. Kourkouta, “History of Nursing,” Pasxalidis Publications, Athens, 2010.

[2] World Health Organization, available at <<https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 >>, [Accessed 02 June 2021].

[3] M. Abdar, W. Ksia̧zek, U. R. Acharya, R. S. Tan, V. Makarenkov, and P. Pławiak, “A new machine learning technique for an accurate diagnosis of coronary artery disease,” Computer Methods and Programs in Biomedicine,vol. 179, p. 104992, 2019.

[4] J. H. Joloudari, “Coronary artery disease diagnosis; ranking the significant features using a random trees model,” International Journal of Environmental Research and Public Health, vol. 17, no. 3, pp. 1–24, 2020.

[5] M. Durairaj and V. Revathi, ‘‘Prediction of heart disease using back propagation MLP algorithm,’’ Int. J. Sci. Technol. Res., vol. 4, no. 8, pp. 235–239, 2015.

[6] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, ‘‘Prediction of heart disease using machine learning,’’ in Proc. 2nd Int. Conf. Electron, Commun. Aerosp. Technol. (ICECA), pp. 1275–1278, Mar. 2018.

[7] A. Rajdhan, A. Agarwal, M. Sai, D. Ravi and P. Ghuli, “Heart Disease Prediction using Machine Learning,” International journal of engineering research & technology (IJERT), Volume 09, April 2020.

[8] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, M. A. Moni, “Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison,” Computers in Biology and Medicine, Volume 136, 104672, September 2021.

[9] S. Mohan, C. Thirumalai, G. Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,” IEEE Access, Volume 7, Page 81542 – 81554, 18770357, June 2019.

[10] P. Rani, R. Kumar, Nada M. O. S. Ahmed and A. Jain, “A decision support system for heart disease prediction based upon machine learning,” Journal of Reliable Intelligent Environments, Volume 7, Page 263–275, January 2021.

[11] U. Nagavelli, D. Samanta and P. Chakraborty, “Machine Learning Technology-Based Heart Disease Detection Models,” Journal of Healthcare Engineering, Volume 2022, Article ID 7351061, February 2022.

[12] S. I. Ansarullah, S. M. Saif, P. Kumar and M. M. Kirmani, “Significance of Visible Non-Invasive Risk Attributes for the Initial Prediction of Heart Disease Using Different Machine Learning Techniques,” Computational Intelligence and Neuroscience, Volume 2022, Article ID 9580896, , February 2022.

[13] H.M. Le, T.D. Tran and L.A.N.G. Van Tran, “Automatic heart disease prediction using feature selection and data mining technique,” Journal of Computer Science and Cybernetics, Volume 34, Page 33–47, November 2018.

[14] C. A. Cheng and H. W. Chiu, ‘‘An artificial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide database,’’ in Proc. 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), page 2566–2569, July 2017.

[15] Ali L, Niamat A, Khan JA, Golilarz NA, Xingzhong X, Noor A, Nour R and S. A. C. Bukhari, “An optimized stacked support vector machines based expert system for the effective prediction of heart failure,” IEEE Access, Volume 7, Page 54007 - 54014, 18632125, April 2019.

[16] G. Paragliola and A. Coronato, “An hybrid ECG-based deep network for the early identification of high-risk to major cardiovascular events for hypertension patients,” Journal of Biomedical Informatics, Volume 113, 103648, January 2021.

[17] F. J. Abdeldjouad, M. Brahami and N. Matta, “A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques,” The Impact of Digital Technologies on Public Health in Developed and Developing Countries, ICOST 2020, Volume 12157, June 2020.

[18] B. Martins, D. Ferreira, C. Neto, A. Abelha, and J. Machado, “Data mining for cardiovascular disease prediction,” Journal of Medical Systems, Volume 45, January 2021.

[19] C. Beulah Christalin Latha and S. Carolin Jeeva, “Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques,” Informatics in Medicine Unlocked, Volume 16, 100203, July 2019.

[20] A. Rajdhan, A. Agarwal, M. Sai, D. Ravi and P. Ghuli, “Heart Disease Prediction using Machine Learning,” International journal of engineering research & technology (IJERT), Volume 09, April 2020.

[21] P. Motarwar, A. Duraphe, G. Suganya, and M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning," In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Page 1-5. IEEE, April 2020.

[22] R. Atallah, and A. Al-Mousa, “Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method,” In 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), Page 1-6. IEEE, December 2019.

[23] B. T. Jijoand and A. M. Abdulazeez, “Classification based on decision tree algorithm for machine learning.” Journal of Applied Science and Technology Trends 2, Page 20-28, March 2021.

[24] G. Manogaran and D. Lopez, “Health data analytics using scalable logistic regression with stochastic gradient descent.” International Journal of Advanced Intelligence Paradigms, 10(1-2), Page 118-132, June 2018.

[25] Y. Liu, Y. Wang, J. Zhang, “New machine learning algorithm: Random Forest.” International Conference on Information Computing and Applications, Page 246-252, September 2012.

[26] A. S. ElDen, M. A. Mustafa, H. M. Harb and A. H. Emara “AdaBoost ensemble with simple genetic algorithm for student prediction model.” International Journal of Computer Science & Information Technology, April 2013.

[27] J. Yang and J. Guan, “A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm.” Information, Information, 13(10), Page 475, October 2022.

[28] J. J. Levy and A. J. O’Malley “Don’t dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning.” BMC medical research methodology, Page 1-15, December 2020.

[29] G. Zeng, “On the confusion matrix in credit scoring and its analytical properties,” Communications in Statistics-Theory and Methods, 49(9), 2080-2093, 2020.

[30] A. Akermark & M. Hallefält, “Churn Prediction,” June 2019.

[31] S. K. Arjaria, A. S. Rathore and J. S. Cherian, “Kidney disease prediction using a machine learning approach: A comparative and comprehensive analysis,” In Demystifying Big Data, Machine Learning, and Deep Learning for Healthcare Analytics, page 307-333, 2021.

PERSONAL KEY INDICATORS OF HEART DISEASE USING MACHINE LEARNING TECHNIQUES

ORIGINALITY REPORT

23%	21%	8%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	6%
2	Submitted to Daffodil International University Student Paper	4%
3	Submitted to Morgan Park High School Student Paper	3%
4	www.ncbi.nlm.nih.gov Internet Source	1%
5	Submitted to Jacksonville University Student Paper	1%
6	Md. Mamun Ali, Bikash Kumar Paul, Kawsar Ahmed, Francis M. Bui, Julian M.W. Quinn, Mohammad Ali Moni. "Heart Disease Prediction Using Supervised Machine Learning Algorithms: Performance Analysis and Comparison", Computers in Biology and Medicine, 2021 Publication	1%

www.arxiv-vanity.com