Open Access Research Journal of
Engineering and Technology
Journals home page: https://oarjpublication/journals/oarjet/
ISSN: 2783-0128 (Online)

(RESEARCH ARTICLE)

Check for updates

Heart disease prediction using machine learning techniques

Amit Jain [1], Suresh Babu Dongala [2] and Aruna Kama [3,*]

[1] *Department of Computer Science & Engineering, SunRise University, Alwar, Rajasthan, India.*
[2] *Department of Computer Science, SR & BGNR Government College, Khammam, Telangana, India.*
[3] *Department of Computer Science, CKM A&S College, Warangal, Telangana, India.*

Open Access Research Journal of Engineering and Technology, 2022, 03(01), 001–006

Publication history: Received on 07 June 2022; revised on 18 July 2022; accepted on 20 July 2022

Article DOI: https://doi.org/10.53022/oarjet.2022.3.1.0028

Abstract

Heart diseases are commonly caused and when neglected becomes life threatening. So, early detection of the disease is very important and for diagnosis to save lives. There can be many parameters that are to be considered to predict the heart disease. Some of them are like age, cholesterol, blood pressure levels. Etc., here we are going to implement Machine Learning model to predict heart disease.

Keywords: Heart Disease Detection; Naïve Bayes; Decision Tree; Random Forest; K-Nearest Neighbour (KNN)

1. Introduction

We can see that from past ten years the major death cause for humans was occurring because of heart disease. Heart diseases are life threatening and may sometimes cause to death. Heart attack can occur because of either narrow or blocked vessels. Some of the blood vessel diseases are life coronary artery disease and heart rhythm problems. Personal and professional practices, as well as genetic susceptibility are hazard factors for heart disease. To avert death, the well planned, precise and premature medical detection of heart disease plays a deciding part in proceeding benefits. Machine learning help in the exposure and diagnosis of distinct diseases. Several Machine Learning algorithms such as Naïve Bayes, K-Nearest Neighbor, Decision Tree and Random Forest are correlated to find the most precise model. Supervised Learning concerns training on a labelled dataset using techniques to achieve exact awareness applying dependent and independent variable. In this project the algorithm is given with input variables and actual output obtained then algorithm compares between the actual and predicted output to identify errors and modifies the model precisely. The heart disease database is from the UCI repository.

2. Framework

The heart disease patient's data is gathered from UCI repository. This data is availed to discover the disease using machine learning algorithms. The algorithms performance and accuracy outcomes are correlated. In this they mainly focused on Random Forest algorithm. This was proposed in the year 2020 [1]. The discovering pattern is done with Naive Bayes, K-Nearest Neighbor, Decision Tree, Neural Networks, and Genetic Algorithm for dataset. The outcomes are contrasted for execution and precision and the calculations. Electrocardiogram is used to analyze heart cycles with many beginning points [7]. For making prediction of heart diseases in a simple and efficient approach we have used advanced methods. With the help of machine learning and deep learning we can perform various experimentation methodologies which are used in this study [8]. Dimensionality reduction by using two methods Feature Extraction and Feature Selection is proposed by Ramalingam VV et al in the year 2020. Large number of features or attributes can direct to overfitting which results in poor output [4].

*Corresponding author: K.Aruna
Department of Computer Science , CKM A&S College, Warangal, Telangana, India

Copyright © 2022 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

3. Data statistics

3.1. Dataset

The dataset that is used is Cleveland database of UCI repository, collection of heart disease patients. It consists of 303 instances of 14 attributes. The target variable is 0 or 1 showing the prediction of heart disease.

Table 1 Dataset 1

Attribute	Description	Range
age	age in years	For all
Sex	gender	1 means male 0 means female
cp	Type of chest pain	0,1,2,3
chol	Serum cholestrol	200-239 mg/dL
fbs	blood sugar content before eating	1 > 120mg/dL 0 < 120mg/dL
trest bps	blood pressure when we are at rest	120
restecg	Resting electrocadiographic result	0-Normal 1-Abnormal
thalach	achieved maximum heart rate	71-202
exang	exercise included angina	0-not caused 1-angina caused
old peak	Depression of ST segment in ECG	1-3
slope	Peak Exercise ST segment slope	0-flat 1-unsloping 2-downsloping
ca	Count of utmost blood vessels which are colored by fluoroscopy.	0-3
thal	Thalassemia	0-3 normal >7-defective
target	Class attribute	0,1

The chest pain is categorized into four types: 0 means typical angina, 1 means atypical angina, 2 means non-anginal pain, asymplomatic-3. Range of normal cholesterol is from 200-239mg/dL. Fasting blood sugar >120mg/dL is marked 1 and <120mg/dL is marked 0. Resting blood pressure of 120 is considered normal. Normal Resting electrocardiographic result is marked as 0 and abnormal is marked as 1. Angina caused during exercise is represented as 1 and 0 if angina is not caused. Old peak is the depression of ST segment in ECG. Flat slope is represented as 0, upsloping is represented as 1, down sloping is represented as 2. No of major blood vessels colored (ca) by fluoroscopy consists of the values (0-3). Thalassemia (thal) levels of (0-3) is considered normal and above 7 is detective.

4. Methodology

4.1. Data Pre-processing

For performing data pre-processing first, we check type of data. Two types of data: Continuous data, Categorical data.Continuous Data For continuous data removal of outliers and standard scaling is performed.

4.1.1. Removal of outliers

Inter Quartile Range

It is used to detect outliers and remove outliers from dataset. Firstly, the dataset is sorted in ascending order and then it is divided into four different quartiles. Q1, Q2 and Q3 are three quartiles that divide the entire dataset into four equal parts. Now the we try to find the interquartile range by the following formula.

$$IQR = Q3-Q1$$

The formula for the lower-limit and upper-limit are

$$Lower\_limit= Q1-1.5*IQR$$

$$Upper\_limit= Q3+1.5*IQR$$

All the data beneath the lower-limit and over the upper-limit are considered as outliers.

4.1.2. Standard Scaling

Inter Quartile Range

The data values of features are of different scales i.e., they are different units. So, we need to perform standardization of data. Standardization is performed by statistical distribution, where we try to equate mean to 0 and standard deviation to 1, for this we apply the transformation function.

$$z = \frac{x - \mu}{\sigma}$$

Categorical Data

For categorical data correcting wrong data and converting categorical to dummy/indicator variables are done. In correction of data all the wrong values of features i.e., all the values of features are out of range will be replaced with NaN. Next categorical data is converted to dummy values. Dummy variables are used to indicate whether or not a category exists in a categorical variable. Next data preprocessing steps performed are filling missing values with median and deleting duplicate rows.

4.2. Naive Bayes Algorithm

It is supported Bayes theorem of conditional probability. Bayes theorem finds the probability of an occurrence occurring as long as other has already occurred.

$$\mathbf{P(y \mid X)} = \frac{\boldsymbol{P(X \mid y)P(y)}}{\boldsymbol{P(X)}}$$

Here we use Gaussian Naïve Bayes as there's continuous data. The probability density function is given by:

$$P(x_i) = \frac{1}{\sqrt{2\Pi\sigma_y^2}} exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

4.3. Decision Tre

A decision tree could be a sort of supervised machine learning algorithm to categorize or make prediction supported how previous set of questions were answered. By using Decision tree algorithm, we first attempt to create a training model by generating a decision rule. For the development of decision tree two terms are involved. They're entropy and information gain.Entropy gives the measure of purity of split. Entropy values ranges from 0 to 1. Formula for entropy is

$$E(s) = -p_{(+)}\, log\, p_{(+)} - p_{(-)}\, log\, p_{(-)}$$

Less is that the entropy value pure and is that the split.Next term is information gain, information gain is employed to check the entropy values before the split and after the split. Information gain are often calculated. The more is information gain best is that the split. On the idea of entropy and information gain we attempt to determine best split. Information gain and entropy are interdependent on one another.Construction of decision tree is completed supported some steps.

- Starting from the root node, prefer the most effective split at each node keep up the foremost information gain.
- Greedy search is the next step in which we emphasize all features and thresholds.
- At every single node conserve the most efficient split feature and split threshold.
- Frame the tree repeatedly.
- To prevent increase in size of tree assign stopping criteria.
- Reserve the foremost common class label of the leaf node.

4.4. Random Forest

Random forest, a collation of decision trees. Every tree evaluates a classification, and it's called as "vote". Appealingly, we acknowledge every vote from every tree and adopt the utmost voted classification (Majority-Voting).Random Forests produce multiple unique trees.

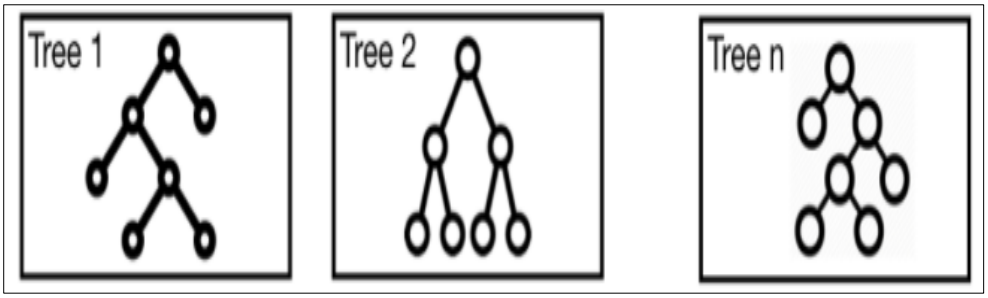

Figure 1 Random Forest Classifer

4.5. K-Nearest Neighbour

The KNN algorithm a supervised learning method, in which classification of objects pivots on closest neighbor. It's a form of memory-based learning. With the help of Euclidian distance, we can find the distance between attribute and its neighbor.

Euclidean distance among two instances

$$d\,(xi, xj)=\sqrt{\sum_{i=1}^{n}\left(x_i - x_j\right)^2}$$

5. Results and discussion

Table 2 Accuracies of ML algorithms

Algorithm	Accuracy
Naïve Bayes	82.45614035087719%
K-Nearest Neighbor	77.19298245614034%
Decision Tree	84.21052631578947%
Random Forest	78.94736842105263%

In the survey research work, Analysis of algorithms is finished based on the Clevel and data sets. The algorithm used in the project for the final model is Random Forest.

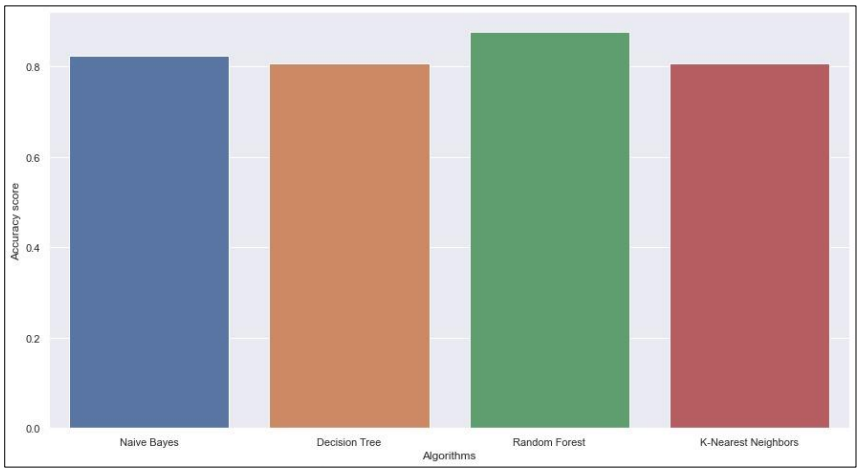

Figure 2 Accuracies Graph of Machine learning algorithms

6. Conclusion

A heart disease prediction model has been created utilizing three ML order demonstrating procedures. This project foresee person with cardiovascular infection by extricating the outpatient clinical records that arouse a lethal heart illness from a dataset that consolidate patients' clinical records, as an example, chest pain, sugar level, cholesterol, etc. The algorithms utilized in this model are machine learning, naive bayes algorithm, decision tree, random forest, KNN etc. By using more training data sets from decision tree, we are able predict the upper chances of whether a patient is stricken by a heart disease or not. The accuracy of our model is 0.9907. At last our ultimate conclusion is that with the help of KNN algorithm we can know the patients who are suffering from heart diseases. From the restrictions of our study, we can say that we need the complicated combined models for attaining enhanced accuracy so that the heart disease is detected earlie.

Compliance with ethical standards

Acknowledgments

This paper is supported by many people, some of whom have a direct role and some of them have an indirect way by publishing their research online which helped to understand this concept easily. We express our deepest gratitude, sincere thanks and deep feeling of appreciation to our Project Guide Prof. Dr. Amit Jain, his presence at any time throughout the Semester, important guidance, opinion, comment, critics, encouragement, and support greatly improved this project work. We are obliged to the college administration for providing the necessary infrastructure and technical support. Finally, we extend our heartfelt gratitude to our fellows and family members.

Statement of conflict of interest

No conflict of interest.

References

[1] Heart Disease Prediction Using Machine Learning Techniques: Galla Siva Sai Bindhika Munaga Meghana, Manchuri Sathwika Reddy, Rajalakshmi.

[2] Heart Disease Prediction using Machine Learning Techniques: Pooja Anbuselvan

[3] Bouali H, Akaichi J. Comparative study of different classification techniques: heart disease use case. In: 2014 13th international conference on machine learning and applications. IEEE. p. 482-86.

[4] Ramalingam VV, Dandapath A, Raju MK heart disease prediction using machine learning techniques a survey. Int J Eng Technol. 2018;7(2.8):684–7.

[5] A. S. Abdullah and R. R. Rajalaxmi, ‘‘A data mining model for predicting the coronary heart disease using random forest classifier,’’ in Proc. Int. Conf. Recent Trends Computation Methods, Commun. Controls, Apr. 2012, pp. 22–25.

[6] H. A. Esfahani and M. Ghazanfari, ‘‘cardiovascular disease detection using a new ensemble classifier,’’ in Proc. IEEE 4th Int. Conf. Knowl. Based Eng. Innov. (KBEI), Dec. 2017, pp. 1011–1014.

[7] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, ‘‘Association rule mining to detect factors which contribute to heart disease in males and females,’’ Expert Syst. Appl., vol. 40, no. 4, pp. 1086–1093, 2013. doi: 10.1016/j.eswa.2012.08.028

[8] D. K. Ravish, K. J. Shanthi, N. R. Shenoy, and S. Nisargh, ‘‘Heart function monitoring, prediction and prevention of heart attacks: Using artificial neural networks,’’ in Proc. Int. Conf. Contemp. Comput. Inform. (IC3I), Nov. 2014, pp. 1–6

[9] H. A. Esfahani and M. Ghazanfari, ‘‘cardiovascular disease detection using a new ensemble classifier,’’ in Proc. IEEE 4th Int. Conf. Knowl. Based Eng. Innov. (KBEI), Dec. 2017, pp. 1011–1014.

[10] Sowjanya, K., & Krishna Mohan, G. (2020). Predicting heart disease using machine learning classification algorithms and along with tpot (Automl). International Journal of Scientific and Technology Research, 9(4), 3202–3210.

[11] Karthick. D & Priyadharshini. B (2018). Predicting the chances of occurrence of CardioVascular Disease (CVD) in people using classification techniques within fifty years of age. Proceedings of the 2nd International Conference on Inventive Systems and Control, ICISC 2018, Icisc, 1182-1186.

[12] Hassoon, M., Kouhi, M. S., Zomorodi-Moghadam, M., & Abdar, M. (2017). Using PSO Algorithm for Producing Best Rules in Diagnosis of Heart Disease. 2017 International Conference on Computer and Applications, ICCA 2017, 306–311.