A MODEL TO PREDICT HEART DISEASE USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the degree of

BACHELOR OF TECHNOLOGY
in
Computer Science and Engineering

by

CHILUMULA SRIJAN: 19K91A0549
MD AFSHA: 19K91A0505
AILA VIGNESH: 19K91A0506
DAMERA RAJU: 19K91A0555

Under the guidance of

DR. CH. B N LAKSHMI
PROFESSOR

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
TKR COLLEGE OF ENGINEERING AND TECHNOLOGY
(AUTONOMOUS)
(ACCREDITED BY NBA AND NAAC WITH 'A' GRADE)
Medbowli, Meerpet, Saroornagar, Hyderabad-500097

DECLARATION BY THE CANDIDATE

We, Mr.**CHILUMULA SRIJAN** bearing Hall Ticket Number: **19K91A0549,** Ms.**MD AFSHA** bearing Hall Ticket Number: **19K91A0505,** Mr.**AILA VIGNESH** bearing Hall Ticket Number: **19K91A0506,** Mr.**DAMERA RAJU** bearing Hall Ticket Number: **19K91A0555** hereby declare that the major project report titled **A MODEL TO PREDICT HEART DISEASE USING MACHINE LEARNING** under the guidance of **DR. MS. CH. B N LAKSHMI, professor** in Department of Computer Science and Engineering is submitted in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering.

CHILUMULA SRIJAN: 19K91A0549
MD AFSHA: 19K91A0505
AILA VIGNESH: 19K91A0506
DAMERA RAJU: 19K91A0555

Place: Meerpet
Date:

CERTIFICATE

This is to certify that the main project report entitled **A MODEL TO PREDICT HEART DISEASE USING MACHINE LEARNING**, being submitted by Mr.**CHILUMULA SRIJAN** bearing Hall Ticket Number: **19K91A0549,** Ms.**MD AFSHA** bearing Hall Ticket Number: **19K91A0505,** Mr.**AILA VIGNESH** bearing Hall Ticket Number: **19K91A0506,** Mr.**DAMERA RAJU** bearing Hall Ticket Number: **19K91A0555** in partial fulfillment of requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering, to the TKR College of Engineering and Technology is a record of bonafide work carried out by him/her under my guidance and supervision.

Signature of the Guide
Dr. CH. B N LAKSHMI

Signature of the HoD
DR. A. SURESH RAO

Place: Meerpet
Date:

TABLE OF CONTENTS

ABSTRACT

In the current state of the globe, it is quite difficult to detect heart disease through early signs. If not treated in a timely manner, this might result in death. An accurate decision support system can play a crucial role in the early-stage diagnosis of heart disease in poor nations when there aren't any heart specialist doctors in remote, semi-urban, and rural locations. According to the clinical characteristics of the patient, the authors of this research have suggested a hybrid decision support system that can help in the early diagnosis of heart disease. The authors' method for handling the missing values is multivariate imputation using chained equations. The choice of appropriate features from the provided dataset was made using a hybridized feature selection approach that combines the Genetic approach (GA) with Fishers Score. SMOTE (Synthetic Minority Oversampling Technique) and common scalar approaches have also been employed for pre-processing the data.

The authors employed support vector machine, Naive Bayes, Logistic Regression, Random Forest, and Ada-boost classifiers in the final stage of creating the suggested hybrid system. With the random forest classifier, the system has been proven to produce the highest accurate results. In the simulation environment created using Python, the suggested hybrid system was evaluated. The Cleveland heart disease dataset from the University of California, Irvine (UCI) machine learning repository was used for testing. Compared to some of the other heart disease prediction methods that can be found in the literature, it has a higher accuracy of 96%.

ACKNOWLEDGEMENTS

The joy and exhilaration that come with completing a design successfully would be absent if we failed to admit the people whose support and guidance made it possible and who have culminated our sweats with success.

I am grateful to my internal guide **Dr.CH B. N. Lakshmi,** Professor in the Department of Computer Science and engineering at TKR College of engineering and technology for his or her assistance and direction during the completion of my thesis or dissertation.

For his assistance and direction during the writing of our thesis and dissertation, department head, **Dr. A Suresh Rao,** Professor of Computer Science and engineering at the TKR College of engineering and technology is also to be tanked.

I would like to express my sincere thanks to **Dr. D. V. Ravi Shankar,** the principle of the TKR College of engineering and technology for allowing me to work on this thesis or dissertation.

Eventually, I would want to express my gratefulness to everyone who helped me complete this thesis or discussion. For their stimulant and support, my family and musketeers earn a special word of thanks.

CHILUMULA SRIJAN: 19K91A0549
MD AFSHA: 19K91A0505
AILA VIGNESH: 19K91A0506
DAMERA RAJU: 19K91A0555

Place: Meerpet
Date:

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Motivation

The primary thing of this study is to offer a heart complaint vaticination model for the auguring of heart complaint prevalence. also, the thing of this exploration is to find the stylish bracket system for determining if a case may have cardiac complaint. By conducting a relative exploration and analysis exercising four bracket algorithms, Random Forest, Ada-boost, Logistic Regression, and Support vector Machine, which are employed at colourful situations of assessments, this work is supported. The vaticination of cardiac complaint is a pivotal task taking the topmost position of delicacy, despite the fact that these machine literacy ways are frequently employed. As a result, the algorithms are assessed using a variety of criteria and assessment ways. Since it is vital to read a person's development of heart complaint. As a result of the fact that indeed youthful people currently are affected by heart complaint, the number of heart attacks has been rising encyclopaedically. This work was chosen in order to address the problem by relating cardiac complaint in people before it becomes a serious problem. People in their youth, middle age, and old age will have their lives saved by our endeavourer. By making this cast, we can educate the public about the threat factors that can contribute to heart complaint and other health issues. We also predicate our vaticination on the data we have gathered from the general public. The presence or absence of alcohol use and smoking are also significant factors in prognosticating the development of heart complaint. This will enable scientists and medical professionals to produce a better.

1.2 Problem definition

The detection of heart disease is a major issue. Even if there are technologies that can predict heart disease, they are either unaffordable or useless when it comes to estimating the probability of heart disease individuals. Early detection of cardiac problems can reduce mortality and overall implications. It is not conceivable for a doctor to consult with a patient for 24 hours, even if it is not always possible to monitor patients exactly on a daily basis and because it requires more knowledge, effort, and ability. Due to the abundance of data available nowadays, we may utilize a number of machine learning algorithms to search for hidden patterns in the data.

1.3 Limitations of existing system

Even while the present technique can assess whether a person has heart disease or not, the degree of heart disease cannot be diagnosed using it. In addition, they overlooked a number of critical factors that are required to correctly predict the progression of heart disease. The prediction is not generally available and cannot be accessed by any coder.

Disadvantages:

- There is no mention of severity
- Desired Features are not considered.
- Not all users will find it easy to use

1.4 Proposed System

In the current project, we developed a model to predict the progression of heart disease using a graphical user interface (GUI), in which the user completes a form. The lasso method takes this data and searches for missing values, filling them with the mean value of the property. In order to move further, this processed data is given to the already-built model. The model will determine the degree of a person's heart disease. We have introduced two new variables that ask about the user's drinking and smoking patterns as additional criteria for predicting the severity of heart disease. The overall prediction accuracy was increased with the use of the extra attributes and efficient machine learning techniques.

Advantages:

- Free-to-use GUI was used.
- Smoking and alcohol were included as additional elements to the dataset.
- Accuracy is increased.

Chapter 2

LITERATURE REVIEW

2.1 Review of Literature

Heart disease is a term used to describe any condition that interferes with the heart's capacity to operate normally. It is thought that the most typical cause of heart failure is the narrowing or blockage of the coronary arteries, which provide blood to the heart itself. CVD, or cardiovascular disease, is another name for heart disease

2.1.1

Title: Heart Disease Prediction using Machine Learning Techniques

Published by: Devansh Shah1, Samir Patel1, Santosh Kumar Bharti.

Throughout the world, it is the most common cause of death. Heart attacks in the US are mostly brought on by coronary artery disease (CAD), which is the most prevalent kind of heart disease. At least one of the right coronary arteries (RCA), left circumflex artery's (LCX), and left anterior descending artery's (LAD) arteries is stenotic in patients with CAD. Numerous variables have been identified by clinical research as risk factors for CAD and heart attacks.

These elements may be divided into two groups: risk factors that cannot be adjusted and those that can. Sex, age, and family history are among the characteristics that cannot be altered, however those connected to a subject's lifestyle, such as smoking, high cholesterol, high blood pressure, and physical inactivity, may. The latter are risk factors that may be altered and, in some circumstances, removed by medication and lifestyle modifications. Currently, angiography, which is thought to be the most accurate procedure, is used the most frequently by doctors to diagnose CAD. However, it comes with significant side effects and a hefty price.

Additionally, it is challenging for doctors to diagnose patients when they must consider too many variables, as was described above. The necessity for non-invasive techniques for heart disease screening is prompted by these issues. Additionally, traditional approaches for the diagnosis of cardiac disease are mostly based on the study of pertinent symptoms by a medical professional, analysis of the patient's medical history, and physical examination results. As a result, these techniques frequently result in inaccurate diagnosis because of human error.

As a result, it is necessary to create an automated diagnostic system for heart disease detection based on machine learning in order to circumvent these issues. In the past ten years, a variety of hybrid diagnostic systems based on features pre-processing and ANN have been created with the aim of increasing classification accuracy. The decision-making abilities of doctors diagnosing patients have increased as a result of these diagnostic methods. Our development of an automated diagnostic system based on a two-statistical model and DNN for the enhanced identification of heart disease was also inspired by the study of these automated diagnostic systems.

Associated Work:

Studies have suggested various automated diagnostic systems based on machine learning models for heart disease prediction, including naive Bayes (NB), k-nearest neighbor (KNN), support vector machine (SVM), fuzzy logic, artificial neural network (ANN), and ensembles of ANN. The capacity of ANN-based approaches to handle challenging linear and non-linear problems has led to their widespread use in medical diagnostics, according to a thorough review of these investigations. For learning the values or weights of parameters from training data, the majority of studies that employed ANN for heart disease diagnosis used Levenberg Marquardt (LM), scaled conjugate gradient (SCG), and Pola-Ribeiro conjugate gradient (CGP) methods.

The IBFGS and Adam optimization methods, however, which were very recently proposed, were applied in this work. Additionally, the deep neural network employed in this article has more hidden layers than the ANN used in prior studies, which only had one hidden layer. Deep neural networks are neural networks with several hidden layers that are trained utilizing novel techniques. Recently, Result et al. introduced a neural network ensemble model for the detection of cardiac illness, and it obtained 89.01% accuracy, 80.95% sensitivity, and 95.91% specificity. For the detection of heart disease, Samuel et al. created a unique hybrid system called the Fuzzy-AHP approach based on ANN and fuzzy analytic hierarchy process.

The prediction accuracy for the system built on ANN and Fuzzy-AHP was 91.10%. Most recently, Paul et al. investigated the applicability of adaptive weighted fuzzy system ensemble approach for heart disease detection issue and obtained classification accuracy of 92.31%. The remainder of the document is structured as follows: The dataset and suggested approaches are provided in section II. Evaluation metrics and validation strategies are covered in Section III. whereas the discussion and outcomes of the experiment are covered in section IV. Conclusion and next efforts are covered in the final section.

Implementation & Methodology:

A. DESCRIPTION OF THE DATASET

The Cleveland heart disease dataset, which is public and accessible through the UCI machine learning repository, was used in this study. There are 303 occurrences in the dataset, of which 297 have no missing data and six have missing characteristics. The dataset's original version contains 76 raw features.

B. FORMULATION OF THE PROBLEM AND PROPOSED SOLUTION

A predictive model's primary objective in machine learning is to produce a hypothesis, or h(x), by applying a learning algorithm (or optimization algorithm) to the training data. In other words, by examining the behavior of training data, the model develops a suitable function. By minimizing the error on all of the training cases, the hypothesis is created.

Results:

They have created an automated approach for diagnosing heart disease in this study. The suggested diagnostic system employed a DNN for classification and a 2 statistical model for the fine-tuning of features. Six distinct assessment criteria, including accuracy, sensitivity, specificity, MCC, AUC, and ROC charts, were used to assess the effectiveness of the proposed diagnostic system. Additionally, the effectiveness of the suggested strategy was assessed against that of other well-known machine learning models and against other techniques covered in the literature. The experimental findings allow them to draw the safe conclusion that the suggested diagnostic method can enhance the decision-making process during the diagnosis of heart disease.

2.1.2

Title: An Automated Diagnostic System for Heart Disease Prediction Based on χ 2 Statistical Model and Optimally Configured Deep Neural Network

Published by: Liaqat Ali, Atiqur Ahman, Aurangzeb Khan, Mingyi Zhou, Ashir Javeed, Javed Ali khan.

Over the previous decade, heart disease has been the leading cause of mortality globally. According to the World Health Organization, approximately 17.9 million people die each year as a result of cardiovascular disease, with coronary artery disease and cerebral stroke accounting for 80% of these fatalities. A large number of fatalities are prevalent in poor and middle-income nations. Heart disease is caused by a variety of risk factors, including personal and professional behaviors, as well as hereditary susceptibility. Smoking, excessive alcohol and caffeine use, stress, and physical inactivity, as well as other physiological variables such as obesity, hypertension, high blood cholesterol, and pre-existing cardiac problems, are all risk factors for heart disease.

In order to take action to save death, an early, accurate, and effective medical diagnosis of heart disease is essential. Data mining is the process of extracting necessary information from enormous databases in a variety of sectors, including the medical, business, and educational fields. These algorithms are capable of analyzing vast amounts of data from many different sectors, the medical industry being one of them. By lowering the errors in projected and actual outcomes, it is a replacement for the common prediction modelling technique that uses a computer to analyze complicated and non-linear interactions among many components.

Huge datasets are analyzed using data mining in order to uncover hidden, vital information for making decisions from a collection of historical data for future examination. There is a vast amount of patient data in the medical industry. These data must be mined using different machine learning methods. In order for healthcare practitioners to make appropriate diagnostic decisions, they analyze this data. Clinical assistance is provided by examination of medical data mined using classification algorithms. It evaluates the algorithms for predicting cardiac disease in patients. Extraction of useful information and data from vast databases is known as data mining.

Heart disease is predicted using a variety of data mining approaches, including regression, clustering, association rules, and classification techniques including Naive Bayes, decision trees, random forests, and K-nearest neighbors. The categorization methods are examined side by side.

They used data from the UCI repository for their study. For the purpose of predicting cardiac disease, the classification model is created using classification techniques.

This study compares the current methods and discusses the algorithms used to forecast cardiac disease. The publication also discusses opportunities for growth and more study.

Related Work:

Experts have been using data mining techniques, according to researchers. There are several risk factors for heart disease, including age, sex, chest discomfort, blood pressure, cholesterol, blood sugar, family history of heart disease, obesity, and inactivity. Medical personnel can quickly diagnose heart disease in patients if they are aware of these risk factors. An essential data mining method is naive Bayes.

Chau Let's compare the patient's heart disease diagnosis using the C4.5 and Naive Bayes algorithms. In the diagnosis of patients with heart disease, Rajkumar and Reena contrasted naive bayes, k-nearest neighbor, and decision list . On the dataset for heart illness, Cheung used a naïve bayes classifier. Bayesian classifiers consistently outperform naive bayes, according to comparison research by atanamahatana and Gunopulos. In order to diagnose liver illness, Ramana, Babu, and colleagues used a classification approach that included bagging and boosting.

A novel decision tree in the field of chemometrics relating to the pharmaceutical sector was put up by Dong-Sheng Cao. Utilizing the C5 algorithm and bagging, Liu Ya-Qin conducted research on breast cancer data.

Tan AC's utilized the C4.5 decision tree, bagged the decision tree on the data from the malignant micro array, and compared the forecast. Sitar-Taut et al. investigated using J48 Decision Trees for the identification of coronary heart disease using the weka tool. Tu et al. applied the J48 Decision Tree and the Weka tool to the detection of heart disease.

Methodology & Implementation:

Datasets from the UCI Machine Learning repository were used for this study. It includes an actual dataset of 300 examples of data with 14 different features (13 predictors; 1 class), such as blood pressure, the kind of chest discomfort, the outcome of an ECG, etc. In this study, researchers employed Nave Bayer's, Decision Tree, K-Nearest Neighbor, and Random Forest algorithms to identify the causes of heart illness and develop a model with the highest degree of accuracy.

Conclusion:

The overarching goal is to develop several data mining approaches suitable for accurate cardiac disease prediction. Our objective is to efficiently and accurately forecast with fewer features and tests. They only take into account 14 key characteristics in their study. They used K-nearest Neighbor, Naive Bayes, decision trees, and other four data mining classification approaches. The overarching goal is to develop several data mining approaches suitable for accurate cardiac disease prediction. Our objective is to efficiently and accurately forecast with fewer features and tests. They only take into account 14 key characteristics in their study. K-nearest neighbor, Naive Bayes, decision trees, and random forests were four data mining categorization approaches they used.

Before being employed in the model, the data underwent preprocessing. In this scenario, the algorithms producing the best results are K-nearest neighbor, Naive Bayes, and random forest. Following the application of four methods, they discovered that the K-nearest neighbor algorithm (k = 7) had the greatest accuracy. This research may be expanded upon by combining more data mining methods, such as time series, clustering and association rules, support vector machines, and genetic algorithms. A more complicated model or a model combination must be used in order to increase the accuracy of heart disease early prediction given the constraints of this study.

2.1.3

Title: Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques

Authors: C. Beulah Christalin Latha, S. Carolin Jeeva

Heart disease is a common condition that affects a lot of people in their middle or later years and frequently results in consequences that are deadly. Males are more likely than females to suffer from heart disease. According to estimates from the WHO, heart conditions are thought to be the primary cause of 24% of non-communicable illness fatalities in India. Heart disease is to blame for one-third of all fatalities worldwide. Heart conditions are to blame for 50 percent of fatalities in the United States and other affluent nations. Cardiovascular disease (CVD) is the leading cause of death globally, with over 17 million deaths occurring each year in Asia. Considered to be the de facto database for heart disease is the Cleveland Heart Disease Database (CHDD).

Age, sex, smoking, family history, cholesterol, poor diet, high blood pressure, obesity, inactivity, and alcohol use are all thought to be risk factors for heart disease, and inherited risk factors including diabetes and high blood pressure all contribute to the condition. A few risk factors can be modified. Along with the aforementioned elements, lifestyle choices including eating patterns, physical inactivity, and obesity are also regarded as significant risk factors. There are several distinct forms of heart conditions, including myocarditis, angina pectoris, congestive heart failure, cardiomyopathy, congenital heart disease, and coronary heart disease. The likelihood of developing heart disease based on risk factors is challenging to calculate manually. Machine learning methods, on the other hand, are helpful in predicting the results from the available data. In order to forecast heart disease risk from risk variables, this article uses a machine learning approach called classification. Additionally, it makes an effort to increase prediction accuracy for heart disease risk using an ensemble technique.

Related Work:

Support Vector Machines (SVM), Neural Networks, Decision Trees, Regression, and Naive Bayes classifiers are a few of the methods utilized for these prediction issues. With 92.1% accuracy, SVM was shown to be the best predictor, followed by neural networks (91% accuracy), decision trees (89.6% accuracy), and neural networks. Neural networks, decision trees, Nave Bayes, and associative classification have all been shown to be effective in the analysis of data mining approaches for the prediction of heart disease.

To increase classification accuracy in the forecasting of heart disease, the current research has applied ensemble approaches. The accuracy of feature extraction improved up to 99.97% when genetic algorithms and neural networks based on fuzzy logic were combined. The accuracy of a trained recurrent fuzzy neural network using genetic algorithms to diagnose heart disease was 97.78%. Using a preliminary set-based classification method with a different dataset, classification accuracy for the prediction of heart disease risk reached up to 93%.

Methodology & Implementation:

A. DATASET:

The tests have been conducted using the Cleveland heart dataset from the UCI machine learning library. 14 characteristics and 303 occurrences make up the dataset. Six of the qualities are numerical, while the other eight are category. This dataset contains individuals ranging in age from 29 to 79. A gender value of 1 designates a patient who is male, whereas a gender value of 0 designates a patient who is female. Indicators of cardiac disease include four different forms of chest discomfort.

B. CLASSIFICATION & ENSEMBLE ALGORITHMS:

Classification is a supervised learning technique used to forecast outcomes using historical data. The technique suggested in this work uses classification algorithms to diagnose cardiac disease and use an ensemble of classifiers to increase classification accuracy. The dataset has been split into a training set and a test set, and each classifier is trained using the training dataset using Naive Bayes, Random Forest, C4.5, and Multilayer Perceptron.

Results:

On the Cleveland dataset, a comparison of several categorization techniques has been done. While some algorithms perform well, others have low accuracy. Ensemble methods are used to enhance the performance of the weak classifiers. Ensemble algorithms including bagging, boosting, voting, and stacking have been employed in this study. The Naive Bayes, Random Forest, Bayes Net, C4.5, multilayer perceptron, and PART algorithms work in an ensemble with the Bagging method. The Adaboost.M1 algorithm has been used for boosting.

The findings demonstrate that weak classifiers can function more effectively when they are ensembled. The categorization of the dataset is done using the Weka tool. The accuracy was increased by 3.3% when C4.5 was assembled with the powerful classifiers. The accuracy was increased by 7.26% by assembling PART with the strong classifier set. The accuracy was increased by 3.65% when a multilayer perceptron with a strong classifier set was constructed.

Conclusion:

The accuracy of a classifier ensemble used in this study to predict heart disease is examined. Both training and testing were done using the Cleveland heart dataset from the UCI machine learning library. We conducted tests using the ensemble algorithms bagging, boosting, stacking, and majority voting. Maximum accuracy improvements of 6.92% occurred when bagging was utilized.

2.1.4

Title: A Clinical Decision Support Framework for Heart Disease Prediction Using Majority Vote-Based Classifier Ensemble.

Published by: Saba Bashir · Usman Qamar · Farhan Hassan Khan · M. Younus Javed

Data mining is the process of examining and discovering hidden patterns, correlations, and information from huge databases that was not attainable using conventional methodologies. Recent studies have shown that data mining techniques are particularly useful for identifying several illnesses, including cancer, stroke, diabetes, and heart disease. Adverse outcomes and unsatisfactory outcomes might be the result of poor clinical judgements. Classification and prediction of cardiac disease may be done with high accuracy using an intelligent computer-based information and decision support system. The primary goal of this study is to convert data into knowledge that may guide clinicians in clinical decision-making, lower medical mistakes, and improve patient safety.

Several data mining approaches have become widely used for intelligent heart disease prediction during the past few decades. The use of different machine learning algorithms is essential for clinical decision assistance.

Related Work:

Radial-based function network structure and support vector machine were used by Ghumbre et al. to create a method for predicting cardiac disease. The investigation demonstrates that the outcomes of the support vector machine approach are on par with those of the radial-based function network. The dataset input method utilized for data collecting has an impact on this procedure.

To detect cardiac illness in a patient early on, Chitra and Seenivasagam used a supervised learning method. Hidden neurons cascaded neural network (CNN) is the name of the suggested classifier. While SVM's high sensitivity suggests it has a high likelihood of correctly predicting sickness, CNN's high specificity demonstrates its capacity to predict a healthy individual.

Methodology & Implementation:

Dataset:

There are 270 instances total in the Statlog dataset, with 243 serving as training instances and 27 serving as testing instances. Thirteen attributes are taken from a wider collection of seventy-five attributes to make up the dataset. Aspects can be actual, ordered, binary, or nominal, among other varieties. Within the Ricco database is the Eric heart disease dataset. Instances in the dataset total 209 in number. 188 occurrences are in the training set, but only 21 are in the testing set.

MV5 Approach

Base Classifiers:

1. Naïve Bayes (NB) Classifier
2. Decision Tree Induction Based on Gini Index (DT–GI)
3. Decision Tree Induction Using Information Gain (DT–IG)
4. Memory-Based Learner (MBL)
5. Support Vector Machine (SVM)

Data gathering, pre-processing, classifier training, and an ensemble model are all part of the MV5 technique to predicting heart disease. Data are gathered through the data acquisition process, which also divides the data into separate sets and chooses variables. The processes of pre-processing include missing value imputation, outlier detection, feature selection, and class label identification.

Every dataset is specifically selected for features. As benchmark datasets that have previously been handled by the appropriate publishers, the heart disease datasets used in the proposed research do not include any extraneous information. As a result, for additional analysis, they employed the whole feature set of each dataset. Using the specified feature selection approach, any additional dataset that could have attributes unrelated to the domain will be cleaned up.

Imputation for missing values: Missing values in medical datasets are a severe issue that arises during the first analysis and interpretation of the data. The suggested pre-processing method locates the missing attribute values, which are then changed for continuous and categorical attributes, respectively, to mean and mod values.

Result:

Five separate heart disease datasets, each with a unique set of features, make up the experimental dataset. Test sets are used for experiments, and the outcomes are assessed. For each dataset, tenfold cross validation is used to assess it in order to address the sample deficiency. The labelled data is split into training and test sets using cross validation. Every classifier is trained on a training set before being used on a test set. Decision tree-based classifiers (DT-GI and DT-IG) develop their own unique classification rules. The test data is then classified using these principles. The final three classifiers (NB, SVM, and MBL) are trained on the training dataset before being put to the test on an additional, secret test dataset.

Conclusion:

For intelligent heart disease analysis and prediction, the research study has developed an ensemble classifier. It makes use of five different heterogeneous machine learning classifiers, merges the output from each classifier into an ensemble model, and produces prediction data for heart disease diagnosis. For the training and testing of the ensemble classifier, five separate datasets were employed, each containing a unique collection of various sorts of characteristics, and they were each supplemented by the majority voting technique. Comparing the proposed ensemble model to other state-of-the-art methods, experimental findings using tenfold stratified cross-validation demonstrate that it predicts heart disease patients with a pretty high degree of accuracy. With an accuracy of 88.52%, 86.96% sensitivity, 90.83% specificity, and 88.85% f-measure, MV5 has obtained. Weighted voting-based classifier ensemble and use of the suggested method are two areas for further research.

2.1.5

Title: Comparing different supervised machine learning algorithms for disease prediction

Published by: Shahadat Uddin, Arif Khan, Md Ekramul Hossain and Mohammad Ali Moni

In order to learn from the past and identify meaningful patterns from huge, unstructured, and complicated datasets, machine learning algorithms use a range of statistical, probabilistic, and optimization methodologies. Automated text categorization, network intrusion detection, spam email filtering, credit card fraud detection, consumer purchase behavior detection, industrial process optimization, and illness modelling are just a few of the many uses for these algorithms. The majority of these applications have been developed utilizing supervised as opposed to unsupervised machine learning algorithms. The result of unlabeled samples can be predicted in the supervised form by learning a prediction model from a dataset where the label is known. The focus of this study is focused on the performance evaluation of illness prediction methods employing several supervised machine learning algorithm variations. In recent years, the data science research community has paid a great deal of attention to disease prediction and, in a larger sense, medical informatics. This is partly due to the widespread use of computer-based technology into the healthcare industry in various forms (such as electronic health records and administrative data) and the consequent accessibility of sizable health databases for academics. These electronic data are used in a variety of healthcare research fields, including the analysis of healthcare utilization, monitoring the efficiency of a hospital care network, examining patterns and costs of care, creating disease risk prediction models, monitoring chronic diseases, and comparing the effectiveness of disease-prevalence drugs. Our study focuses on supervised learning techniques that are used in disease risk prediction models (such as support vector machines, logistic regression, and artificial neural networks). Models built using these methods are trained using patient-labeled training data. Patients for the test set are divided into many categories, such as low risk and high risk.

Methodology & Implementation:

Research focuses on supervised learning techniques used in disease risk prediction models (e.g., support vector machines, logistic regression, and artificial neural networks). Models built using these methods are trained using patient-specific tagged training data. Patients are divided into different categories for the test set, such as low risk and high risk. The findings of this study will aid the academics in better understanding current trends and hotspots of illness prediction models employing supervised machine learning algorithms and in setting appropriate research objectives. The majority of the relevant research in the literature that used machine learning to predict a specific illness used one or more of these methods. This study's main objective is to compare the effectiveness of various supervised machine learning algorithms for disease prediction. The supervised machine learning algorithms used in this study are Logistic Regression, Support vector machine, Decision tree, Random Forest, Naive Bayes, Artificial neural network, and K-nearest neighbor.

Result:

The final dataset consisted of 48 articles, each of which used several supervised machine learning algorithm variants to forecast a single ailment. The techniques section previously covered all implemented variations as well as the main popular performance metrics. Based on this, they evaluated the 48 papers that were ultimately chosen in terms of the techniques employed, the performance indicators, and the condition they sought to treat. The illnesses' names and sources are described, along with the supervised machine learning techniques used to forecast them. In all, 49 diseases or disorders were predicted in this analysis from 48 publications (one article predicted two diseases). The 50 algorithms that were discovered to have the best accuracy for these 49 illnesses. One illness had two algorithms (out of five) that had the same higher-level accuracy. In conclusion, 49 illnesses were predicted by 48 publications that were taken into consideration in this study, and 50 supervised machine learning methods were shown to have higher accuracy.

2.1.6

Title: HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System

Authors: Norma Latif Fitriyani, Muhammad Sharfuddin, Gnajar Arfian, Jontage Rhee

Heart disease continue to be the leading cause of mortality worldwide, accounting for over 30% of all fatalities. In 2030, it is predicted that there would be around 22 million deaths worldwide if current trends continue. According to the American Heart Association, 121.5 million American adults—or close to half of all adults—have a CVD. Heart disease was one of the top three killers in Korea in 2018 and was responsible for over 45% of all fatalities. Plaque on artery walls, which can obstruct blood flow and result in a heart attack or stroke, is a condition known as heart disease.

Unhealthy food, inactivity, and excessive alcohol and cigarette use are a few risk factors that might cause heart disease. These risk factors may be reduced by living a healthy lifestyle every day, which includes cutting back on salt in the diet, eating more fruits and vegetables, getting regular exercise, and giving up alcohol and cigarette usage. These lifestyle changes may gradually help lower the chance of developing heart disease. To minimize mortality rates and enhance decision-making for future prevention and treatment, it has typically been advised that high-risk people with heart disease be identified early and their diagnosis improved using a prediction model. To assist doctors in determining the risk of heart disease, a prediction model that is integrated into the clinical decision support system (CDSS) can be employed.

Related Work:

According to several research, machine learning models have been used to create heart disease diagnosis techniques with the goal of improving the performance of HDPMs. It has been popular among academics to assess the effectiveness of prediction models using two publicly accessible heart disease datasets, Statlog and Cleveland. Long et al. (2015) created the chaotic firefly algorithm-based heart disease clinical decision support system (CFARS-AR) using the Statlog dataset. While classifying the illness, the chaotic firefly technique was utilized, and rough sets were used to limit the amount of characteristics. A comparison between the created model and other models, including NB, SVM, and ANN, was then conducted. According to the outcomes, the suggested model performed the best. In their 2015 paper, Nahato et al. introduced the RS-BPNN, which combines rough sets-based characteristics selection with BPNN. The suggested RS-BPNN was accurate up to 90.4% using the chosen criteria. Six machine learning models (ANN, SVM, LR, k-nearest neighbor (KNN), classification tree, and NB) were tested using a variety of performance indicators by Dwivedi (2018). In terms of accuracy, sensitivity, specificity, and precision, the findings revealed that LR outperformed the other models, obtaining up to 85%, 89%, 81%, and 85, respectively. In order to conduct comparative study, Amin et al. (2019) used machine learning models (k-NN, DT, NB, LR, SVM, Neural Network (NN), and a hybrid (voting with NB and LR)) to find key variables. The hybrid approach (voting with NB and LR), according to the experiment's findings, was successful.

Methodology & Implementation:

Given the patients' present health, the suggested HDPM was created to offer high performance prediction of heart disease's existence or absence. Datasets related to cardiac disease are first gathered. The second step is the pre-processing of data for feature selection and data transformation. Third, given the ideal parameter, the DBSCAN-based outlier identification approach is used to discover the outlier data. Fourth, the training dataset is then updated without the outlier data. Fifth, the training dataset is balanced using the SMOTE-ENN approach. Sixth, the HDPM is produced using learning from the training dataset using the XGBoost-based MLA. The suggested model's performance measures are then provided, and the resulting HDPM is subsequently put into practice. In our work, they used the 10-fold cross-validation approach to prevent overfitting. By repeatedly selecting from diverse sets of training data, cross-validation enables the models to learn from these datasets, increasing the amount of data utilized for validation and maybe preventing overfitting. Previous research has shown that 10-fold cross-validation may be utilized to preserve the bias variance trade-off, which finally provides the generalized model and guards against overfitting. The next subsections provide a full breakdown of each stage, including with explanations of the datasets and modules and performance metrics. Additionally, the suggested model's performance in comparison to cutting-edge models is assessed, and the findings are shown in the results and discussion section. By incorporating the HDPM, they finally guarantee the applicability of the suggested model.

Result:

They conducted an analysis to compare the suggested model's performance to existing classification models and findings from earlier research. They also provided the statistical analysis to support the significance of our model in comparison to other models. By reaching an accuracy of up to 95.90% and 98.40% for datasets I and II, respectively, the experimental findings proved that the suggested model performed better than those of state-of-the-art models and prior study results. The statistically based study result also demonstrated the suggested model's notable advancement over the competing models.

Conclusion:

Through the integration of DBSCAN, SMOTE-ENN, and XGBoost-based MLA, they developed an efficient heart disease prediction model (HDPM) for the diagnosis of heart disease. This model's accuracy was increased. To find and eliminate outlier data, DBSCAN was utilized, the uneven training dataset was balanced using SMOTE-ENN, and the prediction model was learned and created using XGBoost MLA. To create the generalized prediction model, two publically accessible datasets on heart disease were used. In comparison to other categorization models and the findings from earlier research, they carried out assessment study on our suggested model. To further support the significance of our approach in comparison to other models, they also gave the statistical evaluation. The experimental findings showed that the suggested model performed better than leading-edge models. A considerable improvement for the suggested model in comparison to the other models was also shown by the statistically based analysis results. In order to accurately and quickly determine the heart disease state of the subjects/patients, they also constructed and developed the intended HDPM into the Heart Disease Clinical Decision Support System (HDCDSS). In order to communicate the patient data and other diagnosis data to a secure online server, the HDCDSS first collected the patient data. The communicated diagnosis information was all then saved in MongoDB, a platform that can efficiently respond in a timely manner to a medical data influx that is rapidly expanding. Following that, the suggested HDPM was loaded to determine the patients' actual heart disease state, which was then sent to the HDCDSS's diagnosis result interface. Clinicians should therefore benefit from the created HDCDSS.

The overall developed and planned HDCDSS in this study might serve as a useful manual for healthcare professionals. Future research will compare different data sampling techniques with model hyper-parameters and larger medical datasets. Additionally, a comparison and analysis research using various outlier identification techniques should be looked into further. Further research on edge computing and edge device ideas might be done with the aim of enhancing the medical clinical decision support system, especially in light of the growing privacy, security, and time-sensitive application problems. They have not yet received any comments from cardiac specialists on this study. When a particular demographic dataset (from Korea) is eventually gathered, the opinions of regional cardiac specialists will be used to verify the dataset and prediction model.

Chapter 3

REQUIREMENTS ANALYSIS

3.1 Functional Requirements

Requirements engineering or requirements analysis is the process of identifying the requirements and expectations for a new product. In order to define expectations, handle difficulties, and record all of the key needs for the product, it necessitates routine communication with the stakeholders and end-users.

The link between the system's input and output is described in the functional requirements, which also specify which output file should be created from a certain input file. A thorough explanation of all data inputs, their sources, and the range of acceptable inputs must be included in every functional requirement

3.2 Non-Functional Requirements

Describe any system components that are visible to users but are not directly connected to how the system functions. Quantitative requirements like as precision (i.e., how precisely the system's numerical responses are) and reaction time (i.e., how quickly the system responds to user requests) are examples of non-functional needs.

- Portability
- Reliability
- Usability
- Time Constraints
- Responsive design should be implemented
- Space Constraints
- Performance
- Standards
- Ethics

3.3 Software Requirement Specifications

3.3.1 Software Requirements

- Operating System: Windows (10/11)/LINUX/MAC OS
- Programming language: Python3
- Vscode

3.3.2 Hardware Requirements

- Processor: i5 intel/Ryzen processor
- Processor RAM: 4GB
- Hard disk: 128 GB
- Laptop/PC

3.4 Software Development Life Cycle

The Systems Development Life Cycle (SDLC) is a concept used in the fields of systems engineering, information systems, and software engineering to describe the process of creating new systems or upgrading existing ones, as well as the concepts and techniques that go into their design.

3.4.1 Requirements Analysis and Design

Analyses are used to compile the system's requirements. This stage comprises a careful analysis of the organizational business needs. The business procedure could change. The focus of design is on high-level design, such as what programmers are required and how they will interact, low-level design (how individual programs will run), interface design (how will interfaces appear), and data design (what data will be required). During these times, the overall architecture of the software is established. Design and analysis play a significant role throughout the whole development cycle. Any design error might be extremely expensive to repair later on in the software development process. Extreme care is used at this phase. Extreme care is used at this phase.

3.4.2 Implementation

This stage involves turning the designs into code. Computer programs can be written in a conventional programming language or with the aid of an application generator. Compilers, interpreters, and debuggers are some of the programming tools used to create the code. Different high-level programming languages are used for coding, including PYTHON 3.6 and Anaconda

Cloud. The type of application determines the best programming language to use. An organized system is created.

3.4.3 Testing

This stage involves testing the system. Typically, software are written as a series of independent modules, each of which is carefully tested individually. After that, the system is tested as a whole. The separate parts are assembled and evaluated together. The system is tested to make sure the interfaces between modules function (integration testing), that it operates on the designated platform and with the anticipated volume of data (volume testing), and that it achieves what the user wants it to do (acceptance/beta testing).

3.4.4 Maintenance

The system will eventually need maintenance. Software will very definitely change after being sent to the user. Several causes have contributed to the transformation. The system may experience change as a result of unexpected input values. Furthermore, changes to the system could have an immediate effect on how the program functions. The program should be able to adapt to any changes that may come along after implementation.

3.4.5 SDLC Methodologies

This document, which specifies every need for the system, is essential to the creation of the system's life cycle (SDLC). It will serve as the basis for testing and be used by programmers. Future standard modifications will need to have explicit change approval. In his 1988 essay "A spiral Model of Software Development and Enhancement." The SPIRAL MODEL was explained by Barry Boehm. Although this model wasn't the first to examine iterative development, it was the first to provide an explanation for why such models are so well-liked.

The intended duration of the iterations was six months to two years. Each phase starts with a design objective and ends with a customer evaluation of the project's status up to that point. Engineering and analysis work are used at every level of the project with an eye towards its ultimate goal.

The following are the processes for the Spiral Model: The new system requirements are stated as completely as is practical.

1. To do this, it is often necessary to interview many customers who represent all of the various user categories.
2. Whether internal or external, as well as other aspects of the existing system.
3.A preliminary design is created for the new system.
4. To make a first prototype of the new system, the preliminary design is applied.
5. Typically, this is a system that has been reduced in size and comes close to the original.
6. Specifications of the finished product.

- Evaluating the advantages, disadvantages, and dangers associated with the original prototype.

- Outlining the demands for the second prototype.

- Organizing the design of the second prototype.

- Constructing and testing a second prototype.

The customer might choose to abandon the project entirely if the risk is thought to be too great. Risk factors include everything that might, in the view of the client, lead to a less-than-satisfactory final product, including overruns in development expenses, mistakes in running costs, or other elements.

- The current prototype is evaluated similarly to how the last prototype was examined, and if required, a new prototype is made utilizing the previously mentioned four-step process.

- The aforementioned steps are continued until the buyer is satisfied that the redesigned prototype faithfully represents the desired final product.

- The finished system is constructed based on the modified prototype.

- A thorough testing and evaluation of the finished system. Routine maintenance is performed often to prevent widespread failures and save downtime.

3.5 Modules

3.5.1 NumPy

The following operations can be carried out by a developer using NumPy:

- actions on arrays that combine logic and mathematics.

- Shape change using algorithms and Fourier transformations.
- Operations in linear algebra are algebraic operations. Linear algebra and random number generation functions are both included in NumPy.
- A Python-based replacement for MATLAB is NumPy.
- Along with SciPy (Scientific Python) and Matplotlib (a charting toolkit), NumPy is commonly used. This combination usually takes the place of MatLab, a well-known technical computing environment.

3.5.2 Pandas

Python's Pandas library is free and open source. High-performance data structures and data analysis tools are offered ready for use.

Pandas is a widely used data science and analytics package that operates on top of NumPy.

Multi-dimensional arrays and a large variety of mathematical array operations are supported by the low-level data structure known as NumPy. The interface for Pandas is more advanced. Additionally, it offers robust time series capability and simplified tabular data alignment.

Pandas primary data structure is the DataFrame. As a 2-D data structure, it enables the storage and manipulation of tabular data.

On the DataFrame, Pandas offers a robust feature set. As an illustration, consider data alignment, data statistics, data slicing, grouping, merging, concatenating, etc.

Beginning the Pandas Installation Process

Pandas module installation requires Python 2.7 or higher. Use the below command to install it if you are using conda.

```
conda install pandas
```

Use the command below to install the pandas module if you are using PIP.

```
pip3.7 install pandas
```

The following line of code should be added to your Python script to import Pandas and NumPy:

```
import pandas as pd.
import numpy as np
```

We must import this dependency as Pandas depends on the NumPy library.

3.5.3 Pickle

The serialisation and deserialization of a Python object structure is accomplished using Python Pickle. Python allows for the pickling of any object to enable disc storage. The object is first serialized by Python Pickle before being transformed into a character stream, which is then filled with all the information needed to recreate the object in another Python script. It should be noted that the documentation claims that the pickle module is not protected from data that has been intentionally or mistakenly created. As a result, never unpickle data that you have obtained from an unauthorized or shady source.

The methods are really ‘straightforward for recovering pickled data. Pickle.load() is the method you must use to do that. The file object you obtain when opening a file in read-binary (rb) mode is the main parameter for the pickle load method. Simple! Surely not. Using the pickle dump code, let's create the code to retrieve the data we pickled.

Chapter 4

DESIGN

We put the suggested system into practice. Our programme was created using Python and a few Python libraries. In general, it is challenging to obtain necessary files from the many distinct files. As a result, this programme enables users to access data while conserving storage.

4.1 UML Diagrams

UML Diagrams:

Unified Modelling Language (UML) is a shorthand for the phrase "Unified Modelling Language." object-oriented software engineering is a discipline. A standardized general-purpose modelling language like UML may exist. The group that built it, the item Management Group, is in charge of quality. The ultimate objective is for UML to overtake other modelling techniques as the industry standard for object-oriented software. In its current form, UML consists primarily of two elements: a meta-model and a notation. In the future, UML may be tied to or introduced to a new approach or process. A standard for describing, building, and documenting software artefacts as well as business modelling and other non-software systems might be established using the Unified Modelling Language. The UML is a collection of tried-and-true engineering best practices.

GOALS:

The UML design's main objectives are as follows:

- Offer users a ready-to-use visual modelling language that is expressive so they can build and share valuable models.
- To enlarge the fundamental ideas, offer instruments for extension and specialization
- Be unfettered by development procedures or programming languages.
- Lay a strong basis on which to understand the modelling language. Encourage the expansion of the 00 tool market
- Support ideas like frameworks, components, patterns, and partnerships at a higher level of development

4.1.1 Architecture:

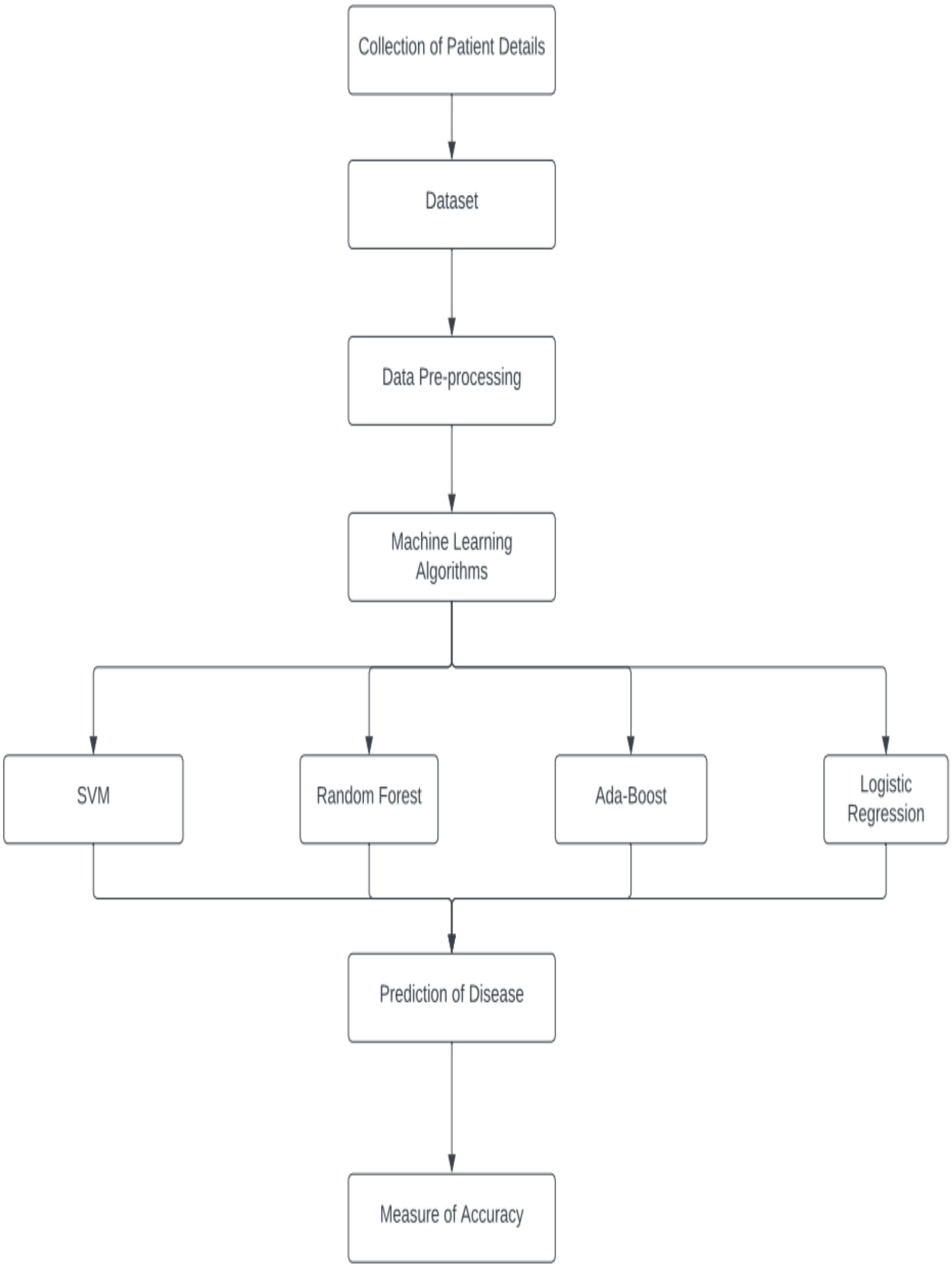


Fig. 4.1 Architecture

4.1.2 Use-Case Diagram:

A use case diagram's goal is to depict the dynamic nature of a system. Here is typically used to gather a system's wishes, including those influenced by both internal and external factors. The main role of a use case diagram is to display the system operations each actor performs. The system's actors' parts are regularly portrayed. The UML may be a crucial tool. component of the software development process, and hence, object-oriented programming. To The UML primarily uses graphical notations to coordinate software projects.

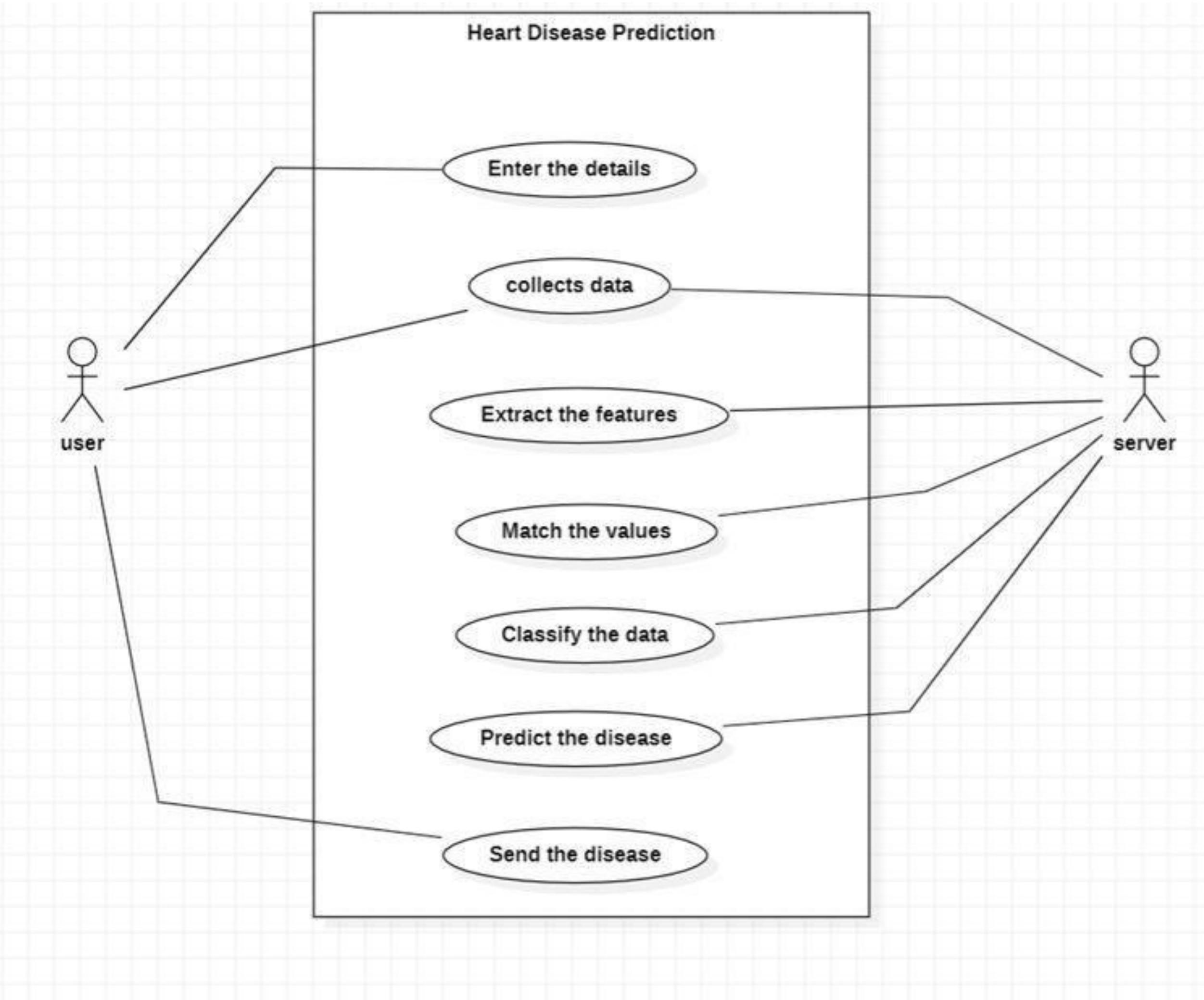


Fig. 4.2 Use-Case Diagram

In this diagram we dicpted the relationship between the two actors (user & server) interactions with the help of different number of use cases which starts with the data abstraction from the user and by using this we mapping the values and classifying the data upon which we will be predicting the disease and notice the user.

4.1.3 Sequence Diagram:

In a sequence diagram, interactions between objects are shown in the order that they happen, or sequentially. This illustration is often referred to as an event scenario or an event diagram. This makes it easier to understand how the method's many elements and objects work together. This features two axes that represent time (vertically) and a number of horizontal objects.

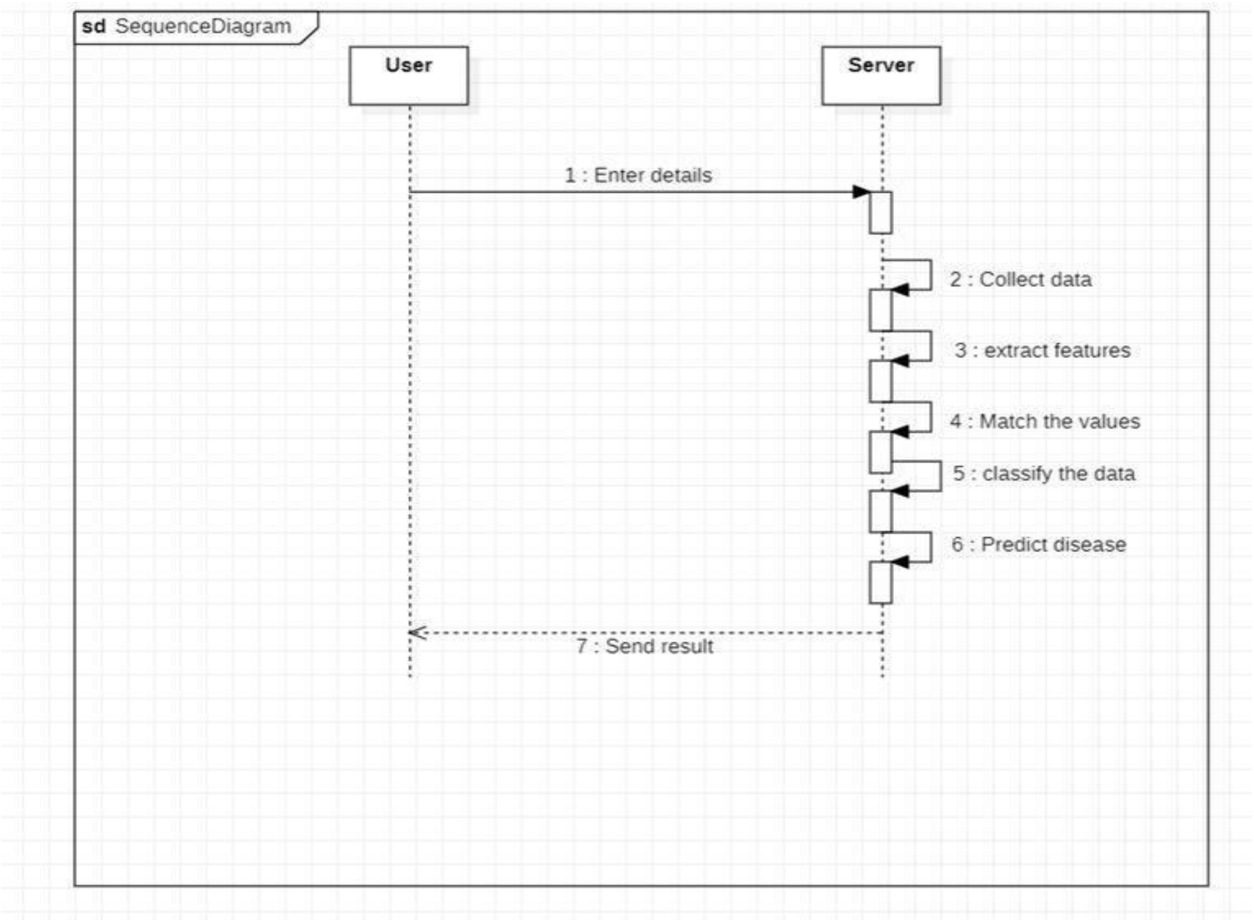


Fig. 4.3 Sequence Diagram

In this diagram we explain the control flow of sequence of actions with the help of life lines which shows the duration of action. The interaction lines show the interaction between the two objects and their activities.

4.1.4 Activity Diagram:

This behavior diagram demonstrates the behavior of a system. It shows how things are controlled from the beginning to the end. Because it doesn't display any messages as a result of one activity leading to another, it is sometimes referred to as a flowchart. They are not flowcharts, despite their resemblance. During the development of a system, activity diagrams are widely used in the Unified Modelling Language to show the operational and business step-by-step workflows of components.

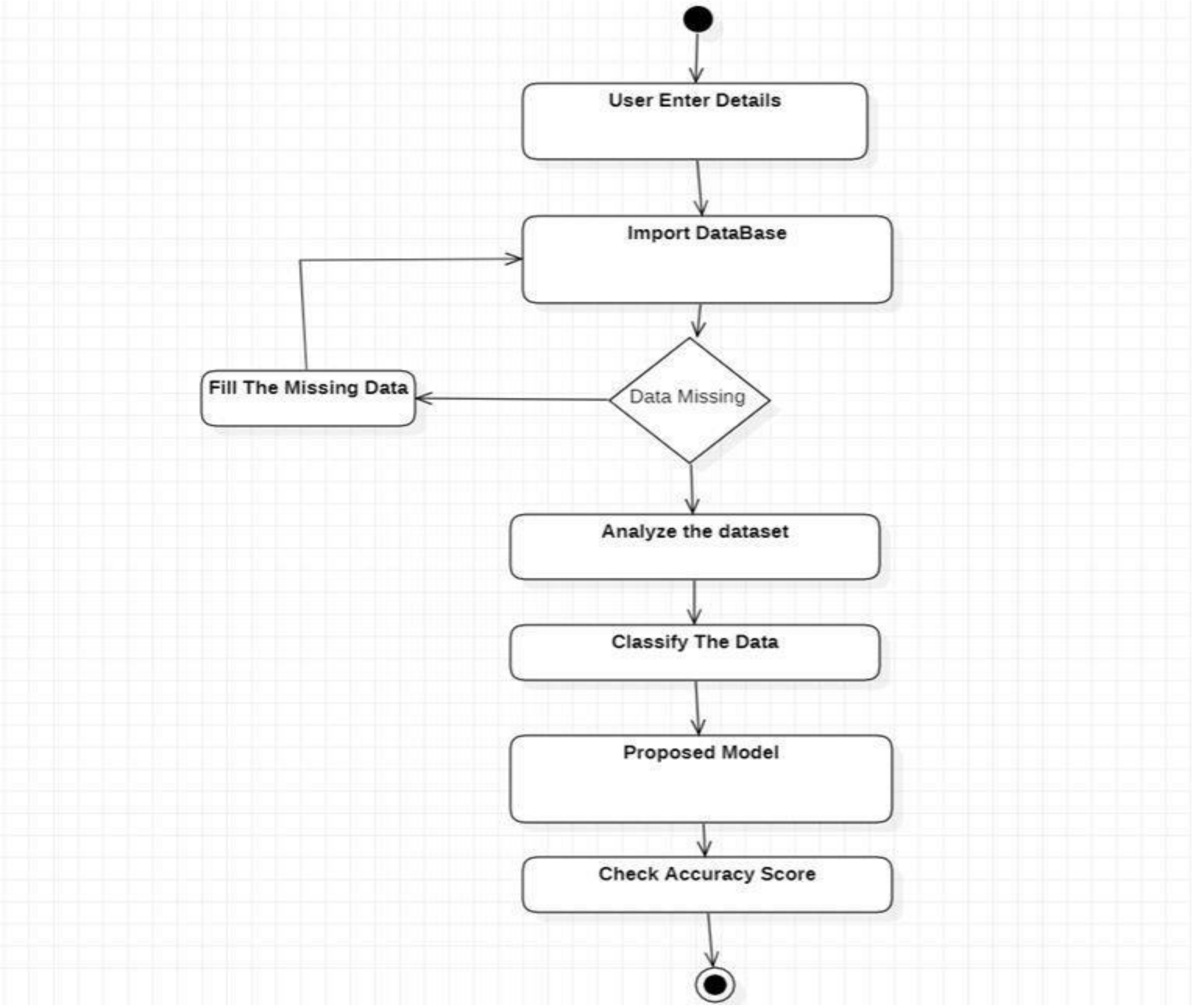


Fig. 4.4 Activity Diagram

In this diagram we depict the start and stop stages of our work in between it and will explain the flow of action through different activity blocks which also consists of condition blocks which verifies the condition.

4.1.5 Dataflow Diagram:

Data Flow Diagram is known by the initials DFD. DFD is a representation of the data flow of a system or process. Additionally, it provides information on the inputs and outputs of each entity as well as the process itself. DFD lacks control flow, loops, and decision rules. Based on the kind of data, a flowchart can describe certain processes.

It is a graphical application that may be used to communicate with users, supervisors, and other staff members. It is helpful for analyzing both current and suggested systems.

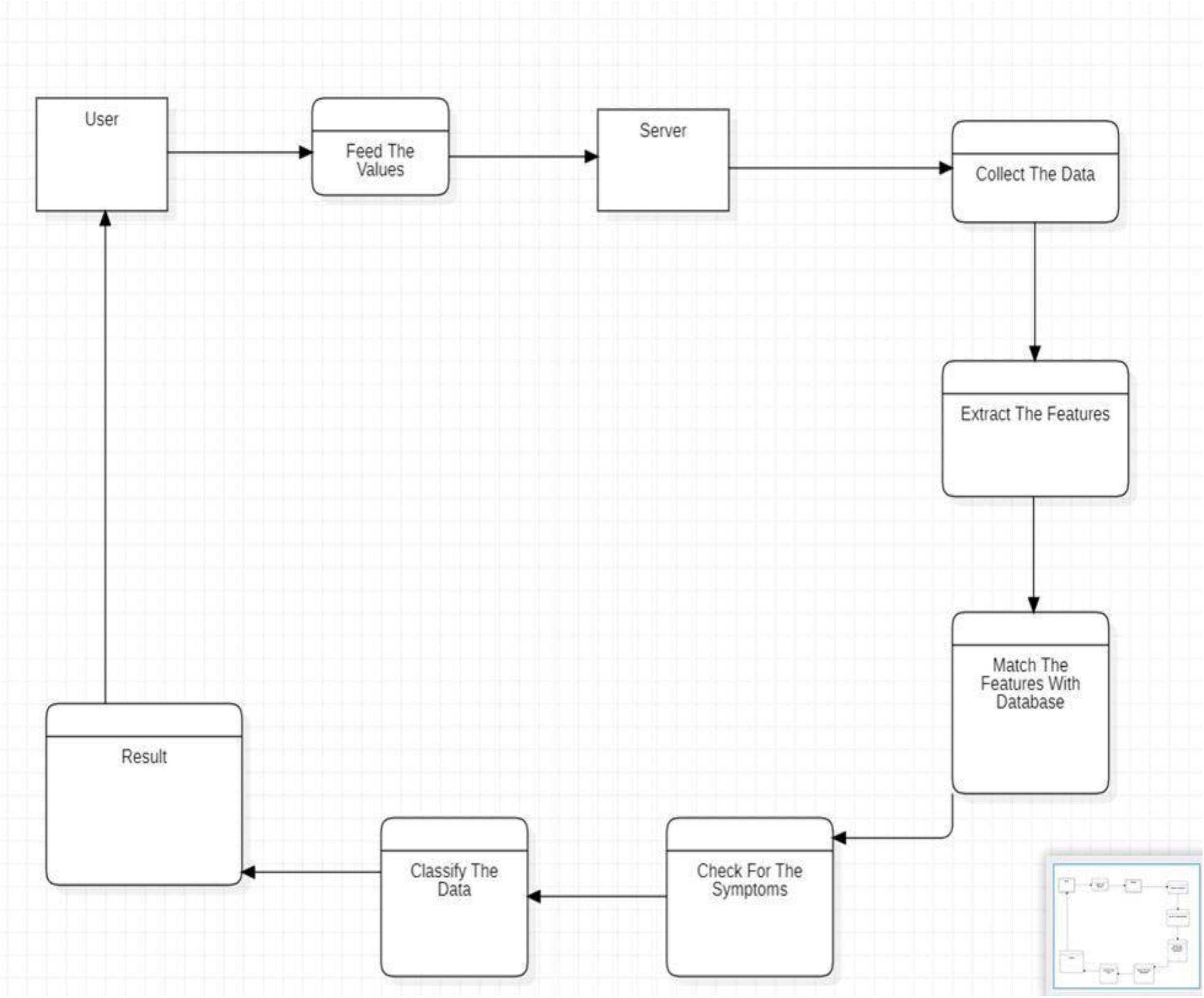


Fig. 4.5 Dataflow Diagram

It shows the interaction between the classes and objects and how the data is being transferred from one class to another and one object to another which the help of interaction lines.

4.2 Algorithm

Step-1: Collect clinical data (Cleveland dataset)

Step-2: Perform some preprocessing techniques in order to ensure the correctness of the data.

Step-3: Implement Multivariate Imputation by Chained Equation (MICE) technique in order to ensure the dataset is without missing values. This technique is implemented by using the LASSO Algorithm.

Step-4: After removing the missing values we will select the desired features from the dataset using the Decision Tree and Genetic algorithm
techniques.

Step-5: For Proper Scaling of the dataset, we used the Standard Scalar Technique which makes the mean of the attribute zero and standard deviation one.

Step-6: After proper scaling of data, we check for the balanced condition of the dataset and validate the dataset by balancing the minority class instances using the Synthetic Minority Oversampling (SMOTE) technique.

Step-7: In this phase, we have built a Machine Learning model using different machine learning algorithms like Random Forest, SVM, Ada-Boost, and Linear Regression.

Step-8: By splitting the dataset into Train and Test Dataset we trained the model which has been built in the previous phase. By using the test dataset, we verified the performance of the model.

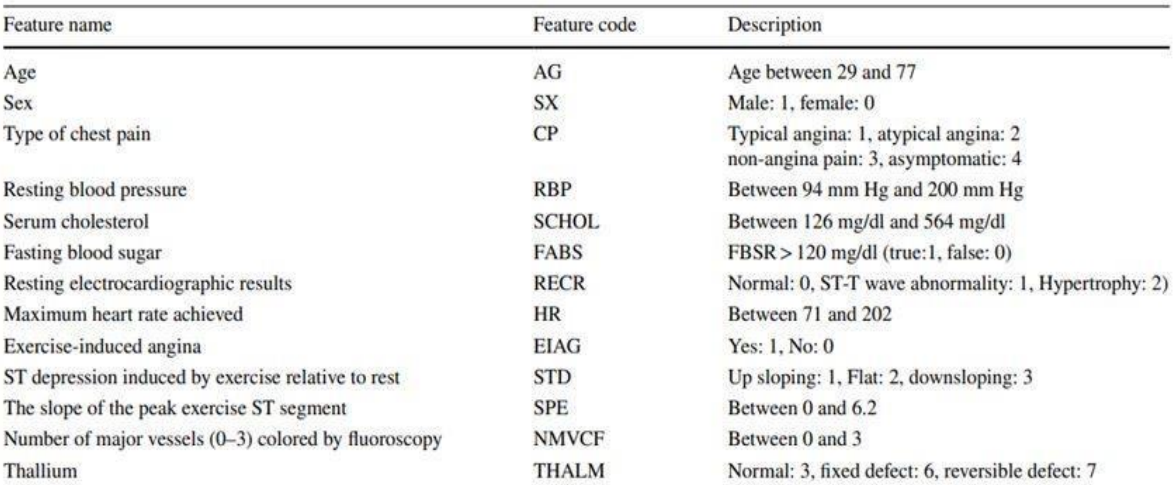
Step-9: The prediction of the heart disease happens by using the Model that we have Built.

4.3 Dataset

1. This dataset contains Heart Disease Data collected obtained from UCI (University of California, Irvine).

2. This dataset is having 14 features out of which eight are categorical features and six are numeric features.

3. Data of patients having age from 29 to 77 are collected in this dataset



Feature name	Feature code	Description
Age	AG	Age between 29 and 77
Sex	SX	Male: 1, female: 0
Type of chest pain	CP	Typical angina: 1, atypical angina: 2 non-angina pain: 3, asymptomatic: 4
Resting blood pressure	RBP	Between 94 mm Hg and 200 mm Hg
Serum cholesterol	SCHOL	Between 126 mg/dl and 564 mg/dl
Fasting blood sugar	FABS	FBSR > 120 mg/dl (true:1, false: 0)
Resting electrocardiographic results	RECR	Normal: 0, ST-T wave abnormality: 1, Hypertrophy: 2)
Maximum heart rate achieved	HR	Between 71 and 202
Exercise-induced angina	EIAG	Yes: 1, No: 0
ST depression induced by exercise relative to rest	STD	Up sloping: 1, Flat: 2, downsloping: 3
The slope of the peak exercise ST segment	SPE	Between 0 and 6.2
Number of major vessels (0–3) colored by fluoroscopy	NMVCF	Between 0 and 3
Thallium	THALM	Normal: 3, fixed defect: 6, reversible defect: 7

Fig. 4.6 Dataset

Chapter 5

CODING

5.1 Code

```
import numpy as np
from pandas._typing import F
import pandas as pd
import pickle
from scipy import stats
from sklearn.linear_model import Lasso
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
df = pd.read_csv('originalData.csv')

X=df.drop('num',axis=1)
y=df['num']
print("check whether the data has null values" , X.isnull().sum())

lr = Lasso()
imp=IterativeImputer(estimator=lr,verbose=2,max_iter=330,tol=1e-
10,imputation_order='roman')
imp.fit(X)
a=imp.transform(X)
my_array = np.array(a)
X = pd.DataFrame(my_array)
print("as a dataframe", X)
print("check whether the data has null values" , X.isnull().sum())

XY=X.drop(0,1)
XY=XY.drop(1,1)
XY=XY.drop(6,1)
```

```
print("check whether the data has null values" , XY.isnull().sum())

from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

def print_score(clf, X_train, y_train, X_test, y_test, train=True):
    if train:
        pred = clf.predict(X_train)
        clf_report = pd.DataFrame(classification_report(y_train, pred, output_dict=True))
        print("Train  Result:\n================================================")
        print(f"Accuracy Score: {accuracy_score(y_train, pred) * 100:.2f}%")
        print("________________")
        print(f"CLASSIFICATION REPORT:\n{clf_report}")
        print("________________")
        print(f"Confusion Matrix: \n {confusion_matrix(y_train, pred)}\n")

    elif train==False:
        pred = clf.predict(X_test)
        clf_report = pd.DataFrame(classification_report(y_test, pred, output_dict=True))
        print("Test Result:\n================================================")
        print(f"Accuracy Score: {accuracy_score(y_test, pred) * 100:.2f}%")
        print("________________")
        print(f"CLASSIFICATION REPORT:\n{clf_report}")
        print("________________")
        print(f"Confusion Matrix: \n {confusion_matrix(y_test, pred)}\n")



from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
scaled_data = scaler.fit_transform(XY)
print(XY)
print("scaled data" , scaled_data)
scaled_data=pd.DataFrame(scaled_data)
print("check whether scaled data has null values",scaled_data.isnull().sum())
scaled_data.mean(axis = 0)
```

```
from sklearn.model_selection import train_test_split
X=XY
X_train, X_test, Y_train, Y_test = train_test_split(X,y,test_size=0.3, stratify=y,
random_state=876)


from imblearn.over_sampling import SMOTE
import collections
counter = collections.Counter(Y_train)
print('Before', counter)

smt=SMOTE()
X_train_sm, Y_train_sm = smt.fit_resample(X_train,Y_train)

print("x",type(X_train_sm))
print("y",type(Y_train_sm))




counter = collections.Counter(Y_train_sm)
print('After', counter)



print('y train sm',Y_train_sm)



########----MODEL BUILDING --- ########
print("########----MODEL BUILDING --- ########")
print("########----MODEL BUILDING --- ########")
print("_______________________________________________________")
print("_________________________________________________________")
print("########----LogisticRegression --- ########")
print("########-----⇩⇩⇩⇩⇩⇩⇩ --########")
```

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
model.fit(X_train_sm, Y_train_sm)

X_train_prediction = model.predict(X_train_sm)



from sklearn.metrics import accuracy_score



train_data_accuracy = accuracy_score(X_train_prediction,Y_train_sm)

print("LogisticRegression training data accuracy" , train_data_accuracy)

X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)




print("LogisticRegression test data accuracy", test_data_accuracy)
print_score(model, X_train_sm, Y_train_sm, X_test, Y_test, train=True)
print_score(model, X_train, Y_train, X_test, Y_test, train=False)




print("________________________________________________")
print("_________________________________________________")
print("########----RandomForestClassifier --- ########")
print("########-----⇩⇩⇩⇩⇩⇩⇩ --########")

from sklearn.ensemble import RandomForestClassifier
modelrf= RandomForestClassifier(n_estimators= 10, criterion="entropy")
```

```
modelrf.fit(X_train_sm, Y_train_sm)
X_train_prediction = modelrf.predict(X_train_sm)



train_data_accuracy = accuracy_score(X_train_prediction,Y_train_sm)

print("RandomForestClassifier training data accuracy" , train_data_accuracy)

X_test_prediction = modelrf.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)


print("RandomForestClassifier test data accuracy" , test_data_accuracy)

print_score(modelrf, X_train_sm, Y_train_sm, X_test, Y_test, train=True)
print_score(modelrf, X_train, Y_train, X_test, Y_test, train=False)

print("______________________________________________________")
print("________________________________________________________")
print("########----AdaBoostClassifier ---########")
print("########-----⇩⇩⇩⇩⇩⇩⇩ --########")
from sklearn.ensemble import AdaBoostClassifier

model= AdaBoostClassifier(random_state=96)

model.fit(X_train_sm, Y_train_sm)
X_train_prediction = model.predict(X_train_sm)



train_data_accuracy = accuracy_score(X_train_prediction,Y_train_sm)

print("AdaBoostClassifier training data accuracy", train_data_accuracy)
```

```
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)



print("AdaBoostClassifier test data accuracy", test_data_accuracy)
print_score(model, X_train_sm, Y_train_sm, X_test, Y_test, train=True)
print_score(model, X_train, Y_train, X_test, Y_test, train=False)

print("---------------------------------------------------------")
print("----------------------------------------------------------")
print("########----SVM --- ########")
print("########-----⇩⇩⇩⇩⇩⇩⇩ --########")
from sklearn import svm
model = svm.SVC(kernel='linear')
model.fit(X_train_sm, Y_train_sm)
X_train_prediction = model.predict(X_train_sm)
######
train_data_accuracy = accuracy_score(X_train_prediction,Y_train_sm)

print("SVM training data accuracy", train_data_accuracy)




X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)



print("SVM test data accuracy", test_data_accuracy)
print_score(model, X_train_sm, Y_train_sm, X_test, Y_test, train=True)
print_score(model, X_train, Y_train, X_test, Y_test, train=False)

filename = 'heart-disease-prediction-knn-model.pkl'
pickle.dump(modelrf, open(filename, 'wb'))
```

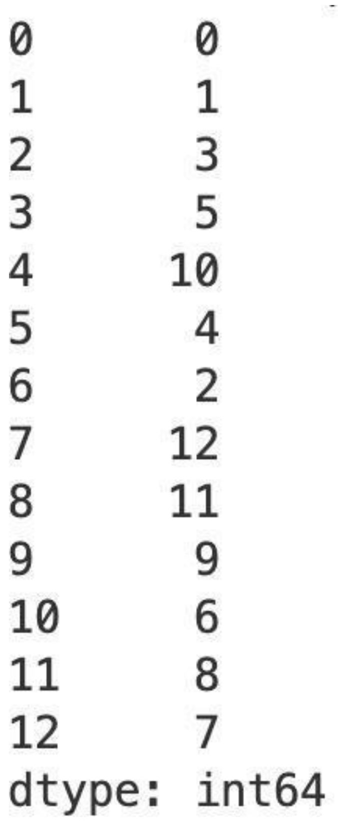
5.1.1 Fisher Score Code

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
import pandas as pd
import numpy as np
from sklearn.experimental import enable_iterative_imputer
from sklearn.linear_model import Lasso
from sklearn.impute import IterativeImputer
from sklearn.model_selection import train_test_split
from skfeature.function.similarity_based import fisher_score
import matplotlib.pyplot as plt

df = pd.read_csv('originalData.csv')
X=df.drop('num',axis=1)
y=df['num']
print(type(X))
X_train, X_test, Y_train, Y_test = train_test_split(X,y,test_size=0.2, stratify=y, random_state=2)

lr = Lasso()

imp=IterativeImputer(estimator=lr,verbose=2,max_iter=330,tol=1e-
10,imputation_order='roman')
imp.fit(X)
a=imp.transform(X)
##type(a)
my_array = np.array(a)
X = pd.DataFrame(my_array)
aX = X.to_numpy()
aY = y.to_numpy()
ranks = fisher_score.fisher_score(aX,aY)
feature_importances=pd.Series(ranks, X.columns[0:len(X.columns)])
print(feature_importances)
```

feature_importances.plot(kind='barh', color='teal')

plt.show()

5.1.2 Fisher Score Output



```
0      0
1      1
2      3
3      5
4     10
5      4
6      2
7     12
8     11
9      9
10     6
11     8
12     7
dtype: int64
```

Fig 5.1 Fisher Score Table result

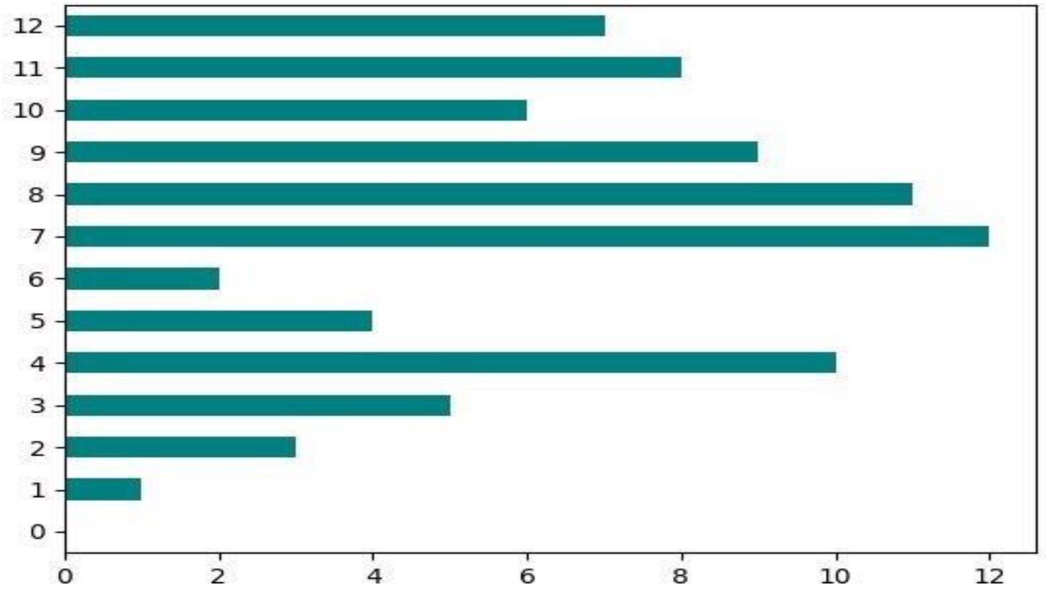


Fig 5.2 Fisher Score Graph result

5.1.3 Genetic Algorithm

```
import numpy as np
from pandas._typing import F
import pandas as pd
import pickle
from sklearn.linear_model import Lasso
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
from genetic_selection import GeneticSelectionCV
from sklearn.tree import DecisionTreeClassifier
df = pd.read_csv('yy.csv')
X=df.drop('num',axis=1)
y=df['num']
print("check whether the data has null values" , X.isnull().sum())
lr = Lasso()
imp=IterativeImputer(estimator=lr,verbose=2,max_iter=330,tol=1e-10,imputation_order='roman')
imp.fit(X)
a=imp.transform(X)
my_array = np.array(a)
X = pd.DataFrame(my_array)
print("as a dataframe", X)
print("check whether the data has null values" , X.isnull().sum())
estimator = DecisionTreeClassifier()
model1 =  GeneticSelectionCV(
    estimator, cv=5, verbose=0,
    scoring="accuracy", max_features=4,
    n_population=100, crossover_proba=5,
    mutation_proba=0.05, n_generations=18,
    crossover_independent_proba=0.5,
    mutation_independent_proba=0.04,
    tournament_size=8, n_gen_no_change=10,
    caching=True, n_jobs=-1)
model2 = model1
model = model2.fit(X,y)
model.transform(X)
my_array = np.array(X)
X = pd.DataFrame(my_array)
print('Features:', X.columns[model.support_==True])
```

5.1.1 UserInterface

```
from flask import Flask, render_template, request
import pickle
import numpy as np
filename = 'heart-disease-prediction-knn-model.pkl'
model = pickle.load(open(filename, 'rb'))
app = Flask(_name_)
@app.route('/')
def home():
        return render_template('main.html')
@app.route('/predict', methods=['GET','POST'])
def predict():
    if request.method == 'POST':

        age = int(request.form['age'])
        sex = request.form.get('sex')
        cp = request.form.get('cp')
        # trestbps = int(request.form['trestbps'])
        chol = int(request.form['chol'])
        fbs = request.form.get('fbs')
        # restecg = int(request.form['restecg'])
        thalach = int(request.form['thalach'])
        exang = request.form.get('exang')
        oldpeak = float(request.form['oldpeak'])
        slope = request.form.get('slope')
        ca = int(request.form['ca'])
        thal = request.form.get('thal')
        smoke = int(request.form['smoke'])
        alcohol = int(request.form['alcohol'])


        data = np.array([[age,sex,cp,chol,fbs,thalach,exang,oldpeak,slope,ca,thal,smoke,alcohol]])
        my_prediction = model.predict(data)

        return render_template('result.html', prediction=my_prediction)
if _name_ == '_main_':
        app.run(debug=True)
```

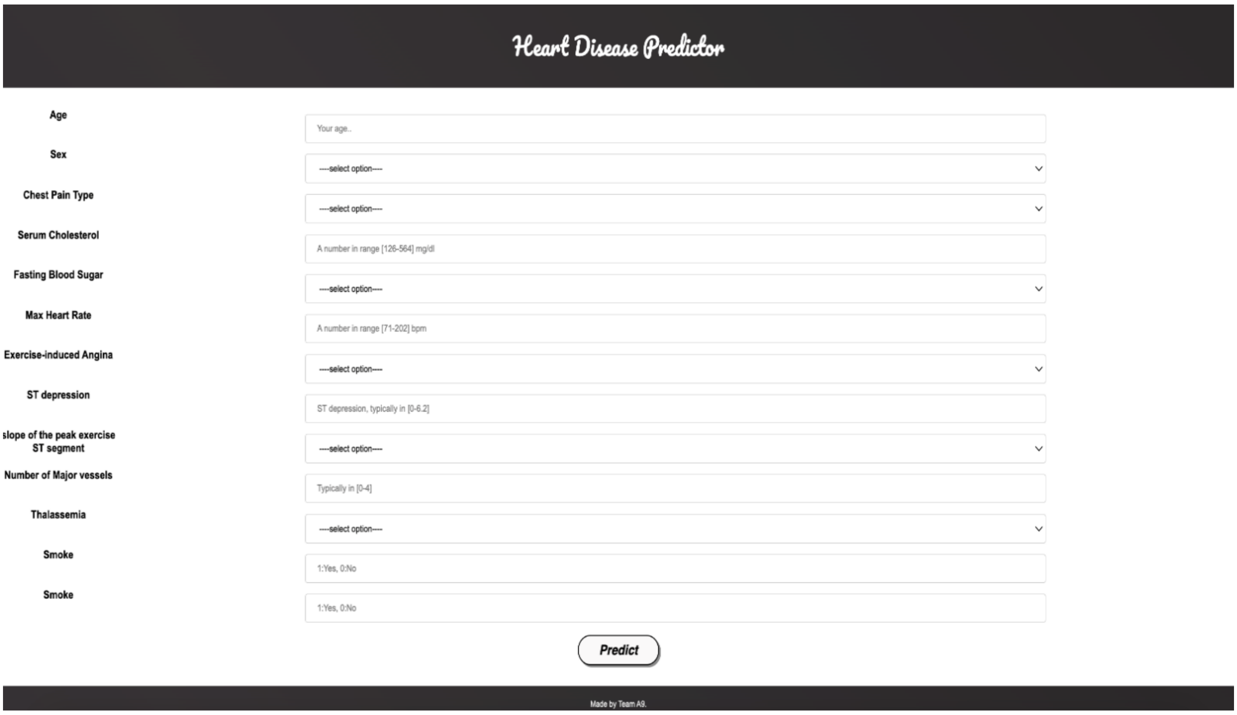
Chapter 6

IMPLEMENTATION and RESULTS

6.2 Method of Implementation

1. Enter the patient data in Heart Disease Predictor.
2. Verify the details properly according to form.
3. Click on Submit Button.
4. Then on the fresh page, you can see prediction.
5. Following Conclusions:
    - Great! You DON'T have a chance of Heart Disease
    - Oops! You have chances of Heart Disease (LOW RISK)
    - Oops! You have chances of Heart Disease (MODERATE RISK)
    - Oops! You have chances of Heart Disease (HIGH RISK)
    - Oops! You have chances of Heart Disease (EXTREMELY RISK)

6.2.1 Input

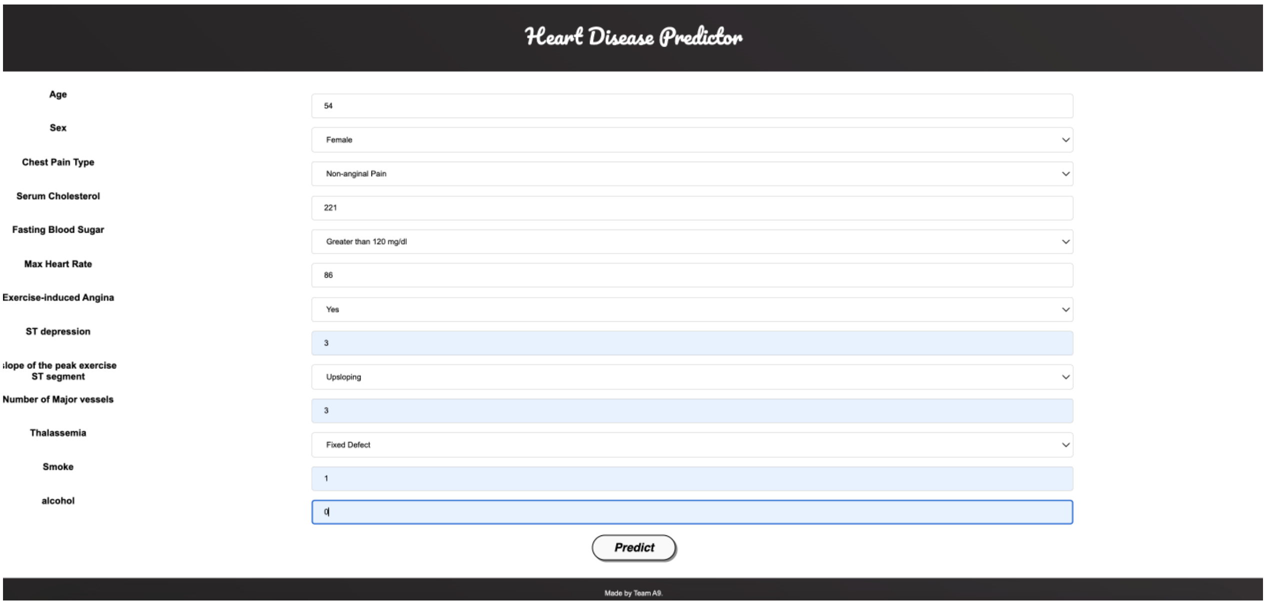


Heart Disease Predictor

Age	Your age..
Sex	----select option----
Chest Pain Type	----select option----
Serum Cholesterol	A number in range [126-564] mg/dl
Fasting Blood Sugar	----select option----
Max Heart Rate	A number in range [71-202] bpm
Exercise-induced Angina	----select option----
ST depression	ST depression, typically in [0-6.2]
slope of the peak exercise ST segment	----select option----
Number of Major vessels	Typically in [0-4]
Thalassemia	----select option----
Smoke	1:Yes, 0:No
Smoke	1:Yes, 0:No

Made by Team A9.

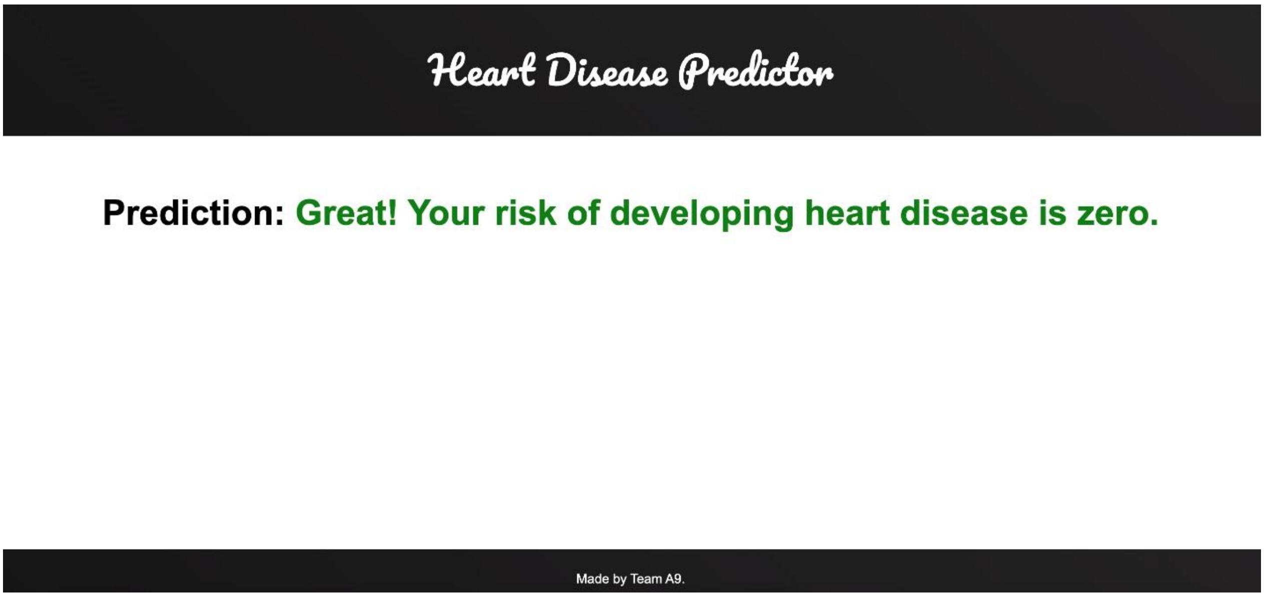
6.2.2 Inputs & Outputs



Heart Disease Predictor

Age	54
Sex	Female
Chest Pain Type	Non-anginal Pain
Serum Cholesterol	221
Fasting Blood Sugar	Greater than 120 mg/dl
Max Heart Rate	86
Exercise-induced Angina	Yes
ST depression	3
lope of the peak exercise ST segment	Upsloping
Number of Major vessels	3
Thalassemia	Fixed Defect
Smoke	1
alcohol	0

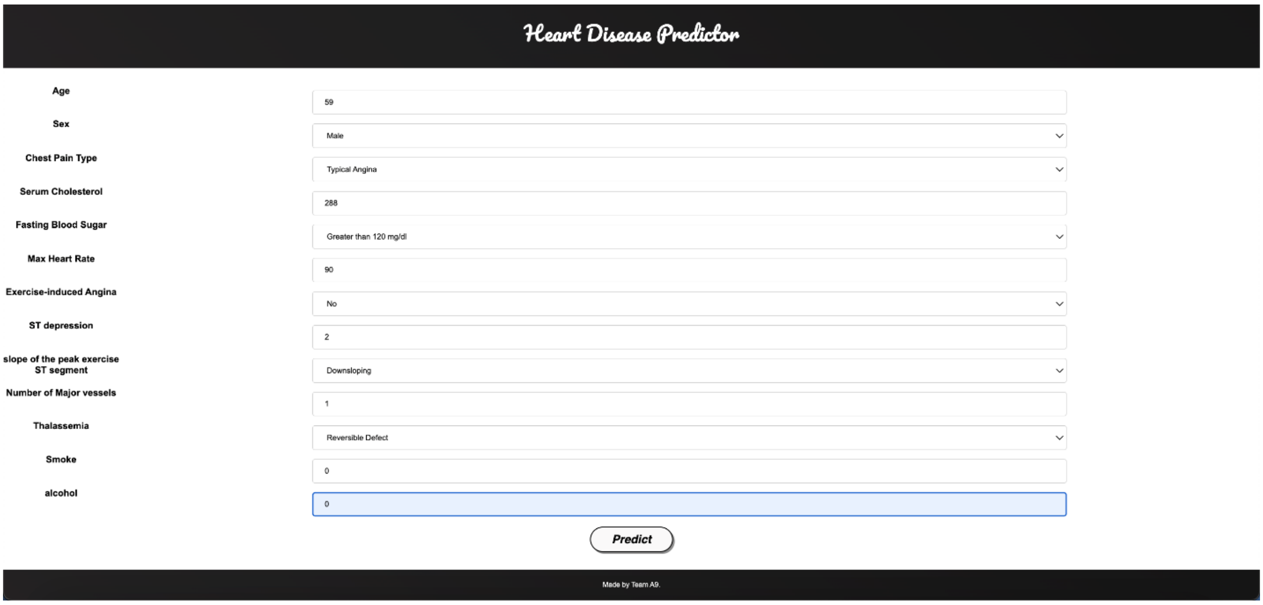
Fig 6.1 Input 1 for No Heart Disease



Heart Disease Predictor

Prediction: Great! Your risk of developing heart disease is zero.

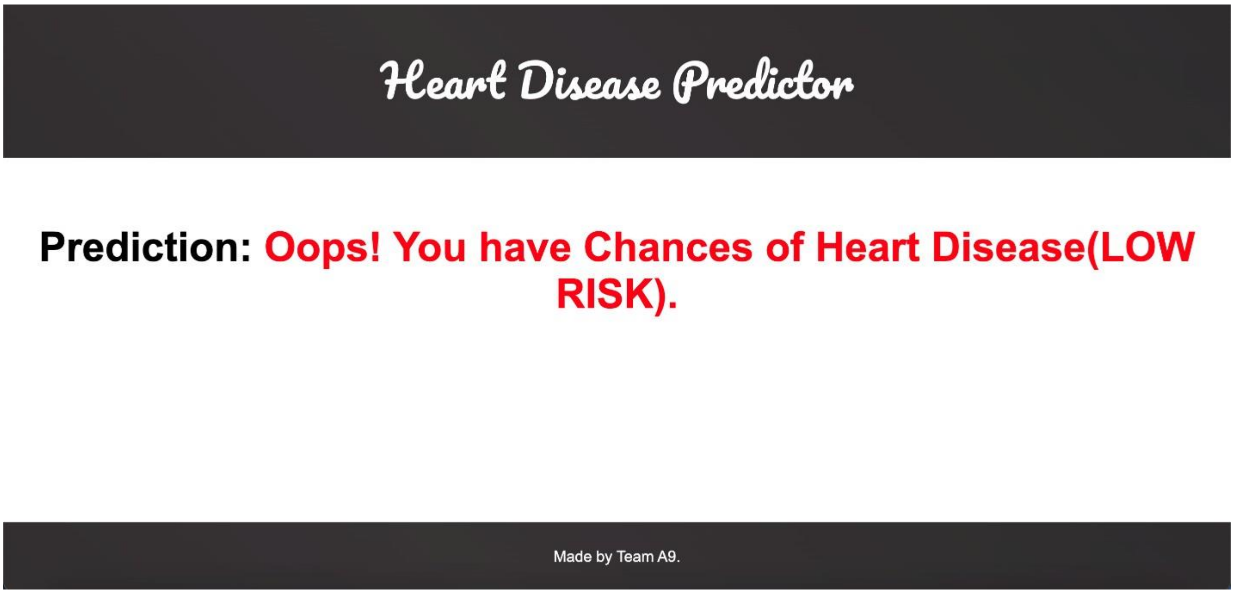
Fig 6.2 Output 1 for No Heart Disease



Heart Disease Predictor

Age	59
Sex	Male
Chest Pain Type	Typical Angina
Serum Cholesterol	288
Fasting Blood Sugar	Greater than 120 mg/dl
Max Heart Rate	90
Exercise-induced Angina	No
ST depression	2
slope of the peak exercise ST segment	Downsloping
Number of Major vessels	1
Thalassemia	Reversible Defect
Smoke	0
alcohol	0

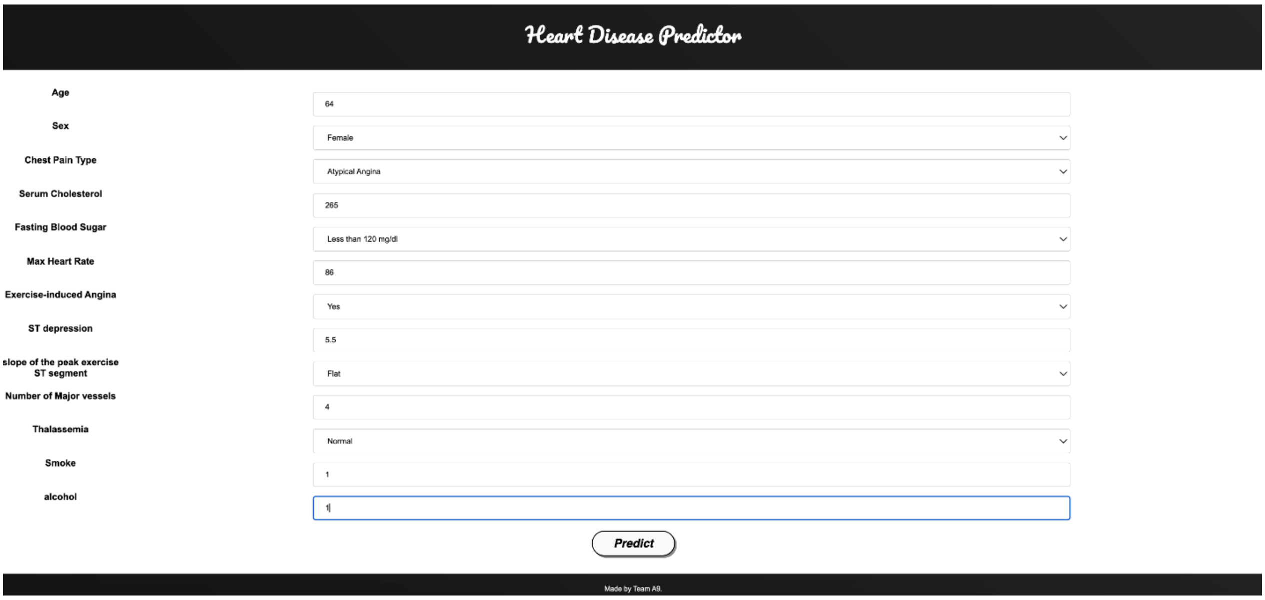
Fig 6.3 Input 2 for Heart Disease with LOW RISK



Heart Disease Predictor

Prediction: Oops! You have Chances of Heart Disease(LOW RISK).

Fig 6.4 Output 2 for Heart Disease with LOW RISK

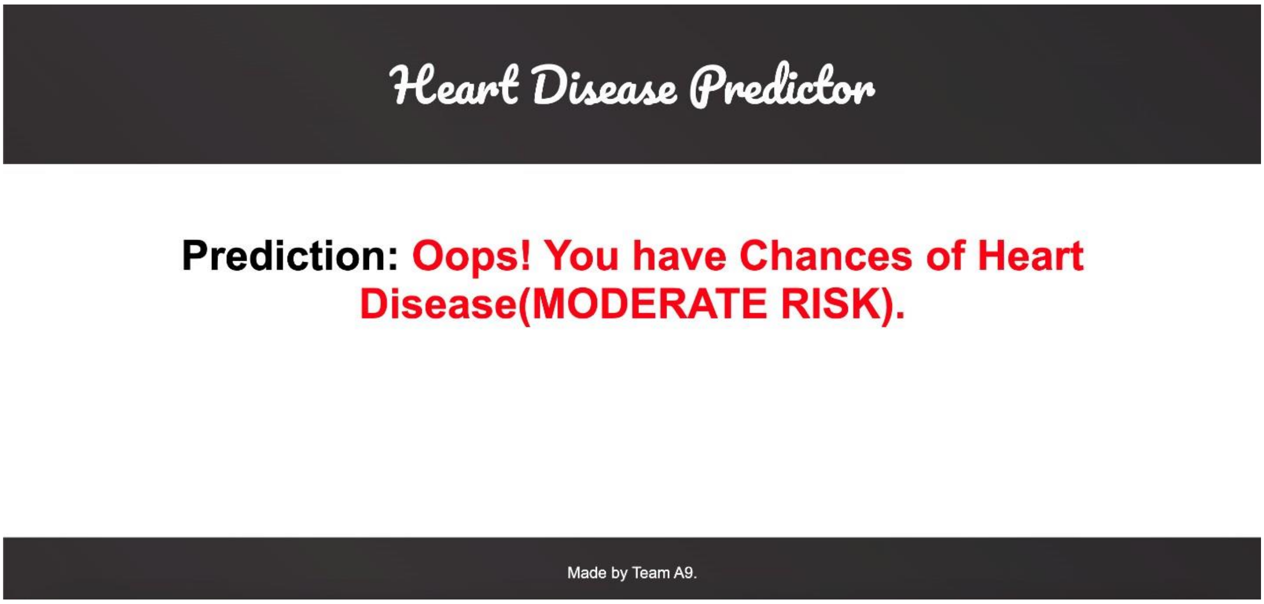


Heart Disease Predictor

Age	64
Sex	Female
Chest Pain Type	Atypical Angina
Serum Cholesterol	265
Fasting Blood Sugar	Less than 120 mg/dl
Max Heart Rate	86
Exercise-induced Angina	Yes
ST depression	5.5
slope of the peak exercise ST segment	Flat
Number of Major vessels	4
Thalassemia	Normal
Smoke	1
alcohol	1

Made by Team A9.

Fig 6.5 Input 3 for Heart Disease with MODERATE RISK

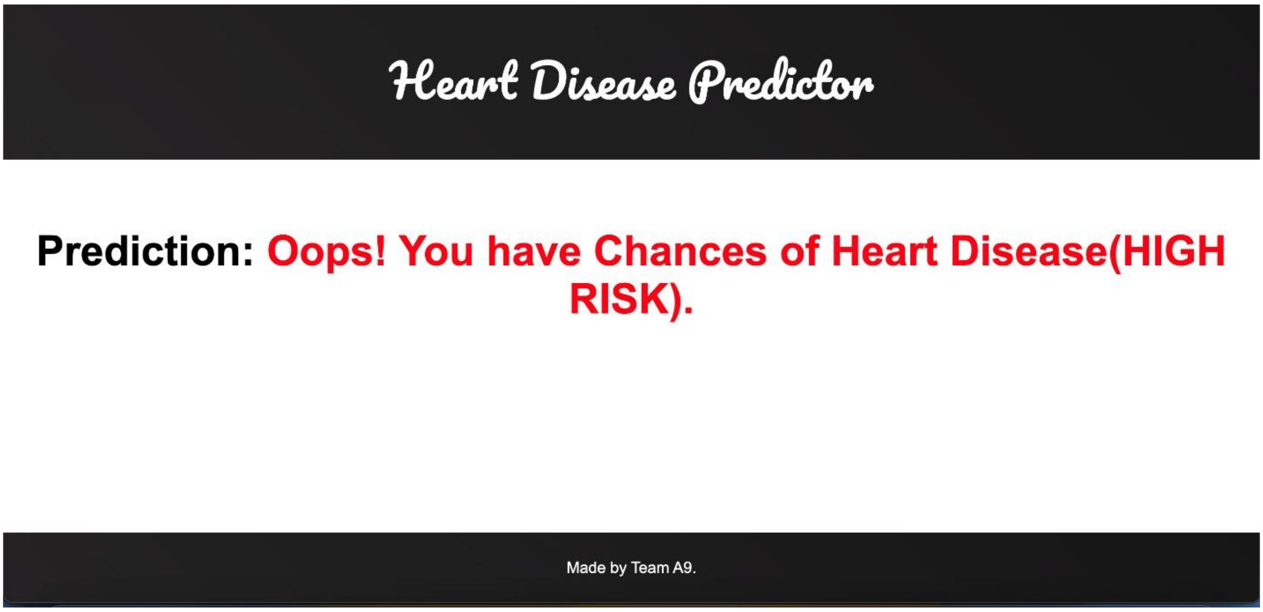


Heart Disease Predictor

Prediction: Oops! You have Chances of Heart Disease(MODERATE RISK).

Made by Team A9.

Fig 6.6 Output 3 for Heart Disease with MODERATE RISK

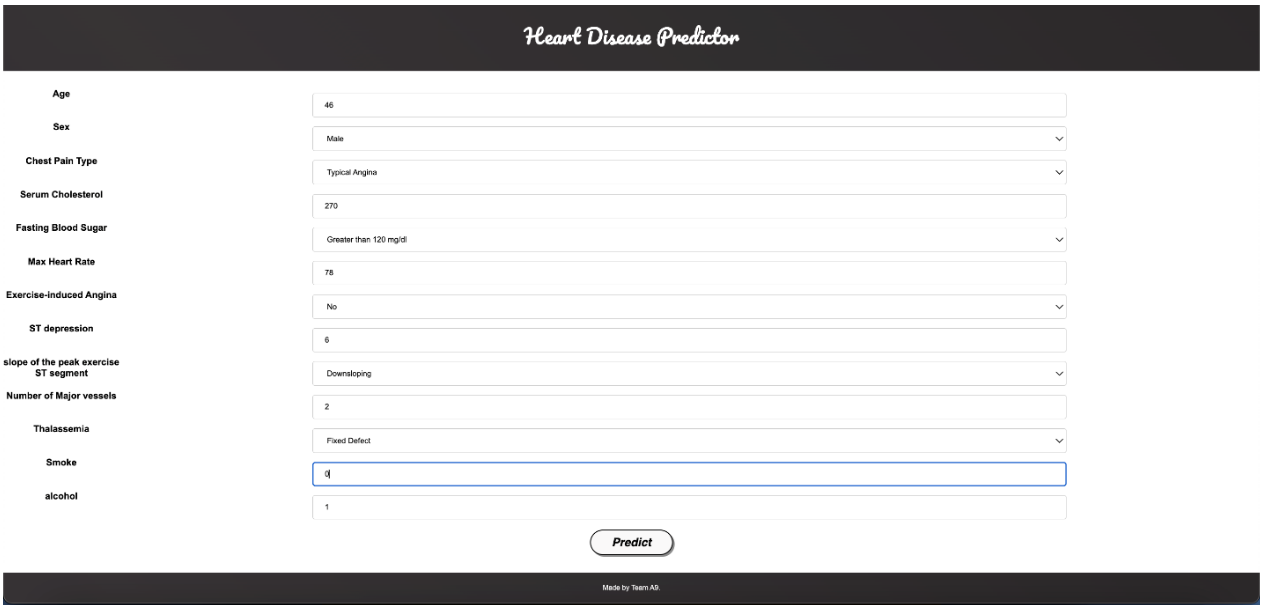


Heart Disease Predictor

Prediction: Oops! You have Chances of Heart Disease(HIGH RISK).

Made by Team A9.

Fig 6.7 Input 4 for Heart Disease with HIGH RISK

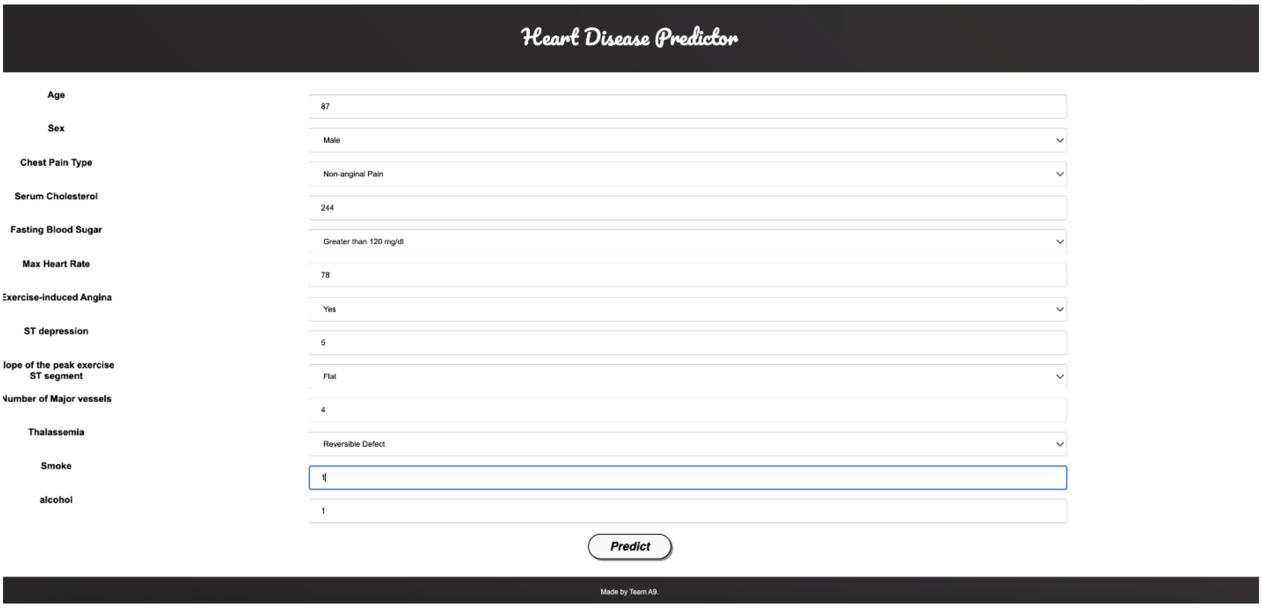


Heart Disease Predictor

Age	46
Sex	Male
Chest Pain Type	Typical Angina
Serum Cholesterol	270
Fasting Blood Sugar	Greater than 120 mg/dl
Max Heart Rate	78
Exercise-induced Angina	No
ST depression	6
slope of the peak exercise ST segment	Downsloping
Number of Major vessels	2
Thalassemia	Fixed Defect
Smoke	0
alcohol	1

Made by Team A9.

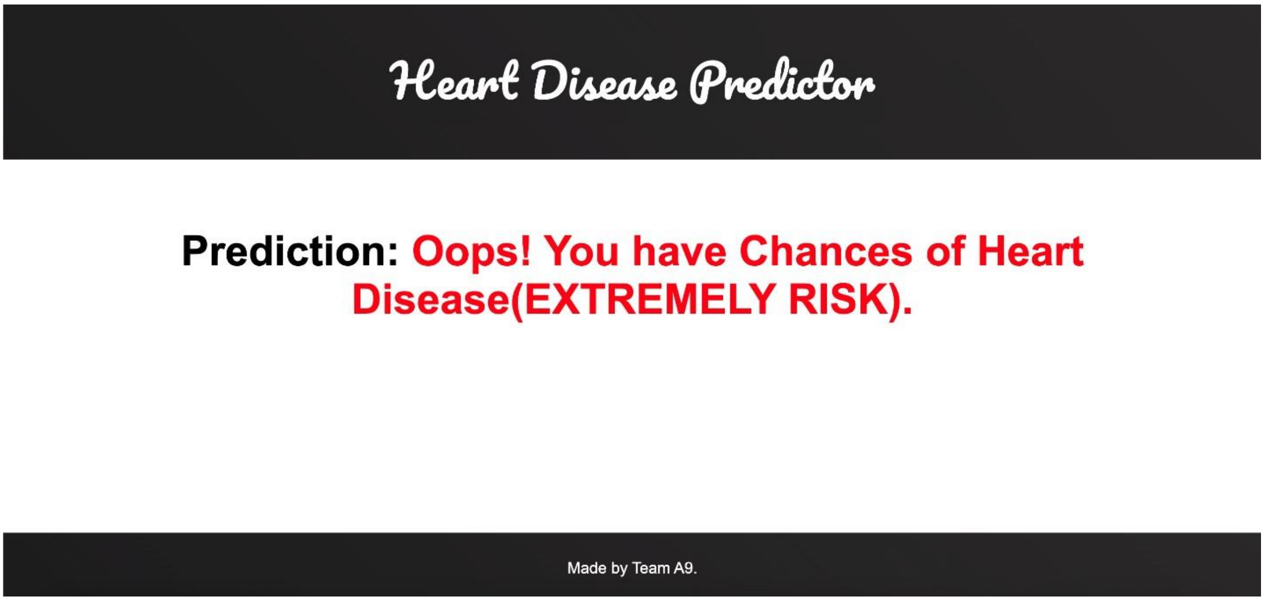
Fig 6..8 Output 4 for Heart Disease with HIGH RISK



Heart Disease Predictor

Age	87
Sex	Male
Chest Pain Type	Non-anginal Pain
Serum Cholesterol	244
Fasting Blood Sugar	Greater than 120 mg/dl
Max Heart Rate	78
xercise-induced Angina	Yes
ST depression	5
lope of the peak exercise ST segment	Flat
umber of Major vessels	4
Thalassemia	Reversible Defect
Smoke	1
alcohol	1

Fig 6.9 Input 5 for Heart Disease with EXTREMELY RISK



Heart Disease Predictor

Prediction: Oops! You have Chances of Heart Disease(EXTREMELY RISK).

Fig 6.10 Output 5 for Heart Disease with EXTREMELY RISK

6.2.3 Result Analysis

We discover that the accuracy of the Random Forest is higher compared to other algorithms after executing the machine learning technique for training and testing. The number count of TP, TN, FP, and FN is supplied, and using the equation for accuracy, value has been determined. It is concluded that Random Forest is the best with 81% accuracy, and the comparison is presented below.

TABLE: Accuracy comparison of Algorithms

Algorithm	Accuracy
Linear Regression	54.95%
Random Forest	97.85%
AdaBoost Classifier	57.14%
Support Vector Machine (SVM)	82.80%

Fig 6.11 Accuracy Table

Chapter 7

TESTING and VALIDATION

7.1 Design of Test Cases and Scenarios

Scenario 1:

We gave input field values into the field domain in the Graphical User Interface(GUI)

Scenario 2:

By using this data, we tried to validate the results by cross checking with the original value.

Scenario 3:

If the result is not desired one then we tried to train the model once again with the machine learning classifiers.

Scenario 4:

Sometimes the input field values may contain Null fields this may cause the inaccuracy of the results. So, we will move the input values to the pre-processing phase where it will bring out the efficient data for making accurate prediction.

Scenario 5:

We Validated the Accuracy of the model with the Dataset of additional features. Which in turn showed high accuracy than the dataset with limited features.

7.2 Conclusion

In all the above scenarios the redundancy, missing values and outliers in the Dataset are eradicated by using the pre-processing techniques and validated the dataset. Then this validated dataset is used for predicting the heart disease by using the Model which is been built by using the machine learning classifiers. In this way we predicted the heart disease of the user by collecting the data using the Graphical User Interface(GUI).

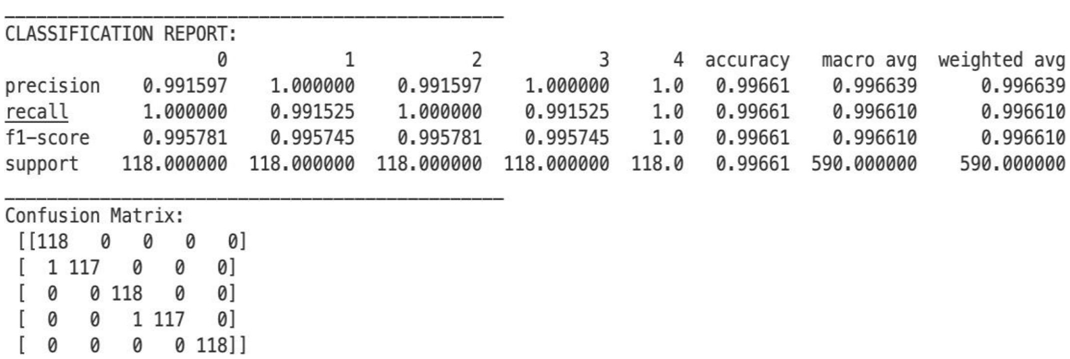
Chapter 8

CONCLUSION

- Since heart disease is a leading cause of death in India and throughout the world, the application of cutting-edge technology, such as machine learning, to the early detection of heart disease, would have a significant influence on society. Early diagnosis of heart disease can help high-risk individuals decide whether to adjust their lifestyles, which will lessen problems and represent a significant advancement in the field of medicine. Each year, there are more and more people developing cardiac illnesses. This necessitates its early detection and treatment. Patients and the medical community might both benefit greatly from the use of appropriate technology help in this area. The four machine learning methods employed in this study to gauge performance are SVM, Random Forest, Linear Regression, Ada-Boost.

- The dataset, which consists of 14 characteristics, comprises the anticipated factors that cause heart disease in patients, and 11 significant aspects that are crucial for assessing the system were chosen from them. If all the features are taken into account, the creator receives a less efficient system. The Fisher's Score is used in attribute selection to improve efficiency. In this case, n characteristics must be chosen in order to evaluate the model that provides more accuracy. Some dataset characteristics have virtually equal correlations thus, they are eliminated. The efficiency significantly declines if all the attributes in the dataset are taken into consideration.

- The below table is classification Report for Random Forest Classification algorithm which gave more accuracy when compared to other Algorithms.



```
_______________________________________________
CLASSIFICATION REPORT:
                    0           1           2           3      4  accuracy   macro avg  weighted avg
precision    0.991597    1.000000    0.991597    1.000000    1.0   0.99661    0.996639      0.996639
recall       1.000000    0.991525    1.000000    0.991525    1.0   0.99661    0.996610      0.996610
f1-score     0.995781    0.995745    0.995781    0.995745    1.0   0.99661    0.996610      0.996610
support    118.000000  118.000000  118.000000  118.000000  118.0   0.99661  590.000000    590.000000
_______________________________________________
Confusion Matrix:
 [[118   0   0   0   0]
 [  1 117   0   0   0]
 [  0   0 118   0   0]
 [  0   0   1 117   0]
 [  0   0   0   0 118]]
```

- The accuracy of each of the four machine learning techniques is evaluated, from which one assessment measures, such as the confusion matrix, accuracy, precision, recall, and f1-score, which accurately predicts the disease. The extreme Random Forest classifier has the best accuracy (97%), when compared to the other four.

REFERENCES

[1] Devansh Shah1, Samir Patel1, Santosh Kumar Bharti, “Heart Disease Prediction using Machine Learning Techniques”, *SN Computer Science,* 2 October 2020, 6 pagesAlred, G. J., Brusaw, C. T. and Oliu, W. E. [2019], *Handbook of technical writing*, Bedford/St. Martin’s Macmillan Learning.

[2] Liaqat Ali, Atiqur Rahman, Aurangzeb Khan, Mingyi Zhou,Ashir Javeed, Javed Ali Khan “An Automated Diagnostic System for Heart Disease Prediction Based on χ2 Statistical Model and Optimally Configured Deep Neural Network”, *IEEE,* 13 March 2019, 8 pages Caxton, P. [1993], ‘The title of the work’, How it was published, The address of the publisher. An optional note.

[3] Abid Ishaq, Saima Sadiq, Muhammad Umer, Saleem Ullah, Seyedali Mirjalili, Vaibhav Rupapara, Michele Nappi, “Improving the Prediction of Heart Failure Patients’ Survival Using SMOTE and Effective Data Mining Techniques”, *IEEE,* 04 March 2021, 10 pages

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

[5] EnsembleSaba Bashir, Usman Qamar, Farhan Hassan Khan, M. Younus Javed, “MV5: A Clinical Decision Support Framework for Heart DiseasePrediction Using Majority Vote Based Classifier Ensemble”, Arabian Journal for Science and Engineering, 27 August 2014, 13 pagesDuzdevich, D., Redding, S. and Greene, E. C. [2014], ‘Dna dynamics and single- molecule biology’, *Chemical reviews* **114**(6), 3072–3086.

[6] Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”, IEEE, 19 June 2019, 13 pages

[7] Nabaouia Louridi, Samira Douzi & Bouabid El Ouahidi, “Machine Learning-Based Identification of Patients with A Cardiovascular Defect”, Springer, 19 October 2021, 12pages

[8] Md Mamun Alia Bikash Kumar Paulabc Kawsar Ahmedbc Francis M.Buid Julian, M.W.Quinne Mohammad AliMoni, “Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison”, ELSEVIER, September 2021, 10 pages

[9] Rahul Katarya & Sunit Kumar Meena , “Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis”, Springer, 19 November 2020, 11 pages

[10] Norma Latif Fitriyani, Muhammad Syafrudin, Ganjar Alfian, Jongtae Rhee, “HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System”, IEEE, 20 July 2020, 17 pages

DrillBit

The Report is Generated by DrillBit Plagiarism Detection Software

Submission Information

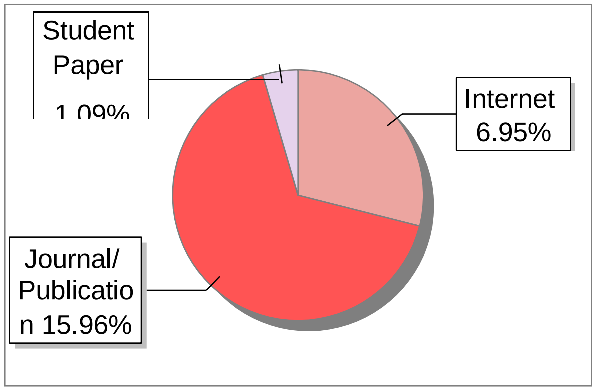
Author Name	A9 BATCH
Title	Detection of Heart Disease
Paper/Submission ID	727910
Submission Date	2023-04-27 15:51:43
Total Pages	54
Document type	Project Work

Result Information

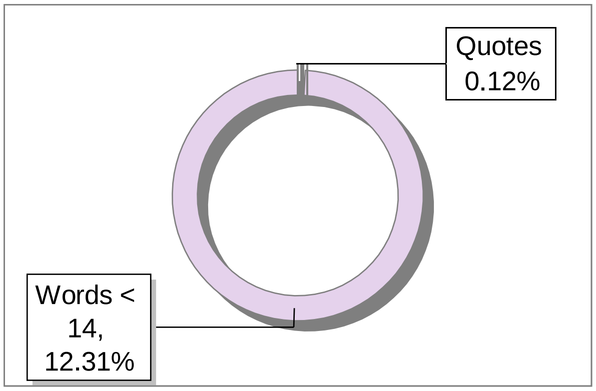
Similarity **24 %**

90

Sources Type



Report Content



Exclude Information

Quotes	Not Excluded
References/Bibliography	Not Excluded
Sources: Less than 14 Words Similarity	Not Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

24 SIMILARITY %	**125** MATCHED SOURCES	**B** GRADE	**A-Satisfactory (0-10%)** **B-Upgrade (11-40%)** **C-Poor (41-60%)** **D-Unacceptable (61-100%)**

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	HDPM An Effective Heart Disease Prediction Model for a Clinical Decision Suppor by Fitriyani-2020	3	Publication
2	Thesis Submitted to Shodhganga Repository	1	Publication
3	An Approach Based on Mutually Informed Neural Networks to Optimize the Generaliz by Ali-2020	1	Publication
4	An Approach of Predicting Heart Disease Using a Hybrid Neural Network and Decisi by Hassani-2020	1	Publication
5	Management of anticoagulation in adult patients with chronic parenchymal or vasc by Ollivier-Hourmand-2020	<1	Publication
6	Multidimensional datasets in database using search Nearest keywo By 16W86D5802 -2018, JNTUH	<1	Student Paper
7	REPOSITORY - Submitted to Jawaharlal Nehru Technological University (H) on 2023-03-12 16-42	<1	Student Paper
8	companyofscientists.com	<1	Publication
9	moam.info	<1	Internet Data
10	dspace.nwu.ac.za	<1	Publication
11	Powered Two-Wheeler Riding Profile Clustering for an In-Depth Study of Bend-Taki by Diop-2020	<1	Publication

12	USAGE OF MACHINE LEARNING FOR STRATEGIC DECISION MAKING AT HIGHER EDUCATIONAL INSTITUTUIONS BY POOJITHA, Yr - 2020	<1	Student Paper
13	asbmr.onlinelibrary.wiley.com	<1	Internet Data
14	Nomograms for two-dimensional echocardiography derived valvular and arterial dim by Cantinotti-2017	<1	Publication
15	oeno-one.eu	<1	Internet Data
16	www.ijmlc.org	<1	Publication
17	Administrative Discretion Presidential Spending Discretion and Congressional by Loui-1972	<1	Publication
18	docplayer.net	<1	Internet Data
19	www.ncbi.nlm.nih.gov	<1	Internet Data
20	Investigation of Machine Intelligence in Compound Cell Activity Classi by Fan-2019	<1	Publication
21	escholarship.org	<1	Publication
22	etd.aau.edu.et	<1	Publication
23	Artificial neural network model-based run-to-run process controller by Wang-1996	<1	Publication
24	ijrte.org	<1	Publication
25	www.ncbi.nlm.nih.gov	<1	Internet Data
26	bura.brunel.ac.uk	<1	Publication
27	moam.info	<1	Internet Data
28	moam.info	<1	Internet Data

29	biomedcentral.com	<1	Internet Data
30	Estimation of HVDC transmission lines shielding failure using LPM meth by Mohammadi-2019	<1	Publication
31	vanderbei.princeton.edu	<1	Publication
32	An enterprise modelling method based on an extension mechanism by Wang-2009	<1	Publication
33	Hierarchical predictive energy management of hybrid electric buses based on driv by Li-2020	<1	Publication
34	The Role of Bone Morphogenetic Proteins in Diabetic Complications by Perera-2019	<1	Publication
35	austlii.edu.au	<1	Internet Data
36	GBoost A novel Grading-AdaBoost ensemble approach for automatic identification by Shastri-2021	<1	Publication
37	ijircce.com	<1	Publication
38	joannalipari.com	<1	Internet Data
39	Predicting Drug-Induced Cholestasis with the Help of Hepatic TransportersAn by Kotsampasakou-2017	<1	Publication
40	quizlet.com	<1	Internet Data
41	rielac.cujae.edu.cu	<1	Publication
42	www.dx.doi.org	<1	Publication
43	www.intechopen.com	<1	Publication
44	www.sciencegate.app	<1	Internet Data
45	blog.ipleaders.in	<1	Internet Data

46	coek.info	<1	Internet Data
47	docview.dlib.vn	<1	Publication
48	e-archivo.uc3m.es	<1	Publication
49	painspecialistsaustralia.com.au	<1	Internet Data
50	www.ijeat.org	<1	Publication
51	www.iosrjournals.org	<1	Publication
52	IEEE 2017 IEEE 8th International Conference on Awareness Science an, by Chou, Tsung-Nan- 2017	<1	Publication
53	bjp.org.br	<1	Internet Data
54	coek.info	<1	Internet Data
55	dochero.tips	<1	Internet Data
56	dspace.univ-bouira.dz 8080	<1	Publication
57	Intrusion Detection with Unsupervised Techniques for Network Management Protocol by Veg-2020	<1	Publication
58	Probability Rule base Clustering Approach for Heart Disease Risk - www.ijcaonline.org	<1	Publication
59	Reducing womens cardiovascular disease risk profile by Kurth-2015	<1	Publication
60	repositorioslatinoamericanos	<1	Publication
61	www.dx.doi.org	<1	Publication
62	www.freepatentsonline.com	<1	Internet Data
63	A2b Adenosine Receptor Regulates Hyperlipidemia and Atherosclerosis by Koupenova-2012	<1	Publication

64	abnews-wire.blogspot.com	<1	Internet Data
65	An analysis of the factors that influence sugarcane yield in Northern by Dieg-2009	<1	Publication
66	Article Published in www.conferenceworld.in	<1	Publication
67	arxiv.org	<1	Publication
68	aushealthit.blogspot.com	<1	Internet Data
69	dl.gi.de	<1	Publication
70	dochero.tips	<1	Internet Data
71	escholarship.org	<1	Publication
72	france-surgery.com	<1	Internet Data
73	moam.info	<1	Internet Data
74	qdoc.tips	<1	Internet Data
75	repositorio.uam.es	<1	Publication
76	towardsdatascience.com	<1	Internet Data
77	www.mdpi.com	<1	Publication
78	moam.info	<1	Internet Data
79	moam.info	<1	Internet Data
80	sites.google.com	<1	Publication
81	www.freepatentsonline.com	<1	Internet Data
82	www.ncbi.nlm.nih.gov	<1	Internet Data

83	academicjournals.org	<1	Publication
84	asbmr.onlinelibrary.wiley.com	<1	Internet Data
85	asset-pdf.scinapse.io	<1	Publication
86	docplayer.net	<1	Internet Data
87	IEEE 2020 IEEE 2nd Eurasia Conference on Biomedical Engineering, Healthcare an	<1	Publication
88	moam.info	<1	Internet Data
89	Thesis Submitted to Shodhganga, shodhganga.inflibnet.ac.in	<1	Publication
90	Thesis Submitted to Shodhganga Repository	<1	Publication
91	www.alliedacademies.org	<1	Publication
92	www.dx.doi.org	<1	Publication
93	www.dx.doi.org	<1	Publication
94	www.dx.doi.org	<1	Publication
95	www.findyoursoulshine.com	<1	Internet Data
96	www.foodunfolded.com	<1	Internet Data
97	www.ijiet.org	<1	Publication
98	adoc.pub	<1	Internet Data
99	Article Published by International Research Journal of Engineering and Technology (IRJET) - www.irjet.net	<1	Publication
100	astesj.com	<1	Internet Data
101	atlassian.swoogo.com	<1	Internet Data

102	austlii.edu.au	<1	Internet Data
103	cardio.jmir.org	<1	Internet Data
104	clutejournals.com	<1	Publication
105	coek.info	<1	Internet Data
106	discovery.ucl.ac.uk	<1	Publication
107	ijece.iaescore.com	<1	Publication
108	ijircce.com	<1	Publication
109	journals.sbmu.ac.ir	<1	Internet Data
110	moam.info	<1	Internet Data
111	moam.info	<1	Internet Data
112	Monosodium glutamate (MSG) intake is associated with the prevalence o, by Insawang, Tonkla S- 2012	<1	Publication
113	openlibrarypublications.telkomuniversity.ac.id	<1	Publication
114	Speaking the same language The World Allergy Organization Subcutaneous Immunoth by Cox-2010	<1	Publication
115	Surprising SES Gradients in Mortality, Health, and Biomarkers in a Lat by Rosero-Bixby-2009	<1	Publication
116	Thesis Submitted to Shodhganga, shodhganga.inflibnet.ac.in	<1	Publication
117	Thesis Submitted to Shodhganga Repository	<1	Publication
118	www.ijtsrd.com	<1	Publication
119	www.nature.com	<1	Publication

120	www.ncbi.nlm.nih.gov	<1	Internet Data
121	www.ncbi.nlm.nih.gov	<1	Internet Data
122	www.researchgate.net	<1	Internet Data
123	www.sciencedirect.com	<1	Internet Data
124	American Institute of Aeronautics and Astronautics 51st AIAA Aerospa	<1	Publication
125	IEEE 2019 International Conference on Sustainable Engineering and Cr	<1	Publication