An Analysis of Heart Disease Prediction using Machine Learning

Saladi Novya Sree [1], K.Balaji Pranav Reddy [2], Bangarulakshmi Mahanthi[3], Dr. S S Nandini[4]

[1,2] Student, CSE, GITAM School of Technology, Visakhapatnam, India
[3,4]Assistant Professor, GITAM School of Technology, Visakhapatnam, India

Abstract

Heart disease is a major cause of death worldwide, making early diagnosis and prevention essential. Predictive models have gained significant attention in recent years, with several algorithms being employed to develop these models. However, there are challenges in implementing heart disease prediction models, including data quality, model accuracy, ethical concerns, and limited data. Therefore, this project aims to develop a heart disease prediction model and analyze different algorithms used in disease prediction. In order to increase the predictive accuracy of machine learning algorithms, this study compares six algorithms, including KNN (K-Nearest Neighbor), Decision Tree, Random Forest, Support Vector Machines, Logistic Regression, and Neural Network. 13 attributes, including age, sex, and cholesterol, are used, and ensemble methods like boosting and bagging are used. The accuracy, recall, f1 score, and precision of each algorithm are calculated to determine the most accurate model. Additionally, this study identifies the limitations of heart disease prediction models and their implications for patient diagnosis and treatment, by developing and analyzing heart disease prediction models. In conclusion, while heart disease prediction models have the potential to be financially feasible and be useful in the future, their current limitations and challenges mean that they cannot be relied upon as the sole means of diagnosis or treatment decisions.

Keywords: Predictive Models, Challenges, Analyze, Ensemble Methods, Diagnosis

1. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally, accounting for approximately 18 million deaths annually, which is about 32% of all global deaths. According to the World Health Organization (WHO), approximately three-quarters of CVD deaths occur in low- to middle-income countries. These statistics underscore the pressing need for early detection and prevention of heart disease, which can significantly reduce mortality rates.

In recent years, there has been increasing interest in developing predictive models for early detection of heart disease. Due to the availability of large amounts of health data and advancements in machine learning, there has been a growing interest in utilizing machine learning algorithms for the purpose of early detection of heart disease. Various algorithms, such as logistic regression, KNN (K-Nearest

Neighbors), decision tree, support vector machines, random forest, and neural network, have been employed to develop these models.

Despite the potential benefits of machine learning algorithms in predicting heart disease, several challenges remain. These include data quality, model accuracy, ethical concerns, and limited data. For instance, several important risk factors such as family history, lifestyle, smoking status, drinking status, and medication history cannot be fully accounted for by current algorithms.

To address these limitations, research efforts are underway to improve the accuracy and effectiveness of machine learning algorithms in predicting heart disease. One approach involves incorporating additional data sources, such as genetic and environmental factors, to improve the predictive power of the models. Another approach involves developing interpretable machine learning models that can provide insight into the underlying mechanisms of heart disease.

The results suggest that while machine learning algorithms have the potential to be financially feasible and useful, there are significant limitations to their current capabilities, including their inability to account for certain risk factors such as family history, lifestyle, smoking status, drinking status, and medication history. As such, while these models may be useful for calculating the probability of heart disease, they should not be used solely for diagnosis or treatment decisions. The findings of this project highlight the need for continued research and development in this field, with a focus on addressing the current limitations of machine learning algorithms in predicting heart disease.

2. Literature Review

Hiteshwar Singh et al.,2021[1] study looks at data mining and ML techniques for predicting CVD. It uses 13 attributes and obtains some highly favorable outcomes. It was discovered that the Random Forest Algorithm functioned better than the Naive Bayes and KNN as it uses an ensemble technique. There are some drawbacks to this study, but we should be able to overcome them with time; we can use a variety of combinational approaches to yield more significant results.

Apurb Rajdhan et al.,2020[2] study examines data mining techniques and ML for predicting CVD. This model gives input to the Machine Learning algorithms like Random Forest, Logistic Regression, Naive Bayes classification, and Decision Tree. The result obtained from this study reveals that the most efficient algorithm among the above is the Random Forest algorithm, which gives an accuracy of 80.16%.

Chaimaa Boukhatem et al.,2022[3] study shows data mining and ML techniques in predicting cardiac events. It uses specific health measurements on four different algorithms: Naive Bayes, Support Vector Machine, Neural Networks, and Random Forest to obtain the output. The collected data was cleaned and preprocessed. The SVM algorithm gave the best result with a 91.67% accuracy.

In their 2020 study, Archana Singh[4] and Rakesh Kumar used the Cleveland heart disease dataset to predict heart disease using four machine learning algorithms including Naïve Bayes. The authors evaluated the models based on accuracy, sensitivity, specificity, and area under the curve (AUC) of the receiver operating characteristic (ROC) curve. The results showed that the decision tree algorithm performed the best in all measures. The study also suggests that using feature selection techniques and ensemble methods could further improve the performance of heart disease prediction models.

The paper by Wang et al.[5] (2020) provides a comprehensive review of recent studies that have used deep learning methods for predicting cardiovascular disease. The authors discuss the various approaches and techniques used in these studies, as well as the strengths and limitations of each method. The paper also highlights the potential of deep learning for improving the accuracy and efficiency of cardiovascular disease prediction, and the challenges that need to be addressed in this field. Overall, this review provides valuable insights into the current state of research on the use of deep learning for predicting cardiovascular disease.

"Predictive modeling of cardiovascular disease using machine learning techniques: A systematic review" by Prabhakar et al.[6] (2021) provides an overview of studies that use machine learning techniques for predicting cardiovascular disease. The authors conducted a systematic review of the literature to identify relevant studies, and found that machine learning models can effectively predict cardiovascular disease risk using various types of health data, including medical history, lifestyle factors, and genetic data. However, the authors also note that there are challenges in implementing these models in clinical practice, including the need for high-quality data, ethical concerns, and limitations in interpretability. The paper concludes with recommendations for future research in this area, including the need for further studies to evaluate the clinical utility of these models and the development of standardized approaches for data collection and analysis.

3. Problem Statements and Objectives

PROBLEM STATEMENT:

Cardiovascular disease (CVD) is a major health challenge globally, and early detection is crucial for effective treatment and prevention of complications. However, current methods for predicting heart disease are either expensive or not efficient enough, posing significant challenges for diagnosis. The development of a reliable and feasible heart disease prediction model using machine learning algorithms is essential to address these challenges. This project aims to conduct a comprehensive analysis of the challenges and limitations associated with the development and deployment of heart disease prediction models.

OBJECTIVE:

The main aim of this project is to assess the effectiveness of predicting heart disease using six machine learning algorithms with a selected dataset. The analysis will involve assessing the feasibility and reliability of using these models as a predictive tool for heart disease diagnosis. The top three performing algorithms will be used to build a heart disease prediction model. The study will also analyze the challenges and limitations associated with the development and deployment of these models, including limited data, incomplete information, changing risk factors, and uncertainty. The final output will be determined by taking a majority vote from the top three algorithms, and the reliability of the predictor will be evaluated to determine its usefulness for heart disease diagnosis in current clinical practice.

4. System Methodology

DESIGN:

The proposed system is a heart disease prediction application that users can access by entering their personal health information. Users can enter details such as fasting blood sugar, cholesterol, chest pain, and so on, and the model will collect the data, extract its features, match the values, classify the data and finally predict the disease and display a report to the user.

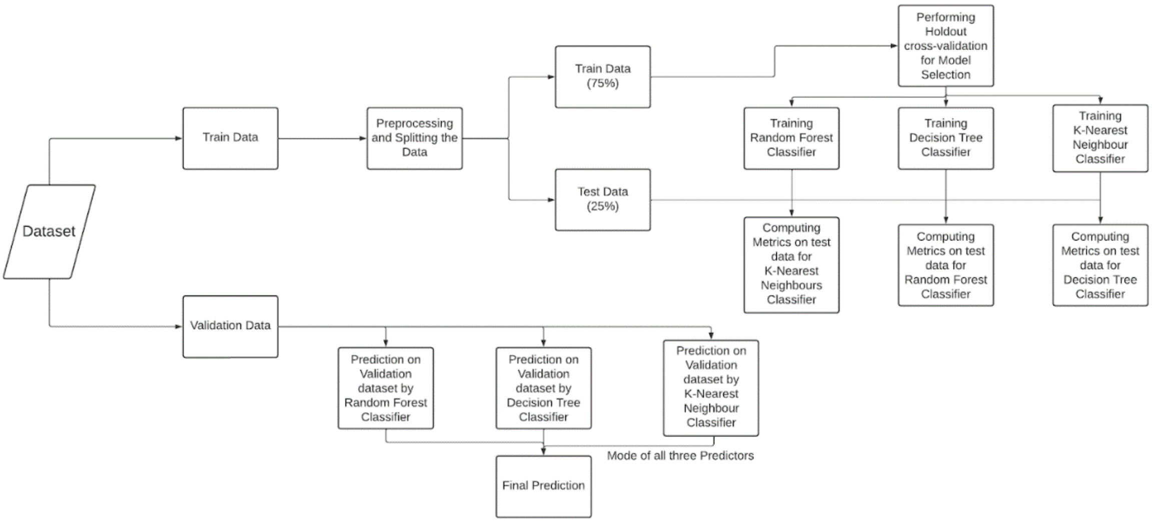

Dataset
Train Data
Preprocessing and Splitting the Data
Train Data (75%)
Test Data (25%)
Performing Holdout cross-validation for Model Selection
Training Random Forest Classifier
Training Decision Tree Classifier
Training K-Nearest Neighbour Classifier
Computing Metrics on test data for K-Nearest Neighbours Classifier
Computing Metrics on test data for Random Forest Classifier
Computing Metrics on test data for Decision Tree Classifier
Validation Data
Prediction on Validation dataset by Random Forest Classifier
Prediction on Validation dataset by Decision Tree Classifier
Prediction on Validation dataset by K-Nearest Neighbour Classifier
Mode of all three Predictors
Final Prediction

Fig. 1: Architecture of Model

METHODOLOGY:

Data Collection and Preprocessing:

To predict the model, data collection and pre-processing are essential. In this project, we used 75% of the dataset for training and 25% for testing. Data cleaning involved removing duplicates, handling outliers, and scaling features using libraries like pandas and scikit-learn. The dataset had 14 features with no null values. Feature scaling was done using MinMaxScaler. The data was visualized using scatter plots. Preprocessing was necessary for machine learning algorithms like Random Forest, which cannot handle null values. Overall, the steps taken improved the quality of the data for the machine learning model.

Exploratory data analysis (EDA):

The EDA performed on the 14 attributes of the heart disease prediction model involves:

- Exploring the relationships between different attributes.
- Identifying trends or patterns in the data.
- Understanding the distribution of individual attributes.

This helps in understanding the dataset and preparing the data for use in the prediction model.

A heatmap is generated using the sns.heatmap() function to visualize the correlation between each pair of attributes in the dataset. This provides an initial insight into which attributes are most correlated with each other, which can be useful in selecting relevant attributes for the prediction model.

A histogram is generated using the plt.hist() function to visualize the distribution of each attribute among individuals with and without heart disease. This helps identify potential trends or patterns between the attributes and heart disease. The same approach is applied to all 13 attributes in the dataset, along with the target attribute, to understand the data better and potentially identify any useful insights.

A pair plot is generated using the sns.pairplot() function to visualize the pairwise relationships between all attributes in the dataset. This helps to identify potential patterns or trends between attributes.
Histograms of all attributes are generated in the dataset using the df.hist() function. This helps to identify the distributions of individual attributes in the dataset.

These techniques help to understand the data and find patterns, correlations, or outliers that may be important for the heart disease prediction model.

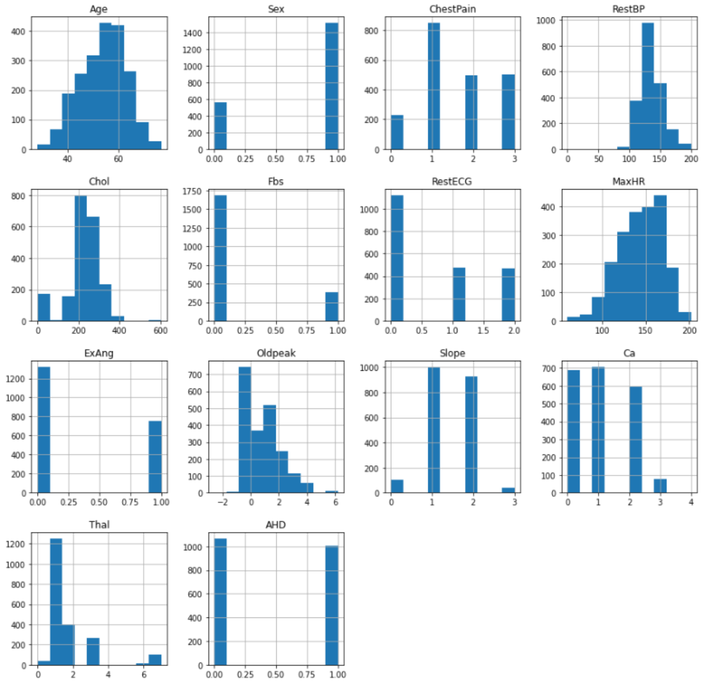


Fig. 2: Histograms of all attributes

Algorithm selection:

Three non-parametric algorithms, KNN, decision tree, and random forest, were chosen for heart disease prediction based on their ability to handle high-dimensional data and non linear data. Parametric algorithms such as SVM, logistic regression, and neural networks were not selected due to their requirement to make assumptions about the underlying distribution of the data. Hold-out validation was used to compare the performance of the three chosen algorithms using performance metrics.
The KNN algorithm relies on the similarity between instances to make predictions, while the decision tree algorithm partitions the feature space into a tree-like structure for interpretability. The random forest algorithm uses multiple decision trees to make predictions and is suitable for high-dimensional data. On the other hand, neural networks were not considered due to the small size of the dataset, and SVM, logistic regression, and neural networks were not selected due to their parametric nature.

The performance of the selected algorithms was evaluated using accuracy, precision, recall, and F1-score. The random forest algorithm performed the best, with an accuracy of 0.82, precision of 0.82, recall of 0.80, and F1-score of 0.81. The decision tree algorithm also performed well, with an accuracy of 0.75, while the KNN algorithm had an accuracy of 0.74. SVM, logistic regression, and neural networks performed poorly compared to the other algorithms.

TABLE 1. PERFORMANCE METRICS OF THE ALGORITHMS

Algorithm	Accuracy	Precision	Recall	F1-Score
KNN	0.74	0.74	0.71	0.73
Decision Tree	0.75	0.75	0.75	0.74
Random Forest	0.82	0.82	0.80	0.81
SVM	0.52	0.00	0.00	0.00
Logistic Regression	0.65	0.66	0.60	0.62
Neural Networks	0.76	0.76	0.78	0.76

Ensemble Methods:

Ensemble learning is a widely used technique to improve the predictive accuracy of machine learning algorithms by combining multiple models. Boosting and Bagging are two popular ensemble methods. Boosting involves training models sequentially and focusing on misclassified data points. Bagging trains models on different subsets of the data and combines their outputs. Both methods were used with the selected algorithms - KNN, Decision Trees, and Random Forests. The accuracy of boosted and bagged models was compared to the single model, and results showed improved accuracy. Overall, the final prediction is made by aggregating the predictions of all the models through voting.

Web application development:

Flask was used to build a web application that predicts the risk of heart disease based on various medical parameters. Pre-trained machine learning models were loaded using the pickle library, and Flask routes were defined using the @app.route decorator. HTML templates were rendered using the render_template function, and user input was retrieved using the request object from HTML forms. Majority voting was applied to combine the predictions of three different machine learning models (K-Nearest Neighbors, Decision Tree, and Random Forest). The return statement was used to display the predicted risk of heart disease to the user.

In summary, Flask created a user-friendly interface that allows users to input their medical parameters and receive a predicted risk of heart disease. The application provides a more accurate and reliable prediction by combining the predictions of the three machine learning models. We used Flask to build a simple yet effective web application for healthcare prediction.

Testing:

Software testing includes unit testing, integration testing, functional testing, user acceptance testing, performance testing, and regression testing. These tests ensure that the application components work as expected, collaborate properly, align with user requirements, are easy to use, perform efficiently and effectively under varying workloads, and still function as expected after changes. We manually tested our application using test cases and scenarios based on user requirements and expectations.

5. Overview of Technologies

ALGORITHMS:

1)KNN is abbreviated as K- Nearest Neighbor. It is one of the well-known machine learning algorithms. Classification and regression problems can be solved using this. The 'K' symbol is used to denote the nearest Neighbors beside an unknown variable to be predicted. It is also called a distance-based approach as it locates all the nearest Neighbors beside an unknown variable to know which class it belongs to.

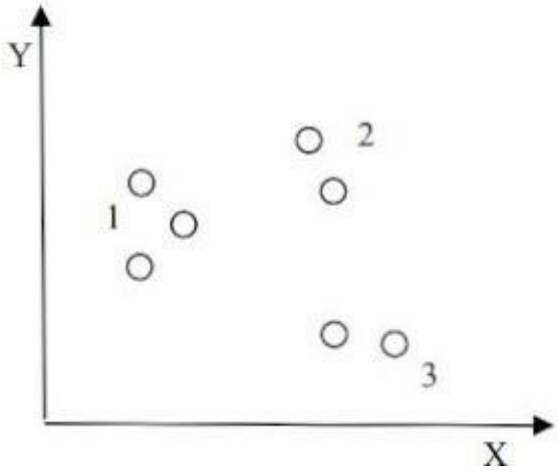


Fig 3: KNN where k=3

2)The decision Tree Algorithm is a supervised ML algorithm. Regression and classification problems can be solved using this. A model used to predict class is created using this algorithm. To predict a class label, always get started from the tree's base. Values of the root attribute and records attribute are compared. After this phenomenon, the next node will be processed.

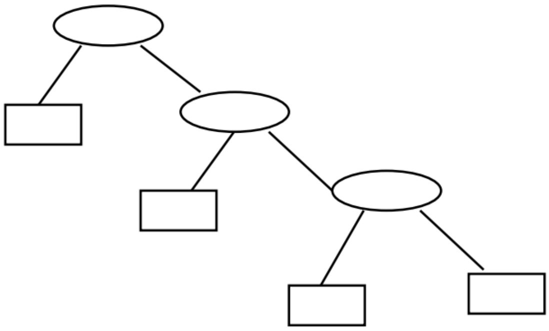


Fig 4: Decision Tree

3)Random forest is an ML algorithm used to solve various regression and classification problems. It displays the majority vote and average for classification and regression, respectively. The primary function of the random forest algorithm is to control the set of data that includes continuous and categorical variables for regression and classification, respectively.

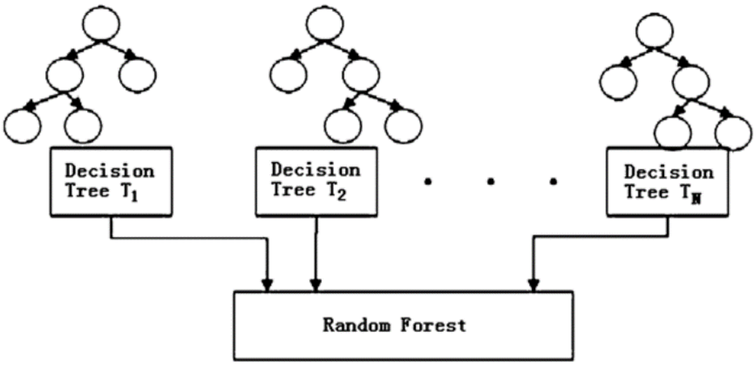


Fig 5: Random Forest

4)SVC stands for Support Vector Machine (SVM) Classifier. SVMs are used in classification and regression analysis to differentiate data into different classes. Essentially, they seek out the hyperplane that best separates the data. SVM is robust and can be used for classification and regression problems.

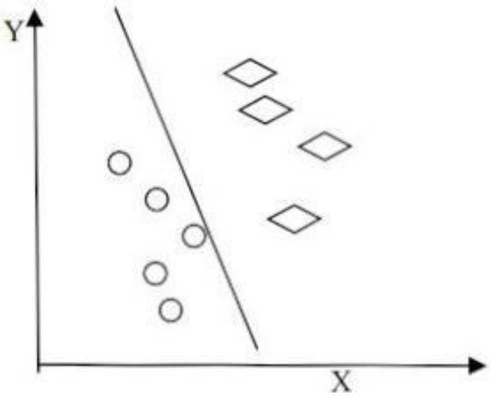


Fig 6: Support Vector Machine (Linear Regression)

5)Logistic Regression models the correlation between the input features and the output variable that produces binary outcomes (0/1) by estimating the probability of the output being 1 for a given set of inputs. Logistic Regression uses the sigmoid function to map input features onto the probability of the output variable. It estimates the model parameters by maximizing the likelihood of observed data given the model.

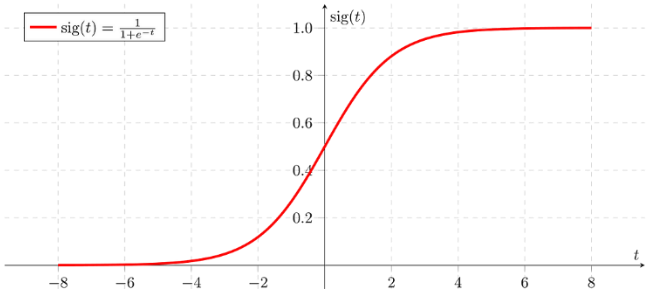


Fig 7: Logistic Regression (Sigmoid Activation Function)

6)Neural Networks are made up of linked nodes or neurons that process and send information through layers of computing and learn complicated patterns and correlations in the data. Neural Networks can be used for various tasks such as classification, Regression, and clustering. They are particularly effective in handling complex and high-dimensional data.

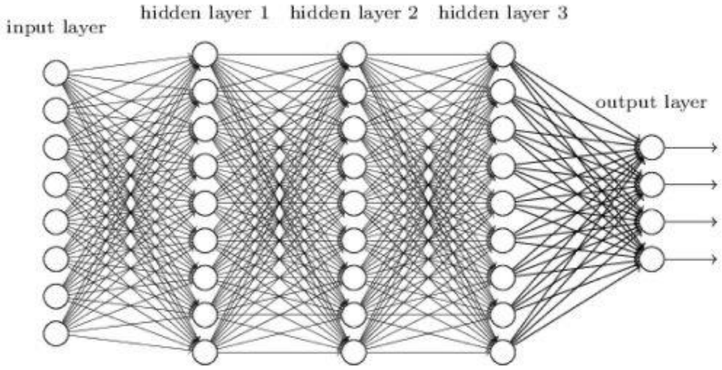


Fig 8: Neural Networks

ENSEMBLE METHODS:

Ensemble methods are used in machine learning to combine multiple models to improve the predictive accuracy of the algorithm. Bagging and Boosting are two popular ensemble methods used in machine learning.

Boosting:

Boosting is an ensemble method that combines multiple weak models to create a more robust model. In this technique, models are trained sequentially, and the misclassified data points from the previous model are given higher weightage while training the next model. This approach helps to focus on the difficult-to-classify data points and create a more robust overall model.

Bagging:

Bagging is an ensemble method that combines multiple models to create a more robust model by training each model on different subsets of the data. In this technique, the training data is randomly sampled with replacement to create multiple subsets. A model is trained on each subset, and the final model is created by combining the output of all the models.

FRONTEND TECHNOLOGIES:

1) HTML, or HyperText Markup Language, is the primary markup language for web documents viewed in a browser. It can be augmented with scripting languages like JavaScript and technologies such as Cascading Style Sheets to enhance a webpage's appearance and functionality. Its purpose is to define the structure of a webpage.It comprises various components. HTML components instruct browsers on how to display content.

2) The creation of style sheets that specify how a document is presented in a markup language, such as HTML or XML, is done using the Cascading Style Sheets language. The WWW's foundational technologies, along with JavaScript and HTML, include CSS. It is a style sheet language that is used to control the layout on a screen, printed materials, or other media. It allows for the management and

uniformity of the design of multiple web pages simultaneously. CSS files can be stored as external style sheets.

3) FLASK: Python-based Flask is a lightweight web application framework. It is designed to make it easy to develop web applications quickly and with less code than other frameworks. Flask provides a simple and intuitive interface for building web applications, including features such as a built-in web server, support for secure cookies, URL routing, and request handling.Flask can be used for many web applications, including simple static websites, dynamic web applications, RESTful APIs, and more. It is highly customizable and extensible, with many third-party libraries and plugins available to add additional functionality.

6. Results and Discussions

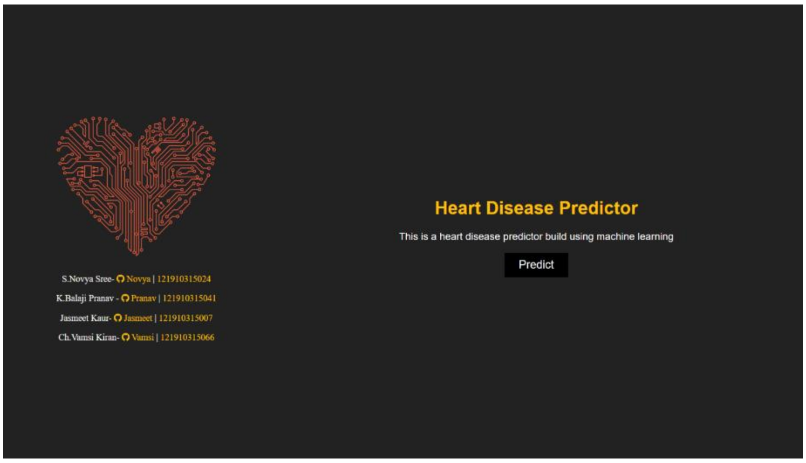
Results:

This project analyzed the feasibility and reliability of using machine learning models as a predictive tool for heart disease diagnosis. The study involved evaluating the performance of six algorithms on a dataset containing patient attributes. Ensemble methods like boosting and bagging were employed to enhance the performance of the models. The analysis identified several challenges and limitations associated with the development and deployment of heart disease prediction models using machine learning algorithms, including limited data, incomplete information, changing risk factors, and uncertainty.

Using the top three algorithms, a final heart disease prediction model was built and tested on a dataset of 6,232 patient records. The model achieved an accuracy of over 82%, which is a promising result indicating the potential for accurately predicting the likelihood of a patient having a certain medical condition. The project's findings point to the potential of machine learning algorithms as a diagnostic tool for cardiac disease, but further study is required to address the problems and restrictions that have been found.

UI OUTPUT:

HOME PAGE:



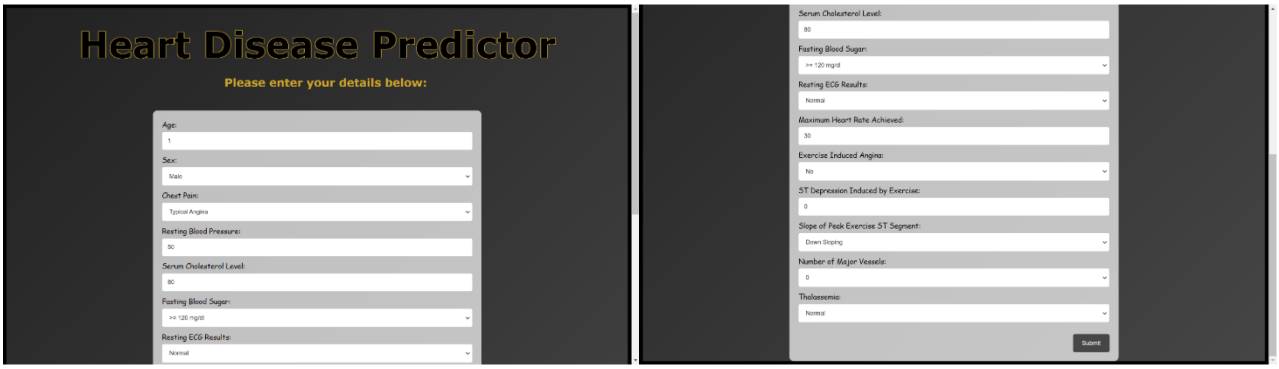
S.Novya Sree- Novya | 121910315024
K.Balaji Pranav - Pranav | 121910315041
Jasmeet Kaur- Jasmeet | 121910315007
Ch.Vamsi Kiran- Vamsi | 121910315066

Heart Disease Predictor

This is a heart disease predictor build using machine learning

Fig 9: Landing page

FORM PAGE:



Heart Disease Predictor

Please enter your details below:

Age:
1

Sex:
Male

Chest Pain:
Typical Angina

Resting Blood Pressure:
50

Serum Cholesterol Level:
80

Fasting Blood Sugar:
>= 120 mg/dl

Resting ECG Results:
Normal

Serum Cholesterol Level:
80

Fasting Blood Sugar:
>= 120 mg/dl

Resting ECG Results:
Normal

Maximum Heart Rate Achieved:
30

Exercise Induced Angina:
No

ST Depression Induced by Exercise:
0

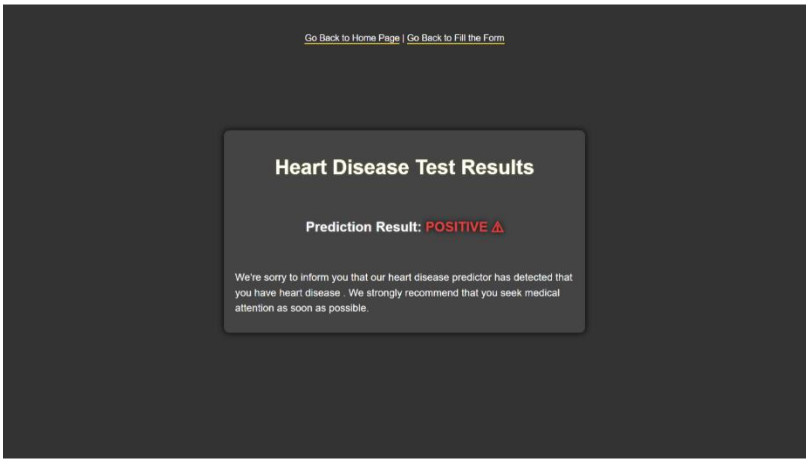
Slope of Peak Exercise ST Segment:
Down Sloping

Number of Major Vessels:
0

Thalassemia:
Normal

Fig 10: Input Form

PREDICTION PAGE:

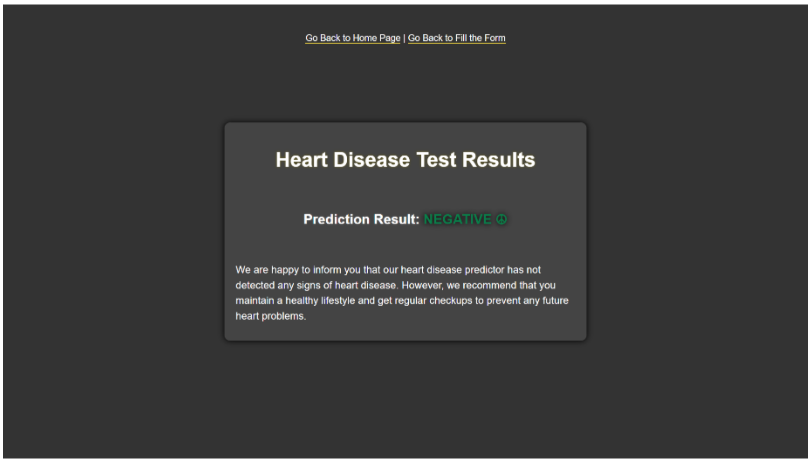


Heart Disease Test Results

Prediction Result: POSITIVE ⚠

We're sorry to inform you that our heart disease predictor has detected that you have heart disease . We strongly recommend that you seek medical attention as soon as possible.

Fig 11: Output page 1 (Positive)



Heart Disease Test Results

Prediction Result: NEGATIVE ☮

We are happy to inform you that our heart disease predictor has not detected any signs of heart disease. However, we recommend that you maintain a healthy lifestyle and get regular checkups to prevent any future heart problems.

Fig 12: Output page 1 (Negative)

Discussion:

The high accuracy achieved by the final model in this project indicates that it can be a useful tool in diagnosing medical conditions. However, it is important to emphasize that the model is not intended to replace medical professionals, but rather to aid them in making diagnoses. The accuracy is also dependent on the quality and quantity of the data used to train it, and further research may be necessary to improve its performance on diverse patient populations or different medical conditions.

Our project highlights the potential of machine learning to enhance medical diagnosis and support healthcare providers in their decision-making. Future work could involve refining the model with additional data or exploring other applications of machine learning in healthcare. By leveraging the power of machine learning, we can advance medical research and improve patient outcomes.

7. Conclusion and Future Scope

In conclusion, our study demonstrates that machine learning models have the potential to accurately predict the likelihood of a patient having a particular medical condition. Our model achieved an overall accuracy of over 82%, which is promising for future applications in healthcare.

However, it is important to note that these models should not replace the expertise of trained medical professionals. Rather, they can serve as a helpful tool for assisting doctors in making informed decisions and improving patient outcomes.

As for future scope, we believe that further research and development of more advanced machine learning models incorporating larger and more comprehensive datasets, and incorporating more features such as genetic information, will be necessary. These developments will enable machine learning models to accurately predict heart disease risk with high accuracy, even in areas where medical professionals are not readily available. This would be similar to the evolution of blood sugar tests, which have become more cost-effective and can now be done in-home. The potential for machine learning models to predict heart disease risk in-home could revolutionize healthcare and benefit both medical professionals and patients.

8. References

Journals/Articles :

1. Chaimaa Boukhatem,Heba Yahia Youssef,Ali Bou Nassif. “Heart Disease Prediction Using Machine Learning”2021.
2. Apurb Rajdhan,Milan Sai.” Heart Disease Prediction Using Machine Learning”2020.
3. Hiteshwar Singh,Tushar Gupta,Jagpreet Sidhu. “Heart Disease Prediction Using Machine Learning”2022.
4. Archana Singh,Rakesh Kumar. “Heart Disease Prediction Using Machine Learning Algorithms”2020.
5. "Predicting cardiovascular disease with deep learning: A review" by Wang et al. (2020).
6. "Predictive modeling of cardiovascular disease using machine learning techniques: A systematic review" by Prabhakar et al. (2021).

e-websites / downloads :

1. Kaggle, https://www.kaggle.com/
2. Course Hero, www.coursehero.com