IJARASEM

ISSN: 2395-7852

International Journal of Advanced Research in Arts, Science, Engineering & Management

Volume 11, Issue 6, November - December 2024

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.583

Predictive Analytics for Heart Disease Using Machine Learning

L.Saroj Vamsi Varun[1], M.Saivardhini[2] , P.Srivarsha[3], Dr.VVSS Balaram[4]

UG Student, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India[*1,2,3]

Assistant Professor, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India[*4]

ABSTRACT: Heart disease is a major challenge for global health, along with high morbidity and mortality. The earlier it is diagnosed, the better the outcome of the patient given timely intervention. This project employs a form of machine learning to train and create a risk assessment model of heart disease from the user-submitted data. The model employs the Random Forest algorithm, one of the most accurate robust algorithms available. We will use a dataset having patient records, such as age, gender, blood pressure, and others. Other medical indicators. Therefore, data preprocessing and feature selection will be applied to enhance the model performance. The main idea is to design an interactive web application which can provide easy access to predict heart disease risk assessment. Users input health parameters, and the application, developed with Streamlit, indicates an immediate prediction for heart disease risk. The interface can be accessed by anyone without technical backgrounds. After the submission, it will process the data through the machine learning model, and risk assessment will be shown. This immediately outcome gotten from this type of risk assessment may prove to be an early warning for patients as they are encouraged to seek medical evaluation in appropriate times. This project, therefore, aims to bridge the gap that exists between advanced machine learning and an actual setting of practical healthcare where users can be well cared for proactively with their hearts and in an effort to ease the burden of heart diseases through early detection and intervention.

KEYWORDS: Machine Learning, Health Parameters, web Application, Data Preprocessing.

I. INTRODUCTION

The rate of heart diseases has surged in the last couple of years to become one of the deadliest conditions around the globe. Predictive analytics, especially through the application of machine learning techniques, provides a sound opportunity for identifying individuals with high risks and thereafter lowering the risk levels before the occurrence of severe symptoms. The current project therefore assists in this vital cause by designing an easy-to-use, web-based application for predicting the risk of heart disease based on the data regarding health furnished by a user. The application is meant to provide immediate risk assessments in the most accurate form possible, using a model of machine learning, such as a Random Forest algorithm, while also alerting users to seek advice from a doctor when necessary. In doing so, it bridges the gap between advanced medical predictions and daily accessibility by ensuring such crucial health information is easily accessible to an individual and can act proactively in managing heart health.

II. RESEARCH METHODOLOGY

This project predicts heart disease using machine learning, specifically a random forest algorithm. The dataset used in this study is sourced from Kaggle. This section gives a detailed description of the methods used for heart disease prediction. Methodology This section outlines the steps followed, which include data processing, model training, and evaluation.

2.1 Data Collection and Pre-processing: The data collections involve the acquisition of relevant medical attributes in heart disease prediction for this study, which include age, sex, cholesterol levels, and maximum heart rate. The target variable used the presence in form of 1 or absence in the form of 0 of heart disease. Handling missing values was the main process in data preprocessing. There were missing values within the given dataset, and imputation has to be carried out as a method in handling missing values using mean/median for continuous variables and mode for categorical variables. Since it is continuous features age, blood pressure, cholesterol and maximum heart rate, it is necessary to uniformly scale them using StandardScaler. Then, the categorical features involving gender and chest pain type experienced were encoded into numeric ones by one-hot encoding.

2.2 Model Development: The RF model was trained in the basis of predictive of heart disease taking relevant features from the dataset. The method of Random Forest is an ensemble learning method, and it constructs multiple decision trees during the training process by relying on the power of diverse models. Individual trees are trained by a randomly selected subset of data alongside the features so that variability is introduced to prevent overfitting. The final prediction is done through majority voting where each tree feeds into the output, therefore the model should be resilient and robust in capturing complex relations in data. To fine-tune the optimal performance of the model, hyperparameters of the model, number of trees, and maximum depth of each tree were optimized using a grid search cross-validation method. This approach systematically controlled both bias and variance thereby enhancing predictive accuracy in determining the presence of heart disease.

2.3 Training and Testing a Model: The data were divided into two subsets, a training set with a ratio of 80%, and the remaining 20% was set to use for testing purposes. This allowed the model to train on the content of the training set without using the other part, reserved for testing the model. The training of the Random Forest model was done using the training data set. The performance of the model was tested using the test set. The accuracy is the general ability of the Random Forest model to classify instances relevant to heart disease, in which this measure is represented as the proportion of times the model was correct. Precision is a measure of the confidence level of the model's positive predictions; that is, the ratio of true positives to the sum of true positives and false positives. Recall measures how good the model is to identify true positives with regard to the actual cases of heart disease. The F1 Score is the balance between precision and recall, especially when your data are unbalanced. It is one value that reflects the trade-off between the two. To avoid overfitting, there's hyperparameter tuning applied and then cross-validation; after that, it guarantees that the model generalizes well to unseen data, thus not attempting to learn the noise in the training set. In any case, all metrics guide tuning in the model's predictive capabilities.

2.4 Web Application Development: To deploy the model, a web application was developed in Streamlit-a Python-based framework for creating interactive web apps. It begins with the header that introduces the app and subsequently includes a sidebar for user input where details such as age, sex, cholesterol levels and other relevant attributes of an individual are entered. The interface loads the model's expected columns and preprocesses the input data by carrying one-hot encoding out for categorical variables. Then, the pre-trained machine learning model is loaded and predictions are made using the processed data. The web app will show whether a person is predicted to have heart disease along with its probability. It also calculates the probability of heart disease along with their risk which could be high, moderate, or low, and recommends the results to the users accordingly. This enables a readable interface to review heart disease probabilities through medical record reviews.

III. THEORY AND CALCULATION

It makes use of diverse types of machine learning algorithms, which predict heart disease based on the patterns within the health data, are informed predictions regarding the well-being of a patient. The main reason for using Random Forest includes an ability to handle nonlinear relationships well, besides offering robustness. This is a style of ensemble learning where it constructs several decision trees during training. Every decision tree is determined by the model based on a random subset of the training data as well as the specific features. This enables the model to exploit an extremely wide range of diversified patterns as well as interactions among the features. The outputs from multiple decision trees are then merged to provide Random Forest, which enhances predictive accuracy and allows control over overfitting, especially useful for complex datasets such as those concerning heart disease prediction.

Actually, the building block of the ensemble is a decision tree-the core theoretical idea behind Random Forest. In simple words, a decision tree basically works exactly like a tree-like model that splits data into different categories based on the values of features so as to classify instances accordingly. The tree has each node as some feature of interest and a branch represents a decision rule leading to a leaf node containing the final prediction outcome. In addition to that, Random Forest utilizes a technique called Bootstrap Aggregation, normally referred to as bagging. The use of bagging helps to train individual decision trees on a randomly chosen subset of training data. This reduces the variance generally and will work toward improving the overall performance of the whole model. This approach only increases the robustness of predictions as well as reduces the overfitting possibilities, thus ensuring that the model would generalize nicely to unseen data. After training the model, its performance is accessed via accuracy, precision, recall, and the F1-score, all together indicating how good the fit is.

Mathematical Expression and Symbols

Several math expressions and symbols are used to describe a number of processes related to model training and evaluation.

- Random Forest, is an ensemble method that combines multiple decision trees. The prediction is the majority vote or average from all decision trees. Mathematically, this can be represented as:

$$h(X)=1/T\sum_{t=1} ht\ (X)$$

- After training, the model's performance is evaluated using accuracy, precision, recall, and F1-score. Accuracy is calculated as:

$$\text{Accuracy} = \frac{Number\ of\ Correct\ Predictions}{Total\ Predictions}$$

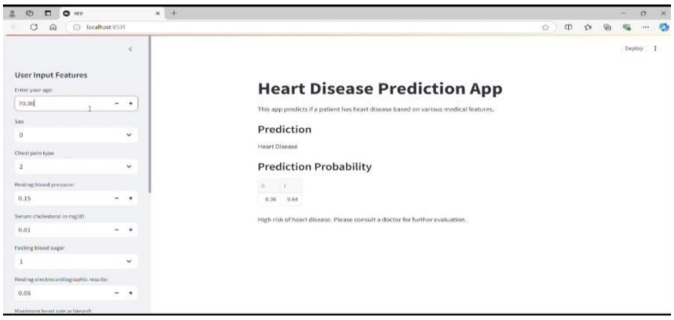
- Precision and recall are calculated based on true positives (TP), false positives (FP), and false negatives (FN):

$$\text{Precision} = \frac{TP}{TP+FP} \ \text{ and Recall} = \frac{TP}{TP+FN}$$

- The F1-score, which balances precision and recall, is given by:

$$\text{F1 Score} = 2 * \frac{Precision*Recall}{Precision+Recall}$$

IV. RESULTS AND DISCUSSION



User Input Features

Enter your age: 70.00

Sex: 0

Chest pain type: 2

Resting blood pressure: 0.15

Serum cholestoral in mg/dl: 0.01

Fasting blood sugar: 1

Resting electrocardiographic results: 0.06

Maximum heart rate achieved:

Heart Disease Prediction App

This app predicts if a patient has heart disease based on various medical features.

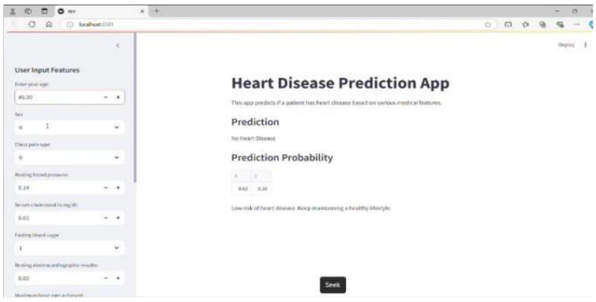
Prediction

Heart Disease

Prediction Probability

0	1
0.36	0.64

High risk of heart disease. Please consult a doctor for further evaluation.



User Input Features

Enter your age: 45.00

Sex: 0

Chest pain type: 0

Resting blood pressure: 0.14

Serum cholestoral in mg/dl: 0.01

Fasting blood sugar: 1

Resting electrocardiographic results: 0.00

Maximum heart rate achieved:

Heart Disease Prediction App

This app predicts if a patient has heart disease based on various medical features.

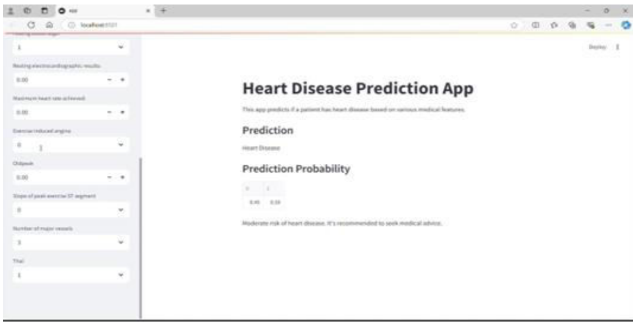
Prediction

No Heart Disease

Prediction Probability

0	1
0.62	0.38

Low risk of heart disease. Keep maintaining a healthy lifestyle.



User Input Features

Fasting blood sugar: 1

Resting electrocardiographic results: 0.00

Maximum heart rate achieved: 0.00

Exercise induced angina: 0

Oldpeak: 0.00

Slope of peak exercise ST segment: 0

Number of major vessels: 3

Thal: 1

Heart Disease Prediction App

This app predicts if a patient has heart disease based on various medical features.

Prediction

Heart Disease

Prediction Probability

0	1
[illegible]	[illegible]

Moderate risk of heart disease. It's recommended to seek medical advice.

This model of heart disease prediction, basically with the use of the Random forest algorithm, produced results that were strong enough to classify patients based on their attributes into different classes such that it could be able to differentiate between individuals with and those without heart disease, with an accuracy of about 90%. Further incremental increases in precision and recall values gave very promising results with good ratios of true positives against most true cases, and low counts of false positives. Such is a kind of more important performance because such performance indicates that the model actually can make informed predictions, which may be very crucial in this clinical decision-making.

F1 score wise, the evaluation metric indicates that besides the good performance of the model of Random Forest at a high level of overall accuracy, it also happens to be in good balance in the precision and the recall. Models containing an aggregate of various decision trees through ensemble learning help a model handle nonlinear relationships, hence minimize the chances of over-fitting. This becomes very useful in the manipulation of high-dimensional medical data where interactions of variables could be complex leading to heart disease outcomes. Bagging within the random forest application translates to bootstrap aggregation contributing to predictions by an application that has aspects of creating robustness and generalization. Such findings have high implications for healthcare professionals.

This model for heart disease prediction can be very resourceful in the early detection and evaluation of risk factors, leading to timely intervention and care for each patient. Such an assessment of features of a patient above in relation to age, cholesterol levels, and blood pressure can thus make use of this model in an all-round examination of risk factors of heart disease. However, the more it should be tried out on larger and far more expansive databases which then should make it quite robust and reliable for populations with diverse demographics. Overall, this study underlines the possibility of machinery learning to help access better management of cardiovascular health and proactive healthcare practices.

V. CONCLUSION

Indeed, the study succeeds in demonstrating the feasibility of machine learning techniques, for example, the Random Forest algorithm, to the problem of heart disease prediction using critical medical features. The model realized high accuracy that reflects its capability to discriminate between patients with and without heart disease, thus making it a valuable clinical decision-making instrument. By using ensemble learning and bootstrap aggregation techniques, the model minimized overfitting and yielded enhanced robustness on high-dimensional data. Precision and recall evaluation metrics also pointed toward the answer, with the model performing well to have minimal false positives but identify actual cases. This research contributes to early detection strategies but also better personal patient care within cardiovascular health. Future work will include adding additional variables to the dataset and comparing several different machine learning algorithms to make predictions even more accurate. In general, these results suggest major empirical potential for data-driven approaches to advance healthcare outcomes and optimize patient management for heart disease.

VI. DECLARATIONS

6.1 Study Limitations:
The study on heart disease prediction using machine learning faced limitations such as model's accuracy depends on input data quality, with Random Forest requiring careful hyperparameter tuning.

6.2 Funding Source:
None.

6.3 Acknowledgements:
I would like to thank my guide and the Head of the Department at Anurag University, for encouragement and invaluable guidance throughout this piece of research work, whose motivation and insights have been instrumental in bringing this project successfully to completion.

6.4 Informed Consent:
All respondents have been given permission to participate in these research-based studies with their informed consent regarding the purpose of the study, the procedures followed and the application of their data for publication in this work.

REFERENCES

1. Rohith R, Senthilnayaki B, Joshua Dayalan M, "Cardio Care: A Predictive Model for Heart Disease Detection", 2024 Third International Conference on Smart Technologies and Systems for Next Generation Computing September(2024).https://ieeexplore.ieee.org/document/10671202
2. Karthigeyan S, R Bhuvaneswari, "Cardiovascular Disease Prediction based on Decision Tree", 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), July (2024). https://ieeexplore.ieee.org/document/10602257
3. Ramalingam, V V & Dandapath, Ayantan & Raja, M. "Heart disease prediction using machine learning techniques: A survey." International Journal of Engineering & Technology. March(2018) https://www.researchgate.net/publication/325116774_Heart_disease_prediction_using_machine_learning_techniques_A_survey
4. R. Chauhan, P. Bajaj, K. Choudhary and Y. Gigras, "Framework to predict health diseases using attribute selection mechanism," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom) March(2015). https://ieeexplore.ieee.org/document/7100571
5. Apurb Rajdhan , Avi Agarwal , Milan Sai , Dundigalla Ravi, Dr. Poonam Ghuli, " Heart Disease Prediction using Machine Learning," INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, April(2020) https://www.ijert.org/heart-disease-prediction-using-machine-learning
6. Modepalli, Kavitha & Gnaneswar, G. & Dinesh, R. & Sai, Y. & Suraj, R.. (2021). :Heart Disease Prediction using Hybrid machine Learning Model.". 6th International Conference on Inventive Computation Technologies (ICICT) January (2021) https://www.researchgate.net/publication/349660671_Heart_Disease_Prediction_using_Hybrid_machine_Learning_Model

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| Mobile No: +91-9940572462 | Whatsapp: +91-9940572462 | ijarasem@gmail.com |

www.ijarasem.com