# Heart Disease Prediction using Machine Learning

Pabitra Kumar Bhunia
Department of Computer Science and Engineering
JIS College of Engineering
Kalyani, Nadia, West Bengal, India

Arijit Debnath
Department of Computer Science and Engineering
JIS College of Engineering
Kalyani, Nadia, West Bengal, India

Poulami Mondal
Department of Computer Science and Engineering
JIS College of Engineering
Kalyani, Nadia, West Bengal, India

Monalisa D E
Department of Computer Science and Engineering
JIS College of Engineering
Kalyani, Nadia, West Bengal, India

Kankana Ganguly
Department of Computer Science and Engineering
JIS College of Engineering
Kalyani, Nadia, West Bengal, India

Pranati Rakshit
Department of Computer Science and Engineering
JIS College of Engineering
Kalyani, Nadia, West Bengal, India

***Abstract*— Heart-disease (HD) is one of the most common diseases nowadays, and for people who provide health care, it is very necessary to work with them to take care of their patients' health and save their life. In this paper, different classifiers were analyzed by performance comparison to classify the Heart Disease dataset to classify it correctly and or to Predict Heart Disease cases with minimal attributes.**

**Large amounts of data that contain some secret information were collected by the healthcare industries. This data collection is useful for making effective decisions. Some advanced data mining techniques are used to make proper results and making effective decisions on data. In this case, a Heart Disease Prediction System (HDPS) is developed using Logistic Regression, K Nearest Neighbor, Decision Tree, Random Forest Classifier, and Support Vector Machine algorithms to predict the heart disease risk level.**

**The results reveal that the Random Forest Classifier and Support Vector Machine obtained the highest accuracy of 90.32%, whereas 87.09%, 70.96%, and 83.87% accuracy scores are obtained by logistic regression, KNN classifier, and decision tree respectively.**

***Keywords*— Machine learning, Logistic regression, Heart disease, Support vector machine, accuracy***

## I. INTRODUCTION

Data mining is the process by which we can find usually unknown scriptures, patterns, and ongoing trends in databases and it uses that piece of information to structure prognostic models. Data mining technology combines analysis based on statistics, machine learning algorithm, and database technology management system to generate disclosed patterns and establish relationships from huge databases.

The World Health statistics 2012 highlights the issue that every one in three adult age group showed prone to high blood pressure- a situation that results in half of the deaths from heart issues and strokes. Disease-related to the heart, also known as cardiovascular disease (CVD), discusses various conditions that affect the heart not just the disease. This juncture proved fatal for one person in every 34 seconds in the United States.

Heart disease of the coronary arteries, cardiomyopathy, and cardiovascular health issues are certain subdivisions where the blood is pumped and its circulation is made throughout the body. Diagnosis is an important task that has to be performed efficiently. This is mainly done under a doctor's guidance. This causes unsatisfactory results & excessive medical costs of treatments provided to patients. So, we conclude that an automated medical diagnosis and prediction system would prove extremely favorable.

## II. LITARATURE REVIEW

Numerous studies have been done that have focused on the diagnosis of heart disease. They have applied different data mining techniques for diagnosis & achieved different probabilities for different methods.

This system evaluates those parameters using the data mining classification technique. The datasets are evaluated in python using two main Machine Learning Algorithms: The decision Tree Algorithm and the Naive Bayes Algorithm which shows the best algorithm between these two in terms of the accuracy level of heart disease [1].

Aditi Gavhane et al. predicted heart attack for early diagnosis to reduce the count of deaths. For this problem Machine Learning plays a major role in this paper. This prediction takes people from the danger zone of their life. In this paper, we use the KNN algorithm and Random Forest algorithm to predict the heart attack in advance [2].

Senthil Kumar et al. introduced a prediction model with different combinations of features, and several known classification techniques. It produced an enhanced performance level with an accuracy level of 88.7%

through the prediction model for heart disease with Hybrid Random Forest with Linear Model (HRFM) [3].

Himanshu Sharma et al. stated and proved that machine learning algorithms and deep learning opens new door opportunities for precise prediction of a heart attack. Paper provides a lot of information about state of art methods in Machine learning and deep learning. An analytical comparison has been provided to help new researchers working in this field [4].

M. Nikhil Kumar et al. worked with 8 algorithms including Decision Tree, J48 algorithm, Logistic model tree algorithm, Random Forest algorithm, Naïve Bayes, KNN, Support Vector Machine, Nearest Neighbor to predict heart diseases. The accuracy of the prediction level is high when using more attributes [5].

Amandeep Kaur et al. stated that Data mining is an important stage of the KDD process that can be used for disease management, diagnosis, and prediction in healthcare organizations. This paper discusses reviews on different methods and approaches in data mining that have been used to predict heart disease [6].

Pahulpreet Singh Kohli developed an Enhanced New Dynamic Data Processing (ENDDP) Algorithm to predict the early stages of heart disease. The results prove the performance of the proposed system [7].

## III. DATA SET INFORMATION

The name of the dataset is heart.csv. There are 303 instances in this dataset, where the cases are either people having heart disease or they are healthy. Among 303, 165 (54.45%) cases are people with heart disease and 138 (45.54%) are people without heart disease. The number of attributes is 14. There are no missing values in the data set nor any null values.
Features include age, sex, chest-pain type, rest BP, cholesterol, blood sugar level, ECG result, maximum heart rate achieved, exercise-induced angina, ST depression, the slope of peak exercise ST segment, number of major vessels, and defect in heart as of 3-normal, 6-fixed defect and 7-reversible defect. Bar graph (Fig.1) showing the positive and negative cases (1=positive, 0=negative) Scatter plot (Fig.2) showing the positive and negative cases depending on age.

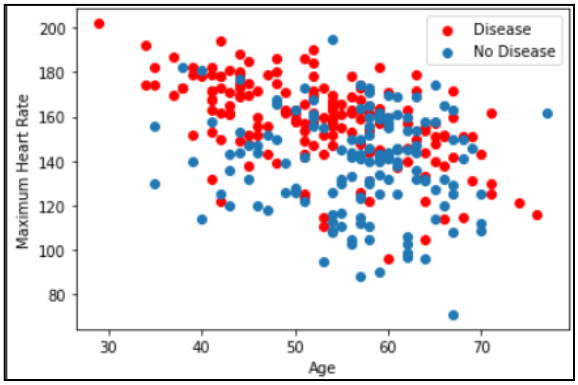


Fig.1 Positive and negative cases

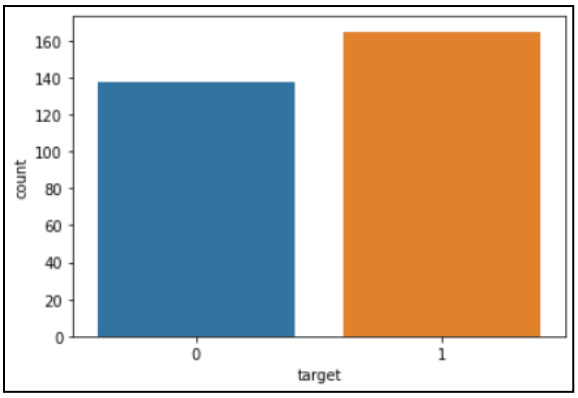


Fig.2 Positive and negative cases depending on age

| | age | sex | cp | trestbps | chol | fbs | ... | exang | oldpeak | slope | ca | thal | target |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 63 | 1 | 3 | 145 | 233 | 1 | ... | 0 | 2.3 | 0 | 0 | 1 | 1 |
| 1 | 37 | 1 | 2 | 130 | 250 | 0 | ... | 0 | 3.5 | 0 | 0 | 2 | 1 |
| 2 | 41 | 0 | 1 | 130 | 204 | 0 | ... | 0 | 1.4 | 2 | 0 | 2 | 1 |
| 3 | 56 | 1 | 1 | 120 | 236 | 0 | ... | 0 | 0.8 | 2 | 0 | 2 | 1 |
| 4 | 57 | 0 | 0 | 120 | 354 | 0 | ... | 1 | 0.6 | 2 | 0 | 2 | 1 |
| .. | ... | ... | .. | ... | ... | ... | ... | ... | ... | ... | .. | ... | ... |
| 298 | 57 | 0 | 0 | 140 | 241 | 0 | ... | 1 | 0.2 | 1 | 0 | 3 | 0 |
| 299 | 45 | 1 | 3 | 110 | 264 | 0 | ... | 0 | 1.2 | 1 | 0 | 3 | 0 |
| 300 | 68 | 1 | 0 | 144 | 193 | 1 | ... | 0 | 3.4 | 1 | 2 | 3 | 0 |
| 301 | 57 | 1 | 0 | 130 | 131 | 0 | ... | 1 | 1.2 | 1 | 1 | 3 | 0 |
| 302 | 57 | 0 | 1 | 130 | 236 | 0 | ... | 0 | 0.0 | 1 | 1 | 2 | 0 |

Fig.3 Data set description

## IV. METHODOLOGY

### A. *Data set information*

The main objective of this research is to develop a heart disease prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set heart disease prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases.

### B. *Training and testing*

The training phase extracts the features (independent variables) from the dataset and the testing phase (containing dependent variables) is used to determine how the appropriate model behaves for prediction. We have divided the dataset into two sections. These are the training and testing phases. We have split the dataset into 90% training and 10% testing phase. And we have taken the random state as 1. For initializing the fixed internal random number generator, we use the random state parameter which will decide the splitting of data into train and test indices. Setting a random state will guarantee a fixed value that the same sequence of random numbers will be generated each time the code is being run. Setting random state, a fixed value will guarantee that the same sequence of random numbers is generated each time we run the code. Then we scaled the data using Standard scattered and fitted the training and testing data using 'fit. transform'.

```
from sklearn.model_selection import train_test_split
X = dataset.iloc[:, :-1].values
Y = dataset.iloc[:, -1].values
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size= 0.1, random_state = 1)   #90% training, 10% testing
```

*C. Classsification used*

- Logistic regression

Logistic Regression is an analytical modeling technique. It is used for analyzing a dataset in which there are one or more independent variables that decide a result. Logistic Regression was imported with a random state of 0. And then the training model was fitted. The testing accuracy was 87.09%

- KNN Classifier

K-nearest neighbor algorithm is utilized for grouping and used in pattern recognition. It is widely used in predictive analysis. On the arrival of new data, the K-NN algorithm [8] identifies existing data points that are nearest to it. From 'sklearn.neighbors', 'KNeighbors Classifier' was imported with n_neighbors = 1. Then the training model was fitted. The testing accuracy was 70.96%

- Support vector machine

Support Vector Machine or SVM is one of the popular Supervised Learning algorithms in machine learning. The benefits of the SVM algorithm is that it creates the best suitable line or decision boundary that can separate a n-dimensional space into classes so that we can easily verify and put the new added data points in the correct category in the future. From 'sklearn', 'svm' was imported and we kept the kernel as linear and gamma as auto and C = 2. And the training model was fitted. The testing accuracy was 90.32%.

- Random forest

Random forest classifier is a powerful supervised classification tool. RF generates a forest of classification trees from a given dataset, rather than a single classification tree. Each of these trees produces a classification for a given set of attributes. From 'sklearn.ensemble', 'Random Forest Classifier' was imported. The n_estimators is kept at 10 and random state at 0. Then the training model was fitted. The testing accuracy was 90.32%.

- Decision Tree

The testing accuracy was 90.32%. A Decision tree is a tree shape-like diagram, where the internal nodes represent a test on an attribute, each branch denotes the outcome of the test, each leaf node denotes a class label. Decision Tree was imported where the random state was kept as 0 and then the training model was fitted. The testing accuracy was 83.87%. 6. Results Amongst all classification techniques, testing accuracy was best in the case of the random forest and SVM approach with an accuracy of 90.32%.

## V. RESULT

Amongst all classification techniques [Table.1], testing accuracy was best in the case of the random forest and SVM approach with an accuracy of 90.32%.

Table. 1 Comparison of performances difference classifier

| Sl no | Algorithm | Testing accuracy |
|---|---|---|
| 1 | Logistic regression | 87.09% |
| 2 | K nearest neighbour | 70.96% |
| 3 | Random forest classifier | 90.32% |
| 4 | Support vector machine | 90.32% |
| 5 | Decision tree | 83.87% |

## VI. CONCLUSION

This heart disease prediction model with an accuracy of 90.32% will help people especially medical professionals to scale different scenarios. They will have a good understanding of a person's health and they can easily understand age related health risk and thus they can warn a patient beforehand. Patients on the other hand can also consult a doctor beforehand and go through checkup and thus can prevent the occurrence of any heart disease. Thus, this model helps to build trust and develops a sense of security among people.

## VII. REFERENCE

[1]. "Prediction of Heart Disease using Machine Learning Algorithms" Krishnan J Santhana and S Geetha ICIICT |Year :2019| Conference Paper | Publisher: IEEE.

[2]. "Prediction of Heart Disease using Machine Learning". Aditi Gavhane, Gouthami Kokkula, Isha Panday and Kailash Devadkar, Proceedings of the 2nd International conference on Electronics Communication and Aerospace Technology(ICECA) |Year :2018| Conference Paper | Publisher: IEEE.

[3]. "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" Senthil Kumar, Mohan Chandrasegar Thirumalai and Gautam Srivastva |Year :2019| Conference Paper | Publisher: IEEE.

[4]. "Prediction of Heart Disease using Machine Learning Algorithms: A Survey" International Journal on Recent and Innovation Trends in Computing and Communication vol. 5 no. 8, Himanshu Sharma and M A Rizvi |Year :2019| Conference Paper | Publisher: IEEE .

[5] "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools" International Journal of Scientific Research in Computer Science Engineering and Information Technology IJSRCSEIT , M. Nikhil Kumar K. V. S. Koushik and K. Deepak|Year :2019| Conference Paper | Publisher: IEEE.

[6] "Heart Diseases Prediction using Data Mining Techniques: A survey" Amandeep Kaur and Jyoti Arora International Journal of Advanced Research in Computer Science IJARCS |Year :2019| Conference Paper | Publisher: IEEE.

[7] "Application of Machine Learning in Diseases Prediction", Pahulpreet Singh Kohli and Shriya Arora ,4th International Conference on Computing Communication And Automation(ICCCA) . |Year :2018| Conference Paper | Publisher: IEEE.

[8] "Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm" M. Akhil B. L. Deekshatulu and P. Chandra ,Procedia Technology. vol. 10 pp. 85-94|Year :2013| Conference Paper | Publisher: IEEE.