JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Heart Disease Prediction Using Machine Learning

[1]Aadarsh Singh, [2]Rohini Bhosle

[1]Student at Sheth L.U.Jhaveri. & Sir M.V. College of Arts, Science and Commerce, [2]Asst. professor at Sheth L.U.Jhaveri. & Sir M.V. College of Arts, Science and Commerce

ABSTRACT: Every day, cases of heart disease rise faster and are very important, and we are concerned that such diseases are predicted in advance. This diagnosis is a difficult task. It must be performed accurately and efficiently. This project report focuses on the potential for heart disease by training six machine learning algorithms. Using data from the Kaggle website, you can analyze and compare logistic regression models, Naive Bayes, K-Nearest Neighbors, support vector machine, decision trees, and random forests. The most robust model that determines the key features of the model. Using a very useful approach, we adjusted the methods of improving the accuracy of heart attack prediction for each individual using the model. The strength of the proposed model is satisfactory and using decision trees and random forests that showed good accuracy compared to previously used classifiers such as Naive Bayes, to provide a specific individual. We were able to demonstrate that heart disease was used.

IndexTerms - Naive bayes, K-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine, Machine Learning and Logistic Regression.

I. INTRODUCTION

Heart disease is a common concept that involves many types of hearts problem. It is also known as cardiovascular disease Heart and vascular disease. Heart disease is leadership The cause of death for large population groups, but there is an opportunity for that It prevents and treats many types of heart disease. There is Many different factors that can develop more Heart disease. Some of these factors are age, family history, and Genetics, lifestyle, illness, etc. Medical diagnosis It's a difficult and long process to do things that are perfect and effective. You need a kind of automated system that can be analyzed The parameters provide a complete and accurate result. Doctors use their own experience and knowledge Diagnosis of disease, but can be analyzed for missing parameter. This absence can be covered using different things technique. This paper focuses on the implementation of 6 Various algorithms Naive bayes, K- Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine and Logistic Regression Identify heart disease with heart-related parameters. Important parameters are used and the algorithm is compared Accuracy conditions.

II. LITERATURE REVIEW

There has been a lot of research being done in medical centers on disease prediction systems using various data mining techniques and machine learning algorithms.

"Purushottam et al, [1]" proposed an efficient prediction system for heart disease using data mining. This system helps doctors to make effective decisions based on specific parameters. By testing and training specific parameters, an accuracy of 86.3% is achieved in the testing phase and 87.3% in the training phase.

“Sumit Sharma, Mahesh Parmar et al, [2]” In this paper applied some classification algorithms (like – K-NN, SVM, Hyper-parameter optimization) on Heart diseases data set. By testing and training the Hyper-parameter optimization (Talos) has a better accuract then the other algorithms with 90.78% accuracy.

“Prof. Dr. R. Sandhiya, A. Shivani, J. Shalini, K. Tejasvi et al, [3]” This project explores Data mining techniques and the algorithms used to present this project are KNN, Naive Bayes, Decision Tree, Random Forest, and other algorithms are used to design the method based on logistic regression. Random Forest produces better outcomes and supports domain experts and even medical professionals in planning for a better and earlier diagnosis for patients.

“Animesh Hazra, Arkomita Mukherjee, Amit Gupta, Asmita Mukherjee et al, [4]” explores the use of algorithms such as Naive bayes, Decision list and k-NN with accuracy of Naive Bayes 52.33% Decision list 52% and k-NN 45.67%. The Decision list shows a better accuracy among the other algorithms.

“S. Seema et al. [5]” focus on techniques that can predict chronic diseases by evaluating data contained in past health records using Naive Bayes, Decision Trees, Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs). A comparative study of

classifiers is conducted to measure better performance in terms of accuracy rate. From this experiment, it is found that SVM provides the highest accuracy rate while Naive Bayes provides the highest accuracy rate in the case of diabetes.

c

III. REQUIREMENTS AND ANALYSIS

3.1 Data Source

Kaggle, a well-known platform for data science and machine learning, serves as a valuable source for datasets across various domains. The dataset used in this research was obtained from Kaggle, ensuring accessibility, reliability, and a structured format suitable for analysis it has total 1026 records. Kaggle hosts a diverse collection of datasets contributed by researchers, organizations, and data enthusiasts, often accompanied by metadata and documentation that enhance usability.

3.2 Dataset Features

FEATURES	DESCRIPTION
Age	Age of the patient
Sex	1: Male, 0: Female
CP (Chest Pain Type)	1: Typical angina, 2: Atypical angina, 3: Non-anginal pain, 4: Asymptomatic
Trestbps	Resting blood pressure (mm Hg)
Chol	Serum cholesterol (mg/dl)
FBS	Fasting blood sugar > 120 mg/dl (1 = True, 0 = False)
Restecg	Resting electrocardiographic results (0,1,2)
Thalach	Maximum heart rate achieved
Exang	Exercise-induced angina (1 = Yes, 0 = No)
Oldpeak	ST depression induced by exercise relative to rest
Slope	Slope of the peak exercise ST segment
CA	Number of major vessels (0-3) colored by fluoroscopy
Thal	3 = Normal, 6 = Fixed defect, 7 = Reversible defect

Table 1: Dataset Feature Description

3.3 Data Preprocessing

• Data Cleaning: Implementing methods to handle missing values, remove duplicates, and correct inconsistencies.

• Data Transformation: Normalize or standardize numerical features and encode categorical variables using techniques like one-hot encoding or label encoding.

• Feature Selection: Identifying and selecting relevant features using methods such as correlation analysis, recursive feature elimination, or machine learning-based feature importance.

3.4 Model Selection

Training various machine learning models to determine the most effective one for predicting heart disease. Candidates include: Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Naive Baye's.

3.5 Training Validation

Splitting the data into two part that is the testing data and training the data. The training data value is of 820 data and the testing data value is 205 data.

IV. METHODOLOGY

4.1 Logistic Regression

Logistic Regression is Often used for binary classification (e.g., disease present or not), logistic regression is a simple and interpretable model that performs well when there is a linear relationship between features.

4.2 Support Vector Machines

Effective for both linear and non-linear classification tasks. SVMs can create complex boundaries between classes to enhance prediction accuracy. SVM performs well with data where the number of features (dimensions) is large compared to the number of samples.

4.3 Decision Trees

Decision trees provide clear interpretability by creating a tree- like model of decisions. Decision trees closely resemble human decision-making processes. The structure is intuitive and easy to visualize. Decision trees can handle both types of data effectively without the need for complex data transformations.

4.4 Random Forests

Random forests, a collection of decision trees, improve accuracy by reducing overfitting. By combining the predictions of multiple decision trees, random forests often achieve higher predictive accuracy than individual decision trees. Random forests are much less prone to overfitting than decision trees.

4.5 k-Nearest Neighbors

A non-parametric algorithm that classifies a data point based on the majority label of its nearest neighbors, useful in cases where relationships between data points are not complex. k- NN is easy to understand and implement. It requires minimal mathematical knowledge to apply, making it one of the most accessible machine learning algorithms.

4.6 Naive Baye's

Naive Bayes is a popular machine learning algorithm based on the Bayesian theorem. It is particularly useful for classification tasks where the assumption of conditional independence between features holds. Despite its simplicity, Naive Bayes is effective in many real-world applications, such as spam filtering, sentiment analysis, and disease prediction.

4.7 Flowchart of predicting heart disease

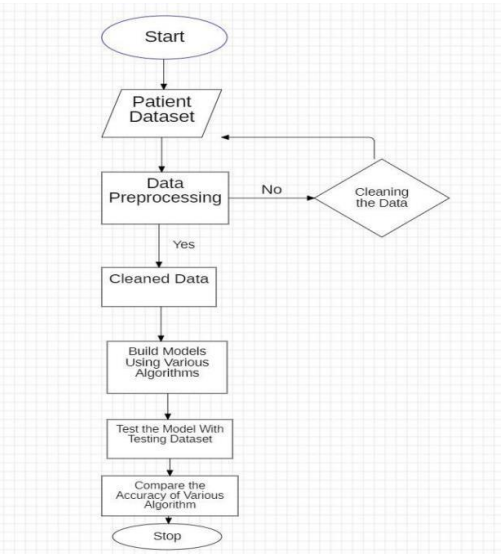

Fig 1: Flowchart of predicting heart disease

This flowchart outlines the process of handling a patient dataset for model development. It begins with collecting the dataset, followed by data preprocessing. If the data is not clean, it undergoes a cleaning process before returning to preprocessing. Once the data is cleaned, it is used to build models using various algorithms. These models are then tested with a testing dataset to evaluate their

performance. Finally, the accuracy of different algorithms is compared to determine the most effective model before the process concludes.

V. EXPERIMENTAL RESULTS AND ANALYSIS

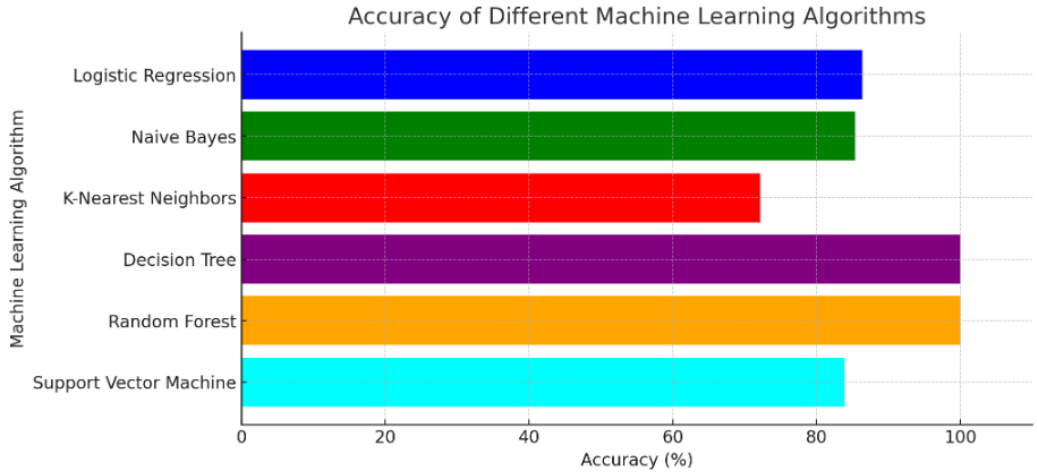

Fig 2: Comparison Among Six Algorithm

The proposed work is implemented in Python 3.6.4 with libraries scikit-learn, pandas, matplotlib and other mandatory libraries. Fig 2 shows that the all the algorithms are performing good except k-NN on heart disease dataset. The accuracy of Logistic Regression and Navie Bayes are almost same. Decision Tree and Random forest has the best accuracy among all the other algorithm.

VI. CONCLUSION

In this research, we explored the application of machine learning techniques for heart disease prediction. By preprocessing patient datasets, cleaning the data, and applying various machine learning algorithms such as Naive bayes, K- Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine and Logistic Regression, we were able to develop predictive models that can assist in early diagnosis. The performance of these models was evaluated using a testing dataset, and the accuracy of different algorithms was compared. The results indicate that machine learning has significant potential in improving heart disease prediction, enabling early intervention and better patient outcomes. Future research can focus on optimizing feature selection, integrating deep learning techniques, and utilizing larger datasets to enhance model performance further. With continued advancements in AI- driven healthcare, machine learning can play a crucial role in reducing the global burden of heart disease.

REFERENCES

[1] Purushottam, Prof. (Dr.) Kanak Saxena, Richa Sharma, “Efficient Heart Disease Prediction System”, 2016

[2] Sumit Sharma, Mahesh Parmar,” Heart Diseases Prediction using Deep Learning Neural Network Model”,2020

[3] Prof. Dr. R. Sandhiya , A. Shivani , J. Shalini , K. Tejasvi, “Heart Disease Prediction using Machine Learning”, 2021

[4] Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee, “Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review”, 2017

[5] Seema Shedole Kumari Deepika, “Predictive analytics to prevent and control chronic diseases”, 2016