VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY
Department of Computer Engineering

Project Report on

Heart Disease Prediction System

In partial fulfillment of the Fourth Year, Bachelor of Engineering (B.E.) Degree in Computer Engineering at the University of Mumbai

Academic Year 2023-24

Submitted by
Harsh Bakhtar - D17B - 04
Deepali Duseja - D17C - 19
Sahil Nagdev - D17C - 38
Jatin Ochani - D17C - 40

Project Mentor
Prof. Sunita Suralkar

(2022-23)

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

Department of Computer Engineering

Certificate

This is to certify that **Sahil Nagdev(D17C, 38), Jatin Ochani (D17C, 40), Deepali Duseja (D17C, 19), Harsh Bakhtar (D17B, 04*)*** of Fourth Year Computer Engineering studying under the University of Mumbai have satisfactorily completed the project on "**Heart Disease Prediction System** " as a part of their coursework of PROJECT-II for Semester-VIII under the guidance of their mentor **Prof. Sunita Suralkar** in the year 2023-24 .

This project report entitled **Heart Disease Prediction System** by ***Sahil Nagdev, Jatin Ochani, Deepali Duseja, Harsh Bakhtar*** is approved for the degree of **B.E. Computer Engineering.**

Programme Outcomes	Grade
PO1,PO2,PO3,PO4,PO5,PO6,PO 7, PO8, PO9, PO10, PO11, PO12 PSO1, PSO2	

Date:
Project Guide:

Project Report Approval
For
B. E (Computer Engineering)

This project report entitled **Heart Disease Prediction System** by ***Sahil Nagdev, Jatin Ochani, Deepali Duseja, Harsh Bakhtar*** is approved for the degree of **B.E. Computer Engineering.**

Internal Examiner

External Examiner

Head of the Department

Principal

Date:
Place: Mumbai

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Sahil Nagdev(38)

Jatin Ochani(40)

Deepali Duseja(19)

Harsh Bakhtar(04)

Date:

ACKNOWLEDGEMENT

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to **Prof. Sunita Suralkar** (Project Mentor) for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.

We are deeply indebted to Head of the Computer Department **Dr. (Mrs.) Nupur Giri** and our Principal **Dr. (Mrs.) J. M. Nair ,** for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their guidance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is a great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement several times.

Computer Engineering Department
COURSE OUTCOMES FOR B.E PROJECT

Learners will be to,

Course Outcome	Description of the Course Outcome
CO 1	Able to apply the relevant engineering concepts, knowledge and skills towards the project.
CO2	Able to identify, formulate and interpret the various relevant research papers and to determine the problem.
CO 3	Able to apply the engineering concepts towards designing solutions for the problem.
CO 4	Able to interpret the data and datasets to be utilized.
CO 5	Able to create, select and apply appropriate technologies, techniques, resources and tools for the project.
CO 6	Able to apply ethical, professional policies and principles towards societal, environmental, safety and cultural benefit.
CO 7	Able to function effectively as an individual, and as a member of a team, allocating roles with clear lines of responsibility and accountability.
CO 8	Able to write effective reports, design documents and make effective presentations.
CO 9	Able to apply engineering and management principles to the project as a team member.
CO 10	Able to apply the project domain knowledge to sharpen one's competency.
CO 11	Able to develop a professional, presentational, balanced and structured approach towards project development.
CO 12	Able to adopt skills, languages, environment and platforms for creating innovative solutions for the project.

Index

Chapter No.	Title	Page No.
1	**Introduction**	12
1.1	Introduction to the project	12
1.2	Motivation for the project	12
1.3	Problem Definition	13
1.4	Existing Systems	14
1.5	Lacuna of the Existing Systems	14
1.6	Relevance of the Project	15
2	**Literature Survey**	16
A	Brief overview of Literature Survey	16
2.1	Research Papers a. Abstract of the research paper b. Inference drawn from the paper	16
2.2	Patent search a. Title of the patent and year of the patent b. Summary of the patent c. Link 1. European Patent: http://worldwide.espacenet.com/, Link 2. Google Patents: https://patents.google.com/	20
2.3	Inference Drawn	22
2.4	Comparison with the Existing Systems	22
3.	**Requirement Gathering for the proposed System**	23
3.1	Introduction to Requirement Gathering	23
3.2	Functional Requirements	24
3.3	Non-Functional Requirements	24
3.4	Hardware, Software, Technology and tools utilized	25
3.5	Constraints	26
4.	**Proposed Design**	27
4.1	Block diagram of the system	27
4.2	Modular design of the system	27

4.3	Detailed Design (Flowchart)	28
4.4	Project Scheduling & Tracking using Timeline / Gantt Chart	30
5.	**Implementation of the Proposed System**	31
5.1	Methodology employed for development	31
5.2	Algorithms and flowcharts for the respective modules developed	32
5.3	Datasets source and utilization	34
6.	**Testing of the Proposed System**	36
6.1	Introduction to testing	36
6.2	Types of tests Considered	36
6.3	Various test case scenarios considered	37
6.4	Inference drawn from the test cases	38
7.	**Results and Discussions**	39
7.1	Screenshots of User Interface (UI) for the respective module	39
7.2	Performance Evaluation measures	42
7.3	Input Parameters / Features considered	44
7.4	Comparison of results with existing systems	45
7.5	Inference drawn	46
8.	**Conclusion**	47
8.1	Limitations	47
8.2	Conclusion	48
8.3	Future Scope	49
	References	49
	Appendix	51
1	**Paper I**	52
a	Paper I	52
b	Plagiarism Report of Paper I	57
c	Project review sheet	58

List of Figures:

Fig no.	Heading	Page no.
4.1.1	Block Diagram	27
4.2.1	Modular Diagram	27
4.3.1	DFD	28
4.3.2	Flow Chart of the system	28
4.3.3	Flow Chart of the model	29
4.4.1	Gantt Chart	30
5.2.1	Working of SVM Algorithm	34
5.3.1	Storage in Firebase	35
5.3.2	Database storage view	35
5.3.3	Details of User for Prediction	35
7.1.1	Screenshot for Intro page	39
7.1.2	Screenshot for SignUp page	40
7.1.3	Screenshot for Login page	40
7.1.4	Screenshot for Home page	41
7.1.5	Screenshot for BMI Calculator page	41
7.1.6	Screenshot for Enquiry page	42
7.1.7	Screenshot for ChatApp page	42

7.2.1	Confusion Matrix of our Model	43
7.2.2	Precision, recall and F-score of SVM Model	44
7.3.1	Login Page and Signup page input parameters	44
7.3.2	BMI Page input parameters	45
7.3.3	Enquiry Page Input Parameters	45

List of Tables:

Table no.	Heading	Page no.
2.4.1	Comparisons	22
3.1.1	Requirements of the system	23
7.4.1	Comparison between existing system and our system	46
7.5.1	Inferences	46

Abstract

Heart disease remains one of the leading causes of mortality worldwide, emphasizing the critical need for effective prediction systems to aid in early diagnosis and intervention. In this study, we propose a heart disease prediction system utilizing Support Vector Machine (SVM), a powerful machine learning algorithm known for its capability to handle high-dimensional data and nonlinear relationships.The system begins with the collection of essential patient data, including demographic information, medical history, and clinical measurements such as blood pressure, cholesterol levels, and electrocardiogram readings. This data is preprocessed to handle missing values, normalize features, and encode categorical variables, ensuring its compatibility with the SVM algorithm.Next, the SVM model is trained using the preprocessed data to learn patterns and relationships between the input features and the presence or absence of heart disease. The SVM algorithm aims to find the optimal hyperplane that best separates the data points belonging to different classes, maximizing the margin between them while minimizing classification errors.Evaluation of the SVM model is conducted using appropriate performance metrics such as accuracy, precision, recall, and F1-score on a separate validation dataset to assess its predictive capability and generalization ability. The model's performance is compared with other machine learning algorithms to validate its effectiveness in heart disease prediction.The deployed SVM-based heart disease prediction system provides a user-friendly interface where individuals can input their relevant health information, and the trained model generates predictions regarding their likelihood of developing heart disease. Additionally, the system allows healthcare professionals to utilize the predictive model as a supplementary tool for early risk assessment and personalized patient care.Through comprehensive experimentation and validation, our results demonstrate the efficacy and reliability of the SVM-based heart disease prediction system in accurately identifying individuals at risk of heart disease. This system holds significant potential for improving preventive healthcare strategies and reducing the burden of cardiovascular diseases on global public health.

Chapter 1: Introduction

1.1. Introduction:

In the relentless pursuit of improving public health outcomes, the development of a Heart Disease Prediction System emerges as a pivotal endeavor. Amidst the escalating global burden of cardiovascular diseases, this system stands as a beacon of hope, offering a proactive approach towards mitigating the prevalence and impact of heart-related ailments. By harnessing the formidable potential of artificial intelligence and machine learning, this innovative system endeavors to revolutionize the landscape of cardiovascular healthcare.

Key Components of the Heart Disease Prediction System:

At its core, the Heart Disease Prediction System integrates multifaceted components to achieve its noble objectives. Through meticulous data collection and integration, it compiles a comprehensive repository of relevant information spanning demographic details, medical history, lifestyle factors, and crucial clinical parameters. This rich dataset serves as the cornerstone for the deployment of advanced machine learning algorithms, which meticulously analyze intricate patterns and relationships within the data. Through iterative model training, validation, and refinement processes, the system achieves unparalleled predictive accuracy, enabling the precise identification of individuals at heightened risk of developing heart diseases.

Benefits and Implications:

The implications of the Heart Disease Prediction System are profound and far-reaching. By empowering healthcare professionals with timely and accurate risk assessments, it enables proactive intervention strategies aimed at averting the onset or progression of cardiovascular ailments. Moreover, the system facilitates personalized healthcare delivery, tailoring interventions and treatment plans to suit the unique needs and risk profiles of individual patients. Ultimately, through its amalgamation of data-driven insights and clinical expertise, the Heart Disease Prediction System heralds a new era in preventive medicine, where early detection and targeted interventions converge to safeguard cardiovascular health on a global scale.

1.2. Motivation:

The motivation for creating a Heart Disease Prediction System arises from the pressing need to combat the widespread impact of cardiovascular diseases. Despite medical advancements, heart ailments remain a

leading cause of mortality worldwide. Harnessing artificial intelligence and machine learning, this system aims to revolutionize healthcare by proactively identifying individuals at risk of heart diseases. Behind this innovation lies a shared vision among researchers and healthcare professionals: to empower individuals with personalized interventions that prevent heart-related complications. Additionally, the human element of compassion propels this endeavor forward, driven by a desire to alleviate suffering and preserve lives. Every statistic represents a human story, reinforcing the urgency to develop effective preventive measures. Ultimately, the motivation transcends technology, embodying a commitment to saving lives and fostering healthier communities for future generations.

1.3. Problem Definition:

The problem addressed by the Heart Disease Prediction System revolves around the need for early detection and prevention of cardiovascular diseases. Despite advances in medical science, heart-related ailments continue to pose a significant threat to global health, with mortality rates remaining high. The challenge lies in identifying individuals at risk of developing heart diseases before symptoms manifest, enabling timely interventions to mitigate adverse outcomes.

Key aspects of the problem include:

Prevalence of Cardiovascular Diseases: Heart diseases, including coronary artery disease, heart failure, and arrhythmias, are widespread and contribute significantly to morbidity and mortality worldwide.

Lack of Early Detection: Traditional healthcare approaches often rely on symptomatic presentation, leading to delays in diagnosis and treatment initiation. Early warning signs may be subtle or absent, making it challenging to identify high-risk individuals.

Personalized Risk Assessment: Cardiovascular risk factors vary among individuals based on genetics, lifestyle, and medical history. A one-size-fits-all approach to risk assessment may overlook crucial indicators and fail to provide targeted interventions.

Resource Allocation: Limited healthcare resources necessitate strategic allocation to maximize impact. Prioritizing individuals at higher risk of heart diseases for preventive measures can optimize resource utilization and improve overall health outcomes.

Empowerment and Prevention: Empowering individuals with knowledge of their cardiovascular risk enables proactive lifestyle modifications and adherence to treatment plans, fostering a culture of prevention rather than reactive treatment.

1.4. Existing Systems:

Existing systems for heart disease prediction encompass traditional risk assessment tools like the Framingham Risk Score and machine learning models that analyze diverse patient data to forecast future cardiovascular events. Mobile applications and wearable devices track health metrics for personalized risk assessments, while telemedicine and remote monitoring systems enable proactive intervention. Genetic testing allows for personalized risk assessments based on genetic predispositions. Population health management platforms aggregate data to identify trends and tailor interventions. These systems aim to identify individuals at risk and implement preventive measures to reduce the burden of cardiovascular diseases. Continued research and innovation promise to enhance the accuracy and effectiveness of heart disease prediction systems, ultimately leading to improved patient outcomes and reduced mortality rates.

1.5. Lacuna of the Existing System:

The existing heart disease prediction systems face several challenges:

Limited Data Integration: They often overlook valuable data sources like genetics and socio-economic factors.

Generalization Issues: Machine learning models may not accurately predict risk for diverse demographic groups.

Static Risk Assessment: Traditional tools provide a one-time risk assessment, ignoring dynamic changes in health.

Lack of Longitudinal Data: They lack data to track disease progression and intervention effectiveness over time.

Interoperability Challenges: Fragmented healthcare systems hinder seamless data integration.

Ethical and Privacy Concerns: Using sensitive health data raises privacy and discrimination concerns.

Limited Access and Adoption: Accessibility barriers hinder widespread adoption, especially in marginalized communities.

Addressing these challenges requires improved data integration, model generalization, longitudinal data collection, interoperability standards, ethical guidelines, and enhanced accessibility strategies. Collaboration and continued research are crucial for overcoming these obstacles.

1.6. Relevance of the Project:

The Heart Disease Prediction System project is crucial as it:

- Predicts cardiovascular disease risk, enabling early intervention.
- Improves health outcomes by tailoring treatments based on risk assessment.
- Optimizes resource allocation in healthcare by targeting high-risk individuals.
- Empowers patients with knowledge to make informed health decisions.
- Shifts focus towards preventive medicine, reducing disease burden.
- Saves lives by identifying individuals at risk before symptoms appear.
- Enhances healthcare efficiency and reduces costs through proactive intervention.
- Addresses disparities in access to preventive care by targeting high-risk groups.
- Aligns with global health goals by promoting proactive management of chronic diseases.
- Represents a transformative approach to healthcare delivery, prioritizing prevention over treatment.

Chapter 2: Literature Survey

A.Overview of literature survey:

The literature survey provides a comprehensive overview of existing research and studies relevant to the Heart Disease Prediction System project. It summarizes findings, methodologies, and advancements in heart disease prediction, encompassing traditional risk assessment tools, machine learning models, mobile applications, genetic testing, and population health management strategies. By synthesizing this body of knowledge, the survey identifies gaps, challenges, and opportunities for further research, guiding the development of the project. Additionally, it offers insights into best practices, emerging trends, and areas for innovation in cardiovascular healthcare. Overall, the literature survey serves as a foundation for informed decision-making, ensuring that the project builds upon existing knowledge and contributes meaningfully to the field.

2.1. Research Papers :

1. Devansh Shah, Samir Patel and Santosh Kumar Bharti. "Heart Disease Prediction using Machine Learning Techniques" in 2020.

Reference: https://link.springer.com/article/10.1007/s42979-020-00365-y

Abstract:

Heart disease, alternatively known as cardiovascular disease, encases various conditions that impact the heart and is the primary basis of death worldwide over the span of the past few decades. It associates many risk factors in heart disease and a need of the time to get accurate, reliable, and sensible approaches to make an early diagnosis to achieve prompt management of the disease. Data mining is a commonly used technique for processing enormous data in the healthcare domain. Researchers apply several data mining and machine learning techniques to analyze huge complex medical data, helping healthcare professionals to predict heart disease. This research paper presents various attributes related to heart disease, and the model on the basis of supervised learning algorithms as Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithm. It uses the existing dataset from the Cleveland database of UCI repository of heart disease patients. The dataset comprises 303 instances and 76 attributes. Of these 76 attributes, only 14 attributes are considered for testing, important to substantiate the performance of different algorithms. This research paper aims to envision the probability of developing heart disease in the patients. The results portray that the highest accuracy score is achieved with K-nearest neighbor.

Inference:
The authors have used a dataset from the UCI Machine Learning repository to analyze various attributes related to heart disease and predict the disease using supervised learning algorithms such as Naïve Bayes, decision tree, K-nearest neighbor, and random forest. The paper also discusses the use of other data mining techniques such as time series, clustering, association rules, support vector machine, and genetic algorithm to improve the accuracy of the model.

2. Md.Julker Nayeem, Sohel Rana,and Md. Rabiul Islam. " Prediction of Heart Disease Using Machine Learning Algorithms" in 2022.
Reference: https://www.ej-ai.org/index.php/ejai/article/view/13/13

Abstract:
Heart disease has become one of the alarming issues of death. It is accountable for fatty plaques in the arteries. If this fatal condition can be identified early, we can preserve many people's arteries. Different types of supervised machine learning algorithms are applied in our research paper in order to predict heart disease existence in patient body.Besides this, we have focused on an efficient way to improve the performance of our applied classifiers.Imputing mean value technique is applied to handle null values present in our dataset.The features which are unnecessary are removed by using the info-gain feature selection technique.In order to calculate prediction accuracy, K-Nearest Neighbors (KNN), Naive Bayes and Random Forest are applied to the heart disease dataset.Accuracy, precision, recall, F1-score,and ROC are calculated which help us to compare the performance of the classification models. Handling null values on a particular column by imputing mean values of that column and our applied info-gain feature selection technique has aided us in improving the accuracy of our prediction models.Random Forest among all has given the best classification accuracy which is 95.63% with precision, recall, F1-score and ROC are 0.93, 0.92, 0.92and 0.9,respectively.

Inference:
The dataset for this study was obtained from Kaggle. There are 65535 records and predict the disease using supervised learning algorithms such as Naïve Bayes, decision tree, K-nearest neighbor, and random forest. The paper also discusses the use of other data mining techniques such as time series, clustering, association rules, support vector machine, and genetic algorithm to improve the accuracy of the model.The KNN model has the highest accuracy (88.52%), compared to other kinds of prediction models.They predicted heart disease using machine learning algorithms.They achieved the highest accuracy from Random Forest (RF) and is 90.16%.They predicted heart disease using machine learning algorithms.They achieved the highest accuracy from the KNN model which is 90.789%.They obtained an accuracy of 87%, which is the maximum the KNN model could provide.A feature selection strategy is used in this study info-gain to increase the

classification model's accuracy. Logistic regression has produced the best accuracy, which is 92.76%. Using MLP and SVM, predicted cardiac disease. They obtained a two-class classification accuracy from SVM of 92.45%, which is the highest. used machine learning methods to forecast hepatitis disease. They obtained a Random Forest accuracy of 92.41%, which is the highest.

3.C.M.Bhatt, Parth Patel, Tarang Ghetiya, Pier Luigui Mazzeo. " Effective Heart Disease Prediction Using Machine Learning Techniques" in 2022.

Reference: https://www.mdpi.com/1999-4893/16/2/88

Abstract:

The diagnosis and prognosis of cardiovascular disease are crucial medical tasks to ensure correct classification, which helps cardiologists provide proper treatment to the patient. Machine learning applications in the medical niche have increased as they can recognize patterns from data. Using machine learning to classify cardiovascular disease occurrence can help diagnosticians reduce misdiagnosis. This research develops a model that can correctly predict cardiovascular diseases to reduce the fatality caused by cardiovascular diseases. This paper proposes a method of k-modes clustering with Huang starting that can improve classification accuracy. Models such as random forest (RF), decision tree classifier (DT), multilayer perceptron (MP), and XGBoost (XGB) are used. GridSearchCV was used to hypertune the parameters of the applied model to optimize the result. The proposed model is applied to a real-world dataset of 70,000 instances from Kaggle. Models were trained on data that were split in 80:20 and achieved accuracy as follows: decision tree: 86.37% (with cross-validation) and 86.53% (without cross-validation), XGBoost: 86.87% (with cross-validation) and 87.02% (without cross-validation), random forest: 87.05% (with cross-validation) and 86.92% (without cross-validation), multilayer perceptron: 87.28% (with cross-validation) and 86.94% (without cross-validation). The proposed models have AUC (area under the curve) values: decision tree: 0.94, XGBoost: 0.95, random forest: 0.95, multilayer perceptron: 0.95. The conclusion drawn from this underlying research is that a multilayer perceptron with cross-validation has outperformed all other algorithms in terms of accuracy. It achieved the highest accuracy of 87.28%.

Inference:

There is the use of various machine learning algorithms in a study related to predicting heart disease. The study utilized a dataset from the Cleveland Clinic Foundation and implemented various ML algorithms, such as decision tree, logistic regression, random forest, naive Bayes, and support vector machine (SVM), to develop prediction models. A 10-fold cross-validation approach was employed during the model development process. The results indicated that the decision tree algorithm achieved the highest accuracy in predicting heart disease, with a rate of 93.19%, followed by the SVM algorithm at 92.30%. Another study mentioned in the paper compared multiple algorithms to identify the most efficient feature selection approach for anticipating cardiovascular illness. The three well-known feature selection methods (filter, wrapper, and embedding) were first taken into account, and then a feature subset was recovered from these three algorithms using a Boolean process-based common "True" condition. This technique involved retrieving feature subsets in two stages.

4. Muntasir Mamun, Afia Farjana, Miraz Al Mamun, Md Salim Ahammed, Md Minhazur Rahman. "Heart failure survival prediction using machine learning algorithm: am I safe from heart failure?" in 2022.

Reference: https://ieeexplore.ieee.org/abstract/document/9817303

Abstract:

Heart Failure (HF) is a prevalent ailment worldwide, and despite significant medical advancements in the past few decades, cardiovascular disease is still the leading cause of death. Although HF itself is a critical risk for patient survival, other coexisting pathophysiological conditions can present a significant threat to patient survival. Because so many elements contribute to a patient's survival in heart failure, predicting the chances of survival without using a computational technique can be difficult for cardiac doctors, eventually preventing the patient from receiving correct care. Fortunately, categorization and prediction models exist, which can assist cardiologists in designing proper treatment schemes using relevant medical data. This study aims to develop prediction models for patient survival in HF conditions. In this paper, we analyzed the UCI heart failure dataset containing relevant medical information of 299 HF patients. We applied several machine learning classifiers to predict the patient survival from HF-related pathophysiological parameters and analyzed the features corresponding to the most crucial risk factors using the correlation matrix. Our prediction models used the following machine learning techniques- Logistic Regression, Decision Tree, Support Vector Machine, XGBoost, LightGBM, Random Forest, KNN, and Bagging and were able to find a better result. Also, this paper presents a comparative study by analyzing the performance of different machine learning algorithms. Our analysis indicates that LightGBM achieved the highest Accuracy of 85% and AUC of 93% in predicting patient survival of HF patients compared to other machine learning algorithms.

Inference:

The study compares the performance of several machine learning models, including Logistic Regression, Decision Tree, Support Vector Machine, XGBoost, LightGBM, Random Forest, KNN, and Bagging. The analysis indicates that LightGBM achieved the highest accuracy of 85% and AUC of 93% in predicting patient survival of HF patients compared to other machine learning algorithms. The paper also provides an overview of related work and discusses the dataset and attribute feature information. The study has potential applications in the healthcare industry and can provide valuable insights into heart failure prediction.

5.Keshav Srivastava, Dilip Kumar Choubey. "Heart Disease Prediction using Machine Learning and Data Mining" in 2020.

Reference:

https://www.researchgate.net/profile/Dilip-Choubey-2/publication/342210798_Heart_Disease_Prediction_using_Machine_Learning_and_Data_Mining/links/5ee8cb6592851ce9e7e7f585/Heart-Disease-Prediction-using-Machine-Learning-and-Data-Mining.pdf

Abstract:

Weighing only 300 grams, Heart is declining the mortality rate at a rapid pace from decades. The major factors that contribute to it are smoking, drinking, unbalanced diet, and many more. Even with these more technical advancements the analysis of the clinical data is a critical challenge. With the use of Machine Learning techniques, it is possible to analyze the data and interpret the cause that lead to heart diseases such as Coronary Heart Disease, Arrhythmia, and Dilated Cardiomyopathy. Many researchers are developing IoT enabled hardware to predict these diseases using various ML and DM techniques. In this study, we propose a novel method to determine the disease using Cleveland Heart Disease Dataset by combining the computational power of various ML and DM algorithms and concluded that among all the algorithms, K-Nearest Neighbors gives the highest accuracy of 87%. Along with this, a web app is developed using flask in python with which the user can enter the attributes and predict the heart disease.

Inference:

Algos are: Decision tree,K-Nearest Algo,Super Vector Machine,Random Forest

The Decision Tree algorithm operates by dividing and conquering, constructing trees in a top-down manner. Pruning eliminates non-significant samples, and entropy is a key concept, calculated as DT_Entropy = (- ∑ij pij * log2(pij)). Euclidean Distance function dis(xi, xj) gathers information through majority votes from k-nearest neighbors. For Support Vector Machines (SVM), the hyperplane equation f(x) = (wTx + b) is found using mathematical computation like (Minw,b),(Ei 1/2 * w^2 + C * Σi=1^n Ei), subject to (yi(wTxi + b) ≥ 1 - Ei) where 'w' and 'b' are coefficients. Bagging involves building decision trees for optimal outcomes, where N trees are averaged for predictions, yielding an error measurement as the standard deviation of trees' predictions.

2.2. Patent Search :

1. METHOD AND SYSTEM FOR PREDICTION OF CORONARY ARTERY DISEASE BY USING MACHINE LEARNING(KR102081055B1)

Inventor: Choi Byeong Geol

The present invention relates to a method and system for predicting coronary artery disease using machine learning. The coronary artery disease prediction method includes a machine learning process and a prediction process using learning data. In the machine learning process, a plurality of

machine learning algorithms are trained using training data to generate first-order prediction models, and a plurality of machine learning algorithms are reinforced by using a combination of prediction results obtained by using the first-order prediction models. Second-order prediction models are generated again, and the most accurate one of the second-order prediction models is set as the measurement prediction model. The prediction process obtains a prediction result set by applying real data about the patient to the first prediction models, and performs the prediction by applying the obtained prediction result set and the real data to the measurement prediction model. By the above-described procedures, it is possible to accurately predict coronary artery disease..

2.CORONARY COMPUTED TOMOGRAPHY CLINICAL DECISION SUPPORT SYSTEM(US20180253531A1)

Inventors: Puneet Sharma, Lucian Mihai Itu,Thomas Flohr, Dorin Comaniciu

A CT-based clinical decision support system provides coronary decision support. With or without CT-FFR, a machine learnt predictor predicts the clinical decision for the patient based on input from various sources. Using the machine learnt predictor provides more consistent and comprehensive consideration of the available information. The clinical decision support may be provided prior to review of coronary CT data by a radiologist and/or treating physician, providing a starting point or recommendation that may be used by the radiologist and/or treating physician.

3. A KIND OF HEART DISEASE RISK FORECASTING SYSTEM(CN109377470A)

Inventors: Ren Haoxing, Liu Yan, Zheng Hui

The invention discloses a kind of heart disease risk forecasting systems, computer vision assembly line including handling cardiac ultrasonic, the computer vision assembly line of the processing cardiac ultrasonic include the probability that view identification classification, the U-Net convolutional neural networks for chamber structures identification, myocardium shape tracking, myocardium shape characteristic vector pickup, the extraction of Characteristics of electrocardiogram data, the extraction of Clinical symptoms data, the deep learning network architecture and data acquisition and training obtain prediction；It assists automatic ultrasound image to distinguish and diagnose by artificial intelligence, calculate multidimensional cardiac muscle SPECKLE TRACKING and the prediction of ischemic heart failure, medical history; clinical data; the comprehensive a large amount data of biomarker realize the prediction to disease on machine learning platform; establish the accurate of first artificial intelligence assistance; it is sensitive; high efficiency, automation can the cardiovascular disease screening of scale realize low cost, the heart disease forecasting and early diagnosis system of high accuracy.

4. SYSTEM AND METHOD FOR HEART FAILURE PREDICTION(US20110119078A1)

Inventors: Gadi CotterBeth, Davison Weatherley

A method for monitoring the health status of a human subject includes the capturing of medical data concerning the health of the subject at defined intervals using a questionnaire. The questionnaire provides a standard script for data capture. Part of the captured data is constrained to a Likert scale while other data is on a visual analog scale. The captured data further includes an assessment by a physician of health symptoms of the subject. The captured data is input into a computer, provided to an algorithm that is configured to assess a risk of acute heart failure. The risk of acute heart failure is computed using the algorithm and the captured data from a plurality of the defined intervals. In one method, the health status of the subject as being either improved or worsening is output to the physician as a function of a value of the computed risk. In another method, a survival function outcome for the subject is predicted using the output of the algorithm.

2.3. Inference Drawn:

- Personalized Risk Assessment: While these systems utilize various data sources, they may lack personalization in risk assessment tailored to individual patient characteristics, lifestyle factors, and genetic predispositions.
- Integration with existing healthcare workflows: Seamless integration of these systems into existing clinical workflows may be challenging and require significant effort.

2.4. Comparison with the Existing Systems:

Other System	Our System
Small Data Integration due to which accuracy rate is low.	Large Data Integration and strong Modeling due to which Accuracy rate is high.
ChatBot is not provided in urgent need.	ChatBot is provided to users in real time so that they can easily interact to solve their doubts.
No Advisor is provided.	Provision of Advisor in case the user wants to clear some questions, differences and doubts.
No Visualizations are provided to get clarity.	No. of Graphs are provided to users which gives clear ideas about heart health awareness.

Fig. 2.4.1 Comparisons

Chapter 3: Requirement Gathering for the Proposed System

In this chapter we are going to discuss the resources we have used and how we analyzed what the user actually needs and what we can provide. We will also discuss the functional and non-functional requirements and finally the software and hardware used.

3.1. Introduction to Requirement Gathering:

The Requirement Gathering is a process of requirements discovery or generating list of requirements or collecting as many requirements as possible by end users. It is also called as requirements elicitation or requirement capture.

The requirements gathering process consists of six steps :

- Identify the relevant stakeholders
- Establish project goals and objectives
- Elicit requirements from stakeholders
- Document the requirements
- Confirm the requirements
- Prioritize the requirements

USE CASE	DESCRIPTION
Register and Login	The patients and users register on the Heart Disease Prediction website and can Login to use features provided by the website to predict heart disease.
Prediction	User is supposed to enter the details of BMI, Age, Systolic Blood Pressure, Diastolic Blood Pressure which in turn tells the user if he or she is suffering from any heart disease or not.
ChatBot	It is a real time chatting bot which can be used by users to clear their queries in real time.
ChatApp	Users can chat with the respective healthcare professionals in the industry
Enquiry	Users can send their query/Feedback in brief if they are unable to interact on chatbot easily.
BMI Calculator	Users can calculate their BMI by entering their height and weight
Send Email	Email is received to the respected authority in reply to which the they will send the answer to the enquiry/feedback

Table No: 3.1.1 Requirements of the system

3.2. Functional Requirements:

- User registration and authentication for healthcare professionals and patients.
- Input forms for collecting patient data including demographics, medical history, symptoms, and test results.
- Provide an Advisor for users to clear their doubts.
- Implementation of heart disease prediction algorithms and risk assessment models.
- Visualization of prediction results and risk factors.
- Secure storage and access control mechanisms for patient data.
- Provide ChatBot to interact in real time.
- Provides an enquiry Page to send their query and feedback in brief.

3.3.Non-Functional Requirements:

- User interface should be simple and easy to interpret.
- The dataset used for the prediction model is highly accurate and precise.
- The dashboard should show the accurate information.
- The trained data should be more accurate and reliable.
- It should be platform independent.
- The application is made more scalable with an advanced search feature that filters and delivers requested data from a huge amount of data.
- Performance requirements: Response time, scalability, and system reliability.
- Security requirements: Compliance with medical data privacy regulations (such as HIPAA), encryption of sensitive data, and access control measures.
- Usability requirements: Intuitive user interface, accessibility features, and support for multiple devices and browsers.
- Compatibility requirements: Compatibility with existing healthcare systems or platforms.
- Maintenance and support requirements: Procedures for system updates, bug fixes, and user support.

3.4.Hardware, Software, Technology and Tools Utilised:

A. Hardware Requirements:-

a. Minimum 8 GB RAM

b. Core I5 7th Gen processor

c. Disk space of 4GB

B. Software Requirements:-

a. Python

b. Nodejs

c. Reactjs

d.Flask

e. Bootstrap(Front-end Framework)

f. Google Colab/Jupyter Notebook

Techniques:-

- **Python:-** Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation.
- **Nodejs:-** Node.js is an open-source, cross-platform, back-end JavaScript runtime environment that runs on a JavaScript Engine and executes JavaScript code outside a web browser, which was designed to build scalable network applications.
- **Reactjs:-** React (also known as React.js or ReactJS) is a free and open-source front-end JavaScript library for building user interfaces based on UI components. It is maintained by Meta (formerly Facebook) and a community of individual developers and companies. React can be used as a base in the development of single-page, mobile, or server-rendered applications with frameworks like Next.js.
- **Bootstrap:-** Bootstrap is a free and open-source CSS framework directed at responsive, mobile-first front-end web development. It contains HTML, CSS and 20 (optionally) JavaScript-based design templates for typography, forms, buttons, navigation, and other interface components.

- **Flask:-** Flask is a web framework, it's a Python module that lets you develop web applications easily. It has a small and easy-to-extend core: it's a microframework that doesn't include an ORM (Object Relational Manager) or such features.It does have many cool features like url routing, template engine. It is a WSGI web app framework.

Tools:-

- **Vscode:-**Visual Studio Code is a streamlined code editor with support for development operations like debugging, task running, and version control. It aims to provide just the tools a developer needs for a quick code-build-debug cycle and leaves more complex workflows to fuller featured IDEs, such as Visual Studio IDE.
- **Google Colab:-** Colaboratory, or "Colab" for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs.

3.5. Constraints:

- Internet Access is required.
- May vary in Accuracy
- Data Privacy and Security

Chapter 4: Proposed Design

4.1. Block Diagram of the proposed system:

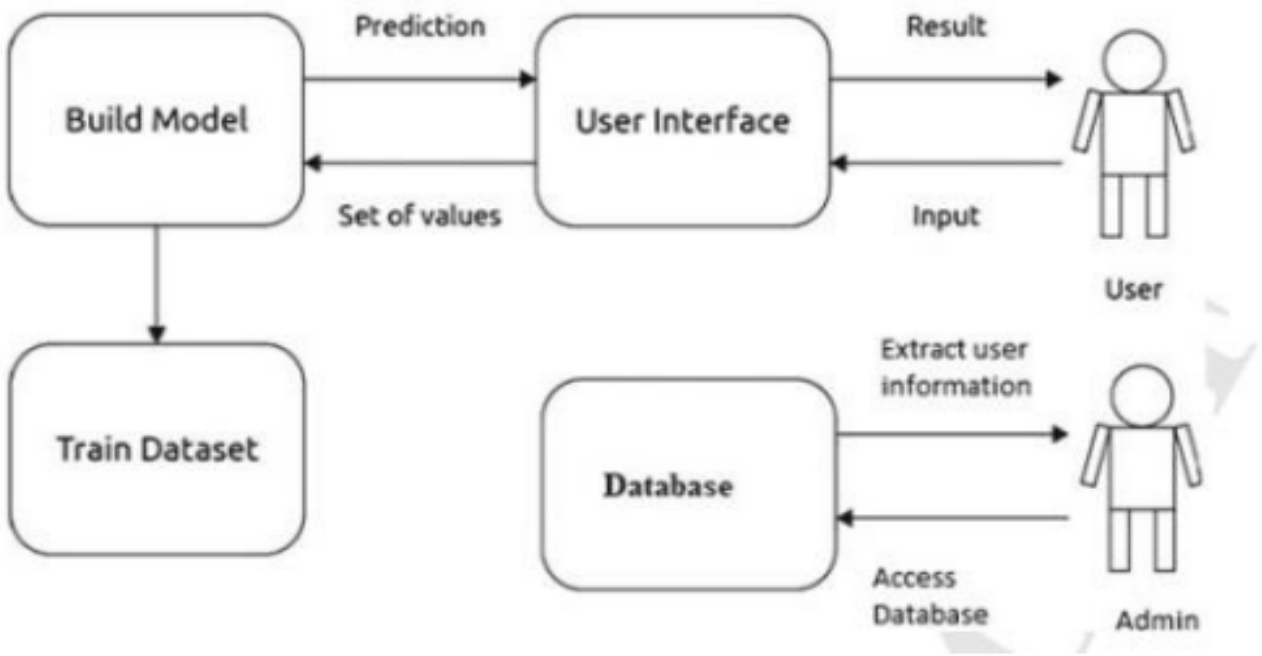


Fig 4.1.1: Block Diagram

4.2. Modular diagram of the system:

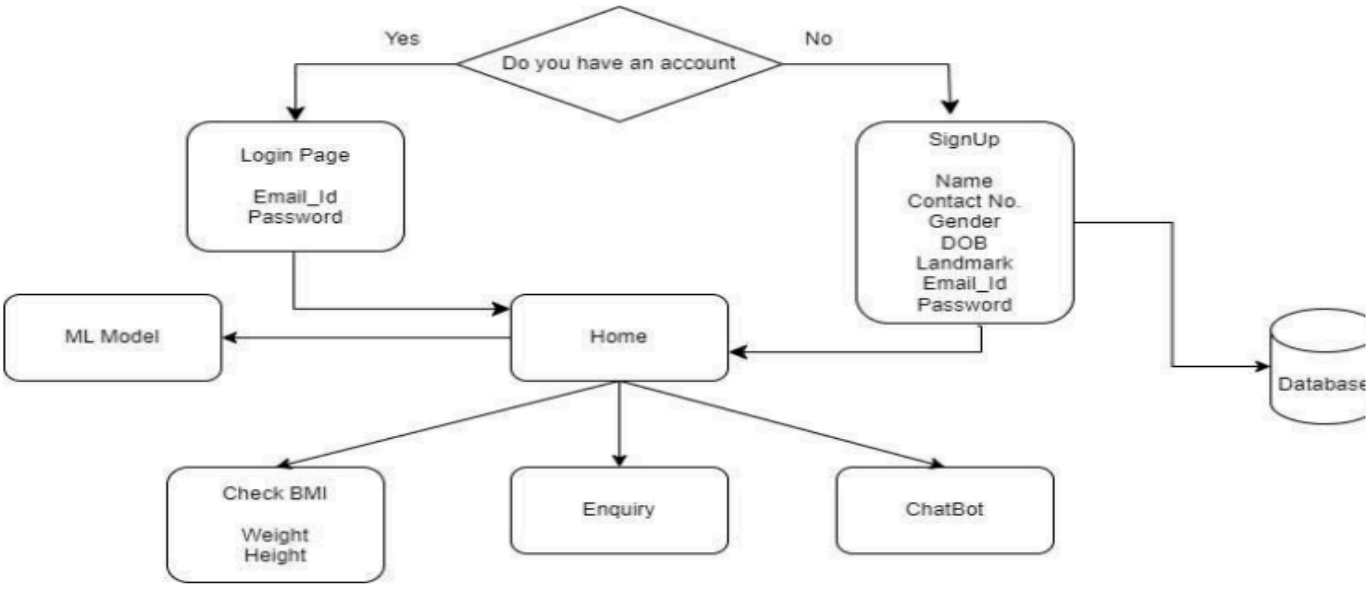


Fig 4.2.1: Modular Diagram

4.3. Detailed Design

DFD Level 0:

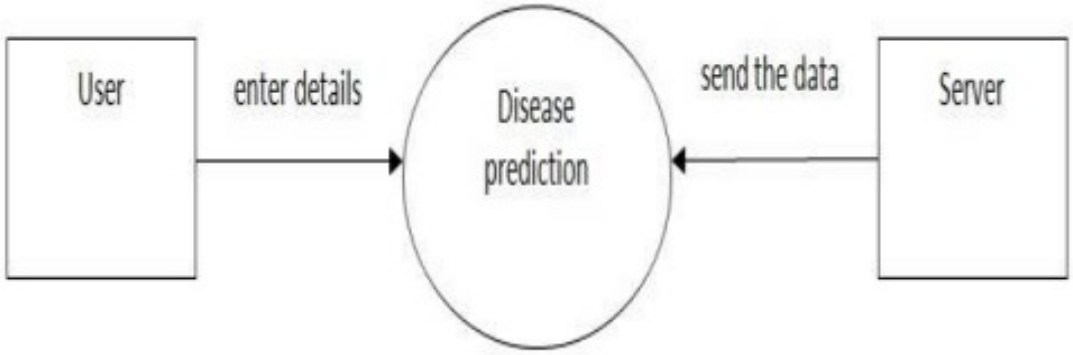


Fig 4.3.1: DFD-Level 0

Flowchart for the proposed system:

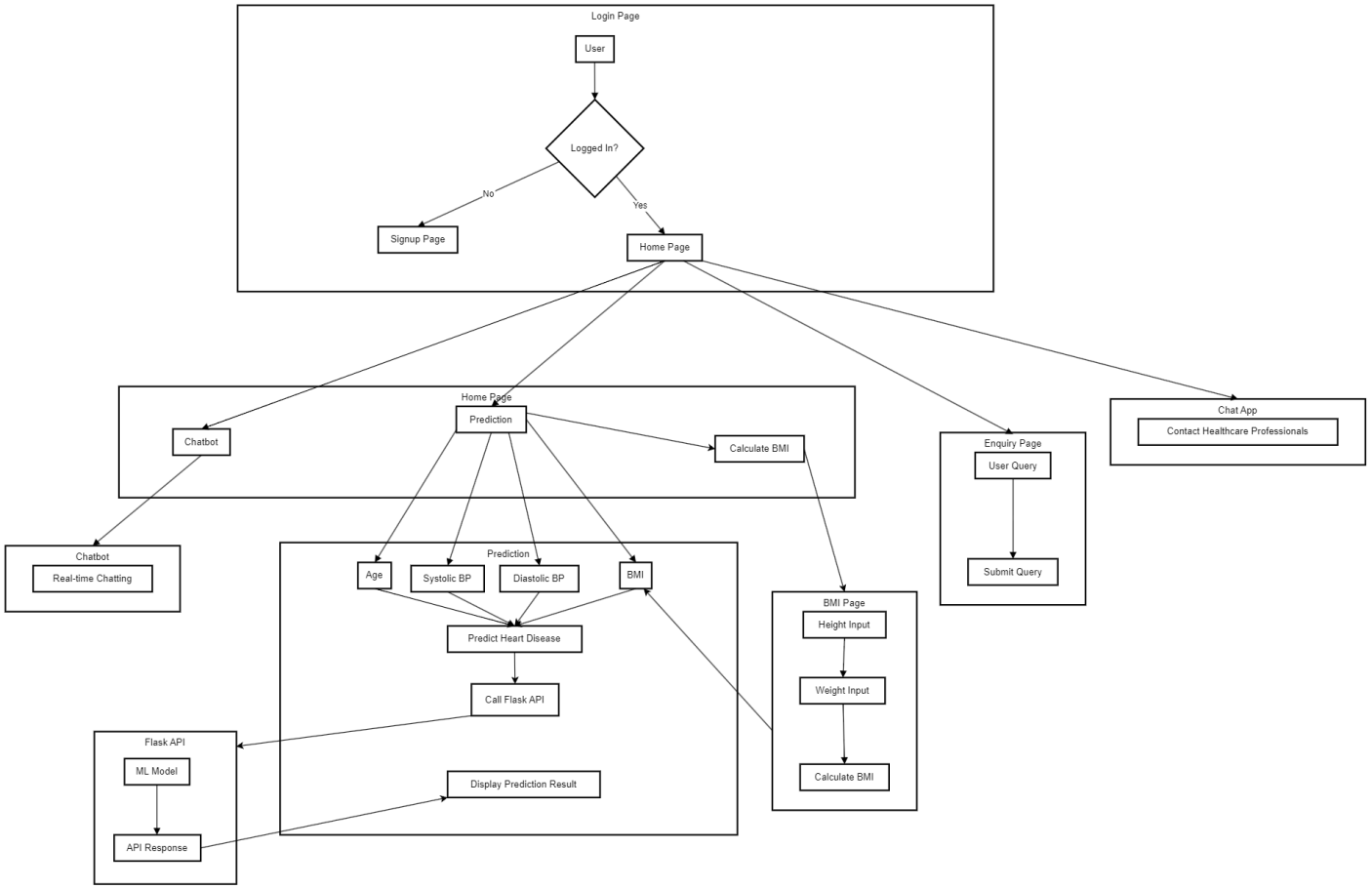


Fig 4.3.2 : Flowchart of System

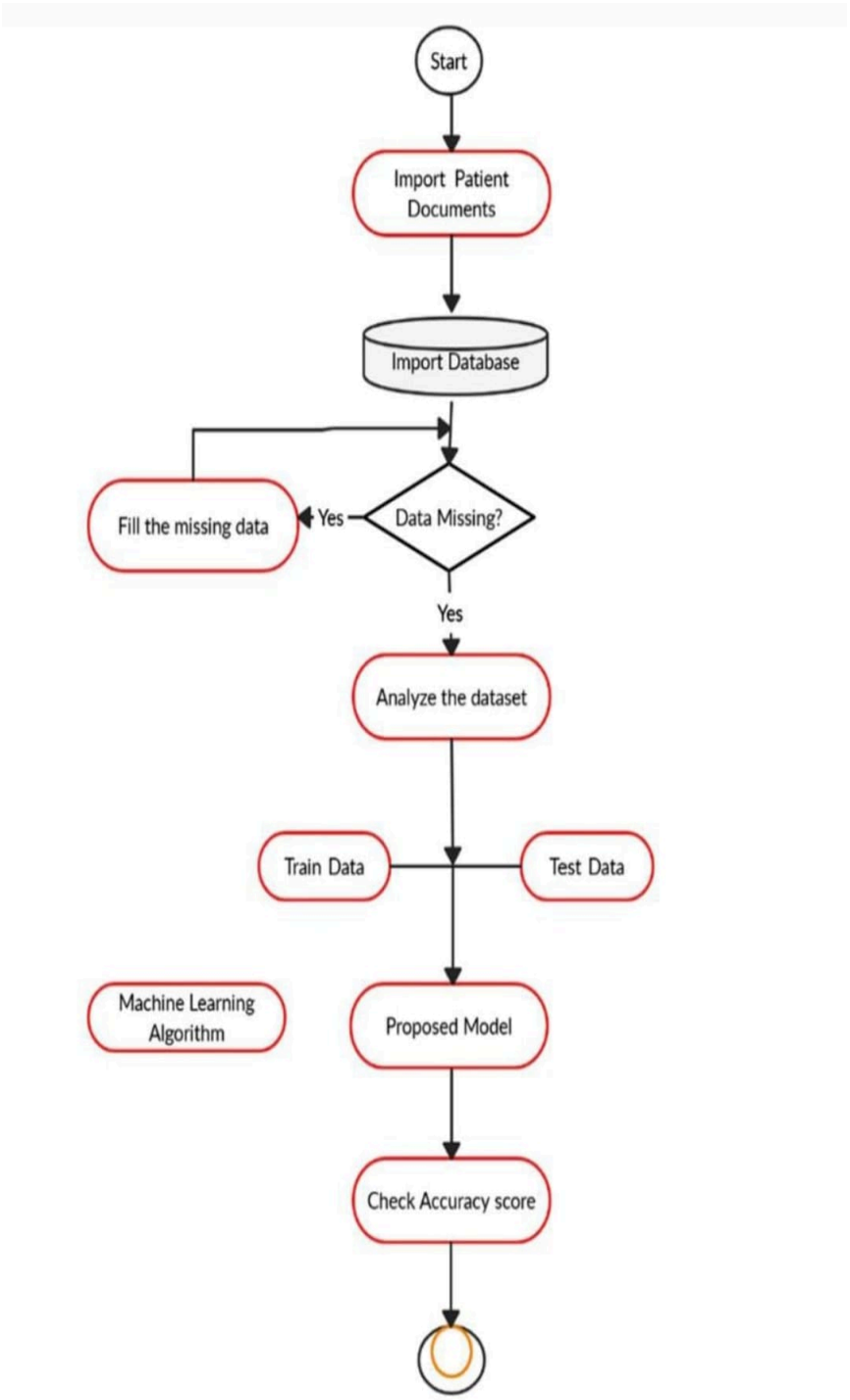


Fig 4.3.3 : Flowchart of Model

4.4. Project Scheduling & Tracking using Time line / Gantt Chart:

The Gantt chart of our project where we worked for the whole semester to create this model is shown in a timeline pattern. It is the most important part to think and design the planning of your topic and so we planned our work like the gantt chart shown.

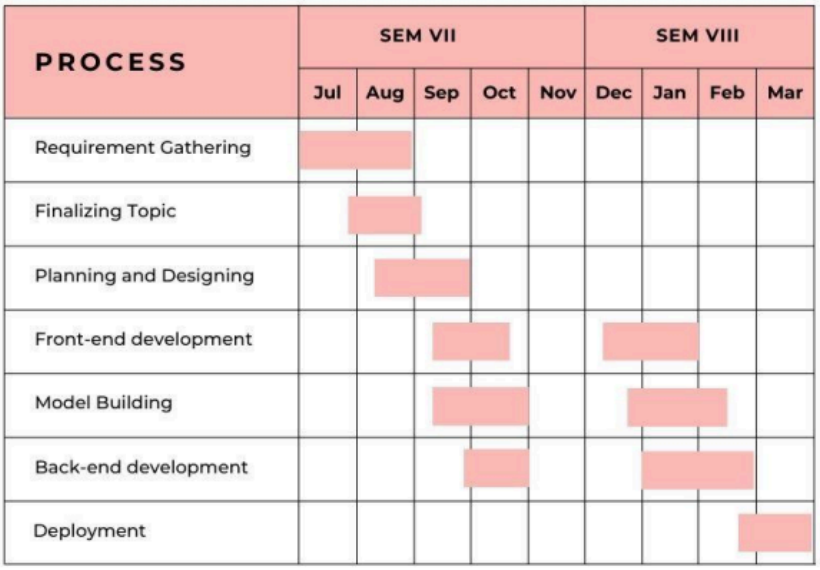


Fig 4.4.1 : Gantt Chart

Chapter 5: Implementation of the Proposed System

5.1.Methodology employed for development:

Data Collection and Preprocessing:

In the data collection and preprocessing stage, the focus is on gathering relevant information from patients and preparing it for analysis. This involves designing a registration form for patients to fill out, which typically includes essential demographic and health-related details. In this case, the registration form requires mandatory fields such as Body Mass Index (BMI), Age, Systolic Blood Pressure, and Diastolic Blood Pressure. Once the data is collected, it needs to be preprocessed to ensure its quality and compatibility with machine learning algorithms. Preprocessing steps may include handling missing values, normalizing or standardizing numerical features, encoding categorical variables, and potentially performing feature engineering to derive new informative features from the raw data.

Model Selection and Evaluation:

The next step involves selecting appropriate machine learning algorithms for heart disease prediction and evaluating their performance to determine the most suitable model. Commonly considered algorithms in this context include Random Forest, Support Vector Classification (SVC), K-Nearest Neighbors (KNN), and Naive Bayes Classification. Evaluation metrics such as accuracy, precision, recall, and F1-score are typically used to assess the models' performance on a validation dataset. By comparing the performance of different algorithms, the most effective model for heart disease prediction can be identified.

Model Training and Website Development:

Once the most suitable model is selected, it needs to be trained using the processed data to enhance its prediction accuracy. This involves feeding the preprocessed data into the chosen model and adjusting its parameters to optimize its performance. After the model is trained and validated, it can be integrated into a user-friendly interface, such as a website, to make it accessible to users. The website should provide an intuitive and easy-to-use platform where users can input their relevant health information, and the trained model can then predict the likelihood of heart disease based on this input. The integration of the model into the website typically requires backend development to handle data processing and model inference, as well as frontend development to create the user interface and ensure a seamless user experience.

Testing and Quality Assurance:

- Perform functionality testing, compatibility testing, and security testing to ensure the website meets quality standards
- Seek feedback from potential users or domain experts to validate the accuracy and usefulness of the prediction system.

Deployment and Maintenance:

- Monitor website performance and user interactions, addressing any issues that arise post-deployment.
- Regularly update the prediction model and website features based on new research findings, user feedback, and technological advancements.

5.2.Algorithms and Flowcharts for the respective modules developed:

We have mainly used SVM algorithm for our project:

a) Support Vector Machine (SVM) :

Support Vector Machine (SVM) is a supervised machine learning algorithm primarily used for classification tasks, although it can be adapted for regression as well. It's particularly effective in cases where the data is linearly separable or can be transformed into a higher-dimensional space where it is separable. SVM aims to find the optimal hyperplane that best separates data points belonging to different classes while maximizing the margin between them.

1.Objective:

The main objective of SVM is to find a hyperplane in an N-dimensional space (where N is the number of features) that distinctly classifies the data points into different classes. If the data is linearly separable, SVM finds the hyperplane that maximizes the margin, i.e., the distance between the hyperplane and the nearest data points from each class.

2.Linear Separability:

For a binary classification problem, if the data is linearly separable, SVM tries to find the hyperplane equation:

$$w \cdot x + b = 0$$

where w is the weight vector perpendicular to the hyperplane, x represents the input features, and b is the bias term.

3.Margin:

The margin is defined as the distance between the hyperplane and the nearest data points from each class. SVM aims to maximize this margin. The points closest to the hyperplane are called support vectors. They are crucial in defining the decision boundary.

4.Optimization Problem:

SVM solves an optimization problem to find the optimal hyperplane. It involves minimizing the classification error while maximizing the margin. This can be formulated as a constrained optimization problem, often using Lagrange multipliers.

5.Non-Linear Separability and Kernel Trick:

In cases where the data is not linearly separable in the original feature space, SVM uses the kernel trick to map the input features into a higher-dimensional space where they become linearly separable. Commonly used kernels include linear, polynomial, radial basis function (RBF), and sigmoid kernels. This allows SVM to handle non-linear decision boundaries effectively.

6.Regularization Parameter (C):

SVM introduces a regularization parameter (\(C\)) to control the trade-off between maximizing the margin and minimizing the classification error. A higher value of \(C\) allows for fewer misclassifications but may lead to overfitting, while a lower value of \(C\) emphasizes a wider margin, potentially sacrificing some classification accuracy.

7.Prediction:

Once the optimal hyperplane is determined during the training phase, SVM can predict the class label of new data points by evaluating which side of the hyperplane they fall on.

8.Multi-Class Classification:

SVM inherently supports binary classification. For multi-class classification, strategies like one-vs-one or one-vs-all can be employed, where multiple binary classifiers are trained, each distinguishing between two classes.

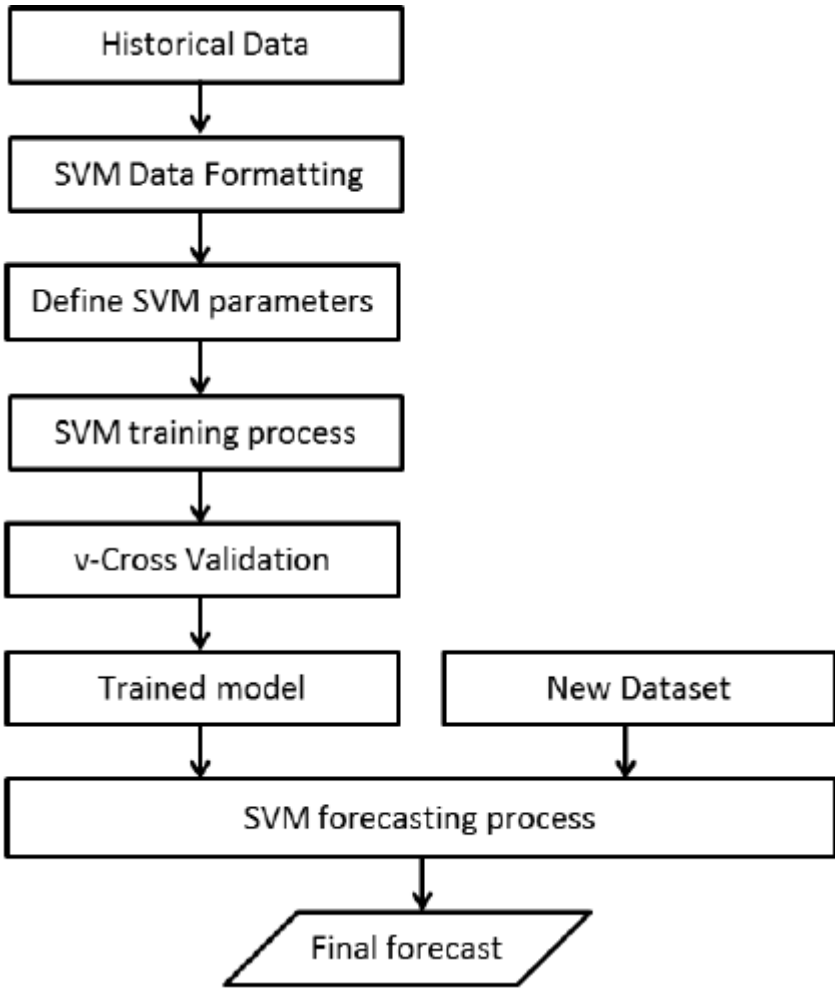


Fig.5.2.1 Working of SVM Algorithm

5.3.Datasets source and utilization:

We have used the dataset "Cardiovascular Disease dataset" from Kaggle as training data for the SVM model. The dataset consists of 70000 records of patients data, 11 features + target.

The features are categorized as follows:

- Age
- Height
- Weight
- Gender
- Systolic blood pressure
- Diastolic blood pressure
- Cholesterol
- Glucose
- Smoking
- Alcohol intake
- Physical activity
- Presence or absence of cardiovascular disease

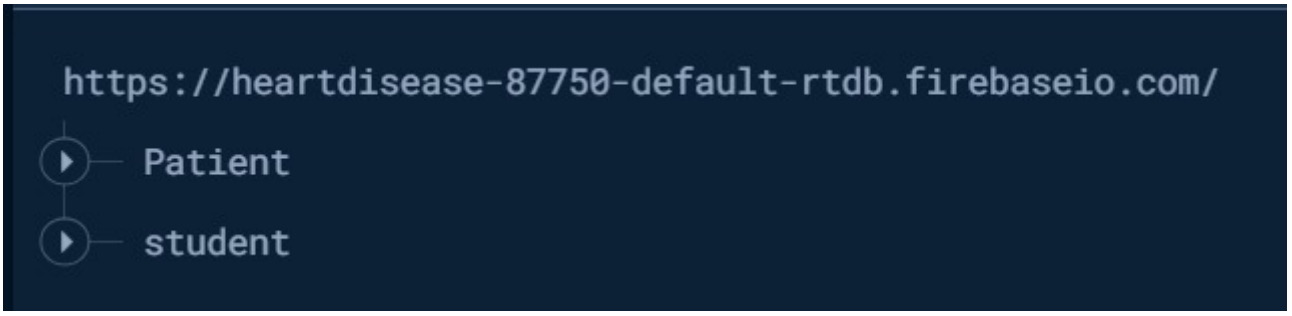


Fig.5.3.1 Storage in Firebase

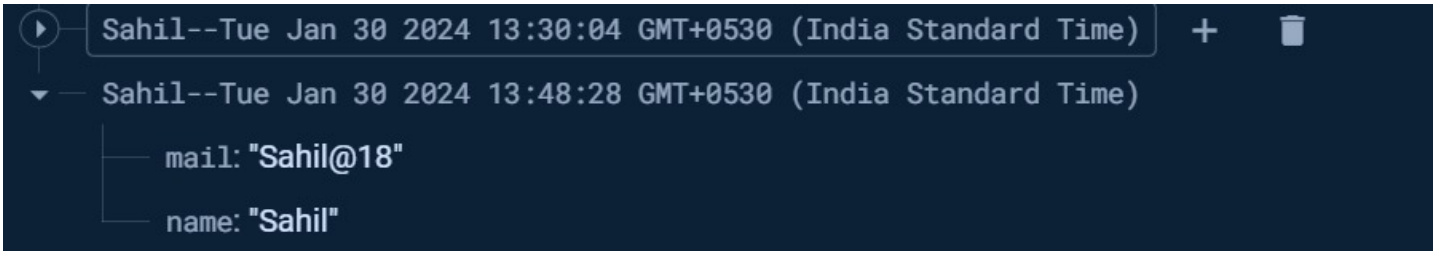


Fig 5.3.2 Database storage view

Fig.5.3.3 Details of User for prediction

Chapter 6: Testing of the Proposed System

6.1.Introduction to Testing :

Testing a heart disease prediction system involves a multifaceted approach to ensure its accuracy and reliability. It begins with unit testing of individual components like data preprocessing and prediction algorithms, followed by integration testing to confirm seamless interaction between modules. Functional testing assesses the system against specified requirements, while regression testing ensures recent changes haven't introduced new bugs. Performance testing evaluates the system's responsiveness under various conditions, while security testing safeguards against potential threats. Usability testing ensures an intuitive interface, and compatibility testing verifies consistent behavior across platforms. Validation testing compares predictions with known medical data, and documentation testing ensures completeness. Error handling testing assesses the system's resilience, while end-to-end testing confirms functionality from input to output. These testing strategies collectively ensure the heart disease prediction system is robust, accurate, and ready for deployment.

6.2.Types of tests Considered:

A. Pre testing phase

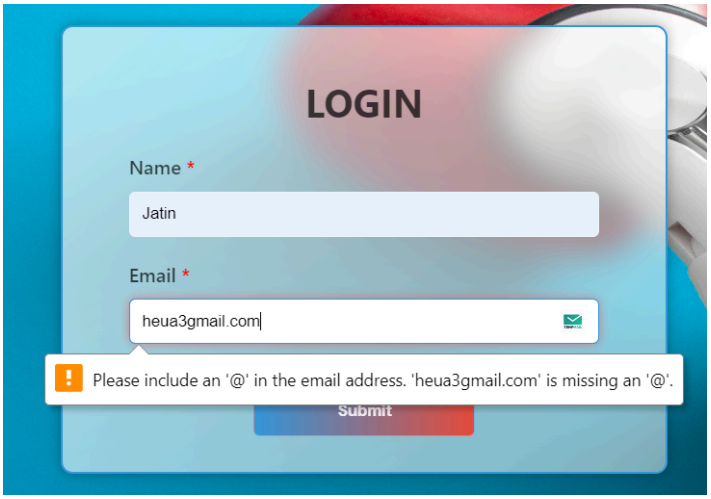
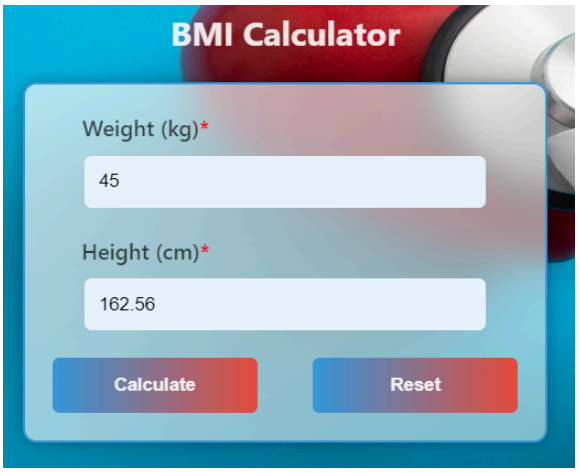
During the design and implementation stage of the system, Nutritional Sciences students developed a questionnaire designed to “pre-test” the concepts underlying the recommendation system. Pre-test questions were developed to elicit information about consumers' notions about making healthy choices in fast food restaurants, and whether and what type of nutrition information or labelling might help consumers select specific foods. Ten volunteers answered questions related to healthy eating at fast food restaurants and nutrition labeling, and findings were used to refine the app's structure, features, and function and to inform the procedures and questions used during the beta testing phase.

B. Beta-Testing Phase

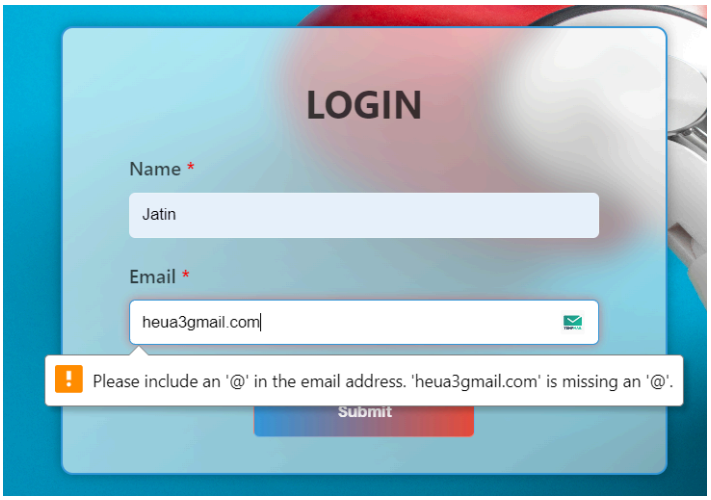
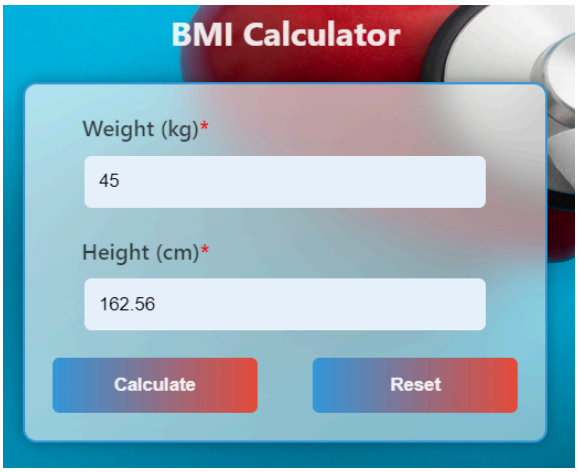
In the pre testing phase of a heart disease prediction system, it's crucial to lay a solid foundation for comprehensive testing. This involves thorough documentation of system requirements, including data sources, algorithm specifications, and intended functionality. Additionally, establishing clear testing objectives and strategies is paramount. This phase also encompasses setting up testing environments and acquiring necessary datasets for validation. Collaborating with domain experts, such as cardiologists, to refine the system's requirements and ensure clinical relevance is essential. By meticulously preparing in this pre testing phase, including defining test cases and scenarios, selecting appropriate testing tools, and

establishing communication channels, teams can streamline the subsequent testing process and increase the likelihood of detecting and addressing issues effectively.

6.3.Various test case scenarios considered:

	Test Cases
Case 1: The input field for 'email' should only accept valid email.	
Case 2: The input field for all input fields must be filled before clicking on the calculate button	
Case 3: The form field should get cleared after hitting the reset button	

6.4.Inference drawn from the test cases:

	Test Cases
Case 1: The input field for 'email' should only accept valid email. The user should not enter invalid format where it doesn't contain "@" or ".com, .in" etc. This email id will then be used by the user for login	
Case 2: The system aims to prevent any calculations based on incomplete or insufficient data, which could lead to inaccurate results or errors.	
Case 3: The reset button helps prevent errors that may arise from accidental submission of incorrect data or confusion due to retained values from previous calculations.	

Chapter 7: Results and Discussions

7.1.Screenshot of Use Interface(UI) for the system:

Fig 7.1.1: Screenshot for Intro page

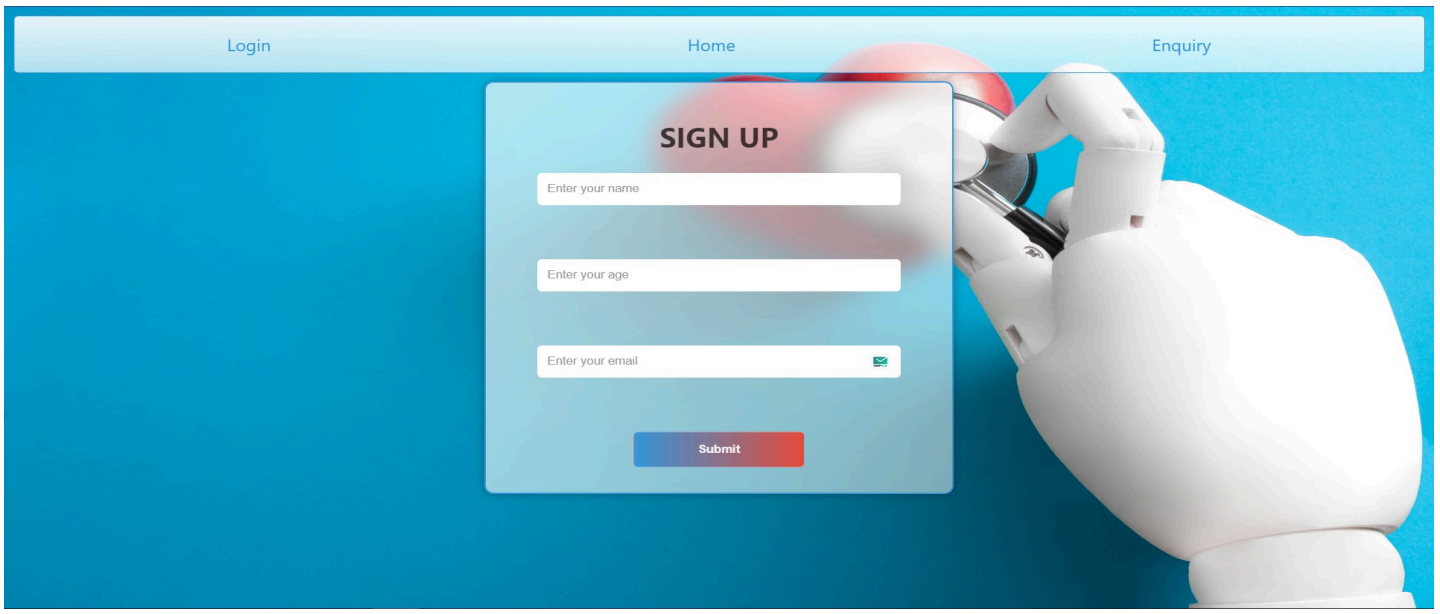


Fig 7.1.2: Screenshot for SignUp page

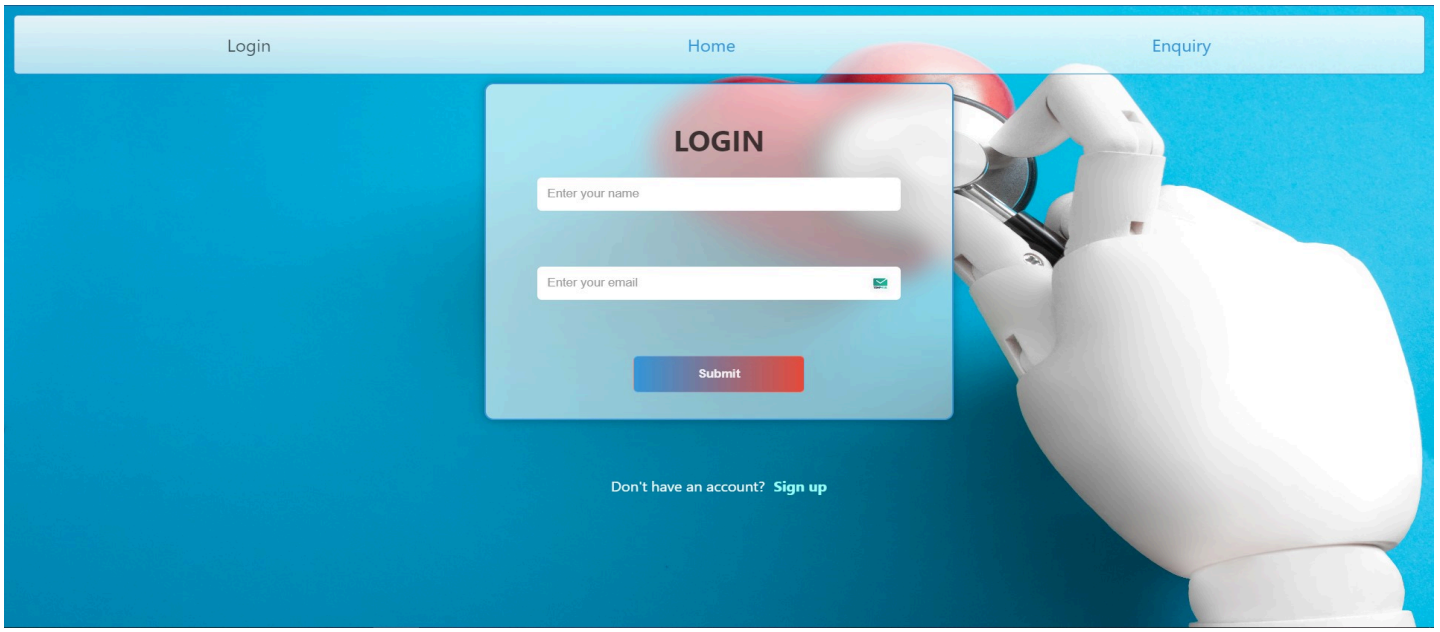


Fig 7.1.3: Screenshot for Login page

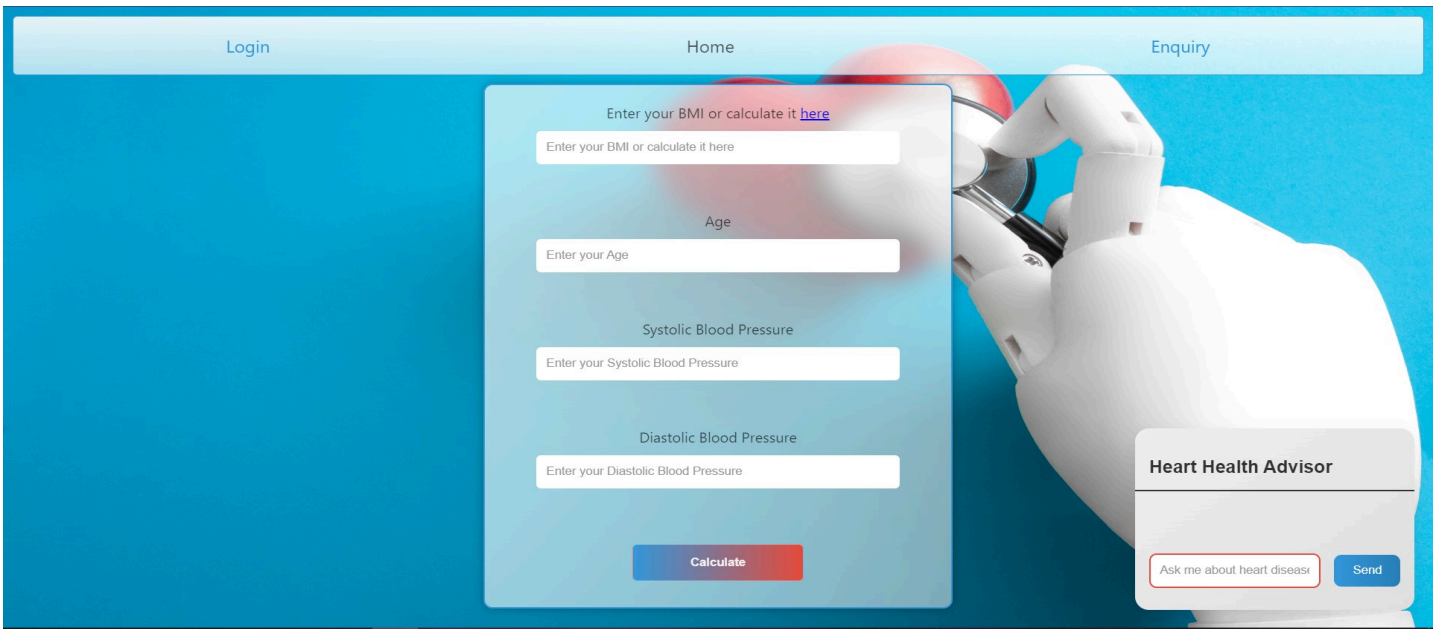


Fig 7.1.4: Screenshot for Home page

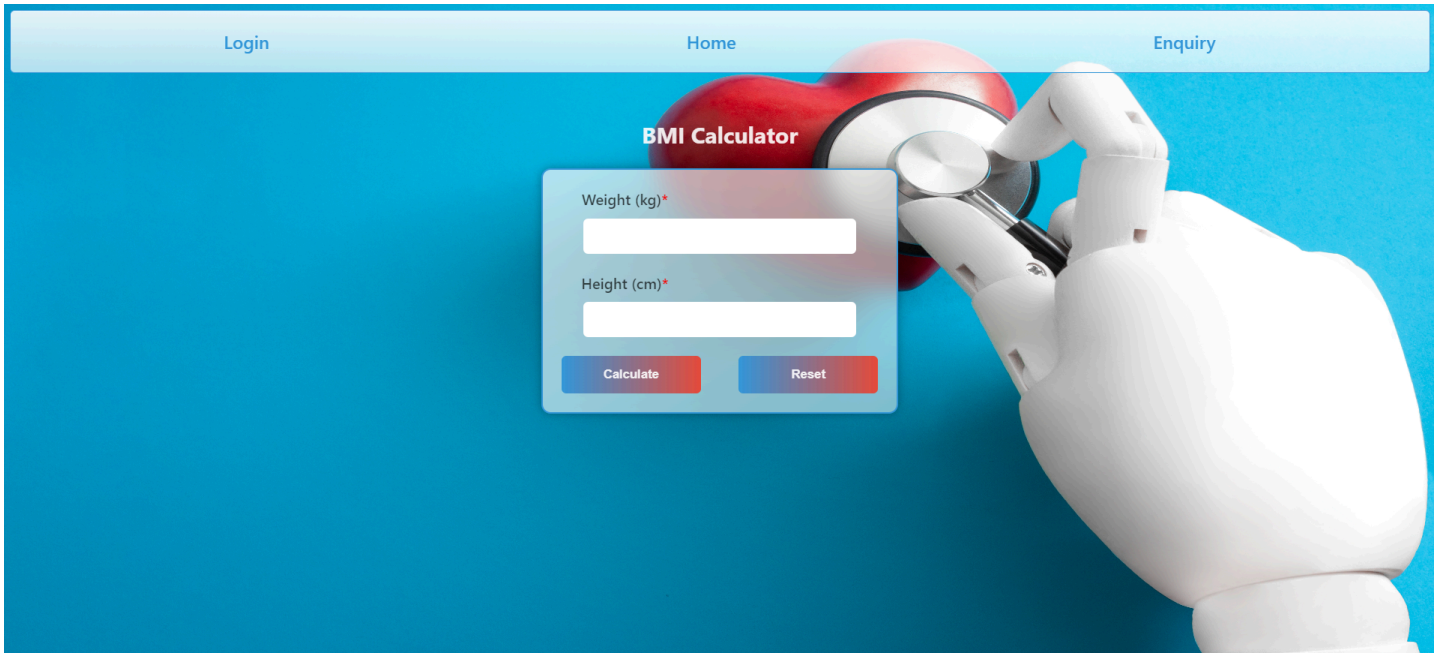


Fig 7.1.5: Screenshot for BMI Calculator page

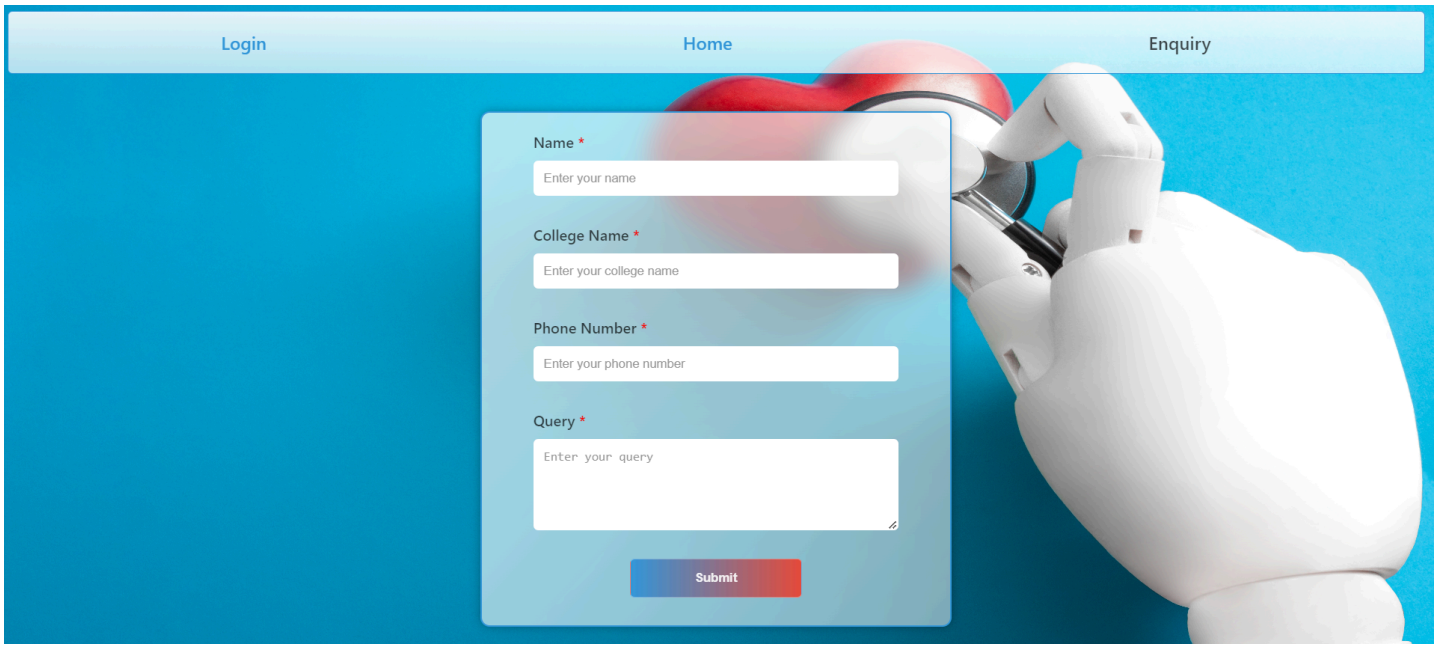


Fig 7.1.6: Screenshot for Enquiry page

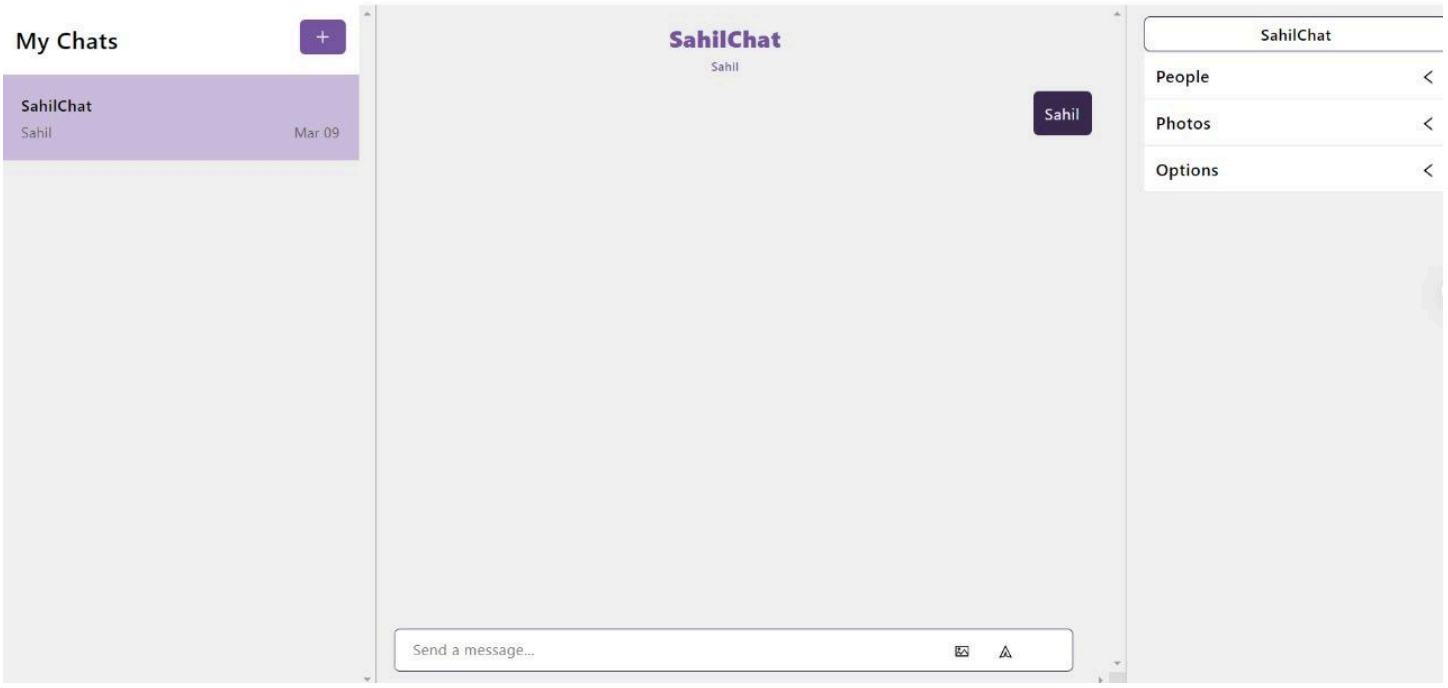


Fig 7.1.7: Screenshot for ChatApp page

7.2. Performance Evaluation Measures:

1. **Precision:** Precision is one indicator of a machine learning model's performance – the quality of a positive prediction made by the model. Precision refers to the number of true positives divided by the total number of positive predictions (i.e., the number of true positives plus the number of false positives). The formula is:

$$Precision = \frac{TP}{TP + FP}$$

where:

TP = True Positives,

FP = false Positives.

2. Recall: The recall is calculated as the ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive samples. The recall measures the model's ability to detect positive samples. The higher the recall, the more positive samples detected. The formula is:

$$Recall = \frac{TP}{TP + FN}$$

where:

TP = True Positives,

FN = false Negatives.

3. F-Score: The F-score (also known as the F1 score or F-measure) is a metric used to evaluate the performance of a Machine Learning model. It combines precision and recall into a single score. The formula is:

F-1 Score

F1=2*((Precision*Recall)/(Precision+Recall))

F1=(T.P)/[(T.P)+1/2(F.P+FN)]

T.P:- True Positive
F.P:- False Positive
F.N:-False Negative

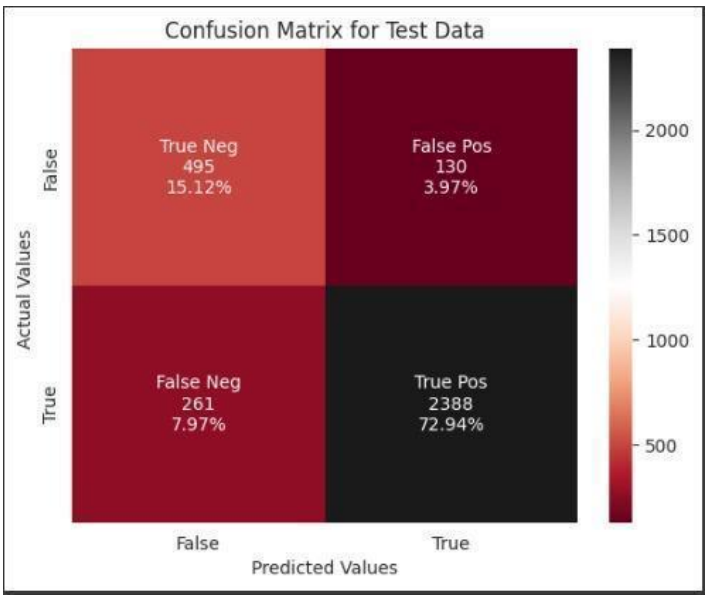


Fig 7.2.1 Confusion Matrix of our Model

```
Test Result:
================================================
Accuracy Score: 88.06%
_____________________________________________
CLASSIFICATION REPORT:
              0       1  accuracy  macro avg  weighted avg
precision  0.65    0.95      0.88       0.80          0.89
recall     0.79    0.90      0.88       0.85          0.88
f1-score   0.72    0.92      0.88       0.82          0.88
support  625.00 2649.00      0.88    3274.00       3274.00
_____________________________________________
Confusion Matrix:
 [[ 495  130]
 [ 261 2388]]
```

Fig 7.2.2:Precision, recall and F-score of SVM Model

7.3. Input Parameters/Features considered:

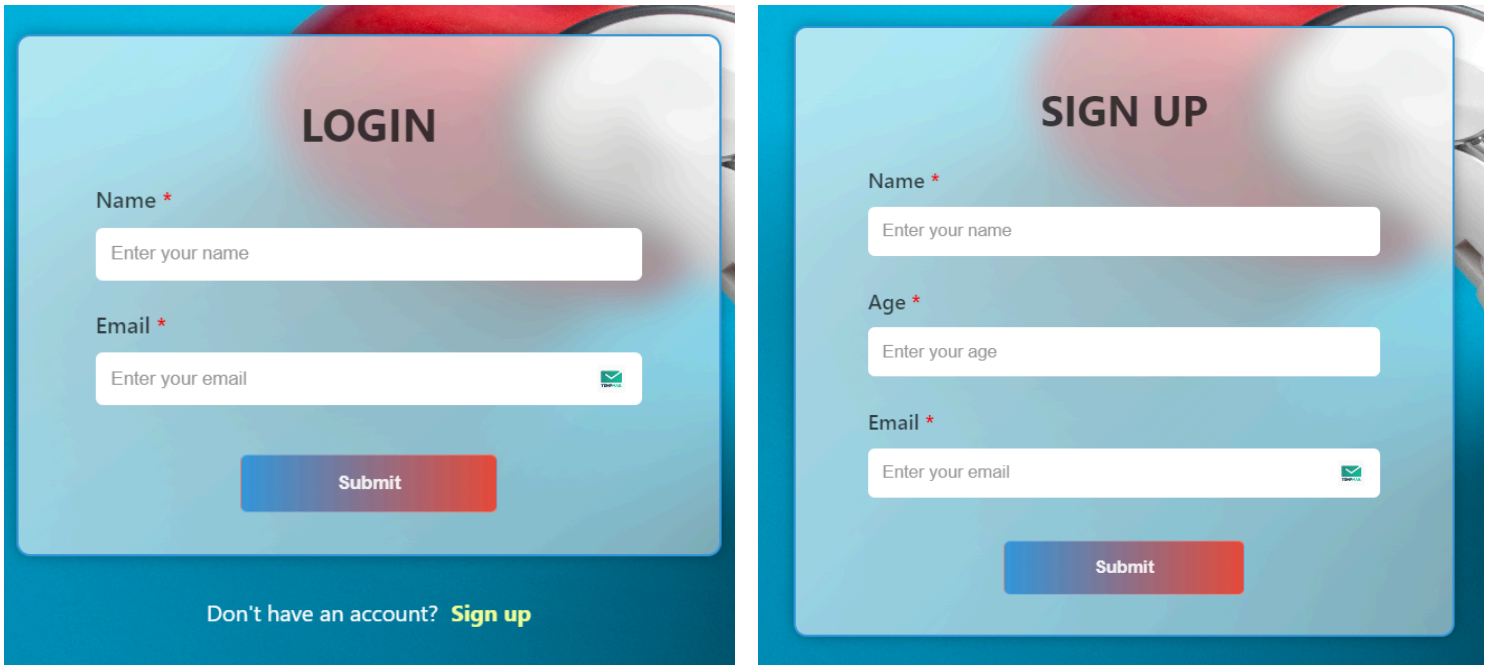


Fig 7.3.1: Login Page and Signup page input parameters

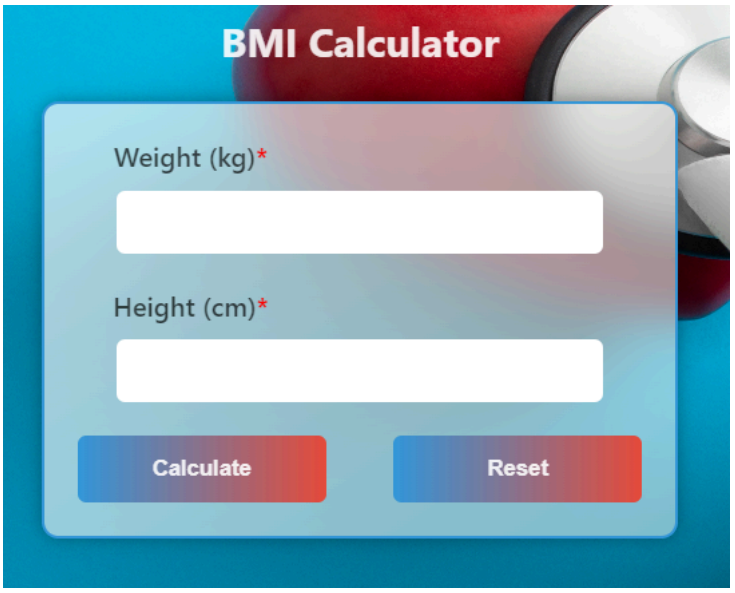


Fig 7.3.2 : BMI Page input parameters

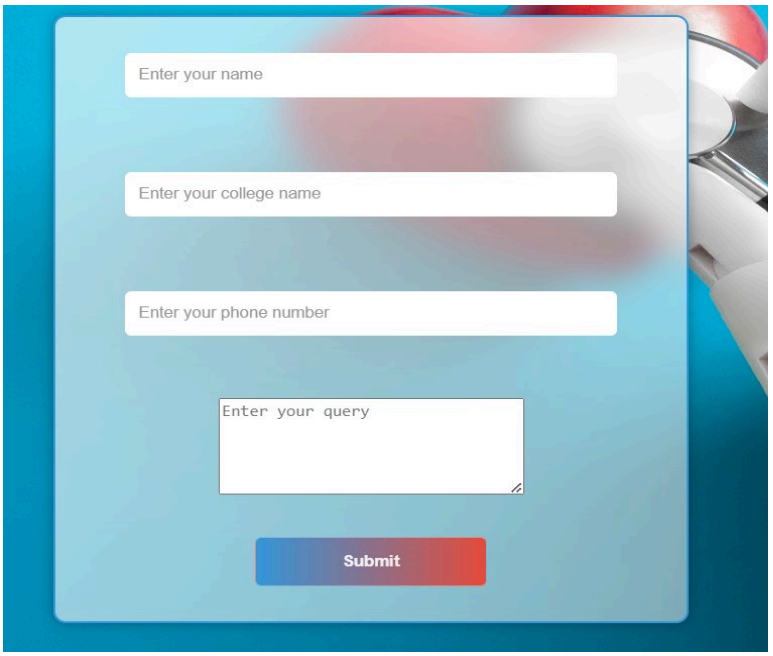


Fig 7.3.3 : Enquiry Page input parameters

7.4. Comparison of Results with Existing System:

Other System	Our System
Small Data Integration due to which accuracy rate is low.	Large Data Integration and strong Modeling due to which Accuracy rate is high.
ChatBot is not provided in urgent need.	ChatBot is provided to users in real time so that they can easily interact to solve their doubts.
No Advisor is provided.	Provision of Advisor in case the user wants to clear some questions, differences and doubts.
No Visualizations are provided to get clarity.	No. of Graphs are provided to users which gives clear ideas about heart health awareness.

Table:7.4.1 Comparison

7.5. Inference Drawn:

In the existing framework, there has been a notable absence of a chatbot feature to address users' immediate inquiries, leaving them without timely responses to pressing concerns. Furthermore, the unavailability of advisors has resulted in users navigating their queries independently. Additionally, there has been a lack of emphasis on promoting health awareness, particularly concerning heart health.

Feature	Existing System	Our System
ChatBot Provision	Not provided in urgent need	ChatBot provided in real-time for user interaction
Advisor Availability	Not provided	Provision of an advisor for clearing doubts and questions
Health Awareness	No specific focus	Integration of graphs to enhance heart health awareness

Table:7.5.1 Inferences

Chapter 8: Conclusion

8.1.Limitations:

While heart disease prediction systems can be valuable tools for identifying individuals at risk and guiding preventive interventions, they also have limitations that need to be considered. Here are some common limitations:

1.Data Quality and Availability: The quality and availability of data can significantly impact the performance of prediction models. Incomplete or inaccurate data, as well as data biases, can lead to unreliable predictions.

2.Limited Feature Set : Prediction models rely on a set of features or variables to make predictions. However, important risk factors for heart disease may not be included in the dataset, potentially leading to incomplete or inaccurate predictions.

3.Assumption of Causality: Prediction models often assume that the features used for prediction are causally related to the outcome (heart disease). However, correlation does not imply causation, and there may be underlying factors or confounding variables that are not accounted for in the model.

4.Model Generalization: Prediction models trained on one dataset may not generalize well to other populations or settings. Differences in demographics, healthcare practices, or risk factors between populations can affect the performance of the model.

5.Overfitting: Overfitting occurs when a model learns to fit the training data too closely, capturing noise rather than underlying patterns. This can result in poor generalization performance when applied to new data.

6.Ethical and Privacy Concerns: Healthcare data used for heart disease prediction often contain sensitive information about individuals' health status and medical history. Ensuring privacy and confidentiality while using this data is crucial and may limit access to certain datasets.

7.Interpretability: Complex machine learning models such as deep neural networks may provide high prediction accuracy but lack interpretability. Understanding how the model makes predictions is important for gaining insights into risk factors and informing clinical decision-making.

8.Temporal Dynamics: Risk factors for heart disease can change over time, and prediction models may not account for temporal dynamics or changes in risk factors over an individual's lifespan.

9.False Positives and False Negatives: Prediction models may produce false positives (predicting heart disease when it is not present) or false negatives (failing to predict heart disease when it is present), which can lead to unnecessary interventions or missed opportunities for preventive care.

10.Regulatory and Legal Considerations: Healthcare prediction systems may be subject to regulatory requirements and legal constraints, particularly regarding patient consent, data sharing, and liability for errors or adverse outcomes.

8.2.Conclusion:

This website-based heart disease prediction system holds immense potential to contribute to preventative healthcare by empowering individuals to assess their personal risk. While acknowledging limitations surrounding model accuracy, explainability, and generalizability, the system offers several promising advantages:

- Early Detection: By prompting users to consider key risk factors and providing accessible risk assessments, the system encourages proactive health awareness and potentially facilitates early detection of heart disease.
- Personalized Risk Assessment: Tailoring predictions based on individual data enhances engagement and empowers users to take informed decisions regarding their health.
- Accessibility and Scalability: The website format allows for widespread access, potentially reaching individuals who might not otherwise seek preventive measures.

8.3.Future Scope:

1. **API Creation**: We'll design a user-friendly API based on the website features and user requirement.

2. **Integration:** We'll seamlessly connect our model developed for the website into API infrastructure, enabling real-time predictions.

3. **Test API Functionality:** Conduct comprehensive testing of the API to verify its functionality, reliability, and performance.

4. **Deploy and Monitor API:** Deploy the API to a production environment, such as a cloud platform or dedicated server.

5. **Promote API Adoption:** Market the API to potential users and developers who could benefit from integrating the heart disease prediction functionality into their applications or systems

References

[1] Devansh Shah, Samir Patel and Santosh Kumar Bharti. " Heart Disease Prediction using Machine Learning Techniques" in 2020
https://link.springer.com/article/10.1007/s42979-020-00365-y

[2] Md.Julker Nayeem, SohelRana,and Md. Rabiul Islam. "Prediction of Heart Disease Using Machine Learning Algorithms"in 2022
https://www.ej-ai.org/index.php/ejai/article/view/13/13

[3] C.M.Bhatt, Parth Patel, Tarang Ghetiya, Pier Luigui Mazzeo. "Effective Heart Disease Prediction Using Machine Learning" in 2022
https://www.mdpi.com/1999-4893/16/2/88

[4] Muntasir Mamun, Afia Farjana, Miraz Al Mamun, Md Salim Ahammed, Md Minhazur Rahman.
"Heart failure survival prediction using machine learning algorithm: am I safe from heart failure?"in 2022
https://ieeexplore.ieee.org/abstract/document/9817303

[5] K. Karthick, S. K. Aruna. "Implementation of a Heart Disease Risk Prediction Model Using Machine Learning"in 2023
https://www.hindawi.com/journals/cmmm/2023/9764021/

[6] Mohammed Khalid Hossen. "Heart Disease Prediction Using Machine Learning Techniques" in 2022
https://www.researchgate.net/publication/362150416_Heart_Disease_Prediction_Using_Machine_Learning_Techniques

[7] Apurv Garg , Bhartendu Sharma and Rijwan Khan. "Heart disease prediction using machine learning techniques"in 2022
https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012046/meta

[8] V.V. Ramalingam*, Ayantan Dandapath, M Karthik Raja. "Heart disease prediction using machine learning techniques : a survey"in 2022
https://www.researchgate.net/publication/325116774_Heart_disease_prediction_using_machine_learning_techniques_A_survey

Appendix

1] Paper I details :-

a.Paper I :-

Heart Disease Prediction System

Sahil Nagdev
Dept of Computer Engineering,
Vivekanand Education Society's
Institute Of Technology,
Mumbai, India
sahil.r.nagdev2017@gmail.com

Jatin Ochani
Dept of Computer Engineering,
Vivekanand Education Society's
Institute Of Technology,
Mumbai, India
jatinochani04@gmail.com

Deepali Duseja
Dept of Computer Engineering,
Vivekanand Education Society's
Institute Of Technology,
Mumbai, India
dusejadeepali4@gmail.com

Harsh Bakhtar
Dept of Computer Engineering,
Vivekanand Education Society's
Institute Of Technology,
Mumbai, India
bakhtarharsh2002@gmail.com

Prof. Sunita Suralkar
Assistant Prof., Department of
Computer Engineering,
Vivekanand Education Society's
Institute Of Technology,
Mumbai, India

***Abstract*—Heart disease is a leading global cause of death, and treating it involves advanced technologies. However, a common issue in medical centers is the variability in medical professionals' knowledge and expertise, leading to suboptimal decisions and poor outcomes. To address this, machine learning algorithms have been employed to predict heart disease, aiding in automated diagnosis within hospitals.Prediction of heart disease relies on analyzing various health parameters, such as Age, Gender, smoke,glucose,Blood Pressure (BP),(low and high). In our study, we used an open-source dataset and implemented 5 different algorithms: Naive Bayes, KNN, Decision Tree, Support Vector Machine, and Random Forest.**

Our study aimed to identify the most accurate technique for predicting the occurrence of heart disease by analyzing these health parameters. Our experiments revealed that Support Vector Machine achieved the highest accuracy, with an impressive rate of 88%.

Keywords*—*Random Forest, k Nearest Neighbor (KNN), Logistic Regression, Decision tree, Support Vector machine, Heart disease

I. Introduction

Heart disease is a global epidemic, claiming more lives than any other cause, with approximately 12 million annual fatalities. In the United States, an individual falls victim to heart disease approximately every 34 seconds, frequently as a result of restricted flow of blood to the brain or heart. The evaluation of different factors such as blood pressure, glucose levels, lipids, and stress is involved in the process of identifying individuals who are in danger. These parameters can be conveniently monitored at home.

Heart diseases encompass various conditions affecting the heart and blood vessels, categorized as coronary heart disease, cardiomyopathy, and cardiovascular disease. These conditions impact blood circulation and can lead to severe consequences, including disability and death.

Diagnosing heart disease is a complex and vital aspect of medicine. Automating this process proves invaluable, especially in regions with limited specialized expertise and resources. Data mining plays a pivotal role in uncovering hidden patterns and knowledge essential for informed decision-making by healthcare professionals.

Efforts to support healthcare providers with varying expertise levels are crucial. However, current methods have limitations in drawing precise conclusions. In our approach, we employ machine learning algorithms like Decision Tree, k Nearest Neighbor (KNN), Decision Tree, Support Vector Machine(SVM), and Random Forest. These methods aim to predict heart disease based on specific health parameters, offering a potential solution to enhance accuracy and healthcare quality.

II. Literature Review

[1] In this research paper, diverse attributes associated with heart disease are explored, utilizing supervised learning algorithms such as Naïve Bayes, decision tree, K-nearest neighbor, and random forest. The study leverages an established dataset sourced from the Cleveland database of the UCI repository, featuring information on 303 instances and 76 attributes related to heart disease patients. Specifically, 14 carefully selected attributes are considered for testing to evaluate the efficacy of different algorithms. The primary objective of this research is to assess the likelihood of heart disease development in patients. The findings reveal that the K-nearest neighbor algorithm achieves the highest accuracy score among the tested models.

[2] The study utilized a dataset sourced from Kaggle, consisting of 65,535 records, to predict diseases employing supervised learning algorithms including Decision tree, K-nearest neighbor, Random forest, and Naïve Bayes. The paper also discusses the use of other data mining techniques such as time series, clustering, association rules, support vector machine, and genetic algorithm to improve the accuracy of the model. The highest accuracy (88.52%) is observed in the K-nearest neighbor (KNN) model, surpassing other prediction models. Heart disease prediction was conducted using machine learning algorithms, and the Random Forest (RF) model yielded the highest accuracy at 90.16%. Additionally, the KNN model achieved the top accuracy at 90.789%. The maximum accuracy attainable for the KNN model was 87%. To enhance the classification model's accuracy, a feature selection strategy employing info-gain was implemented. Logistic regression exhibited the best accuracy, reaching 92.76%. Prediction of cardiac disease using MLP and SVM resulted in a two-class classification accuracy of 92.45%, the highest among methods employed. Machine learning methods were utilized to forecast hepatitis disease, with the Random Forest model achieving the highest accuracy of 92.41%.

[3] Presented in this paper is a technique proposing k-modes clustering with Huang starting, aiming to enhance classification accuracy. Various models, including XGBoost (XGB), Multilayer Perceptron (MP), Random Forest (RF), and Decision Tree classifier (DT), were employed. GridSearchCV was utilized for hyperparameter tuning to optimize model performance. The proposed model was tested on a Kaggle dataset comprising 70,000 instances, with training data split in an 80:20 ratio. The accuracy achieved for decision tree was 86.37% (with cross-validation) and 86.53% (without cross-validation), XGBoost recorded 86.87% (with cross-validation) and 87.02% (without cross-validation), random forest showed 87.05% (with cross-validation) and 86.92% (without cross-validation), and multilayer perceptron exhibited 87.28% (with cross-validation) and 86.94% (without cross-validation). The proposed models were evaluated based on AUC (area under the curve) values: decision tree (0.94), XGBoost (0.95), random forest (0.95), and multilayer perceptron (0.95). The research concludes that the multilayer perceptron with cross-validation outperforms other algorithms, achieving the highest accuracy at 87.28%.

[4] The study compares the performance of several machine learning models, including Analyzing the prediction of patient survival for heart failure (HF) patients, a variety of machine learning algorithms, including Bagging, Decision Tree, KNN, LightGBM, Logistic Regression, Random Forest, Support Vector Machine, and XGBoost, were evaluated. The findings reveal that LightGBM achieved the highest accuracy of 85% and an AUC of 93%, surpassing other algorithms in the study. The paper also provides an overview of related work and discusses the dataset and attribute feature information. The study has potential applications in the healthcare industry and can provide valuable insights into heart failure prediction.

[5] In the heart disease prediction model employing machine learning algorithms such as Naive Bayes, Support Vector Machine (SVM), XGBoost, LightGBM, and Random Forest, statistical methods are utilized for visualization, and the area under the curve (AUC) serves as an appropriate metric for assessing the discrimination between classes. Moreover, it is observed that males are more susceptible to heart disease compared to females.

[6] In the comparison of five algorithms, the Logistic Regression algorithm stands out for its

high-performance accuracy. The Logistic Regression model achieves an accuracy rate of 95%, signaling its potential as a pre-defined tool for detecting heart diseases in the near future. Additional statistics, including f1-score, recall, and precision rate, are also computed for Logistic Regression, all registering at 95%. These calculated values underscore the algorithm's exceptional accuracy.

[7] KNN (K Nearest Neighbour) gives higher accuracy than Random Forest Algorithm in Heart Disease Prediction System and is a classification technique which has to select a point and to classify a category and to determine the majority category among the K-Neighbours .

[8] Based on Bayes Theorem and assuming feature independence, Naïve Bayes is a classification technique. On the Cleveland dataset, it achieved an accuracy of 84.1584% with selected features and 83.49% with all attributes. Random Forest, an ensemble algorithm utilizing multiple decision trees, showcased remarkable performance, attaining 91.6% accuracy on the Cleveland dataset and 97% on People's Hospital data, with an f-measure of 0.86. Particularly effective for large, high-dimensional datasets, Random Forest demonstrated its robustness. K-Nearest Neighbor (KNN), a nonparametric classification technique predicting target values by identifying k nearest data points, is also employed in the analysis. It reached 83.16% accuracy with k=9 using cross-validation. With Ant Colony Optimization, it achieved 70.26% accuracy and an error rate of 0.526. An efficiency of 87.5% was reported by Ridhi Sainietal. Decision Tree, a supervised learning algorithm, is suited for classification with continuous and categorical attributes. It partitions data based on significant predictors using information gain or entropy, leading to a tree-like structure.

[9] The algorithms considered are Support Vector Machine, Random Forest, Decision Tree, and K-Nearest Neighbor. The Decision Tree algorithm functions through a divide-and-conquer approach, constructing trees in a top-down manner.Pruning eliminates non-significant samples, and entropy is a key concept, calculated as DT_Entropy = (- ∑ij pij * log2(pij)). Euclidean Distance function dis(xi, xj) gathers information through majority votes from k-nearest neighbors. For Support Vector Machines (SVM), the hyperplane equation f(x) = (wTx + b) is found using mathematical computation like (Minw,b),(Ei 1/2 * w^2 + C * Σi=1^n Ei), subject to (yi(wTxi + b) ≥ 1 - Ei) where 'w' and 'b' are coefficients. Bagging involves building decision trees for optimal outcomes, where N trees are averaged for predictions, yielding an error measurement as the standard deviation of trees' predictions.

III. DATASET

For predicting heart disease, a built-in dataset from kaggle has been employed.

The database encompasses the following 13 attributes:

1. Age
2. Height
3. Weight
4. BMI
5. Gender
6. Systolic blood pressure
7. Diastolic blood pressure
8. Cholesterol
9. Glucose
10. Smoking
11. Alcohol
12. Physical Activity
13. Cardio

IV. METHODOLOGY

In the data collection and preprocessing stage, the focus is on gathering relevant information from patients and preparing it for analysis. This involves designing a registration form for patients to fill out, which typically includes essential demographic and health-related details. In this case, the registration form requires mandatory fields such as Body Mass Index (BMI), Age, Systolic Blood Pressure, and Diastolic Blood Pressure. Once the data is collected, it needs to be preprocessed to ensure its quality and compatibility with machine learning algorithms.The next step involves selecting appropriate machine learning algorithms for heart disease prediction and evaluating their performance to determine the most suitable model. Once the most suitable model is selected, it needs to be trained using the processed data to enhance its prediction accuracy. This involves feeding the preprocessed data into the chosen model and adjusting its parameters to optimize its performance. Perform functionality testing, compatibility testing, and security testing to ensure the website meets quality standards

Seek feedback from potential users or domain

experts to validate the accuracy and usefulness of the prediction system.Monitor website performance and user interactions, addressing any issues that arise post-deployment.Regularly update the prediction model and website features based on new research findings, user feedback, and technological advancements.

Algorithm used in our project:

Support Vector Machine

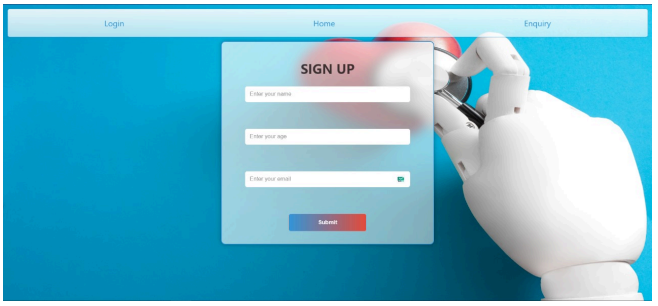
Support Vector Machine (SVM) is a machine learning algorithm widely employed in heart disease prediction systems due to its efficacy in handling high-dimensional data and nonlinear relationships. Initially, datasets comprising pertinent features like age, blood pressure, cholesterol levels, and ECG readings are collected and preprocessed, ensuring uniform scaling and handling of missing values. SVM model training ensues, where the algorithm learns underlying patterns by discerning optimal hyperplanes to separate data points of different classes, like presence or absence of heart disease, while maximizing margins between them. Model optimization involves parameter tuning and feature refinement to enhance performance and generalization. Evaluation metrics such as accuracy, sensitivity, and area under the ROC curve gauge model effectiveness. Post-training analysis aids in understanding decision boundaries and feature contributions, facilitating model interpretation. Integrated into heart disease prediction systems, SVM models offer clinicians actionable insights and accurate predictions, potentially improving diagnosis and treatment outcomes.

Website Development integrating with ML Model in our project:

After training the Machine Learning Model, we had developed a website integrating it with our ML Model.

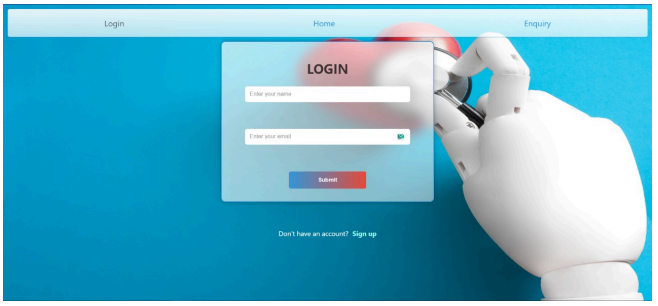
Some of features of our website are shown as below:

1.SignUp/Registration:



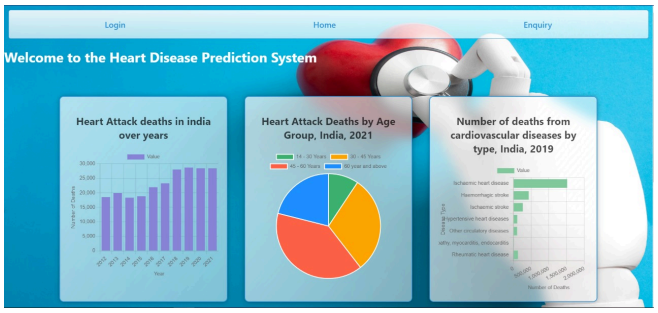
Here, the user is supposed to enter his details of Name, Age, Email address which would be stored in our database. Users can directly Login after registration on our website.

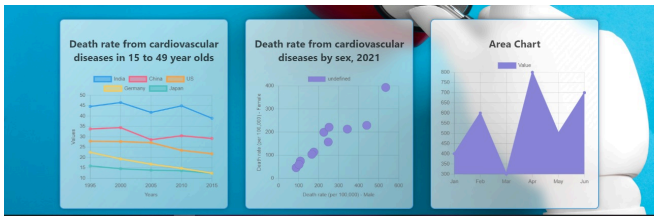
2. Login Page:



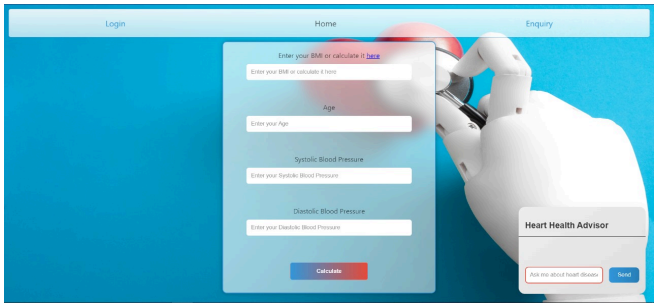
Here, a user can directly Login and can use the website and its features to predict if he/she is suffering from any heart disease.

3.Home Page:

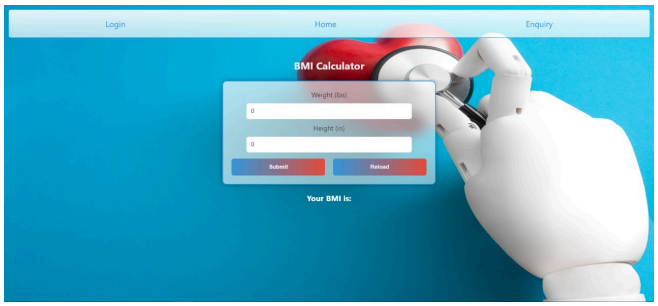




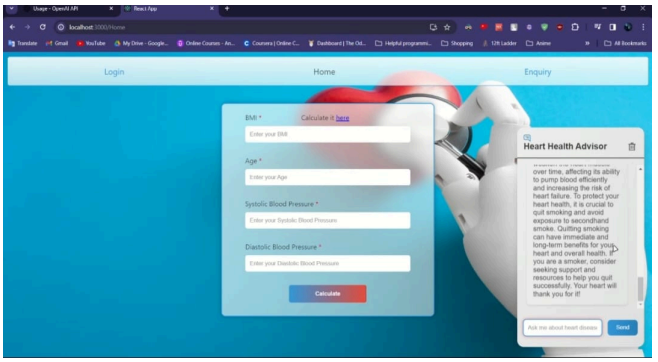
On Login users will have this UI to get all about the current scenario of this world. Then comes:



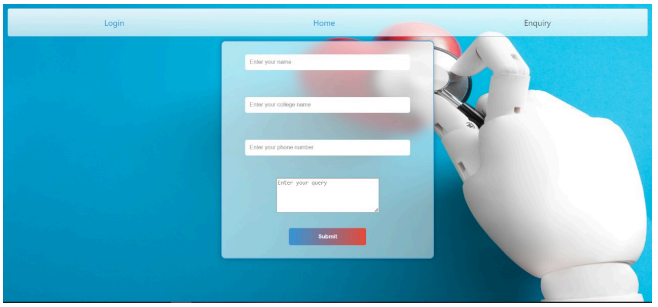
Here, users are supposed to put details of BMI, Age, Systolic Blood Pressure, Diastolic Blood Pressure to predict if he/she is suffering from Heart Disease or not. If users are unknown about BMI, they can calculate at that instant also. Here it is how:



Here, the user is supposed to enter the details of weight and height. Along with this there is a feature of advisor where users can put their query in turn advisor will answer and clear the doubt of users.



4.Enquiry Page:



Here, the user is supposed to enter the details of Name, Phone no, Email address and most importantly about the query.

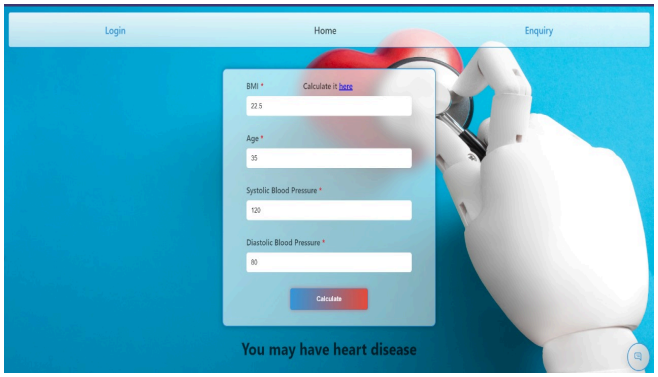
IV. Results And Discussions

	Model	Training Accuracy %	Testing Accuracy %
0	LogisticRegression	87.12	85.00
1	LogisticRegression	87.12	85.00
2	K-nearest neighbors	90.40	85.83
3	Random Forest Classifier	89.65	87.57
4	Decision Tree	89.66	87.72
5	Support Vector Machine	89.76	88.06

Accuracies of different models

Our study primarily focused on leveraging data mining techniques in the healthcare sector, specifically for the early detection of heart disease, a life-threatening condition. We utilized a variety of data mining algorithms, including K-Nearest Neighbors (KNN), Logistic Regression, Decision Trees, Support Vector Machine, and Random Forest, to predict the occurrence of heart disease. To gauge their effectiveness, we examined metrics such as Accuracy, True Negatives (TN), False Positives (FP), False

Negatives (FN), and True Positives (TP). We conducted five experiments on the same dataset and compared the results of these algorithms, which are presented in a tabular format for ease of understanding and comparison. Interestingly, our results indicated that Support Vector Machine outperformed the other algorithms with an accuracy of 88.06%, followed by K-Nearest Neighbors With 85.83%, Random Forest and Decision Tree with 87.72%, and Logistic Regression with 85%.



Thus, we had finally predicted if user is suffering from any Heart Disease or not. Along this many Features such as ChatBot, Advisor are provided on our website which will help the at anytime and in every place.

V. Conclusion and Future Works

In this research endeavor, our primary objective was to harness the power of data mining techniques in the healthcare domain, with a specific focus on the early detection of heart disease. We employed a range of data mining algorithms, namely K-Nearest Neighbors (KNN), Logistic Regression, Decision Trees, Support Vector Machine, and Random Forest, to predict the likelihood of heart disease. The effectiveness of these algorithms was evaluated based on various metrics such as Accuracy, True Negatives (TN), False

Positives (FP), False Negatives (FN), and True Positives (TP). We carried out five separate experiments on the same dataset and juxtaposed the results of these algorithms. The results were tabulated for easy comprehension and comparison. Our findings revealed that the Support Vector Machine surpassed other algorithms with an accuracy rate of 88.06%, followed by K-Nearest Neighbors at 85.83%, Random Forest and Decision Tree at 87.72%, and Logistic Regression at 85%. These results underscore the efficacy of data mining techniques in healthcare for early detection and diagnosis of diseases.

In conclusion, our upcoming endeavor revolves around crafting an innovative application dedicated to predicting heart disease. We are steadfast in our commitment to refining this system, with a primary focus on augmenting its feature set. By expanding the scope of variables considered, we aim to fortify the predictive capabilities of the application, thereby offering users a more comprehensive assessment of their cardiovascular health. Through meticulous attention to detail and continuous iteration, we aspire to deliver a solution that not only detects potential risks with precision but also empowers individuals to take proactive steps towards mitigating them.

References

[1] Devansh Shah, Samir Patel and Santosh Kumar Bharti. “ Heart Disease Prediction using Machine Learning Techniques” . 2020 https://link.springer.com/article/10.1007/s42979-020-00365-y

[2] Md.Julker Nayeem, SohelRana,and Md. Rabiul Islam. “Prediction of Heart Disease Using Machine Learning Algorithms”.2022 https://www.ej-ai.org/index.php/ejai/article/view/13/13

[3] C.M.Bhatt, Parth Patel, Tarang Ghetiya, Pier Luigui Mazzeo “Effective Heart Disease Prediction Using Machine Learning”.2022 https://www.mdpi.com/1999-4893/16/2/88

[4] Muntasir Mamun, Afia Farjana, Miraz Al Mamun, Md Salim Ahammed, Md Minhazur Rahman. “Heart failure survival prediction using machine learning algorithm: am I safe from heart failure?”. 2022 https://ieeexplore.ieee.org/abstract/document/9817303

[5] K. Karthick, S. K. Aruna. “Implementation of a Heart Disease Risk Prediction Model Using Machine Learning”. 2023 https://www.hindawi.com/journals/cmmm/2023/9764021/

[6] Mohammed Khalid Hossen
"Heart Disease Prediction Using Machine Learning Techniques" . 2022
https://www.researchgate.net/publication/362150416_Heart_Disease_Prediction_Using_Machine_Learning_Techniques
[7] Apurv Garg , Bhartendu Sharma and Rijwan Khan
"Heart disease prediction using machine learning techniques". 2022
https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012046/meta
[8] V.V. Ramalingam*, Ayantan Dandapath, M Karthik Raja
"Heart disease prediction using machine learning techniques : a survey". 2022
https://www.researchgate.net/publication/325116774_Heart_disease_prediction_using_machine_learning_techniques_A_survey
[9] Keshav Srivastava, Dilip Kumar Choubey. 2020
"Heart Disease Prediction using Machine Learning and Data Mining". 2020
https://www.researchgate.net/profile/Dilip-Choubey-2/publication/342210798_Heart_Disease_Prediction_using_Machine_Learning_and_Data_Mining/links/5ee8cb6592851ce9e7e7f585/Heart-Disease-Prediction-using-Machine-Learning-and-Data-Mining.pdf

b.Plagiarism Report

c.Project review sheet ;

Project review sheet 1:

Inhouse/ Industry_Innovation/Research: Inhouse

Class: D17 A/B/C

Sustainable Goal:

Project Evaluation Sheet 2023 - 24

Group No.: 38

Title of Project: Heart Disease Prediction System

Group Members: Sahil Nagdev (38), Jatin Ochani (40, D17C), Harsh Bakhtar (04, D17B), Deepali Duseja (19, D17C)

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
3	3	3	2	3	2	2	2	2	2	2	2	2	2	2	34

Comments: - Paper Publication Required

Name & Signature Reviewer1

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
4	4	4	2	3	2	2	2	2	2	2	2	2	2	2	37

Comments:

Date: 10th february, 2024

Name & Signature Reviewer 2

Project review sheet 2

Inhouse/ Industry _Innovation/Research:

Sustainable Goal:

Project Evaluation Sheet 2023 - 24

Class: D17 A/B/C

Group No.: 38

Title of Project: Heart Disease Prediction System

Group Members: Harsh Bakhtar (04), Deepali Duseja (19), Jatin Ochani (40) Sahil Nagdev (58)

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
3	3	3	2	3	2	2	2	2	2	2	2	2	2	2	35

Comments: Conceptual diagram needs to be revised
Integrate the Web Appl & Chat Bot Appl
Needs to be checked Web App UI
Paper - 3rd person
- Summary findings Table Lifne Cs
- implementation & snshots
SVM?

Name & Signature Reviewer1

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
3	3	3	2	3	2	2	2	2	2	2	2	2	2	2	35

Comments: After detecting the heart disease. what next?

Date: 9th March, 2024

Sunita Sunalkar.

Name & Signature Reviewer 2