Heart Disease Prediction using Machine Learning

Amol Bhilare, Harsh Bihani, Tejas Desai, Vedant Bhosle, Aryan Kirange

Vishwakarma Institute of Technology (Pune), Department of Computer Engineering

ABSTRACT

Heart is very vital organ in living beings. Being such significant body part, there is a need for precise diagnosis of heart diseases. Errors in diagnosis can be risky or even fatal. Deaths due to heart diseases are on rise and require immediate attention. To address this issue, accurate predictive system is required for disease awareness at early stage. This can be done using Machine Learning. Machine learning is a branch of Artificial Intelligence (AI) that provides help and assistance by using data and algorithms in anticipating any type of event. The dataset used for training and testing in this system is UCI repository dataset and the algorithms used are Logistic Regression (LR), Support Vector Machine(SVM) and Random Forest (RF). The whole system is implemented in R language using R studio as it is the finest tool for implementing R programming since it has various types of libraries and header files that make the task more exact and precise. Shiny package has been used for UI.

Keywords *— Heart Disease prediction, Logistic Regression, Random Forest, SVM, Machine Learning*

I. INTRODUCTION

The heart is one of the largest and most important organs in the human body, so taking care of it is essential. Most of the diseases are related to the heart so it is necessary to predict the heart disease and for this purpose, in this area we have need for comparative research, nowadays most of the patients who die from their disease are detected at the end stage. Due to imprecision in equipment's, we need the most efficient algorithms for disease prediction.

Machine learning algorithms are most effective for prediction, based on training and testing data. Machine Learning is particular area of artificial intelligence (AI), one of the many fields of learning in which machines mimic human abilities. Machine learning systems are trained how to analyze and utilize data whereas artificial intelligence is the name given to the result of the two technologies working together.

In this project, we use biological parameters as test data, such as cholesterol, blood pressure, gender, and age, and on the basis of these, a comparison is made in terms of the accuracy of algorithms for which in this project we have used algorithms like Logistic Regression, Random Forest, and KNN.

II. LITERATURE REVIEW

[1] The paper named "Predicting Heart Disease at Early Stages using Machine Learning: A Survey" by authors Rahul Katarya, Polipireddy Srinivas tested algorithms Decision tree and Random Forest with accuracy of 86.3% and 93.33% respectively.

[2] The paper named "Heart Disease Prediction using Hybrid machine Learning Model" by authors Dr. M. Kavitha, G. Gnaneswar, R. Dinesh, Y. Rohith Sai, R. Sai Suraj tested algorithms Hybrid Model, Decision tree and Random Forest with accuracy of 79%, 81% and 88% respectively.

[3] The paper named "Heart disease prediction using machine learning algorithms" by authors Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath tested algorithms Naïve Bayes and KNN with accuracy of 85% and 87.5% respectively.

[4] The paper named "Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method" by authors Xiao-Yan Gao, Abdelmegeid Amin Ali, Hassan Shaban Hassan, and Eman M. Anwar tested algorithms KNN, SVM and RF with accuracy of 98%, 84.7% and 97.9% respectively.

[5] The paper named "Using Machine Learning for Heart Disease Prediction" by authors Dhai Eddine Salhi , Abdelkamel Tari , and M-Tahar Kechadi tested algorithms KNN, SVM and Neural Network with accuracy of 85.5%, 90% and 93% respectively.

[6] The paper named "Heart Disease Prediction using Machine Learning Techniques" by author Pooja Anbuselvan tested algorithms Logistic Regression, Naïve Bayes, SVM, KNN, Decision Tree, Random Forest and XGBoost with accuracy of 75.41%, 77.05%, 73.77%, 57.83%, 77.07%, 86.89% and 78.69% respectively.

[7] The paper named "HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS" by authors Gunturu Deepthi, Cherukuri Shivani, Koruprolu Nagavinith and Kesuboyina Hanudeep tested algorithms XG-boost, SVM, LR. RF, Naïve Bayes, Decision Tree,

Adaboost with accuracy of 81.3%, 80.2%, 79.1%, 79.1%, 76.9%, 75.8% and 73.6%respectively.

[8] The paper named "Early and Accurate Prediction of Heart Disease Using Machine Learning Model" by authors B Padmaja , Chintala Srinidhi , Kotha Sindhu , Kalali Vanaja , N M Deepika and E Krishna Rao Patro tested algorithms RF, LR, KNN, SVM, Decision Tree, Gradient Boosting and Naïve Bayes with accuracy of 93.4%, 88.5%, 91.8%, 83.6%, 86.6% and 85.2%, respectively.

[9] The paper named "Heart Disease Prediction Using Various Algorithms of Machine Learning" by author Rati Goel tested algorithms LR, KNN, SVM, Naïve Bayes, Decision Tree and RF with accuracy of 77%, 82%, 86%, 63%, 83% and 83% respectively.

III. METHODOLOGY

IMPLEMENTATION:

- R Studio
- Shiny Package

A. Data Collection: First and most basic step is to understand the problem and collect required data. If data is good, predictions will be better. Our System uses UCI repository dataset.

B. Attribute Selection: In this phase, we find out which are the best subset of attributes among all the attributes in dataset that will be relevant for our system. All the attributes used in our system are mentioned in TABLE 1.

C. Data Cleaning: In this phase, incorrect, corrupted, duplicate or incomplete data is fixed or removed to increase quality of data.

D. Data Transformation: In this phase, particular formats of data are converted to another to organize data and increase its efficiency,

E. Data Visualization: In this phase, we use graphical representations of data to find patterns, trends and correlations. We used Bar Plot [Fig. 3], Histogram [Fig. 4], Box Plot [Fig. 5] and correlated the attributes [Fig. 6] in our system for better understanding of data.

F. Data Split: To ensure good performance of model and better understanding of its characteristics, we need to divide our dataset into training and testing datasets. In our model we split dataset in 70:30 ratio for training and testing dataset respectively.

G. Choosing a Model: In this step, we decide which model to be tested out of various ML models depending on the data. We tested Logistic Regression, Support Vector Machine and Random Forest algorithms on our system.

H. Training a Model: This is the most important phase in Machine Learning where we need to train data to our models to make predictions.

I. Evaluation: In this phase, we verify the accuracy of model by evaluating models on basis of testing data set. For our models, accuracies are mentioned in TABLE 3.

J. Prediction: In this phase, model with best accuracy is used for prediction. We used RF for prediction of heart disease.

ACTIVITY DIAGRAM:

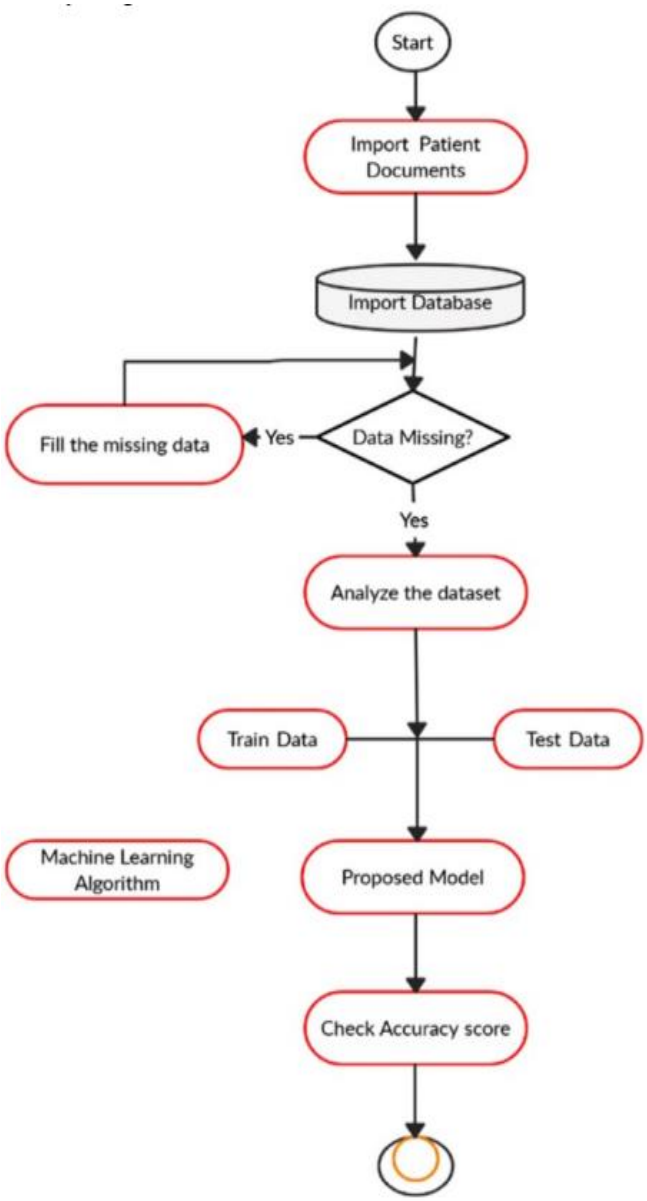

Fig. 2 Activity Diagram

TABLE.1 Attributes of the Dataset

Sr. No.	Attribute	Description	Type
1	slope_of_peak_exercise_st_segment	an electrocardiography read out indicating quality of blood flow to the heart	int
2	thal	results of thallium stress test measuring blood flow to the heart, with possible values normal, fixed_defect, reversible_defect	categorical
3	resting_blood_pressure	resting blood pressure	int
4	chest_pain_type	Chest pain type (4 values)	int
5	num_major_vessels	number of major vessels (0-3) colored by flourosopy	int
6	fasting_blood_sugar_gt_120_mg_per_dl	fasting blood sugar > 120 mg/dl	binary
7	resting_ekg_results	resting electrocardiographic results (values 0,1,2)	int
8	serum_cholesterol_mg_per_dl	serum cholestoral in mg/dl	int
9	oldpeak_eq_st_depression	oldpeak = ST depression induced by exercise relative to rest, a measure of abnormality in electrocardiograms	float
10	sex	0: female, 1: male	binary
11	age	Age in years	int
12	max_heart_rate_achieved	maximum heart rate achieved (beats per minute)	int
13	exercise_induced_angina	exercise-induced chest pain (0: False, 1: True)	binary

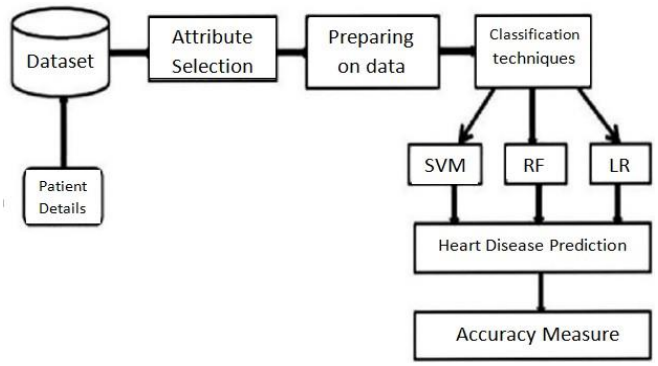

Fig.1 Architecture of Prediction System

IV. MACHINE LEARNING ALGORITHMS

A. Logistic Regression

Machine learning-inspired statistical analysis technique called logistic regression. When the dependent variable is binary or dichotomous, this is utilized. This indicates that there is just one variable and just two possible exits. For instance, whether or not someone survives this accident, whether or not a student passes this exam, etc. Either yes or no is the answer (two outputs). This regression technique, which is comparable to linear regression, can be used to estimate the likelihood of a classification issue.

A statistical model commonly used to model binary dependent variables using logistic functions. Another name for a logistic function is a sigmoid function, which is given as:

$$F(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{e^{x}+1}$$

B. Support Vector Machine

It is a type of machine learning technique that relies on the idea of a hyperplan, which classifies the data by building a hyperplane between it.

(Yi, Xi) is the training sample dataset where i=1, 2, 3,..., n and Xi is the ith vector whereas Yi is the target vector The quantity of hyperplanes determines the kind of support vector; for instance, if a line is employed as a hyperplane, the technique is known as a linear support vector.

C. Random Forest

This supervised machine learning approach called random forest is frequently applied to classification and regression issues. Create a decision tree with various examples, and in the case of regression, make a majority vote on the classification and average.

ACCURACY CALCULATION:

The four values true positive (TP), false positive (FP), true negative (TN), and false negative affect how accurate the algorithms are.

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

Where:

TP = Number of heart disease patients

TN = stands for the number of people with and without heart disease.

FP = The proportion of people without heart disease

FN = The number of people with and without heart disease.

V. RESULT AND DISCUSSION

Data Visualization:

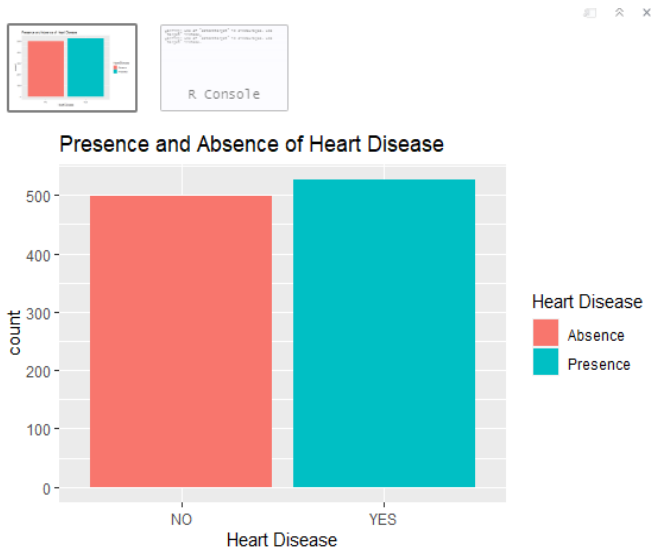

Fig.3 Bar Plot for target (heart disease)

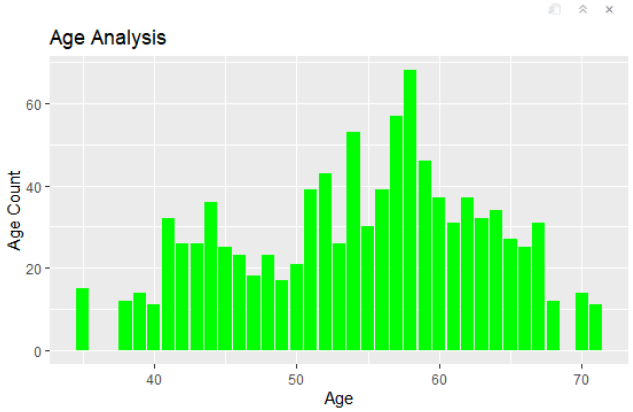

Fig.4 Frequency of age

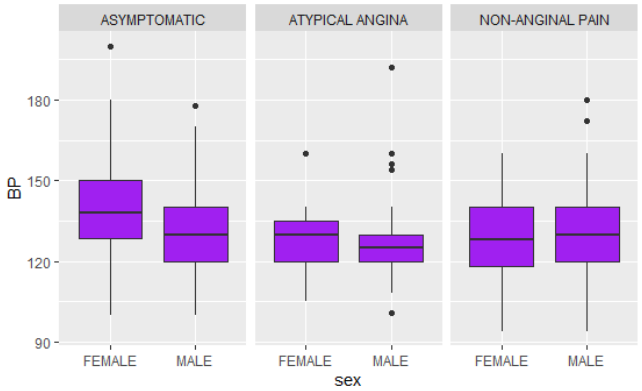

Fig.5 Box Plot (Compare BP across the chest pains)

Correlation:

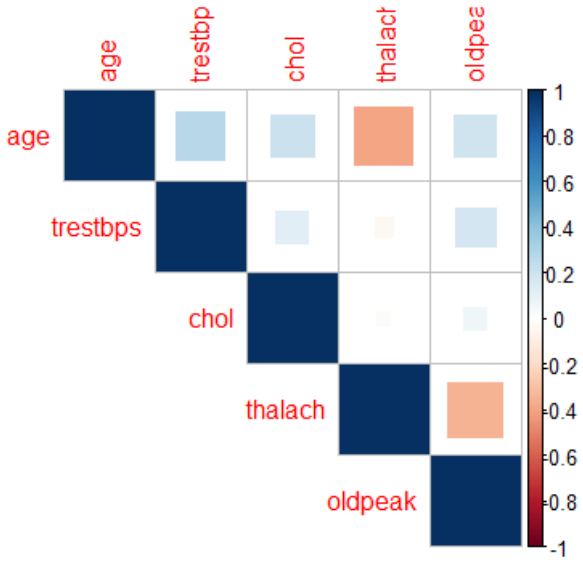

Fig.6 Correlation of Attributes

TABLE.2 Confusion Matrix

Sr. No	Algorithm	True Positive	False Positive	False Negative	True Negative
1.	LR	137	26	20	123
2.	SVM	154	9	3	140
3.	RF	157	4	0	145

TABLE.3 Analysis of Algorithms

Sr. No.	Algorithm	Accuracy
1.	Logistic Regression	84.97%
2.	Support Vector Machine	96.08%
3.	Random Forest	98.69%

After applying machine learning, we found the accuracies of LR, SVM and RF. On comparing all the three, RF is best with accuracy of 98.69%.

Further, we created a UI using Shiny Package where user can provide details of various attributes regarding his health conditions. On Applying RF algorithm on this data, we predicted whether the patient is suffering from heart disease or not.

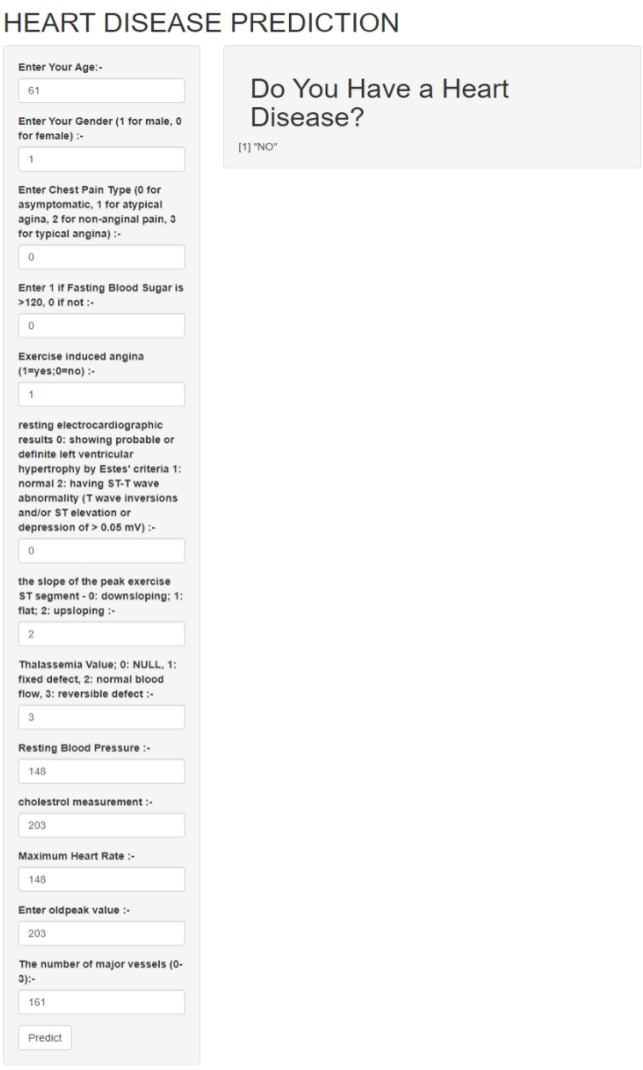

Fig.7 Prediction of Heart Disease

VI. FUTURE SCOPE

More machine learning techniques can be used for better prediction of heart diseases.

VII. CONCLUSION

Prediction about congestive heart failure is a major deal for human beings. Accuracy becomes a major factor in efficient analysis of algorithms to predict heart disease. Accuracy of the Machine learning algorithms rely on attributes of the dataset that is used. When we perform the analysis of algorithms considering dataset whose attributes are shown in TABLE.1 and confusion matrix, we find Random Forest is most accurate compared to other algorithms.

VIII. ACKNOWLEDGEMENT

I would like to thank our mentor, Prof. Amol Bhilare for timely guidance and assistance during the course of this project

REFERENCES

[1] Rahul Katarya, Polipireddy Srinivas, Predicting Heart Disease at Early Stages using Machine Learning, A Survey, Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020) IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4.

[2] Vijeta Sharma, Manjari Gupta, Shrinkhala Yadav, Heart Disease Prediction using Machine Learning Techniques, 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) | 978-1-7281-8337-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICACCCN51052.2020.9362842.

[3] Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath, Heart disease prediction using machine learning algorithms, IOP Conference Series: Materials Science and Engineering, Volume 1022, 1st International Conference on Computational Research and Data Analytics (ICCRDA 2020) 24th October 2020, Rajpura, India.

[4] Xiao-Yan Gao, Abdelmegeid Amin Ali, Hassan Shaban Hassan, and Eman M. Anwar, Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method, Hindawi Complexity Volume 2021, Article ID 6663455.

[5] Dhai Eddine Salhi, Abdelkamel Tari, and M-Tahar Kechadi, Using Machine Learning for Heart Disease Prediction, February 2021, DOI:10.1007/978-3-030-69418-0_7, In book: Advances in Computing Systems and Applications (pp.70-81)

[6] Pooja Anbuselvan, Heart Disease Prediction using Machine Learning Techniques, International Journal of Engineering Research & Technology (IJERT) ISSN:2278-0181 IJERTV9IS110259 Vol. 9 Issue 11, November-2020.

[7] Gunturu Deepthi, Cherukuri Shivani, Koruprolu Nagavinith and Kesuboyina Hanudeep, HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

[8] B Padmaja , Chintala Srinidhi , Kotha Sindhu , Kalali Vanaja , N M Deepika, E Krishna Rao Patro, Early and Accurate Prediction of Heart Disease Using Machine Learning Model, Turkish Journal of Computer and Mathematics Education, Vol.12 No.6 (2021), 4516-4528.

[9] Rati Goel, Heart Disease Prediction Using Machine Learning Algorithms, Conference: 2020 International Conference on Electrical and Electronics Engineering (ICE3).