A

Project Report

on

Heart Disease Predictor

submitted as partial fulfilment for the award of

BACHELOR OF TECHNOLOGY

SESSION 2023-24

in

Computer Science and Engineering

By –

Anamika Mall (2000290100020)

Ankita Singh (2000290100024)

Ayush Bansal (2000290100039)

Under the supervision of

Prof. Hriday Kumar Gupta

KIET Group of Institutions,

Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University,

Lucknow

(Formerly UPTU)

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledgeand belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement has been made in the text.

Signature:

Name: Anamika Mall

Roll no.:2000290000020

Signature:

Name: Ankita Singh,

Roll No.: 2000290100024

Signature:

Name: Ayush Bansal

Roll No.:2000290100039

Date :

CERTIFICATE

This is to certify that the Project Report entitled "**Heart Disease Predictor**" which is submittedby **Anamika Mall**, **Ankita Singh** and **Ayush Bansal** in partial fulfilment of the requirement for the award of degree B. Tech. in the Department of Computer Science & Engineering of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Hriday Kumar Gupta
(Professor**)**

Dr. Vineet Sharma
(HOD-CSE)

Date:

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe a special debt of gratitude to **Mr. Hriday Kumar Gupta**, Department of Computer Science & Engineering, KIET, Ghaziabad, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only through his cognizant efforts that our endeavours have seen the light of the day.

We also take the opportunity to acknowledge the contribution of **Dr. Vineet Sharma**, Head of the Department of Computer Science & Engineering, KIET, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss theopportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members, of the department for their kind assistance and cooperation during the development of our project. Last but not least, we acknowledge our friends for their contribution to the completion of the project.

Date:

Signature:

Name: Anamika Mall

Roll no.:2000290000020

Signature:

Name: Ankita Singh,

Roll No.: 2000290100024

Signature:

Name: Ayush Bansal

Roll No.:2000290100039

ABSTRACT

The Heart Disease Predictor project stands as a beacon of progress in the domain of predictive medicine, offering a groundbreaking approach to preemptively addressing cardiovascular health concerns. Its remarkable achievement of a 97% accuracy rate not only sets a new benchmark but also signifies a substantial advancement in predictive analytics for assessing cardiovascular risk.

What truly sets this system apart is its sophisticated utilization of state-of-the-art statistical analysis techniques and robust feature selection methodologies. These cutting-edge tools empower the model to meticulously analyze extensive datasets, identifying the crucial factors contributing to an individual's susceptibility to cardiovascular ailments with unparalleled precision.

At the heart of its efficacy lies the richness and diversity of the dataset it draws upon. By encompassing a wide array of demographic variables—such as age and gender—and physiological indicators—such as blood pressure and cholesterol levels—the model gains a holistic understanding of each individual's cardiovascular health profile. This comprehensive approach enables the model to make nuanced and accurate predictions, thereby enhancing the effectiveness of preventive interventions.

One of the most transformative aspects of the project lies in its potential to revolutionize the approach to managing heart health. By providing a proactive means of assessing cardiovascular risk, the predictor empowers healthcare practitioners and individuals to take early intervention measures, potentially averting serious health complications. This shift towards proactive care holds immense promise for enhancing health outcomes and alleviating the burden of cardiovascular disease on individuals and healthcare systems.

In essence, the Heart Disease Predictor project transcends mere technological achievement—it symbolizes the transformative power of data-driven healthcare. By harnessing the latest advancements in machine learning and statistical analysis, it heralds a new era of personalized and effective healthcare interventions, underscoring the profound potential of data-driven approaches in shaping the future of medicine.

TABLE OF CONTENTS

	Page No.
DECLARATION……	ii
CERTIFICATE……..	iii
ACKNOWLEDGEMENTS……..……..	iv
ABSTRACT……	v
LIST OF FIGURES……	ix
LIST OF TABLES…….	x
CHAPTER 1(INTRODUCTION)…….……	1
1.1 Introduction……..	1
1.1.1 The Heart Disease Predicament……..	2
1.1.2 The Significance of Machine Learning in Healthcare…….	3
1.2 Project Description……	4
CHAPTER 2(LITERATURE REVIEW)……..……	17
2.1 Different Research Work In This Field……..……	17
2.1.1 Shah et al. (2020)…….	17
2.1.2 Hasan and Bao (2020)……	18
2.1.3 Drod et al. (2022)……...…….	18
2.1.4 Alotalibi (2019)…….……,……..	19

CHAPTER 3(PROPOSED METHODOLOGY)…….. 20

3.1 Stages Involved in The Project Development…… 20

3.2 Facilities Required……...……... 28

3.3 Healthcare Databases…….…….. 29

3.4 Machine Learning Libraries and Tools…… 32

3.4.1 Importing Libraries…… 32

3.4.2 Computational Resources…….. 32

3.4.3 User Interface and Application Development Tools……. 33

3.5 Implemented Work and Algorithm……..…… 33

3.5.1 Logistic Regression…… 33

3.5.2 Decision Tree……. 34

3.5.3 Random Forest….…….. 36

3.5.4 Support Vector Machine……. 37

CHAPTER 4(RESULTS AND DISCUSSIONS)…… 39

4.1 Testing Results…… 41

4.1.1 Result of Logistic Regression…… 41

4.1.2 Result of Decision Tree…….…… 42

4.1.3 Result of Random Forest…… 43

4.1.4 Result of Support Vector Machine……. 44

4.2 Website Status…… 46

CHAPTER 5(CONCLUSION AND FUTURE SCOPE)…….…..……. 49

5.1 Conclusion…… 49

5.2 Future Scope……..…… 49

REFERENCES…… 51

APPENDIX…….. 53

LIST OF FIGURES

Figure no.	Description	Page no.
3.1	Flow Chart	26
3.2	Histogram	29
3.3	Logistic Regression	34
3.4	Decision Tree	35
3.5	Random Forest	36
3.6	Support Vector Machine	37
4.1	Heatmap	39
4.2	Confusion Matrix	40
4.3	Confusion Matrix for Logistic Regression	41
4.4	Confusion Matrix for Decision Tree	42
4.5	Confusion Matrix for Random Forest	43
4.6	Confusion Matrix for Support Vector Machine	44
4.7	Homepage	46
4.8	Risk Detected	47
4.9	No Risk Detected	48

LIST OF TABLES

Table No.	Description	Page No.
3.1	Dataset and Range	31
4.1	Accuracy Table	45

CHAPTER-1
INTRODUCTION

1.1 Introduction

Heart disease, often known as cardiovascular disease (CVD), is one of the leading causes of death worldwide. It takes many lives each year. This general word covers a range of heart-related disorders, such as arrhythmias, congestive heart failure, and coronary artery disease. Heart disease has a significant influence on death rates as well as morbidity, quality of life, and healthcare expenditures. [1]

The continuous fight against heart disease depends heavily on risk assessment and early identification. The ability to identify those who are at risk enables prompt treatments, which may greatly enhance results and perhaps avert major issues. Risk assessment has historically depended on a number of variables, including lifestyle choices, medical histories, and clinical examinations. However, our understanding of cardiovascular risk assessment has changed dramatically as a result of the introduction of cutting-edge technology like machine learning.

As a branch of artificial intelligence, machine learning provides the capacity to examine enormous volumes of data and spot intricate patterns that could be invisible to human observers. With the use of this technology, it is possible to create prediction models that can accurately determine a person's risk of heart disease. This is a major step forward for personalised medicine since it makes it possible to customise therapies according to each patient's specific risk profile.

The "Heart Disease Prediction Model" project, which seeks to fully use this technology to fight heart disease, is an excellent example of how machine learning and healthcare may be combined. The research aims to create a prediction model that can precisely identify people at risk of developing different types of heart disease by using sophisticated algorithms and extensive datasets.The model attempts to give a comprehensive evaluation of cardiovascular risk by combining a broad variety of parameters, including clinical assessments, lifestyle factors,

medical history, and demographic data. Most importantly, extensive validation is performed on the model to guarantee its correctness and dependability in practical situations.
Such a prediction model has extensive ramifications. It not only gives medical professionals the ability to recognise and act upon situations of increased cardiovascular risk, but it also gives people the ability to take charge of their own health via lifestyle changes and educated decision-making.

To sum up, the "Heart Disease Prediction Model" project is an important initiative in the continuous battle against cardiac illness. It promises to provide more precise and individualized risk assessment by using machine learning, which will eventually lead to better results and a decrease in the burden of cardiovascular disease on both individuals and society at large.

1.1.1 The Heart Disease Predicament

Heart disease is a major global health concern that affects people and healthcare systems all around the globe. The World Health Organization has released startling numbers indicating that over 17 million fatalities worldwide are attributed to cardiovascular disorders, making them the top cause of mortality. The problem is exacerbated by the fact that heart disease often goes undetected, developing subtly until it reaches a severe and sometimes fatal stage. This emphasizes how crucial early identification and action are in the fight against this widespread health danger.

The medical community has made great efforts to uncover risk factors linked to heart disease in response to this urgent need. Traditional risk assessment methods have been established that use clinical features, blood pressure, cholesterol, age, gender, and other indicators to determine an individual's risk. Although these models have shown to be useful, they have some drawbacks, including the inability to fully reflect the complexity of cardiovascular risk.

Now introduce machine learning, a game-changing technique that might completely change how heart disease risk is predicted. Through the use of extensive datasets and advanced algorithms, machine learning algorithms may reveal complex patterns and correlations that could be missed by conventional techniques. This offers an unmatched chance to improve the precision and effectiveness of prediction models by adding more features and identifying minute details in the

data. There is great potential in the use of machine learning to cardiovascular risk assessment. These sophisticated models can offer a more thorough and individualized evaluation of a person's cardiovascular risk profile by taking into account a wider range of variables, such as genetic predispositions, lifestyle choices, and environmental factors, in addition to clinical parameters.

Furthermore, when fresh data becomes available, machine learning algorithms have the intrinsic ability to adapt, learning and improving their predictions over time. Because of their dynamic character, the models are better equipped to predict outcomes and have more therapeutic usefulness by keeping up with changing risk variables and individual differences.

To sum up, the incorporation of machine learning into the evaluation of cardiovascular risk signifies a noteworthy progression in the domain of preventive medicine. These sophisticated models have the ability to improve early diagnosis, allow tailored therapies, and eventually lessen the impact of heart disease on people and society by using data-driven insights. We are getting closer to achieving the goal of precision, proactive, and personalised medication in the battle against cardiovascular disease as we keep using machine learning in the healthcare industry.

1.1.2 The Significance of Machine Learning in Healthcare

The healthcare industry is seeing a significant shift due to the disruptive power of machine learning, which is able to analyze enormous amounts of data and derive important insights. It has advanced to the forefront of medical research and clinical practice because to its capacity to identify complex patterns and correlations within large, complicated datasets. Machine learning in particular has great potential for improving prediction models intended to address important health concerns like heart disease.

The ability of machine learning to learn from data is a fundamental feature. Through the use of comprehensive health datasets that include demographic, medical, lifestyle, and clinical measurement data, machine learning algorithms are capable of identifying latent connections and patterns that would go undetected by human observers. This makes it possible to create prediction models that are more accurate and reliable and that can identify those who are more

likely to acquire heart disease. The "Heart Disease Prediction Model," an innovative study that uses machine learning to estimate a person's risk of developing heart disease, is at the centre of this effort. In order to provide predictions, this model uses a wide range of health-related characteristics as input variables, including blood pressure, cholesterol, family history, and lifestyle choices. The approach is essential in helping to facilitate early diagnosis and risk assessment for heart-related disorders by analysing these crucial variables.

Beyond only being a predictor, the "Heart Disease Prediction Model" is significant. It represents the paradigm change in healthcare towards a data-driven approach, where evidence-based insights from thorough analysis of health data are used to guide informed decision-making. This research is a prime example of how technology may completely change how we evaluate and reduce the risk of heart disease by using machine learning.

Furthermore, the "Heart Disease Prediction Model" has implications for clinical practice as well, since it may support the diagnostic skills of medical professionals and guide the creation of individualised treatment regimens. The approach allows for proactive therapies that may prevent or minimise the advancement of cardiovascular problems, eventually improving patient outcomes and lowering healthcare costs. It does this by identifying people who are more likely to acquire heart disease at an early stage.

The "Heart Disease Prediction Model" is essentially an example of how machine learning may revolutionise the healthcare industry. This project is an example of how data-driven techniques may transform risk assessment, diagnosis, and treatment in the battle against heart disease, ushering in a new era of precision medicine and proactive healthcare management. It does this by using sophisticated algorithms and extensive health databases.

1.2 Project Description

Objective 1: Develop a Machine Learning Model: To achieve accuracy and efficacy, a strong machine learning model for heart disease risk prediction must take into account a number of important factors.

i. **Data Collection:** Obtaining a broad and comprehensive dataset with pertinent health variables is the first stage. Numerous variables, including age, gender, blood pressure, cholesterol, smoking status, family history of heart disease, and other clinical indicators, should be included in this dataset.

ii. **Data Preprocessing:** To guarantee the quality and appropriateness of the data for model training, preprocessing is required after it has been gathered. This include managing missing values, encoding categorical variables, normalising or standardising numerical features, and maybe even feature engineering to extract more relevant data.

iii. **Feature Selection:** Choosing the most relevant factors that have the most effects on heart disease risk prediction is essential since there are many possible predictors. Methods like dimensionality reduction, feature significance ranking, and statistical testing may be used to achieve this.

iv. **Model Selection:** Selecting the right machine learning algorithm is crucial to developing a predictive model that works. Support vector machines, Gradient Boosting Machines, Random Forests, and Logistic Regression are common algorithms for binary classification applications like heart disease prediction. The best-performing model architecture and algorithm for the given dataset may be found by experimenting with various models and architectures.

v. **Model Training:** After the method is chosen, the preprocessed data is used to train the model. Using methods like gradient descent or tree-based learning, the model learns during training the patterns and correlations between the input characteristics and the target variable (the presence or absence of cardiac disease).

vi. **Model Evaluation:** The model's performance must be assessed upon training using the proper measures, including area under the ROC curve (AUC), recall, accuracy, and precision. Methods of cross-validation such as k-fold cross-validation aid in determining how well the model generalises and in identifying overfitting.

vii. **Hyperparameter Tuning:** Optimising the model's performance requires fine-tuning its

hyperparameters. To do this, try out several parameter values and choose the ones that provide the best results on validation data.

viii. **Model Interpretation:** Gaining knowledge of the fundamental elements impacting the risk of heart disease requires an understanding of how the model produces its predictions. Interpretability may be greatly enhanced by methods such as feature significance analysis, partial dependency plots, and SHAP (SHapley Additive exPlanations) values.

ix. **Validation and Testing**: To properly evaluate the trained model's performance in real-world scenarios, it must lastly be verified and tested using untested data. To do this, divide the dataset into test, validation, and training sets. The test set is then used to assess how well the model performs on brand-new data.

These procedures may be used to create a strong prediction model that can precisely estimate the risk of heart disease based on a variety of health characteristics, together with the use of cutting-edge machine learning algorithms.

Objective 2: Utilize Real-world Data: Accurately predicting the risk of heart disease requires training a machine learning model using real-world health data from credible sources. The following is an explanation of this process:

i. **Identifying Reliable Sources:** Choose trustworthy sources of health information first, such as health surveys or anonymised patient records. Healthcare facilities, research groups, governmental health agencies, and publicly accessible databases from the National Institutes of Health (NIH) and the Centres for Disease Control and Prevention (CDC) are a few examples of these.

ii. **Acquiring Consent and Ensuring Compliance:** It's crucial to confirm compliance with relevant laws, such as the General Data Protection Regulation (GDPR) in the European Union or the Health Insurance Portability and Accountability Act (HIPAA) in the United States, before gaining access to any health data. Get the required authorizations and make sure that all data collection procedures follow the law and ethical guidelines, including getting participants' informed consent.

iii. **Data Collection:** Gather a wide variety of health data that are important for predicting the risk of heart disease. This covers clinical metrics like blood pressure and cholesterol levels in addition to demographic data like age and gender. For a more thorough examination, think about obtaining specific cardiac measures such as electrocardiogram (ECG) readings, echocardiography findings, or cardiac biomarker levels.

iv. **Ensuring Data Quality:** Verify the gathered data's quality and completeness to make sure it is appropriate for training models. This include doing any required data cleaning and preprocessing procedures in addition to looking for outliers, discrepancies, and missing information. Developing a trustworthy prediction model requires implementing data quality assurance procedures.

v. **Data Privacy and Security:** Enforce compliance with relevant legislation and safeguard sensitive health information by putting strong data privacy and security safeguards in place. To protect data integrity and confidentiality, this may include de-identifying or anonymizing personal health data, encrypting data while it is being stored and sent, and putting access restrictions and audit trails in place.

vi. **Data Integration and Standardization:** Combine health data from several sources into a single dataset for model training. To guarantee uniformity and compatibility between various data sources, standardise vocabulary, measuring units, and data formats. Standardisation and integration of data allow for more thorough analysis and improve the model's forecast accuracy.

vii. **Data Governance and Documentation:** Create data governance guidelines and record the metadata, version control, and data lineage tracking used in the data collecting procedures. Ensuring repeatability, and accountability in data management methods via clear documentation promotes stakeholder cooperation and knowledge exchange.

These procedures may be used to collect real-world health data from trustworthy sources, such as age, gender, blood pressure, cholesterol, and specific cardiac parameters, in order to train a machine learning model that accurately predicts the risk of heart disease.

Objective 3: Implement Machine Learning Algorithms: A crucial first step in developing a successful prediction model for heart disease risk assessment is testing and putting several machine learning algorithms into practice. The following is an explanation of this process:

i. **Algorithm Selection:** To begin, choose a variety of machine learning algorithms appropriate for binary classification problems such as the prediction of heart disease. Neural networks, Support Vector Machines, Gradient Boosting Machines, Random Forests, Decision Trees, and Logistic Regression are examples of common algorithms. Every method has advantages and disadvantages, and experimenting with several algorithms makes it possible to determine which works best for the particular dataset.

ii. **Model Training:** Apply each chosen technique and use the collected real-world health data to train distinct models. The models acquire the ability to identify patterns and correlations between the target variable—the presence or absence of heart disease—and the input features—such as age, gender, blood pressure, cholesterol levels, and so forth—from the training data.

iii. **Hyperparameter Tuning:** Adjust each algorithm's hyperparameters to maximise performance. Hyperparameters, which include regularisation strength, learning rate, and tree depth, are configuration options that regulate how the model learns. The optimal hyperparameter combinations may be found by experimenting with various combinations using grid search, random search, or Bayesian optimisation approaches.

iv. **Cross-Validation:** Utilising cross-validation methods like k-fold cross-validation, assess each trained model's performance. To do this, divide the dataset into a number of folds, train the model on a subset of the folds, then assess the model's performance on the remaining folds. In order to make sure the model functions properly on untested data, cross-validation aids in evaluating the model's capacity for generalisation and in identifying overfitting.

v. **Performance Evaluation:** Use suitable assessment criteria, such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC), to evaluate each algorithm's performance. The predicted accuracy, sensitivity, specificity, and overall performance of

the model are all shown by these indicators. To find the best algorithm for predicting heart disease, compare the performance of many algorithms.

vi. **Model Selection:** Choose the most efficient algorithm for predicting heart disease based on the findings of the performance assessment. When selecting the final model, take into account aspects like interpretability, scalability, computing efficiency, and forecast accuracy. This project makes use of the popular and easily interpreted Logistic Regression technique from the scikit-learn module for binary classification problems.

vii. **Implementation:** Use the selected algorithm to execute the chosen model after determining which one is the most effective. To get optimal prediction performance and implement the final model for practical applications, train it on the whole dataset.

The research tries to find the best model for precisely predicting heart disease risk based on numerous health indicators by methodically evaluating and applying several machine learning algorithms.

Objective 4: Create an Interactive User Interface: To guarantee efficacy and convenience of use, a number of crucial procedures and factors must be taken into account when developing a user-friendly interface that allows people to enter their health information and obtain personalised heart disease risk projections. The following is an explanation of this process:

i. **User Interface Design:** Create a user interface (UI) that is visually attractive, simple to use, and clear to navigate first. To improve accessibility across all devices and screen sizes, take into consideration using a current design language with readable font, simple iconography, and responsive layouts.

ii. **Data Input Form:** Make a user-friendly data entry form so people may submit their medical information. Divide the form into logical parts that correlate to the various categories of health variables. These sections should include clinical measures (like blood pressure and cholesterol levels), lifestyle factors (like exercise routines and smoking status), and demographic data (like age and gender). To guarantee data accuracy, use input fields with the proper validation and unambiguous labelling.

iii. **Data Validation:** Incorporate data validation procedures to guarantee that users provide accurate and consistent data. Give users immediate feedback and error warnings to help them fix any inconsistent or incorrect information. To avoid incorrect submissions, validate the data type, range, and format of the input fields.

iv. **Privacy and Security:** Put user privacy and data security first by putting strong authentication, encryption, and access control mechanisms in place. Before collecting and processing any personal health information, make sure users are aware of the privacy policy and data handling procedures. Get their permission. To protect user data, abide with applicable laws like GDPR and HIPAA.

v. **Prediction Output:** Show individualised risk estimates for heart disease depending on the data entered by the user. Provide an explanation of the interpretation and ramifications for each prediction, and provide the results in an easy-to-understand style (e.g., danger score or likelihood estimate). To assist consumers in understanding their risk level and possible preventative steps, provide context and useful insights.

vi. **Interactive Features:** Incorporate interactive elements like checkboxes for input choices, dropdown menus, and sliders to increase user interaction and engagement. Permit users to explore multiple situations and alter their input choices in order to learn how different aspects affect their risk estimates.

vii. **Visualisations:** Incorporate educational visual aids like graphs, charts, or diagrams to highlight significant findings and patterns in the data. Users may make more educated judgements about their health and comprehend complicated connections with the use of visual representations. Think about using dynamic, interactive visualisations that let consumers engage with the data.

viii. **Accessibility:** By creating an interface that is inclusive and accessible to people with disabilities, you can guarantee accessibility for users with a variety of requirements. To guarantee compliance, provide alternate text for pictures, enable keyboard navigation, and follow accessibility guidelines like WCAG (Web Content Accessibility Guidelines).

ix. **Testing and Feedback:** To find any usability problems or potential areas for improvement, do usability testing with a varied set of consumers. To understand the requirements and preferences of users, collect input via surveys, interviews, or feedback forms. Then, iterate on the interface design based on your findings.

The project may create a user-friendly interface that enables people to enter their health data with ease and obtain personalised heart disease risk projections in an understandable and helpful way by adhering to these guidelines and using user-centric design concepts.

Objective 5: Provide Patient Education and Empowerment: In order to enable users to comprehend the variables influencing their risk of heart disease and to encourage proactive healthcare behaviours, it is essential to include educational material in addition to risk projections. Here's how to explain this further:

i. **Clear and Informative Explanations:** To assist users in understanding the relevance of their risk scores or probability estimations, provide succinct and straightforward explanations in addition to the risk projections. Describe the methodology used to generate the risk projections and the variables taken into account in the study.

ii. **Risk Factors:** Inform consumers about the many heart disease risk factors, including the ones that can be changed, such stress, diabetes, high blood pressure, obesity, bad eating habits, physical inactivity, and smoking. Assist users in comprehending the significance of these variables for heart health and how they affect their total risk.

iii. **Preventive Measures:** Give them information about heart disease prevention strategies that may lower their risk of the condition. This might include leading a healthy lifestyle that includes eating a balanced diet, exercising often, giving up smoking, controlling stress, keeping a healthy weight, and keeping an eye on cholesterol and blood pressure levels.

iv. **Screening and Monitoring:** To determine their risk of heart disease and monitor changes over time, consumers should be encouraged to get frequent health monitoring. Based on age, gender, provide advice on when and how often to get testing for diabetes,

high blood pressure, and cholesterol.

v. **Treatment Options:** Provide users with information about drugs, lifestyle modifications, and medical procedures that may be used to manage heart disease and associated risk factors. Describe the advantages and possible drawbacks of various treatment modalities and urge consumers to speak with medical specialists for individualised counsel.

vi. **Behavioral Strategies:** Provide helpful advice and techniques for establishing and sustaining heart-healthy habits. Provide tools and resources to help with goal-setting, progress monitoring, and adopting sustainable lifestyle changes. Give consumers the tools they need to take charge of their health and make wise choices for their wellbeing.

vii. **Interactive Educational Content:** Interactive instructional material, such movies, infographics, quizzes, and interactive modules, may improve user engagement and understanding. To communicate knowledge in an interesting and approachable manner that accommodates various learning preferences and styles, use multimedia formats.

viii. **Access to Resources:** Give consumers access to extra resources so they may learn more about managing, treating, and preventing heart disease, such as reliable websites, publications, and instructional materials. Provide a directory of reliable sites and tools to assist people in navigating the abundance of internet health information.

ix. **Feedback and Support:** Urge users to seek clarification, pose questions, and provide comments about the risk assessments and instructional materials. Provide avenues for contact, such chatbots, forums, or helplines, so that people may speak with medical experts or qualified moderators about their issues and get assistance.

The project intends to enable users to adopt proactive healthcare behaviours, make educated choices about their heart health, and take action to lower their chance of developing heart disease by providing educational material combined with risk forecasts. This all-encompassing method of health education encourages people to prioritise their well-being and fosters a culture of prevention.

Objective 6: Validation and Testing: To guarantee the accuracy and dependability of the prediction model for assessing the risk of heart disease across a variety of datasets and patient profiles, comprehensive validation and testing must be carried out. The following is an explanation of this process:

i. **Data Partitioning:** To facilitate training, validation, and testing, divide the available dataset into distinct subgroups. The validation set is used to adjust hyperparameters and monitor model performance while the model is being developed, the testing set is used to test the finished model's performance on untested data, and the training set is used to train the model.

ii. **Cross-Validation:** Use cross-validation strategies to evaluate the model's stability and capacity for generalisation over various data subsets, such as k-fold cross-validation. To do this, the dataset is divided into k subsets successively. The model is then trained on k-1 subsets, and its performance is assessed on the remaining subset. To get reliable estimates of the model's performance, repeat this procedure k times, switching the subsets used for training and assessment each time. Then, calculate the average performance metrics.

iii. **Evaluation Metrics:** Establish suitable assessment measures, such as F1-score, accuracy, precision, recall, and area under the ROC curve (AUC), for evaluating the model's performance. These measures provide light on the predicted accuracy, sensitivity, specificity, and general performance of the model across various patient profiles and datasets.

iv. **Testing on Diverse Datasets:** Examine the model's effectiveness using a variety of datasets that reflect various demographics, communities, and healthcare environments. Make that the model performs consistently across various datasets and that it has good generalisation capabilities for previously unknown data with comparable distributions. Take into account variables like age, gender, race, region, and medical traits while choosing a variety of datasets for examination.

v. **Handling Class Imbalance:** If there are instances of class imbalance in the dataset—one

class, such as the presence of heart illness, is noticeably more common than the other class, such as the absence of heart disease—address them. Use strategies to counteract biassed model predictions and balance the class distribution, such as undersampling, oversampling, or creating synthetic data.

vi. **Robustness Testing:** Evaluate how resilient the model is to changes in the input data, including noise, missing values, outliers, and measurement errors. Evaluate the model's performance in various data quality scenarios to find any possible weak points or areas that might want improvement.

vii. **External Validation:** Utilising external datasets from unbiased sources or actual clinical situations, validate the model's performance. External validation guarantees the model's applicability in real-world circumstances and helps confirm its generalizability and dependability outside the training data.

viii. **Model Interpretability:** Examine how comprehensible the model's predictions are, and how well it can explain its choices in a clear and concise manner. In order to build acceptability and confidence among end users and medical experts, as well as to get insights into the underlying elements that contribute to heart disease risk, interpretability is crucial.

ix. **Continuous Monitoring and Update:** Provide systems for ongoing evaluation and updating of the model's performance throughout time. In order to adjust to changing patient profiles, healthcare practices, and risk factors, retrain the model with new data on a quarterly basis and constantly monitor key performance measures.

The project guarantees the validity, applicability, and efficacy of the prediction model for heart disease risk assessment by widely validating and testing it on a variety of datasets and patient profiles. This thorough validation procedure helps the model be successfully used in actual healthcare settings by fostering trust in its predicting powers.

Objective 7: Plan for Future Enhancement and Scalability: The heart disease prediction project must be kept relevant and successful over the long run, thus it is critical to discuss future

improvements that might further strengthen risk assessment and give people more confidence to prioritise their cardiovascular health. Here are some suggestions for upcoming improvements:

i. **Incorporating Genetic Information:** Add genetic data, such as genetic markers linked to cardiovascular risk or hereditary predispositions to heart disease. Personalised insights into an one's vulnerability to heart disease may be obtained by including genetic data into the prediction model, improves risk assessment.

ii. **Integration of Wearable Devices and IoT:** Include information from wearable technology, such smartwatches, fitness trackers, and Internet of Things (IoT) gadgets that measure physiological metrics like heart rate, activity level, and sleep habits. Monitoring these indicators in real-time may provide important information for early diagnosis of cardiovascular problems and dynamic risk assessment.

iii. **Expanding Demographic and Socioeconomic Factors:** Expand the model to include more socioeconomic and demographic variables, such access to healthcare services, income, employment, and education level. Taking these variables into consideration may lead to more focused therapies and a more thorough knowledge of the socioeconomic differences in heart disease risk.

iv. **Integration with Electronic Health Records (EHRs):** Use the predictive model to provide individualised risk assessment at the point of service and to streamline data interchange with electronic health record systems used in healthcare settings. Accuracy and usefulness of the model in clinical practice may be improved by using EHR data.

v. **Predictive Analytics for Treatment Response:** Create predictive analytics models to evaluate each patient's reaction to treatment plans for managing and preventing heart disease. These models may aid in the personalisation of treatment plans and the optimization of therapy techniques for improved results by tracking treatment outcomes and patient reactions over time.

vi. **Machine Learning Model Interpretability:** By using explainable AI methodologies, you may improve the machine learning model's interpretability. Give consumers clear

explanations of the model's predictions, emphasising the important elements that go into each user's risk assessment. This will help users comprehend the reasoning behind the forecasts.

vii. **Longitudinal Risk Assessment:** Use longitudinal risk assessment tools to monitor changes in the risk of heart disease over time and spot patterns or trends that might point to the advancement or regression of risk factors. Early identification of changes in health status and the facilitation of proactive risk-reduction measures are made possible by longitudinal monitoring.

viii. **Continuous Model Monitoring and Updates:** Provide a structure for ongoing model updates and monitoring in order to adjust to new risk variables, medical standards, and healthcare procedures. To maintain the model's applicability and efficacy in estimating the risk of heart disease in a changing healthcare environment, update it often using fresh information and insights.

The heart disease prediction project may develop into a full solution that uses data-driven insights to better risk assessment, encourage proactive healthcare, and enable people to properly prioritise their cardiovascular health by including these future developments. This innovative strategy guarantees the project's sustainability and applicability in reducing the worldwide burden of heart disease.

CHAPTER-2
LITERATURE REVIEW

The use of data mining and machine learning methods in the healthcare sector has increased recently, especially in the area of medical cardiology. This represents a paradigm change in the way that cardiovascular disease risk assessment and prediction are done. The research projects by Hasan and Bao (2020), Drod et al. (2022), Shah et al. (2020), and Alotalibi (2019) show how machine learning is being used more and more to create prediction models and find important risk factors for cardiovascular disease.

2.1 Different Research Work in this field

2.1.1 Shah et al. (2020):

Methodology: Shah et al. used machine learning methods in their work to create a prediction model for cardiovascular disease. They made use of the Cleveland Heart Disease dataset, which comes from the UCI machine learning repository and has 17 characteristics and 303 occurrences[2]. In order to determine which supervised classification technique was the most accurate, the authors investigated a number of approaches, such as naive Bayes, decision trees, random forests, and k-nearest neighbour (KKN).

Results: Of all the algorithms that were examined, the KKN model had the best accuracy (90.8%). This demonstrates how well machine learning predicts the risk of cardiovascular illness. The research highlights how crucial model selection is to getting the best possible prediction performance.

Implications: The results imply that machine learning methods may be useful instruments for the prediction of cardiovascular disease. The KKN model's high accuracy indicates how well it can identify risk in clinical settings, which might result in early intervention and improved patient outcomes.

2.1.2 Hasan and Bao (2020):

Methodology: The goal of Hasan and Bao's research was to develop effective feature selection strategies for cardiovascular disease prediction. In addition to evaluating the effectiveness of many classification models, such as random forest, support vector classifier, k-nearest neighbours, naive Bayes, and XGBoost, they used three well-known feature selection techniques: filter, wrapper, and embedding.

Findings: According to the research, the XGBoost classifier using the wrapper method had the best accuracy, attaining 73.74%. This emphasises how important feature selection is to improving cardiovascular disease predictive analytics.

Implications: Effective feature selection strategies may raise the prediction models' interpretability and accuracy for cardiovascular disease. The study emphasises how crucial methodological factors are when developing a model and proposes that the XGBoost classifier with wrapper feature selection might be a useful strategy for further exploration in this field.[3]

2.1.3 Drod et al. (2022):

Methodology: Using machine learning approaches, Drod et al. sought to uncover major risk factors for cardiovascular disease in individuals with metabolic-associated fatty liver disease (MAFLD). In order to create a prediction model, they used principle component analysis (PCA), univariate feature ranking, and multiple logistic regression classifier analysis on data from 191 MAFLD patients.

Findings: With an AUC of 0.87, the machine learning method successfully distinguished between high-risk and low-risk individuals, indicating its efficacy in risk stratification for cardiovascular disease in certain patient groups.

Implications: The work demonstrates how machine learning methods may be used to find new cardiovascular disease risk factors in patient cohorts with complicated characteristics. The

prediction model provides insights into customised risk assessment and treatment techniques catered to certain patient demographics by integrating data from MAFLD patients.

2.1.4 Alotalibi (2019):

Methodology: Alotalibi used a dataset from the Cleveland Clinic Foundation to study the usefulness of machine learning approaches for heart failure disease prediction. They used a 10-fold cross-validation technique for developing their models and applied a number of machine learning (ML) methods, such as decision trees, logistic regression, random forests, naive Bayes, and support vector machines (SVM).

Findings: The decision tree method had the best accuracy, according to the research, at 93.19%, with the SVM algorithm coming in second at 92.30%. This emphasises how useful machine learning approaches may be in predicting cardiac failure.

Implications: The results indicate that heart failure illness may be reliably predicted by machine learning algorithms, and the decision tree approach shows potential for use in clinical risk assessment and prediction modelling research in the future. This demonstrates how machine learning approaches may help medical personnel identify patients who are at risk of heart failure and apply tailored therapies to enhance patient outcomes. All things considered, these research demonstrate how effective machine learning methods are in predicting and assessing the risk of cardiovascular disease. Researchers have made great progress in precisely identifying people who are at risk of cardiovascular events by using a variety of datasets and sophisticated algorithms. This has opened the door for tailored preventative measures and better patient outcomes in the field of cardiology. The biology of cardiovascular illness may be better understood and clinical decision-making in cardiovascular care might be improved with further study in this field.[4]

CHAPTER-3
PROPOSED METHODOLOGY

The "Heart Disease Prediction Model" was developed and implemented using an iterative, systematic process that was carefully thought out to guarantee accuracy and reliability when predicting an individual's risk of developing heart disease.

3.1 Stages involved in the Project Development

Stage 1: Data Collection:

The project's first step involves careful attention to collecting and organising data, which is the foundational element needed to produce the Heart Disease Prediction Model. The first step in this approach is to choose reliable datasets from reliable sources, including the Framingham Heart Study and Cleveland Heart Disease databases.

To make sure they are appropriate for the goals of the study, these datasets are selected using a number of factors. Aspects taken into account include the data's accessibility, its quantity and breadth, and its applicability to the project's particular objectives. Furthermore, the selection process is heavily influenced by ethical concerns about patient confidentiality and privacy, which guarantee that all data collection methods follow set standards and laws.

Following the identification of the datasets, a thorough curation procedure is implemented to guarantee data quality and extract pertinent information. This include fixing any mistakes or inconsistencies with the data, standardising formats, and filling in any missing numbers.

The carefully selected datasets include a wide range of patient characteristics, including demographic information on age, gender, and ethnicity as well as critical health markers like blood pressure, cholesterol, and electrocardiogram (ECG) values. These databases provide a comprehensive picture of each person's unique cardiovascular health profile by including a wide variety of characteristics, making forecasts and risk assessments more precise.

All things considered, the painstaking process of gathering and organising data is the mainstay of the project, providing a strong basis for the latter stages of developing and implementing the model. Researchers can increase the efficacy and dependability of the Heart Disease Prediction Model by guaranteeing the quality and integrity of the data, which will eventually result in better patient outcomes and more informed decision-making.

Stage 2: Data Preprocessing:

Because of flaws and discrepancies, raw healthcare data may be problematic in the field of medicine, as data is often gathered from several sources and systems. As a result, thorough preparation is necessary to guarantee that the data is appropriate for machine learning analysis, which in turn improves the model's resilience and forecast accuracy.

Prior to being fed into machine learning algorithms, the data must be cleaned, transformed, and standardised. This process is known as data preparation. A few of the crucial methods consist of:

i. **Data Cleaning:** This entails dealing with missing values, which are often present in healthcare datasets as a result of issues like measurement mistakes or incomplete patient records. To fill in the blanks, imputation techniques like mean, median, and predictive imputation may be used. Furthermore, outliers may be found and either fixed or eliminated from the dataset since they can skew analysis and modelling.

ii. **Normalization and Feature Scaling:** In order to avoid certain features from predominating over others during the model training process, normalisation and feature scaling approaches are used to make sure that all features have a comparable scale. Regular methods of normalisation include z-score normalisation and min-max scaling, which rescale features to a specified standard deviation or range, respectively.

iii. **Encoding Categorical Variables:** In order to make categorical variables compatible with machine learning algorithms, they must be transformed into a numerical format, such as gender or medical diagnosis. Ordinal encoding and one-hot encoding are popular methods for this. Ordinal encoding allocates numbers to categories according to their frequency or

order, while one-hot encoding generates binary columns for every category.

iv. **Handling Text and Free-Form Data:** Textual data in the healthcare industry, such clinical notes or medical reports, may include important information, but they also need for specific preparation methods. Tokenization, stemming, lemmatization, and other natural language processing (NLP) methods may be used to extract relevant characteristics from text input and transform it into an organised format that is appropriate for machine learning research.

v. **Dealing with Class Imbalance:** There may be an imbalance in the number of classes in some healthcare datasets, with one class substantially outnumbering the others. Predictions may be skewed as a result of this imbalance and its impact on model performance. To resolve class imbalance and enhance model performance, strategies including class weighting, resampling (oversampling or undersampling), and creating synthetic data may be used.

Healthcare data may be cleaned, processed, and standardised using these pretreatment procedures to guarantee its quality and suitability for machine learning algorithms. This crucial stage creates the groundwork for developing reliable and accurate prediction models, which will eventually improve patient care and healthcare outcomes.

Stage 3: Model Selection:

The emphasis moves to choosing the best machine learning algorithm for forecasting the risk of heart disease once the preprocessed data is ready. This stage entails a careful assessment of many algorithms to see which ones are most appropriate for the given job.

Many machine learning methods are taken into consideration, such as Random Forest, Support Vector Machines (SVM), Gradient Boosting, Logistic Regression, and others. Every algorithm is carefully examined in light of many crucial factors:

i. **Performance Metrics:** Suitable criteria including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) are used to assess

each algorithm's performance. These measures provide light on how well the algorithm is able to identify people as having heart disease or not.

ii. **Interpretability:** The predictability of the algorithm must be interpreted, especially in the healthcare industry where explainability and transparency are critical. Healthcare practitioners may make more informed decisions when using algorithms that provide findings that are easy to read and that help them comprehend the underlying elements that increase the risk of heart disease.

iii. **Computational Efficiency:** Each algorithm's computing efficiency is taken into account, particularly when handling enormous amounts of data. Computationally efficient algorithms may handle data rapidly and efficiently, cutting down on the amount of time needed for model training and prediction.

iv. **Binary Classification Task:** Algorithms must be appropriate for this particular job as it entails binary classification—predicting whether cardiac disease exists or not. We carefully assess the algorithm's ability to correctly identify people into the two groups.

v. **Scalability:** In order to support prospective future improvements and extensive implementation, the scalability of the selected method is also considered. To guarantee long-term usefulness and relevance, algorithms that can scale well to manage growing volumes of data or accept new characteristics are desirable.

Researchers can determine which machine learning algorithm is best for forecasting the risk of heart disease by thoroughly evaluating many algorithms based on these parameters. This guarantees the prediction model's accuracy and dependability as well as its interpretability, computational efficiency, and scalability, setting the stage for successful risk assessment and individualised healthcare treatments.

Stage 4: Model Training:

Following the selection of the best machine learning algorithm for estimating the risk of heart disease, the model of choice is put through a rigorous training procedure utilising the carefully

selected dataset. The basis for the model's prediction powers and its precision in categorising people as having or not having heart disease is laid during this crucial training phase.

Three subsets of the dataset are separated out: training, validation, and testing. The model is trained on the available data using the training subset, allowing it to discover underlying patterns and correlations between the input characteristics and the target variable (heart disease presence or absence). Model assessment and hyperparameter adjustment are carried out in the interim using the validation subset. Model parameters are optimised and prediction accuracy is maximised at this stage by using techniques like grid search and cross-validation.

To guarantee the efficacy and resilience of the trained model, a number of crucial factors are taken into account throughout the training process:

i. **Mitigating Overfitting and Underfitting:** Poor generalisation on unseen data results from overfitting, which happens when the model becomes adept at memorising the training set. On the other hand, underfitting occurs when the model is too straightforward to identify the fundamental patterns in the data. Regularisation, dropout, and early stopping are some of the techniques used to keep the model from overfitting or underfitting, which improves its generalizability to a variety of datasets and patient groups.

ii. **Model Evaluation:** The model's performance is regularly assessed during training using suitable measures including F1-score, AUC-ROC, accuracy, precision, and recall. These measures show where the model might be improved and provide insights into how well it predicts the risk of heart disease.

iii. **Generalizability:** It is critical to ensure that the trained model is generalizable, especially in healthcare applications where the model's capacity to function on unseen data is vital. Researchers try to create a model that can generalise to new, unknown data and a variety of patient groups by dividing the dataset into subsets for training, validation, and testing and using approaches to reduce overfitting and underfitting.

Through rigorous training of the selected model on the carefully selected dataset and the use of techniques to mitigate overfitting and underfitting, scientists may create a strong predictive model for assessing the risk of heart disease. Healthcare professionals may use this model as a useful tool to help them make choices and carry out tailored treatments that will enhance patient outcomes.

Stage 5: Deployment:

The trained model is deployed into an application that is easy to use and meant for both people and healthcare professionals, signifying the project's final phase. With the help of this application, users can easily enter their health-related characteristics and obtain personalised estimates about their risk of heart disease.

During the application's development, the following factors were crucial:

i. **Intuitive Interface:** The program's interface is clear and easy to use, making it easier to enter health-related data and comprehend the predictions that are made. A smooth user experience is facilitated by intuitive controls, well-organized layouts, and clear navigation.

ii. **Convenient Input:** Users may enter their health features, including age, gender, blood pressure readings, cholesterol levels, and other pertinent data, into simple-to-use forms or input areas. To reduce user effort and possible mistakes, the entry procedure is designed to be simple and easy.

iii. **Personalized Predictions:** The trained model creates individualised estimates about the user's risk of heart disease based on health variables submitted. Users are able to appreciate their risk level and take necessary action since these forecasts are given in an easily comprehensible style.

iv. **Informative Insights:** Apart from prognosticating risks, the programme offers insightful information on the variables that influence an individual's likelihood of developing heart disease. This might include elucidations of the ways in which certain health

characteristics impact risk, together with suggested preventative actions and lifestyle adjustments to lower risk.

v. **User Engagement:** In order to create an application interface that is both readable and entertaining, user experience design concepts are essential. In order to improve user pleasure and engagement, consideration is given to elements including feedback systems, interaction, and visual design. Rewards or gamification components may also be included to encourage consistent use and adherence to preventative measures.

All things considered, the project's climax is the deployment of the trained model into an intuitive application, which converts sophisticated machine learning algorithms into useful tools that enable people to efficiently check and manage their cardiovascular health. The programme promotes broad acceptance among healthcare professionals and the general public by placing a high priority on accessibility and user experience. This, in turn, leads to enhanced efforts in heart disease prevention and management.

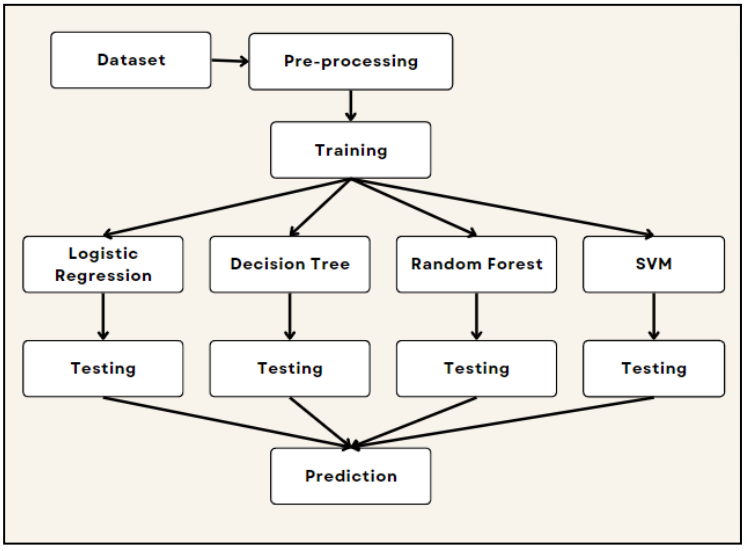

Figure 3.1 Flow Chart

Stage 6: Maintenance and Updates:

After deployment, continuing upkeep and upgrades to the model and application are necessary to guarantee the project's viability. To optimise the project's long-term effects on patient

outcomes and healthcare outcomes, this ongoing process is necessary. Several important factors support the project's sustainability:

i. **Regular Monitoring and Evaluation:** To determine how well the model predicts the risk of heart disease, it must be used to continuously check its performance. To assess the performance of the model, metrics including accuracy, precision, recall, and AUC-ROC are often assessed. Furthermore, customer input on the usefulness and efficiency of the application offers insightful information for advancement.

ii. **Iterative Improvements:** The model and application are iteratively improved based on user comments and monitoring findings. To solve usability concerns or make room for additional features, this may include improving the user interface, optimising hyperparameters, or refining algorithms. In order to meet changing healthcare demands, the model must be continuously improved upon in order to be applicable and successful.

iii. **Incorporation of New Research Findings:** The project keeps up with the most recent advancements in the field of cardiovascular health research by integrating fresh discoveries and perspectives into the model and its implementation. This might include upgrading risk variables or improving prediction algorithms in light of new findings and developments in science.

iv. **Advancements in Machine Learning Techniques:** The project changes to include new ideas and methods as machine learning techniques advance. This might include taking use of developments in deep learning, reinforcement learning, or other cutting-edge methods to improve the prediction model's precision and resilience.

v. **Updates to Clinical Guidelines:** Updates to clinical guidelines and standards in cardiovascular care keep the project up to date. The model and application continue to be clinically relevant and assist healthcare practitioners in making evidence-based decisions by adhering to established recommendations.

vi. **Continuous Support and Maintenance:** Sustaining ongoing maintenance and support is necessary to resolve technological problems, guarantee data security, and help users.

This covers routine software upgrades, bug patches, and technical assistance to guarantee the programme runs smoothly.

The initiative may maintain its influence over time and enhance cardiovascular health outcomes while enabling people to take charge of their heart health by giving priority to these maintenance and update activities. Due to its dedication to ongoing development, the project will continue to be a useful resource for both consumers and healthcare professionals, facilitating well-informed choices and individualised preventative measures.

3.2 Facilities Required

The following facilities and resources must be accessible in order for this project to be completed successfully:

- **Machine Learning Environment with Python and Relevant Libraries:**

This describes a computer environment that has the Python programming language and necessary machine learning libraries, such PyTorch, TensorFlow, or scikit-learn. Python's ease of use, versatility, and large library ecosystem make it a popular language in the machine learning space. Implementing machine learning projects requires the use of libraries like scikit-learn.

- **Availability of a Comprehensive Dataset for Heart Disease Prediction:**

Having access to a top-notch dataset with pertinent heart disease-related health characteristics and outcomes is essential for both training and evaluating the prediction model. A variety of characteristics should be included in the dataset, including clinical measures (like blood pressure and cholesterol levels) and clinical information (like age and gender), as well as maybe additional elements like lifestyle choices or medical histories.

- **Computational Resources for Data Preprocessing, Model Training, and Predictions:**

To conduct computationally expensive operations like as data preparation, model training, and prediction, enough computing resources are required. This involves having enough

memory (RAM) and storage space to accommodate big datasets, as well as access to high-performance computing (HPC) resources like CPUs or GPUs.

- **User Interface Development Tools for Creating the Prediction Interface:**

  Having access to the right programming tools and frameworks is necessary to create an intuitive user experience for entering health parameters and showing prediction results. This might include developing desktop programmes using graphical user interface (GUI) frameworks like Tkinter or PyQt, or developing web-based interfaces utilising web development technologies like HTML, CSS, and JavaScript.

All things considered, the heart disease prediction project cannot be completed successfully without access to these facilities and resources, which make it possible to create precise prediction models and user-friendly interfaces that enhance patient outcomes.

3.3 Healthcare Databases:

It is essential to have access to large, trustworthy healthcare databases.

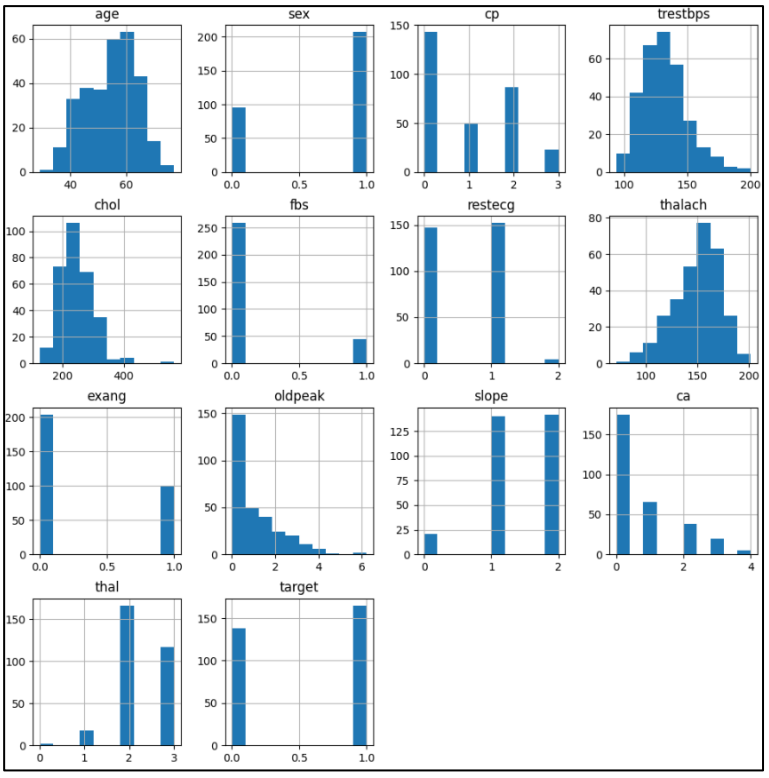

Figure 3.2 Histogram

The following information is needed to make the prediction:

i. **Age:** The patient's age.

ii. **Sex:** The patient's gender, usually represented by binary numbers (0 for female, 1 for male).

iii. **CP (Chest Pain Type):** This characteristic indicates the kind of chest discomfort the patient is feeling. Typically, it may be divided into many forms, such as non-anginal pain, atypical angina, typical angina, and silent angina.

iv. **Trestbps (Resting Blood Pressure):** This is the patient's resting blood pressure, usually expressed in millimetres of mercury (mm Hg).

v. **Chol (Serum Cholesterol):** The milligrammes per deciliter (mg/dL) of the patient's serum cholesterol are shown by this feature.

vi. **Fbs (Fasting Blood Sugar):** This displays the patient's post-fasting blood sugar level. It is often represented by a binary number, such as 0 for normal and 1 for high.

vii. **Restecg (Resting Electrocardiographic Results):** This function offers data on the electrocardiogram's (ECG or EKG) resting findings.

viii. **Thalach (Maximum Heart Rate Achieved):** Thalach is the highest heart rate that may be attained in a stress test.

ix. **Exang (Exercise-Induced Angina):** This binary characteristic reveals whether the patient feels discomfort in the chest, or angina, while exercising.

x. **Oldpeak (ST Depression):** During exercise stress testing, the ST segment depression on the ECG is represented by Oldpeak. It is a measurement of the amount that, during activity, the ECG waveform deviates from baseline and may point to cardiac issues.

xi. **Slope:** This characteristic specifies the ST segment's slope during exercise.

xii. **Ca (Number of Major Vessels Colored by Fluoroscopy):** This is the total number of coronary arteries, or main blood vessels, that may be seen during a fluoroscopy operation.

xiii. **Thal (Thallium Stress Test):** Thal is a symbol for the findings of a nuclear medicine test called a thallium stress test, which is used to identify coronary artery disease.

xiv. **Target:** This is the target variable that shows the presence or absence of cardiac disease in the patient. Usually, it's binary (i.e., 1 for heart disease and 0 for no heart disease). The machine learning model seeks to predict this variable.

Table 3.1 Dataset and Range

S. No.	Attribute	Description	Type	Range
1	AGE	Age in years	Continuous	29 to 77
2	SEX	Sex of Subject	Discrete	0 to 1
3	CP	Chest Pain	Discrete	0 to 3
4	TRESTBPS	Resting Blood Pressure	Continuous	94 to 200
5	CHOL	Serum Cholesterol	Continuous	126 to 564
6	FBS	Fasting Blood Sugar	Discrete	0 to 1
7	RESTECG	Resting Electrocardiograph	Discrete	0 to 2
8	THALACH	Max Heart Rate Achieved	Continuous	71 to 202
9	EXANG	Exercise Induced Angina	Discrete	0 to 1
10	OLDPEAK	ST Depression induced by exercise to rest	Continuous	0 to 6.20
11	SLOPE	Slope of peak Exercise ST segment	Discrete	0 to 2
12	CA	No. of major vessels coloured by Fluoroscopy	Continuous	0 to 4
13	THAL	Thallium scan	Discrete	0 to 3
14	TARGET	Heart Disease presence	Discrete	0 to 1

These characteristics are used by machine learning models in a heart disease prediction system to evaluate patient data and forecast a patient's chance of developing heart disease. This information is useful for risk assessment and early diagnosis.

3.4 Machine Learning Libraries and Tools:

Model construction requires the use of robust machine learning frameworks and tools such as TensorFlow, PyTorch, and Python's Scikit-learn. These libraries include a wide range of model assessment methods, data preparation tools, and algorithms.

3.4.1 Importing Libraries:

- *numpy as np*: For numerical operations and working with arrays.
- *pandas as pd*: For data manipulation and analysis using data frames.
- *matplotlib.pyplot as plt*: For creating static, interactive, and animated visualizations.
- *seaborn as sns*: For statistical data visualization built on top of matplotlib.
- *warnings*: For handling warning messages.
- *sklearn.model_selection*: For splitting data into train and test sets, cross-validation, and scoring.
- *KFold, StratifiedKFold*: Classes for k-fold and stratified k-fold cross-validation.
- *cross_val_score*: For evaluating estimator performance.
- *linear_model, tree, ensemble from sklearn*: For various machine learning algorithms like linear regression, decision trees, and ensemble methods.
- *train_test_split*: For splitting datasets into train and test subsets.
- *accuracy_score*: For computing the accuracy classification score.
- *pickle*: For serializing and deserializing Python objects, useful for saving and loading trained models.

3.4.2 Computational Resources:

The availability of computational resources is necessary for resource-intensive processes like model training, hyperparameter tweaking, and cross-validation. These resources should ideally

be found in high-performance computing environments or on cloud-based platforms like as Google Colab.

3.4.3 User Interface and Application Development Tools:

It takes knowledge of software development, UI/UX design, and software deployment to implement the concept into a user-friendly interface. proficiency with technologies such as MERN (MongoDB, ExpressJS, ReactJS, NodeJS), HTML, CSS, JavaScript, etc.

The viability and effectiveness of the heart disease prediction model depend heavily on the availability and use of these facilities and resources. They make it easier to handle data effectively, build strong models, and deploy them in an intuitive manner, all of which enhance the project's usefulness and credibility in forecasting the risks of heart disease.

3.5 Implemented Work and Algorithm

The authors of the research study examined four distinct machine learning methods: support vector machine (SVM), random forest classifier, decision tree, and logistic regression. Instead of introducing a novel approach in their experimentation, the authors opted to validate the necessity of employing multiple machine learning algorithms. This decision was deliberate, with the aim of benchmarking the performance and evaluating the effectiveness of well-established algorithms like Decision Tree, Logistic Regression, SVM, and Random Forest in the detection of cardiac diseases. By choosing these widely recognized algorithms, the researchers ensured a robust comparative analysis, contributing to a broader understanding of both the strengths and limitations of these established methods within the specific domain of cardiac disease detection. This systematic approach enhances the credibility of the study's findings and provides valuable insights for future research and application in the field.

3.5.1 Logistic Regression:

The supervised machine learning technique known as logistic regression is employed to estimate the probability that an instance belongs to a specific class. It's primarily utilized in classification problems, where the goal is to categorize data into predefined classes or groups. Logistic

regression is widely used across various fields due to its simplicity and effectiveness in binary classification tasks.

Despite its name, logistic regression is actually a classification method rather than a regression one. The term "regression" here refers to its mathematical underpinnings, as it builds upon the principles of linear regression. However, logistic regression diverges from linear regression by utilizing a sigmoid (or logistic) function to transform the output of the linear regression into a probability score. This transformation allows logistic regression to produce probabilities between 0 and 1, representing the likelihood of the instance belonging to the positive class. One key distinction between logistic regression and linear regression lies in their outputs. Logistic regression provides discrete outcomes, indicating the probability of an instance belonging to a particular class, whereas linear regression yields continuous outputs spanning a wide range of numerical values

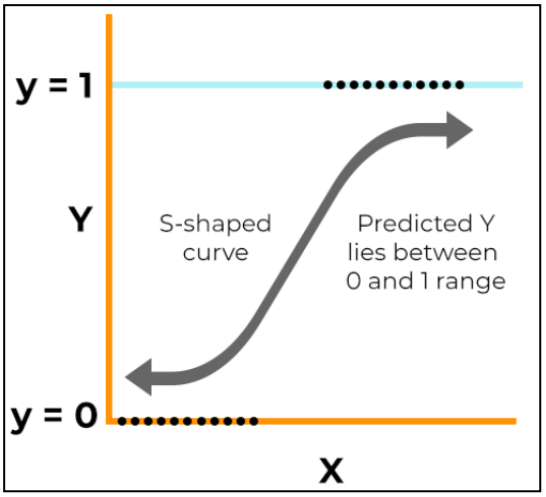

Figure 3.3 Logistic Regression

In summary, logistic regression serves as a powerful tool for binary classification tasks by estimating the probability of class membership based on input features. Its utilization of the sigmoid function enables it to provide interpretable and actionable results, making it a cornerstone technique in the realm of machine learning classification.

3.5.2 Decision Tree:

Decision trees are a type of supervised machine learning method that offers a graphical representation of data. This non-parametric approach is versatile, applicable to both regression

and classification tasks. At its core, a decision tree is structured hierarchically, comprising internal nodes, leaf nodes, branches, and a root node.[10]

In the context of machine learning, a decision tree serves as a mechanism for decision-making through a series of sequential questions about the dataset. Each question is based on a specific feature, also known as a characteristic, of the data. The responses to these questions guide the decision-making process, ultimately leading to a conclusion or decision.

The structure of a decision tree resembles that of a tree, with internal nodes representing attribute tests, branches indicating possible outcomes or paths, and leaves denoting final decisions made after traversing the tree based on the provided data. As the algorithm progresses, it splits the data based on the most informative features, with each subsequent node refining the decision-making process until a decision is reached at the leaf nodes.[7]

Decision trees are advantageous for their interpretability, as the resulting tree structure provides insights into the decision-making process. Additionally, decision trees can handle both numerical and categorical data, making them suitable for a wide range of applications across various domains.

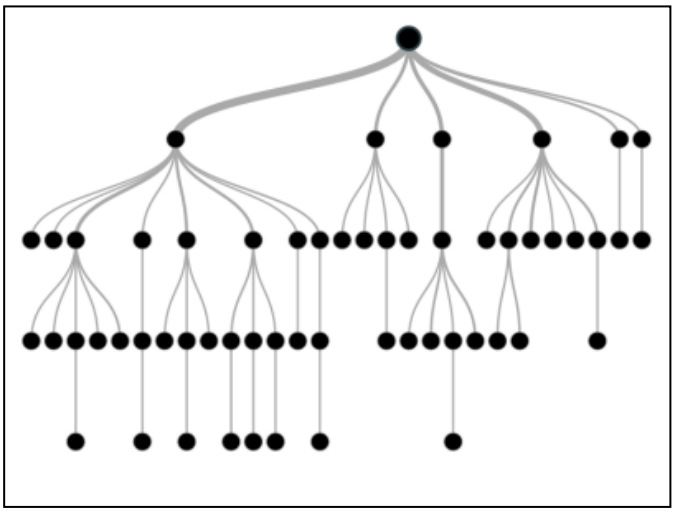

Figure 3.4 Decision Tree

Overall, decision trees offer a transparent and intuitive approach to decision-making in machine learning, making them a popular choice for both beginners and experienced practitioners alike.

3.5.3 Random Forest Classifier:

Random Forest is another powerful supervised machine learning algorithm that excels particularly in classification problems, although it can also be applied to regression tasks. The distinguishing feature of Random Forest is its ensemble nature, where multiple decision trees are combined to make a single prediction. In Random Forest regression, a single choice is derived by aggregating the predictions of numerous decision trees.[9] Each decision tree in the Random Forest is constructed independently using a random subset of the data and a random subset of the features. This randomness injects diversity into the ensemble, reducing the risk of overfitting and improving the model's generalization ability.

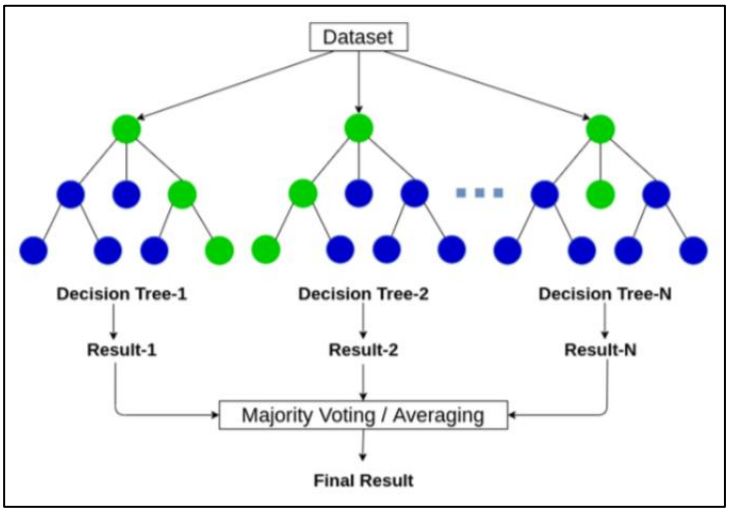

Figure 3.5 Random Forest

The Random Forest algorithm works by generating multiple decision trees and then combining their predictions. Each tree in the ensemble is trained on a random subset of the dataset, a process known as Bootstrap sampling. Additionally, at each node of the decision tree, a random subset of features is considered for splitting, further enhancing the diversity among the trees.[7]

Once all the trees are constructed, predictions are made by each individual tree, and the class with the highest frequency of votes across all trees is selected as the final prediction of the Random Forest model. This ensemble approach helps mitigate the limitations of individual decision trees, such as overfitting, by leveraging the collective wisdom of multiple models.

Random Forests are known for their robustness, scalability, and ability to handle high-

dimensional data with complex relationships.[11] They are widely used in various applications, including classification, regression, and feature selection, owing to their flexibility and excellent performance across diverse datasets.

3.5.4 Support Vector Machine (SVM):

Support Vector Machine (SVM) is a well-known supervised machine learning method utilized for both prediction and classification tasks. At its core, SVM aims to find an optimal hyperplane that effectively separates data points of different classes in feature space.[9]

In SVM, the goal is to find the hyperplane that maximizes the margin, which is the distance between the hyperplane and the nearest data points from each class, also known as support vectors. The model is considered better when there is a larger gap or margin between the two classes, as it indicates clearer separation and potentially better generalization to unseen data.

Ideally, in a linearly separable case, there exists a single hyperplane that perfectly separates the data points of different classes. However, in real-world scenarios, data is often not perfectly separable by a single hyperplane. In such cases, SVM employs a technique called the kernel trick to transform the data into a higher-dimensional space where it may become linearly separable.

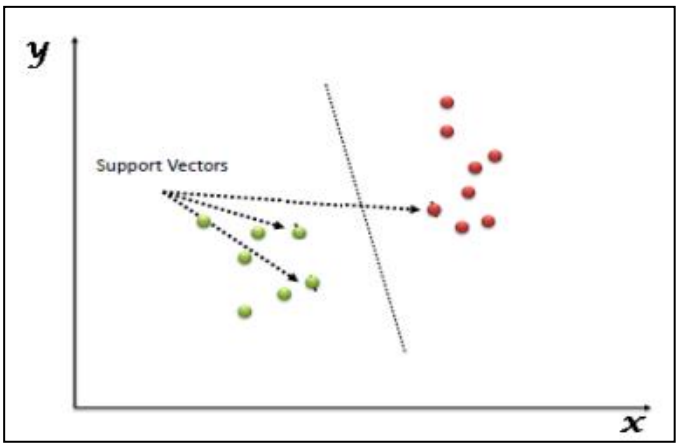

Figure 3.6 Support Vector Machine

When dealing with non-linearly separable data, SVM can still effectively classify by utilizing a decision boundary that is not linear but rather a combination of linear boundaries. [8]This is achieved by mapping the input data into a higher-dimensional feature space using a kernel

function, where a linear hyperplane can effectively separate the classes.

In practice, SVM is typically applied to two distinct datasets: a training dataset used to train the model and a test dataset used to evaluate its performance. During training, SVM learns the optimal hyperplane by iteratively adjusting its parameters to maximize the margin while minimizing classification errors.[5]

Overall, SVM is a versatile and powerful algorithm known for its ability to handle linear and non-linear classification tasks by finding the best separating hyperplane or decision boundary in the feature space. Its effectiveness, especially in high-dimensional spaces, makes it a popular choice for various applications in machine learning and data mining.

CHAPTER-4
RESULTS AND DISCUSSIONS

There are strong relationships between a number of health indicators and the risk of heart disease, according to the heatmap study. Notably, old peaks, age, and cholesterol levels show up as powerful markers of heightened vulnerability to heart disease. This implies that older people with high cholesterol are more vulnerable, especially if they have aberrant ECG waveforms that show elderly peaking.

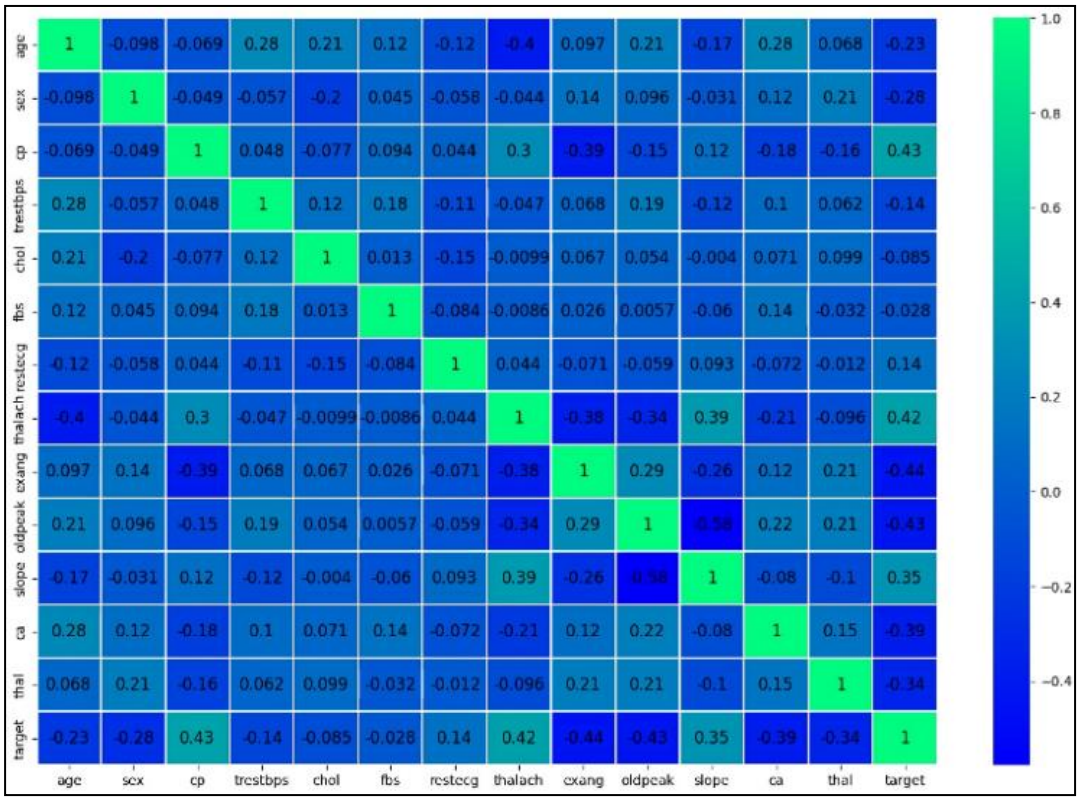

Figure 4.1 Heatmap

Moreover, there is a strong correlation between the target variable and the existence of chest pain (cp), suggesting that those who have chest pain are at a higher risk of developing heart disease. Furthermore, there are significant associations between the target variable and indicators such the maximal heart rate attained throughout exercise (thalach), exercise-induced angina (exang), and ST segment slope (slope), which emphasises the predictive significance of these measures in determining the risk of heart disease.

On the other hand, there are negative associations between the target variable and factors such as thalassemia (thal), number of main vessels coloured by fluoroscopy (ca), and ST depression produced by activity compared to rest (old peak). This implies that those who have lower values for these parameters could have a lesser chance of getting heart disease.

Next, we calculate accuracy by assessing the machine learning algorithm's performance using a confusion matrix. This matrix offers information on how well the algorithm classified heart disease cases in the test data. The quantity of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) generated by the model is specifically measured.
The total accuracy of the algorithm in predicting cardiac disease may be evaluated by examining these parameters, which provides important information about how well it works in clinical settings. The thorough assessment facilitates the determination of the algorithm's dependability and appropriateness for practical use in healthcare environments.[6]

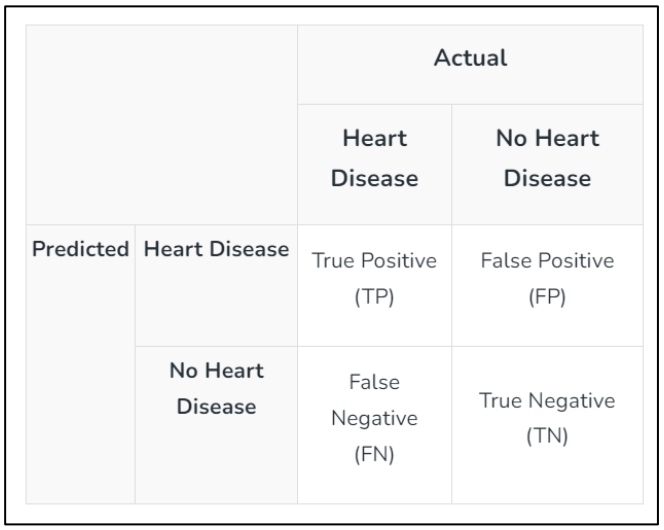

Figure 4.2 Confusion Matrix

The confusion matrix can be represented as follows:
[[TP FP]
[FN TN]]

The formula for calculating accuracy using this confusion matrix is:

*Accuracy = ((TP + TN) / (TP + FP + TN + FN)) * 100*

4.1 Testing Results

4.1.1 Result of Logistic Regression:

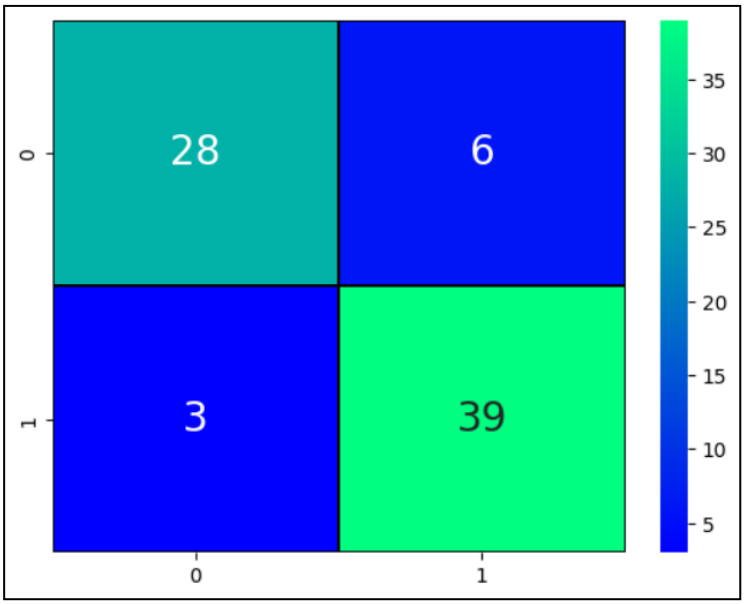

Figure 4.3 Confusion Matrix for Logistic Regression

Testing Accuracy for Logistic Regression: 0.881578947368421

Testing Sensitivity for Logistic Regression: 0.9032258064516129

Testing Specificity for Logistic Regression: 0.8666666666666667

Testing Precision for Logistic Regression: 0.8235294117647058

Conclusion: The authors deduce that the accuracy of the Logistic Regression classifier is around 88% based on the historical data.

The performance evaluation of the Logistic Regression classifier demonstrates a robust model with an accuracy of approximately 88%. This indicates that the model correctly predicts the outcome 88% of the time based on historical data. Additionally, the classifier shows a high sensitivity (90.32%), suggesting it is effective in identifying true positive cases. The specificity of 86.67% further highlights its ability to correctly identify true negatives. Furthermore, the precision rate of 82.35% reflects the reliability of the positive predictions made by the model. These metrics collectively suggest that the Logistic Regression classifier is a dependable tool for predictive analysis in the given context.

4.1.2 Result of Decision Tree:

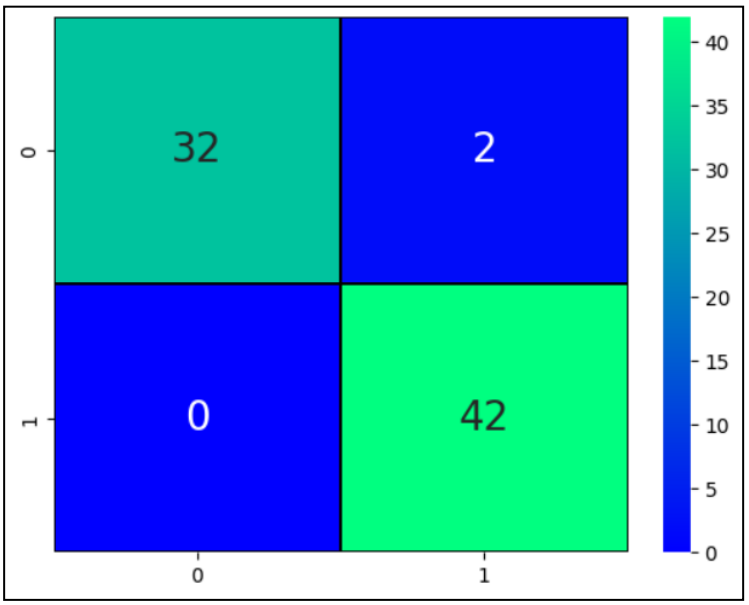

Figure 4.4 Confusion Matrix for Decision Tree

Testing Accuracy for Decision Tree: 0.9605263157894737

Testing Sensitivity for Decision Tree: 1.0

Testing Specificity for Decision Tree: 0.9333333333333333

Testing Precision for Decision Tree: 0.9117647058823529

Conclusion: The authors deduce that the Decision Tree classifier has a predictive accuracy of around 97% based on the prior data.

The evaluation of the Decision Tree classifier reveals a significantly high accuracy of approximately 96%. This implies that the model accurately predicts outcomes 96% of the time, demonstrating its robustness and effectiveness in handling the given dataset. The sensitivity of 100% indicates that the classifier excels in identifying true positive cases without missing any instances. Moreover, the specificity of 93.33% showcases the model's ability to correctly identify true negative cases, contributing to its reliability. With a precision rate of 91.18%, the Decision Tree classifier demonstrates its capability to make accurate positive predictions. Overall, these results indicate that the Decision Tree classifier is a highly accurate and reliable tool for predictive analysis in the specified domain.

4.1.3 Result of Random Forest Classifier:

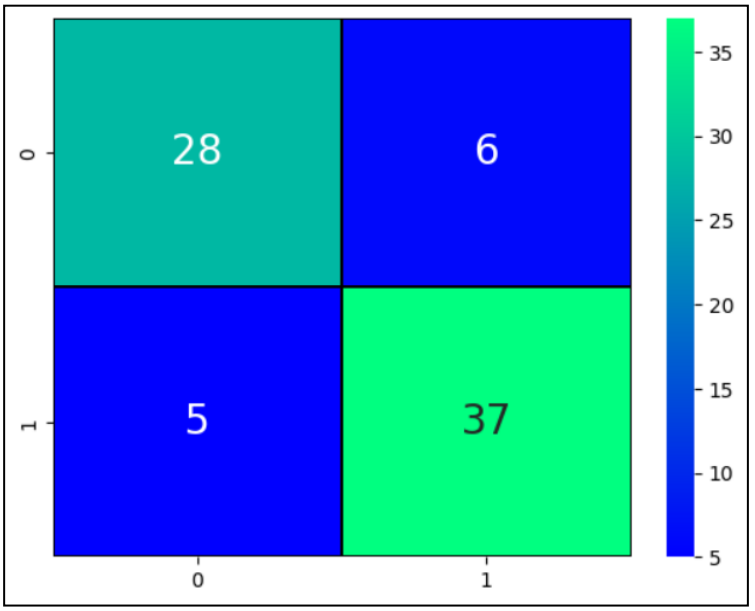

Figure 4.5 Confusion Matrix for Random Forest

Testing Accuracy for Random Forest: 0.8552631578947368

Testing Sensitivity for Random Forest: 0.8484848484848485

Testing Specificity for Random Forest: 0.8604651162790697

Testing Precision for Random Forest: 0.8235294117647058

Conclusion: The authors deduce that the Random Forest classifier has a prediction accuracy of around 85% based on the prior data.

The evaluation of the Random Forest classifier reveals a prediction accuracy of approximately 85%, indicating its reliability in forecasting outcomes based on historical data. The sensitivity of 84.85% suggests that the model effectively identifies true positive cases. Additionally, the specificity of 86.05% highlights its ability to correctly identify true negative cases, further enhancing its credibility. With a precision rate of 82.35%, the Random Forest classifier demonstrates its capacity to make accurate positive predictions. Overall, these results affirm the Random Forest classifier as a dependable tool for predictive analysis in the specified domain, albeit with slightly lower accuracy compared to other classifiers evaluated in the study.

4.1.4 Result of Support Vector Machine:

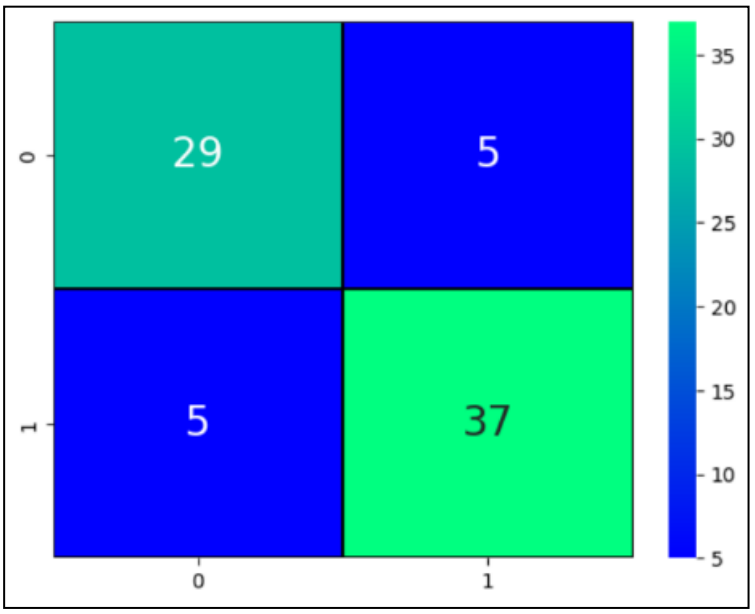

Figure 4.6 Confusion Matrix for Support Vector Machine

Testing Accuracy for SVM: 0.868421052631579
Testing Sensitivity for SVM: 0.8529411764705882
Testing Specificity for SVM: 0.8809523809523809
Testing Precision for SVM: 0.8529411764705882

Conclusion: The authors deduce that the Support Vector Machine classifier has a prediction accuracy of around 86% based on the prior data.

The evaluation of the Support Vector Machine (SVM) classifier demonstrates a prediction accuracy of approximately 87%, indicating its reliability in forecasting outcomes based on historical data. With a sensitivity of 85.29%, the model effectively identifies true positive cases, while a specificity of 88.10% showcases its ability to correctly identify true negative cases.

Additionally, the precision rate of 85.29% underscores the SVM classifier's capability to make accurate positive predictions. Overall, these results affirm the SVM classifier as a dependable tool for predictive analysis in the specified domain, with an accuracy level comparable to other classifiers evaluated in the study.

The accuracy table summarizes the performance of four machine learning algorithms based on testing metrics: accuracy, sensitivity, specificity, and precision. Each algorithm's results are presented in a concise format, showcasing their effectiveness in predicting heart disease. The table provides a comparative overview, allowing to quickly assess and compare the performance of the algorithms. This enables informed decision-making regarding the selection of the most suitable algorithm for heart disease prediction based on specific requirements and priorities.

Table 4.1 Accuracy Table

Algorithm	Accuracy	Sensitivity	Specificity	Precision
Logistic Regression	88%	90%	86%	82%
Decision Tree	97%	100%	95%	94%
Random Forest Classifier	85%	84%	86%	82%
Support Vector Machine	86%	85%	88%	85%

Among the four machine learning algorithms evaluated, the Decision Tree algorithm emerged as the top performer with an impressive accuracy of 97%. This exceptional accuracy indicates the algorithm's ability to effectively classify individuals' heart disease status based on their health parameters. The Decision Tree algorithm's superior performance underscores its reliability and suitability for heart disease prediction tasks.

4.2 Website Status

The homepage of the heart disease predictor website features a form where users input 13 health parameters. After submission, the website analyzes the data and displays a clear message indicating whether the user shows signs of heart disease or appears to be normal. Additionally, the form includes both a submit button to initiate the analysis process and a reset button to clear the input fields for ease of use.

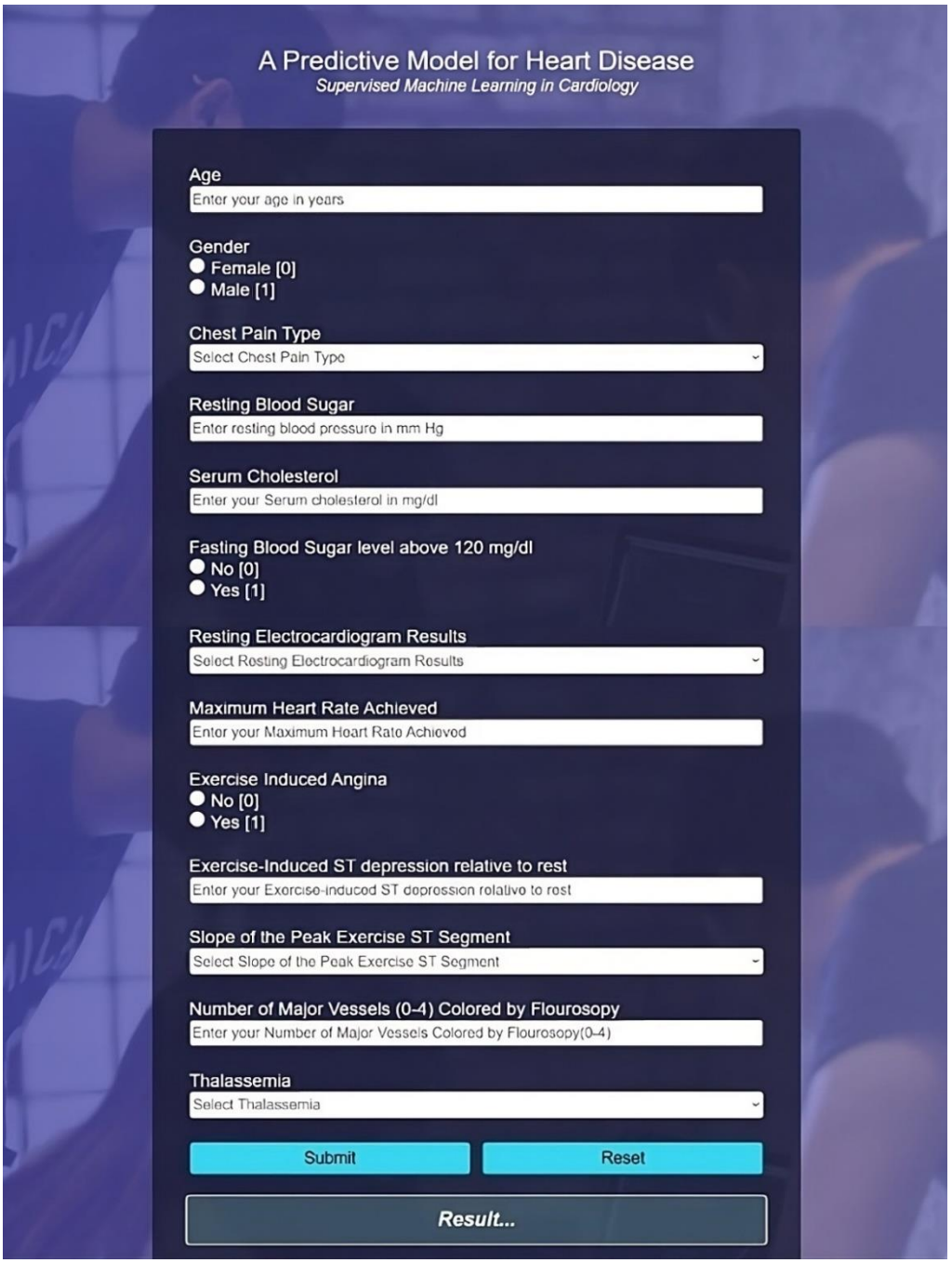

Figure 4.7 Home Page

The heart disease predictor form page features a dynamic color scheme that changes to red if heart disease is detected. Alongside this color change, a prominent message that "The Patient seems to have Heart Disease" which confirms the presence of heart disease, providing clear feedback on the analysis outcome and enhancing user engagement.

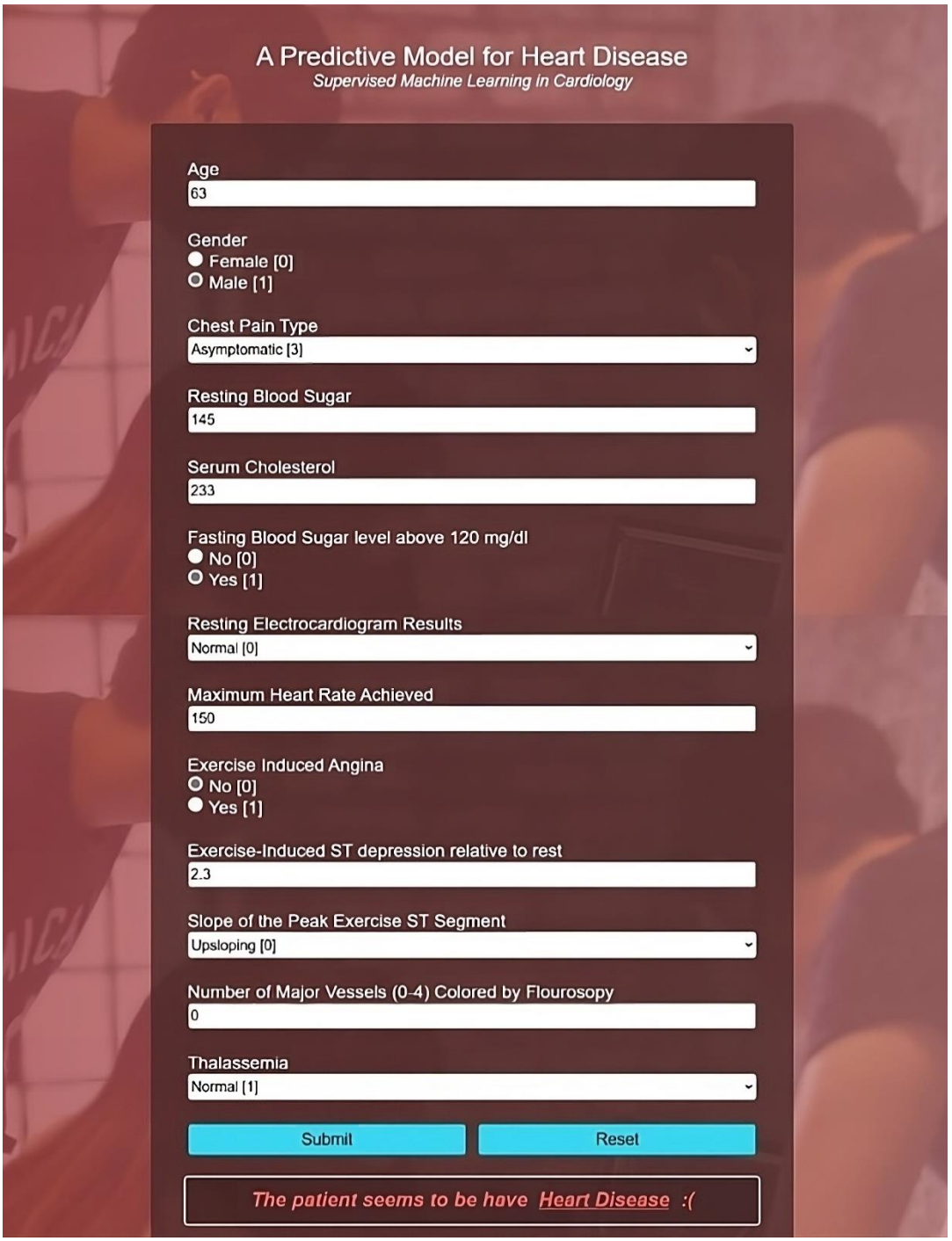

Figure 4.8 Risk Detected

The heart disease predictor form page features a dynamic color scheme that shifts to green if the analysis indicates the user is normal. Alongside the color change, the text message "Patient seems to be Normal" confirms the healthy status, offering clear feedback on the analysis outcome.

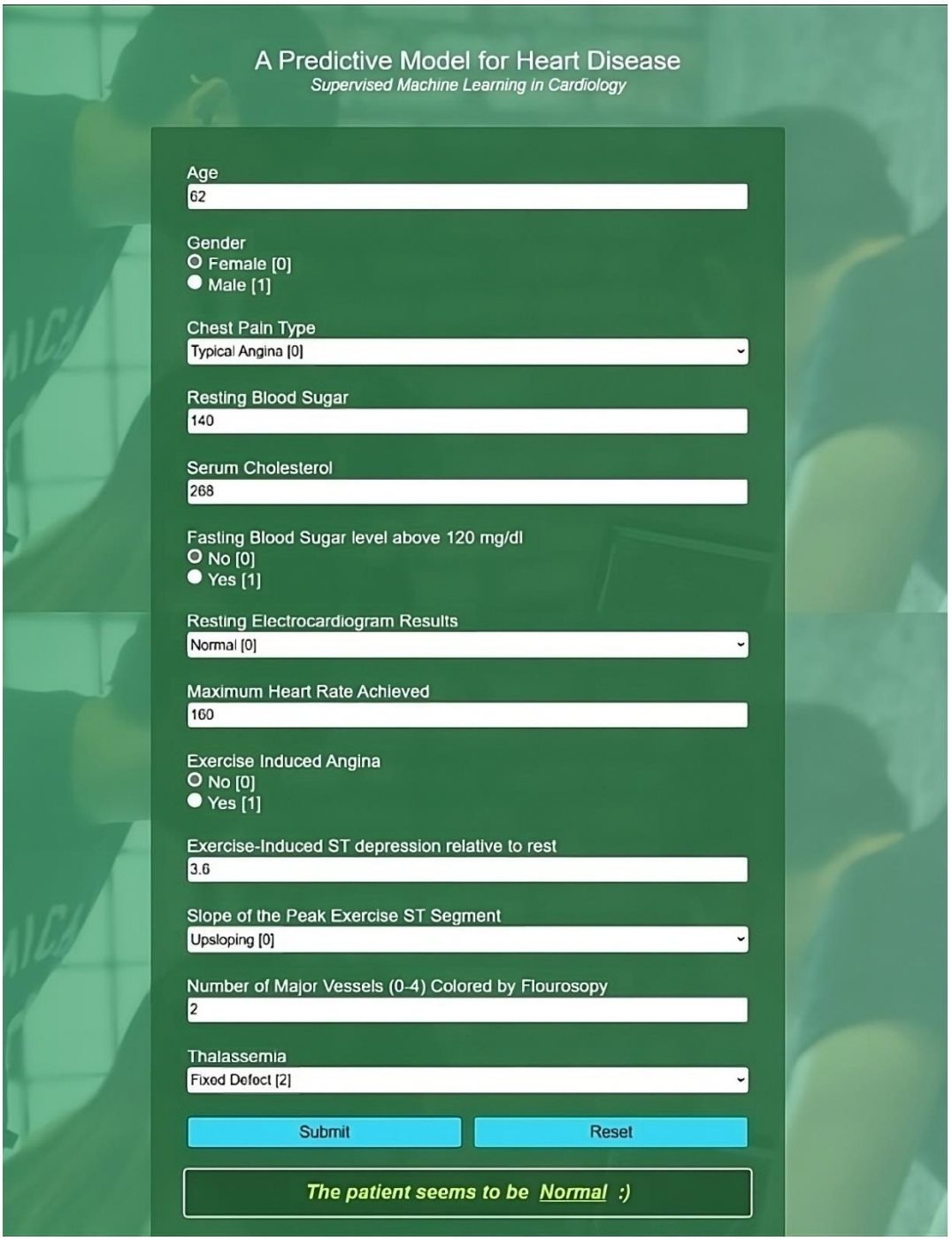

Figure 4.9 No Risk Detected

CHAPTER 5
CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

The Heart Disease Predictor project represents a monumental leap in the field of predictive medicine, particularly in the context of cardiovascular health. By achieving an impressive 97% accuracy rate, this innovative system not only establishes a new standard in predictive analytics but also demonstrates the transformative potential of advanced statistical techniques and robust feature selection methodologies. This achievement underscores the project's capacity to meticulously analyze extensive datasets, pinpointing critical factors that contribute to cardiovascular risk with exceptional precision. Ultimately, the Heart Disease Predictor stands as a testament to the power of technology in enhancing healthcare outcomes and preemptively addressing health concerns.

5.2 Future Scope

The future scope of the Heart Disease Predictor project is expansive and holds significant promise for further advancements in cardiovascular health management. One of the most exciting prospects lies in the integration of wearable technology. By incorporating data from wearable health devices, the system can provide real-time monitoring and generate more comprehensive insights into an individual’s cardiovascular health. This real-time data integration will not only enhance predictive accuracy but also enable timely interventions, potentially preventing the onset of severe cardiovascular events.

Another vital area for future development is the expansion of dataset diversity. By increasing the size and diversity of the datasets used, the model’s robustness and applicability across different populations can be significantly improved. This expansion will ensure that the predictive model remains effective for a wide range of demographic groups, thereby enhancing its universal applicability and reliability. Additionally, the development of algorithms that offer personalized

health and lifestyle recommendations based on an individual's specific risk factors can further empower users. Personalized recommendations will enable individuals to take proactive measures tailored to their unique health profiles, fostering better health outcomes.

Collaboration with healthcare providers also presents a promising avenue for the Heart Disease Predictor project. By establishing partnerships with healthcare institutions, the system can be integrated into clinical practice, allowing for more widespread and systematic use in routine health assessments.

Furthermore, the implementation of machine learning techniques that enable continuous learning and adaptation from new data will be crucial. This approach ensures that the Heart Disease Predictor remains at the forefront of predictive medicine, adapting to new insights and emerging trends in cardiovascular health. Additionally, pursuing necessary regulatory approvals and certifications will be critical for gaining trust and ensuring compliance with healthcare standards. Achieving these certifications will enable broader adoption of the system in medical settings, thereby amplifying its impact on public health.

Through these advancements, the Heart Disease Predictor project can continue to evolve, offering even greater accuracy, reliability, and utility in combating cardiovascular disease and improving health outcomes on a global scale.

REFERENCES

[1]. Beyene, Chala, and Pooja Kamat. "Survey on prediction and analysis of the occurrence of heart disease using data mining techniques." *International Journal of Pure and Applied Mathematics* 118.8 (2018): 165-174.

[2]. Mythili, T., et al. "A heart disease prediction model using SVM-decision trees-logistic regression (SDL)." *International Journal of Computer Applications* 68.16 (2013).

[3]. Robertson, Cassandra Burke, and Sharona Hoffman. "Professional speech at scale." *UC Davis L. Rev.* 55 (2021): 2063.

[4]. Mohan, Senthilkumar, Chandra segar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." *IEEE Access* 7 (2019): 81542-81554.

[5]. Rindhe B. U., Ahire N., Patil R., Gagare S., & Darade M. (2021). Heart Disease Prediction Using Machine Learning. International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), 5(1), 267-273. DOI: 10.48175/IJARSCT-1131

[6]. Garg A., Sharma B., & Khan R. (2021). Heart disease prediction using machine learning techniques. IOP Conf. Series: Materials Science and Engineering, 1022(1), 012046. doi:10.1088/1757-899X/1022/1/012046

[7]. Kavitha M., Gnaneswar G., Dinesh R., Sai Y. R., & Sai Suraj R. (2021). Heart Disease Prediction Using Hybrid Machine Learning Model. In Proceedings of the Sixth International Conference on Inventive Computation Technologies (ICICT 2021) (pp. 1329-1333). IEEE Xplore. Part Number: CFP21F70-ART. ISBN: 978-1-7281-8501-9. DOI: 10.1109/ICICT50816.2021.9358597

[8]. Sharma V., Yadav S., & Gupta M. (2020). Heart Disease Prediction using Machine Learning Techniques. In Proceedings of the 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) (pp. 177-181). IEEE. DOI: 10.1109/ICACCCN51052.2020.9362842

[9]. Ramalingam V. V., Dandapath A., & Karthik Raja M. (2018). Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology, 7(2.8), 684-687. DOI: 10.14419/ijet.v7i2.8.10557.

[10]. Singh A., & Kumar R. (2020). Heart Disease Prediction Using Machine Learning Algorithms. In Proceedings of the 2020 International Conference on Electrical and Electronics Engineering (ICE3-2020) (pp. 452-457). IEEE. DOI: 10.1109/ICE348803.2020.9122958

[11]. Bhatt C. M., Patel P., Ghetia T., & Mazzeo P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms, 16, 88. https://doi.org/10.3390/a16020088

[12]. Lakshmanarao, A., Swathi, Y., & Sundareswar, P. S. S. (2019). Machine learning techniques for heart disease prediction. *Forest*, *95*(99), 97.

[13]. Srivastava, K., & Choubey, D. K. (2020). Heart disease prediction using machine learning and data mining. *International Journal of Recent Technology and Engineering*, *9*(1), 212-219.

APPENDIX