Heart Disease Prediction Using Machine Learning

Sashank Yadav[1], Aman Singh[2], Veena Jadhav[3], Dr. Rohini Jadhav[4]
[1,2]*Student, Bharati Vidyapeeth (Deemed to be) University, College of Engineering, Pune*
[3]*Assistant Professor, Bharati Vidyapeeth (Deemed to be) University, College of Engineering, Pune*
[4]*Associate Professor, Bharati Vidyapeeth (Deemed to be) University, College of Engineering, Pune*

***ABSTRACT* -- Lately, predicting any cardiac disease is the most complex tasks in the health care sector. At present, nearby one person per minute dies of a coronary attack** [1]**. The untimely occurrence turns into life taking scenario. Data science as a domain has an important role in gathering insights from huge amounts of data in the health care sector as predicting any heart disease is a real complex task, there is a need to automate the prediction process in order to avoid risks associated with and alert the patient in-advance. This paper uses the heart disease dataset from the UCI Machine Learning repository** [2] [1]**. The work here predicts the possibility of heart disease by using 7 machine learning algorithms such as the Naive Bayes, Decision Tree, Logistic Regression, KNN (K-Nearest Neighbors), SVM (Support Vector Machine), Gradient Boosting and Random Forest algorithms** [3]**. Therefore, this paper brings up a comparison of performing measures between different machine learning algorithms used in the proposed work. The results acquired from the classification report confirms that the KNN (K-Nearest Neighbors) algorithm achieved a very high accuracy of 85.18% compared to other ML algorithms used. This Algorithmic model is then serialized into a byte stream as a pickle file(pkl) which is unpickled in the web application developed via Flask micro web framework. The application performs predictions over the user inputs via the HTML template and returns the prediction.**

Keywords -- Naive Bayes, Decision Tree, Logistic Regression, KNN (K-Nearest Neighbors), SVM (Support Vector Machine), Gradient Boosting, Random Forest [3]**, Algorithm, Machine Learning**

I. INTRODUCTION

The work proposed in this paper focuses on the various data mining practices used to predict heart disease. The human heart is one of the most important part of the human body pumping oxygen rich blood throughout the body. Any disorder to the normal functioning of the heart valves be classified as cardiopathy consisting of various types [1]. In today's stressful and unhealthy lifestyle, heart disease stands as one of the major and sudden causes of death. They are caused by many factors such as unhealthy lifestyle, smoking and alcohol consumption. Diets with high unhealthy fats causing high blood pressure, cholesterol, artery blockages leading to clots, etc. As per the World Health Organization(WHO), over sixteen million people die from coronary disease each year. A hale and hearty lifestyle and early detection are the only ways to prevent heart-related ailments [1].

Medical practitioners over the time make large sets of data of such patients which can be used [1] to analyse and extract important information from them. Data science techniques are a way to extract important and hidden information from a large amount of available data. Mostly the medical websites contain a variety of information regarding ailments. Therefore, making decisions using various data becomes hectic and difficult. Machine Learning (ML) which is a subset of Artificial Intelligence carries a large scale of well-designed algorithms. The main purpose of this paper is to provide a tool for physicians to diagnose heart disease as a first stage. This will help to provide effective treatment to patients and avoid side effects [1]. ML plays a very important role in discovering different hidden patterns and thus analysing the data provided. After data analysis ML techniques help assist in the diagnosis of heart disease and early diagnosis. This paper presents an analysis of the effectiveness of various ML algorithms and tries to pick the best fit performing model for the used dataset. [1]

II. LITERATURE OVERVIEW

Senthilkumar Mohan et al [4], has used a hybrid machine learning algorithm to predict heart disease. The dataset used is the Cleveland dataset. Columns of age and sex from the database are not used as the authors assume that the information is personal and does not affect the prediction. They developed their own Hybrid Random Forest Linear Method (HRFLM), a combination of both Random Forest (RF) and Linear method (LM) algorithms. The

authors move on to suggest that they have implemented a set of four algorithms. They concluded that the combination of these two algorithms produced better results than individual classifiers. Authors promotes continuous improvement in accuracy by using a combination of various machine learning algorithms.

Nagaraj M Lutimath, et al [5], worked on a prediction model for heart disease using the Naive bayes classification and Support Vector Machine. The performance measurements used in the analysis are Mean Absolute Error, Root Mean Squared Error and Sum of Squared Error [1]. They found that SVM emerged as a much better performer in terms of accuracy than Naive Bayes algorithm. They analysed the algorithms of Random Forest, Decision Tree, Naive Bayes and Logistic Regression classifiers on the basis of accuracy, precision, recall and f score. Then identified the best classification algorithm that can be used in predicting heart disease [1].

Yeshvendra K Singh, et al [6], worked with various machine learning algorithms such as Random Forest, SVM(Support Vector Machine), Linear Regression, Logistic Regression, Decision Tree with 3, 5 and 10 cross validation techniques. The authors used different splits, different tree numbers for each reference and a different number of cross validation fold. In the Random Forest, 85.81% accuracy is achieved by 20 splits, 75 trees and 10 folds.

Fahd Saleh Alotaibi [7] worked on proposing a ML model that compared five different algorithms. Rapid Miner tool has been used which has resulted in higher accuracy compared to the Matlab and Weka tool. In the study, accuracy of the Naive Bayes, Decision Tree, Logistic Regression, Random Forest and Support Vector Machine classifiers are made to compare. The Decision Tree algorithm showed the highest accuracy on the Rapid Miner tool.

III. PROPOSED MODEL

The proposed work predicts heart disease by examining the seven classifying algorithms mentioned above and develop a HTML template to provide interface to the user to test their conditions. Data is entered into a model that predicts the possibility of heart disease. Figure 1 shows the Generic model of the project.

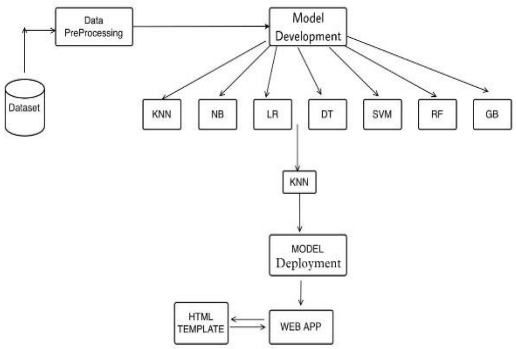


Fig 1: Generic model of Heart Disease Prediction

A. DATA GATHERING AND PREPROCESSING

The dataset used in this project was downloaded from the UCI Machine Learning Repository [5] for heart disease dataset. The dataset consists of 14 distinguishable features as opposed to the general 76 features given in the dataset description. The dataset was already seen to be filtered out with the common 14 features considered to be deterministic. The description and attributes of the following features are shown below in TABLE I. [1] [2]

TABLE I. DATASET FEATURES

Serial Number	Description	Attributes
1	Age	Any positive integer value
2	*Sex*- Gender (0-F, 1-M)	0,1
3	*CP*- Chest Pain severity	0,1,2,3
4	*RestBP*- Resting blood pressure	90-200.
5	*Chol*- Cholesterol levels	126-564.
6	*FBS*- Fasting blood sugar	0,1
7	*RestECG*- Resting Electrocardiography	0,1,2
8	*Thalach* - Max heart beat of patient	71-202
9	Exang- exercise induced angina. yes=1, no =0	0,1
10	*OldPeak*- Describes patient's. depression level	0 - 6.2
11	*Slope*- Patient condition in peak exertion.	1,2,3
12	*CA* – Fluoroscopy value	0,1,2,3
13	*Thal*- Severity of chest pain or trouble breathing. 4 types of values exists for this testing.	0,1,2,3
14	*Target* – This represents the final outcome of whether the patient is positive (2) or negative (1) for any heart disease condition based on the 13 parametrical features above.	1,2

B. DATA MODELLING AND CLASSIFICATION

The features mentioned in the table above are provided to the 7 Machine Learning algorithms used in this particular work such as Naive Bayes, Decision Tree, Logistic Regression, KNN (K-Nearest Neighbors), SVM (Support Vector Machine), Gradient Boosting and Random Forest algorithms [3]. The dataset is split into training and test data, with 30% stored as test data used for validation and performance testing, while training data is used to train the model for prediction

i. LOGISTIC REGRESSION:

Logistic Regression is a split algorithm that is widely used in binary split problems. In this algorithm, instead of a straight line or a high plane, the logistic regression algorithm uses a moving function to compress the output of the line number between 0 and 1.

ii. SUPPORT VECTOR MACHINE:

Support vector machines come in a variety of forms, linear and non-linear. The support vector machine divides data into categories. Common in this context, two different data sets are involved with SVM, training and a test set. In the ideal case, the classes are divided into sequences. In such a case a line can be found, which separates the two classes completely. Yet it is not just one line that separates the database well, but a lot of lines do it. Of these the best lines are selected as the "hyper plane".

iii. RANDOM FOREST:

Random Forest algorithm is a supervised algorithm. This process can be used both retrospectively separation functions but generally performs better in separation functions. It processes multiple decision trees before delivering output. In segregation, it uses a voting system and determines the category, and by regression, it takes a definition of all the results of each decision tree. Works well with large data sets with high magnitude.

iv. DECISION TREE:

A decision tree is a supervised learning method used for both planning and retrospective problems but is often preferred in solving planning problems. It is a tree-shaped divider, where the internal nodes represent the elements of the database, the branches represent the rules of decision and each leaf node represents the result.

v. K-NN (K-NEAREST NEIGHBORS):

The K-NN algorithm captures the similarities between data available and places a new case in the category that closely resembles available categories. This means that wherever the new data comes, it can be categorized into a section using the K-NN algorithm.

vi. NAIVE BAYES ALGORITHM:

Naive Bayes Classifier is one of the most straightforward and most efficient segmentation algorithms that help build faster machine learning models that can perform faster Predictions. Bayes theorem holds on to calculate the posterior probability of event(A) w.r.t to some prior probability event (B) as shown in equation 1: [1]

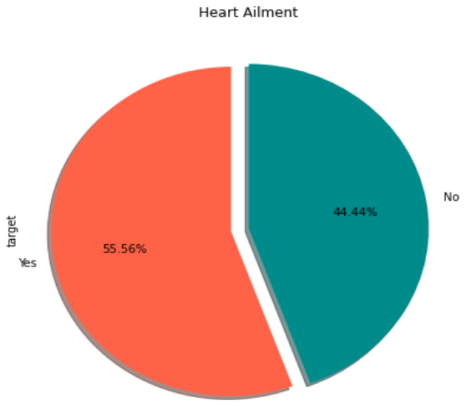
$P(A|B) = P(B|A)\,P(A) / P(B)$ (1) [1]

vii. GRADIENT BOOSTING:

Gradient Boosting is another learning technique used in classification and also regression. The set of rules basically predicts model based totally on the ensemble of weaker models and is said to carry out better than Random Forest [8]

IV. RESULTS AND ANALYSIS

The following dataset when visualised using pie chart(fig:2) shows to have 55.56% positive and 44.44% negative heart patients



In this section we also see the values obtained from the Confusion matrix providing the performance of each seven models used. The analysis of these algorithms are given by their metrics: Accuracy score, Precision (P), Recall (R) and F-measure [1] . These metrics are shown below in table 2 and 3 for comparison.

$$Precision = (TP)/(TP + FP)\ (2)\ [1]$$

$$Recall = (TP) / (TP + FN)\ (3)\ [1]$$

$$F\text{–}\ Measure = (2 * Precision * Recall) / (Precision + Recall)\ (4)\ [1]$$

Here, TP: True Positive , TN: True Negative, FP: False Positive, FN: False Negative [1]

TABLE II. CONFUSION MATRIX OF ALGORITHMS

Algorithm	True Positive	False Positive	False Negative	True Negative
Logistic Regression	38	4	10	29
KNN	38	4	8	31
Decision Tree	33	9	15	24
Random Forest	37	5	10	29
Support Vector Machine	36	6	9	30
Gradient Boosting	36	6	10	29
Naive Bayes Classifier	36	6	9	30

From table 2, KNN (K-Nearest Neighbors) possess the highest True Positive and True Negative values ideal for model deployment.

TABLE III. ANALYSIS OF MODEL PERFORMANCE

Algorithm	Precision (P)	Recall (R)	F-Measure	Accuracy
Logistic Regression	0.79	0.90	0.84	82.71%
KNN	0.83	0.90	0.86	85.18%
Decision Tree	0.69	0.79	0.73	70.37%
Random Forest	0.79	0.88	0.83	81.48%
Support Vector Machine	0.80	0.86	0.83	81.48%
Gradient Boosting	0.78	0.86	0.82	80.24%
Naive Bayes Classifier	0.80	0.86	0.83	81.48%

V. CONCLUSION

Heart related ailments has become one of the major reason for deaths around the world. The untriggered cardiac symptoms if not detected at the earliest turns into life threatening situation. Use of machine learning to predict heart disease is evolved over the time and it's thus expected to produce more accurate predictions based on the historical data. Although there have been many researches in the same domain, the need to improvise is always essential and suggested. The present project aims to improve data modelling with the inclusion of seven machine learning classifiers. [1] with a HTML template acting as an User Interface. Of the seven classifiers used (mentioned above in Table 3), K-Nearest Neighbors emerged out to produce highest accuracy of 85.18% and further this model is deployed into the web application for predictions. In future, work can be done to combine multiple weaker classifiers and ensemble them into a multi prediction model to improve their accuracy.

VI. ACKNOWLEDGEMENT

Firstly, we are thankful to *Bharati Vidyapeeth (deemed to be) University, College of Engineering, Pune* for the opportunity to work on this project. We would like to extend our thanks to Prof. Veena Jadhav, *Assistant Professor, Department of Computer Engineering* for continuous guidance. Also grateful to Dr. Sandeep Vanjale, *Head of Department, Department of Computer Engineering* for the support. We also like to thank our principal Dr. Vidula Sohoni for bestowing trust on us.

REFERENCE

[1] A. A., M. S., D. R. D. P. G. Apurb Rajdhan, "Heart Disease Prediction using Machine Learning," INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT), vol. 09, no. 4, pp. -, 2020.

[2] D. a. G. C. Dua, "UCI Machine Learning Repository: Statlog (Heart) Data Set," University of California, Irvine, School of Information and Computer Sciences, - - 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/statlog+(heart). [Accessed - May 2022].

[3] D. Varghese, "Comparative Study on Classic Machine learning Algorithms," Towardsdatascience.com, 6 December 2018. [Online]. Available: https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222. [Accessed - May 2022].

[4] C. T., G. S. Senthilkumar Mohan, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access 7, pp. 81542-81554, 2019.

[5] C. C. B. P. Nagaraj Lutimath, "Prediction of Heart Disease using Machine Learning," International Journal of Recent Technology and Engineering, vol. 8, no. 2S10, pp. 474-477, 2019.

[6] N. S. S. K. S. Yeshvendra K Singh, "Heart Disease Prediction System Using Random Forest," Advances in Computing and Data Sciences, vol. 721, no. -, pp. 613-623, 2017.

[7] F. S. Alotaibi, "Implementation of Machine Learning to Predict Heart Failure Disease," International Journal of Advanced Computer

Science and Applications (IJACSA), vol. 10, no. 6, p., 2019.

[8] Wikipedia, "Gradient boosting," Wikimedia Foundation, 2022. [Online]. Available: https://en.wikipedia.org/wiki/Gradient_boosting. [Accessed - May 2022].

[9] "What is precision, Recall, Accuracy and F1-score?," Nomidl, [Online]. Available: https://www.nomidl.com/machine-learning/what-is-precision-recall-accuracy-and-f1-score/. [Accessed May 2022].