International Journal of Engineering and Applied Physics (IJEAP)
Vol. 1, No. 2, May 2021, pp. 179~185
ISSN: 2737-8071

Heart Disease Prediction System

Shwetank Mishra[1], Satya Vijay Neurkar[2], Riddhesh Patil[3], and Sharmila Petkar[4]
[1,2,3,4]RamraoAdik Institute of Technology, Nerul, Navi Mumbai, 400706
India

Article Info

Article history:

Received Sep 9, 2020
Revised Mar 20, 2021
Accepted Mar 30, 2021

Keywords:

Artificial Neural Network
Data analysis
Flask
Machine Learning
Random Forest

ABSTRACT

It might have happened many times that you need doctors to facilitate instantly, however, they are not always available, or sometimes it's all about the formalities before check-in thanks for some reason. This project is based on the Web Application for Online Consultancy for people all around the world. This Web App allows us to consult ourselves while sitting at our home. Here we propose a system that enables users to urge instant direction on their health problems through an associate intelligent health care system. This project comprises different classification techniques used for predicting the risk level of each person based on Age, Gender, Blood pressure, Cholesterol, Pulse rate, etc. The system analyses the symptoms provided by the user as input and predicts the occurrence of the disease as an output. Disease Prediction is done by implementing six algorithm techniques such as KNN, Decision Tree, Logistic Regression and Random Forest, SVM, Artificial Neural Network with one hidden layer. This project also provides an intuition on EDA. Further, the web platform is composed of predicting the health risk of the user, providing the users with necessary suggestions depending on their health conditions.

Corresponding Author:

Shwetank Mishra,
RamraoAdik Institute of Technology,
Nerul, Navi Mumbai, 400706
India,
Email: shwetankm141@gmail.com

1. INTRODUCTION

Machine learning at present has come up as one of the best applications of Artificial Intelligence that uses various advanced algorithms and analytics approaches to learn from the data provided and predict the best possible result. It helps to get a smart model which can even be downloaded on our disk. That job is performed by 'pickle' that provides a serialized format. Talking about the project we specifically choose heart disease because it has one of the most dangerous diseases in the world which leads to serious trouble if one does not take precautions in advance. The health of the heart depends on the personal and professional conduct of a person. The cause may be genetic also. Every year around twelve million deaths happen due to the different types of heart diseases and this list includes people around 20-30 years. There are several factors such as lack of sleep, dejection, bad diet, smoking, high BP, and cholesterol. The identification within the time boundary is very important. The most common signs are chest pain, breathlessness, and heart palpitations. Some types of heart disease can lead to a heart attack. Heartburn, gastric upset, and heavy chest are some of the symptoms of a heart attack. Whereas some heart conditions have no signs at all they are generally found in older adults and diabetes patients. In this respect, instant diagnosis becomes a must that.

2. MOTIVATION

Heart Disease has become the number one cause of death within the last two decades. It's estimated that one person dies every thirty-five seconds. Cardiovascular diseases are heart and blood vessel disorders it comprises rheumatic HD, is chaemic HD, etc. 95% of deaths are due to heart attack and stroke. (World Health Organization, 2018) Around 10 million people died due to cardiovascular diseases. Diagnosis has become a complex piece of work in the sanatoriums. Since 1990, average annual mortality of 128.9% per 100,000 people in India, recorded 5.6% per year cardiovascular diseases. This motivated us to put forward a Heart disease Web-app using machine learning and deep learning algorithms. This prediction system can detain problems by performing detailed analysis and accurate prediction. ML has now become one of the best applications of AI because of the various algorithms that predict output based on learned data or analysed data. This application is not only helpful to patients but also the doctors who could see their patients report virtually and then accordingly provide their suggestions.

3. PROBLEM DEFINATION

Many research papers applied ML concepts for prediction over a large dataset using particular ML techniques but they are not the improved ones. Although we come across some optimized ones such as Particle Swarm Optimization. For disconnected attributes, recruitment attributes of mining are selected. We perform feature selection to remove immaterial data and aim to decrease dimensionality hence providing better and precise results. Classification techniques will be used for getting the best model. We will minimize the loss error using hyper parameter tuning and one of the methods to do that is exhaustive search i.e. searching for the correct match. So basically, our aim to get the best algorithm for the given dataset and use that model to deploy on the web using the Flask framework. We understand the features of our dataset categorical and continuous parameters and perform EDA. Understanding the correlation between the features i.e. positive or negative correlation. We analyse the features and then step forward for the data model, we allot thirteen features to x and the output to y resulting in training and test set which is split in the particular ratio and finally normalization. After we get the model, we have to design a web frontend using HTML and CSS where users will be able to check for their predicted outcomes. The user's data will be stored in the database. Users need to first sign up then fill in their login details. We used php myadmin for the database.

4. SYSTEM REQUIREMENTS

The web prediction system is very adaptable with a fundamental interface. It is very systematic with good UI. The UI is very straightforward and easy to acknowledge and use.

A. Hardware Requirements

- Processor: i3/i5/i7 8gen or above
- Processor Speed: 2.16GHZ or above
- RAM: 4 GB RAM or above
- Storage: 20 GB hard disk or above

B. Software Requirements

- Python 3.8 or above
- Anaconda
- Visual Studio Code
- XAMPP server

5. METHODOLOGY

We used the UCI ML repository dataset for our implementation. There are a total of 14 columns where the 14th column represents the output in the form of 1's and 0's. Here '1' refers to the person suffering from heart disease and vice versa.

Description of the dataset can be seen below-

1) age: The person's age in years.
2) sex: The person's sex (1 = Male, 0 = Female)
3) cp: The chest pain experienced (Value 0: Typical Angina, Value 1: Atypical Angina, Value 2: Non-Anginal pain, Value 3: Asymptomatic)
4) trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)
5) chol: Person's cholesterol measurement in mg/dl.
6) fbs: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)

7) restecg: Resting electrocardiographic measurement (0 = Normal, 1 = Having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
8) thalach: Maximum heart rate achieved
9) exang: Exercise-Induced angina (1 = yes; 0 = no)
10) oldpeak: ST depression induced by exercise relative to rest, displays the value which is integer or float.
11) slope: Number of major vessels (0-4) colored by flourosopy : displays the value as integer or float.
12) ca: Peak exercise ST segment;1 = upsloping, 2 = flat, 3 = downsloping
13) thal: displays the thalassemia; 1 = normal, 2 = fixed defect, 3 = reversable defect
14) target: Heart disease (0 = no, 1 = yes)

TABLE I DATASET

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

A. Exploratory Data Analysis

We check for any missing values or any repetition in the dataset. It's better to remove any particular if its presence does not influence the result. Further, we go for sampling to enlarge the production of the algorithm. Our dataset didn't have any missing value although we had shown how to deal with it when it's the case. When your data has any missing value it's advisable to swap it with the mean value in case there is no outlier and replace it with median if there is an outlier. And if we have any categorical column, we must swap it with mode value. If there are any repetitive values in the dataset, we drop that row from the dataset. We must drop the outlier value if any present in the dataset because it affects the mean value.

Outlier value refers to that value that is far from the collection of values that are nearby in the dataset. Then another feature is Bivariate Analysis which means analysing two variables. We get to know about the relationship between the two data. Our dataset consists of 3 types of variables and is Continuous, Categorical, and Ordinal variables. On the other hand, bivariate analysis in our dataset can be done in 3 ways-

- Numerical vs Numerical - Scatter, Line, Heatmap for correlation.
- Categorical vs Numerical - Bar chart, Violin plot.
- Two Categorical variables - Bar chart.

Another important point is Feature scaling although we did not use this because our dataset values were not so varying from each other. Thus, to systematize the ranges we perform scaling. For categorical data for example in our dataset chest pain has 4 categories but we don't have to worry because each category has been assigned a numerical value in the dataset so that it gets a unique value in the nominal feature column. This method is called one-hot encoding.

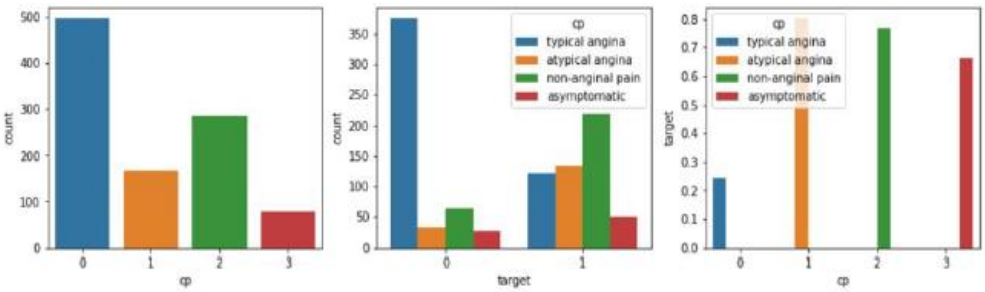

FIGURE 1. Categorical feature

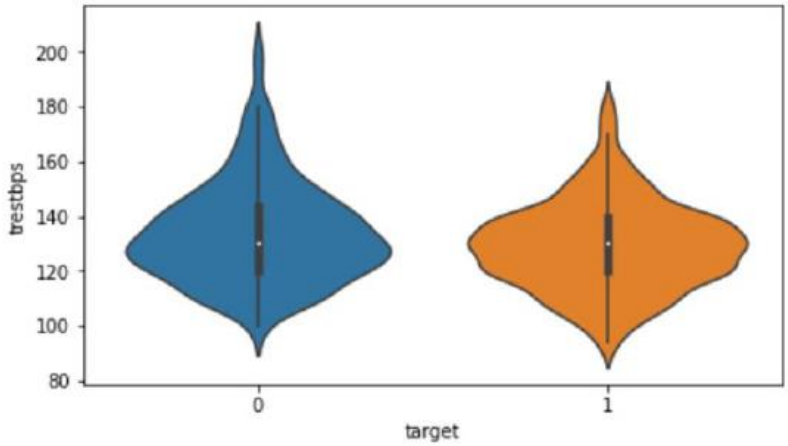

FIGURE 2. Continuous feature (Violin plot)

B. Preparing the model

The ruling aim of our project is to identify the best algorithm that gives the highest accuracy with this dataset and then deploy that on a web that is accessible to any user across the world. We used six (classification and regression) algorithms and used the best accurate model for further development. After the EDA, we split our dataset in training and testing set in 80% and 20% respectively.

6. IMPLEMENTATION

A. Logistic Regression:

It is one of the most common ML classification algorithms. A supervised learning algorithm that calculates the result when the dependent variable is categorical. Using the sigmoid function, it plots any number into the 0 and 1 value range. And that 1 and 0 in our project work represent persons suffering from heart disease and vice-versa. The figure shown below.

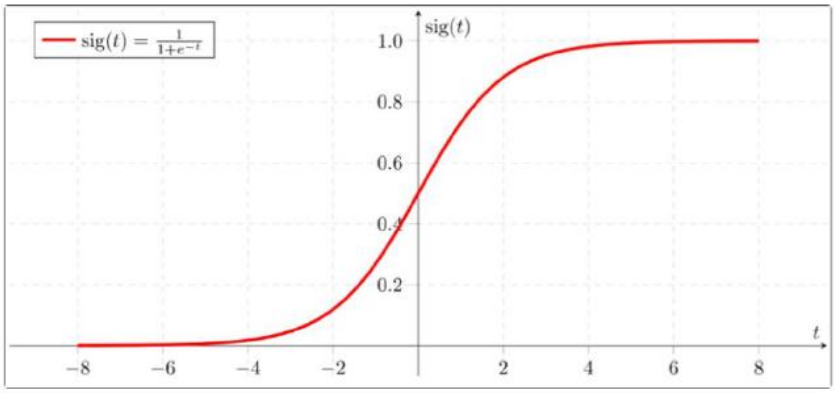

FIGURE 3. Sigmoid function

B. Decision Forest:

This algorithm is used mostly for classification problems. Tree in the structure where the leaf node represents the result and interior nodes represent parameters of the dataset. It uses the Classification and Regression algorithm to build a tree. The algorithm starts from the root node and starts comparing values with the dataset values and then further makes a branch. From set to subsets, it generates a decision tree node. After a certain stage, no classification is possible and then we get a leaf node.

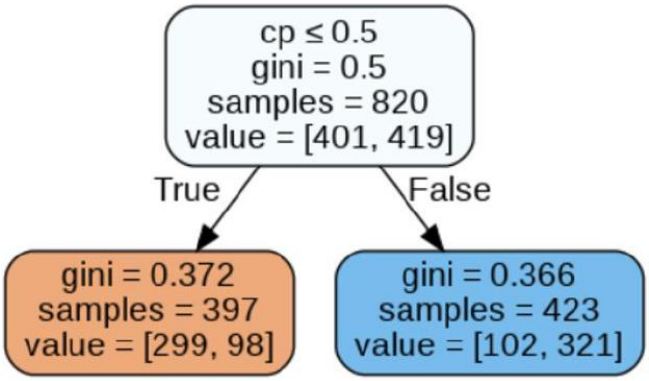

Figure 4. Decision tree with max_depth=1

C. Random Forest:

It is a classification as well as regression supervised learning algorithm. It utilizes the idea of ensemble learning which uses the procedure of merging many classifiers to solve complex problems and bring out the enhanced performance of the model. It uses many decision trees on different subsets of the dataset and performs its average to bring an overall good accuracy. This algorithm takes less training time as compared to others. It maintains accuracy even if a large portion of data is missing. The work can be described by selecting arbitrary X data points from the training dataset then set up a decision tree with respect to selected points. Choose no P for the decision tree that you desire to form. The over fitting problem does not occur with this algorithm. Starting with the data pre-processing step we continue with the fitting stage. We fit the RF algorithm to the training set using Random Forest Classifier. There are two parameters first is n estimators which refers total no of trees in a random forest. Another parameter is random state which allows controlling random alternatives. We got the best possible precision and recall from this algorithm which is one. And therefore, the F Score is also one.

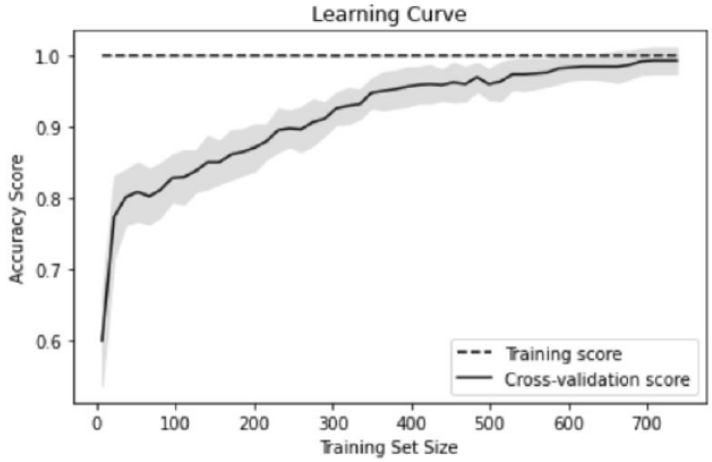

Figure 5. Learning curve of Random forest

D. K-Nearest Neighbors:

This algorithm classifies new data based on affinity. It's a very slow algorithm because it stores the dataset during the training period and then classifies the new data according to the stored data based on the similarity between them. Firstly, we decide the K number then calculate the Euclidean distance of the k number of neighbors. We assign new data to that particular category for which neighbors number is maximum. Generally, we prefer the value for K equal to 5. KNN is good for fluctuated values if it's there in the dataset and is produced when the training data is large.

E. SVM:

Support Vector Machine aims to distinguish two categories by creating a margin between them. That margin or line is known as a hyperplane. SVM chooses extreme points to create hyperplane and that extreme cases are called support vectors. We need to create a hyperplane that classifies the categories in the best way. Distance between the hyperplane and the data points (support vectors) should be maximum. There

are two types Linear and Non- Linear SVM. So for some cases, we need to add a dimension to segregate the categories.

F. ANN:

Our last algorithm is Artificial Neural Network with one hidden layer. It is used in the same way the human brain examines and operates information. It contains 3 layers Input, Hidden, and Output layer. Input layers can take numbers, words, images, and then in the hidden layer mathematical computation is performed. Other than this there are various hyper parameters that we use in the algorithm that affect the performance of the model. There is some weight for every node. Transfer function calculates the weighted sum and then finally activation function gives the output. Activation function like Sigmoid, ReLU, tanh is used for better learning. To reduce the loss error, we perform Back propagation. We used the Sigmoid and ReLU activation function and Adam optimizer. Optimizers are used for solving optimization problems by reducing losses.

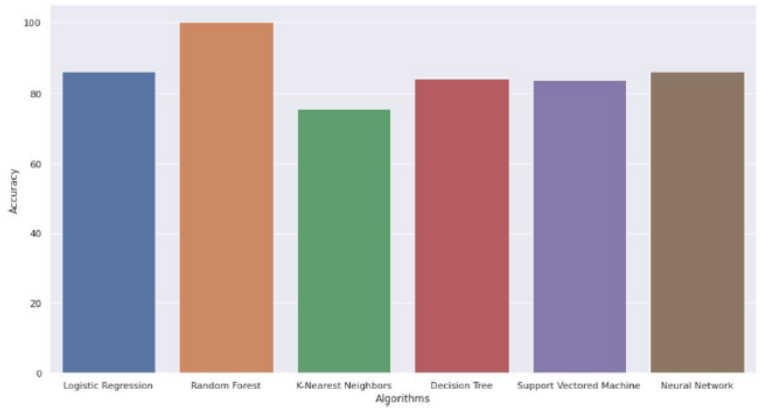

Figure 6. Comparison between algorithms accuracy

To implement our project for end-users we made a web app using flask. Firstly, we need a pickle file and since Random Forest gives the best accuracy, we used that algorithm to create a pickle file. We can download the pickle file model on our system and can use it in the flask code. We used HTML, CSS, Bootstrap, MySQL, php myadmin from the home page to all connected HTML pages. We used XAMPP local web server that allows storing the user's credentials in the database and that credentials will be available to the user also on the home page in case if he forgets his credentials. The website starts with a Login page if there is a new user then he/she can register by entering his information details.

Pop up box when a new user is registered further he/she can use that credentials to log in to the home page. Later on the home page, we have all the parameters required to be filled by the end-user and after entering all the 13 parameters the user gets the prediction stating that 'you are suffering / not suffering from heart disease'. We thought of adding some feedback or a report based on his/her prediction like to have exercise or have control on the cholesterol level, etc. And there is another feature added on the predicted page like if someone needs the report page printed then that can be done.

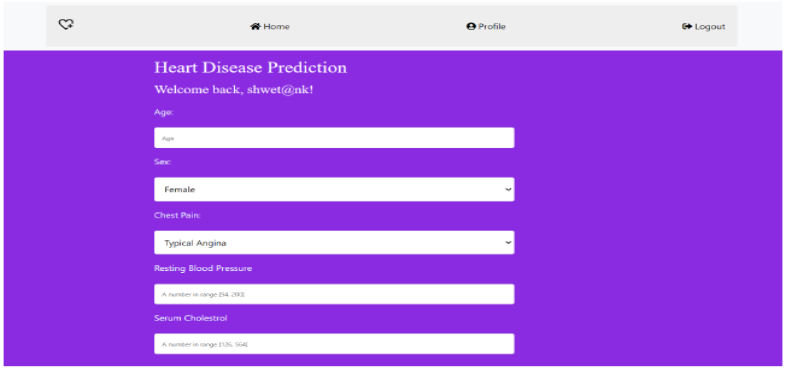

Figure 7. Home Page in Web Application

7. CONCLUSION

The overall objective of our project is to predict the presence of heart disease accurately and provide the user a good platform with smart features at any point in time. In this project, thirteen attributes are considered which form the primary basis for tests and give accurate results. Random forest gives the best accuracy. This system performs practically well even without reskilling. This project could respond to uncertainties with respect to the ease of the model explication. And with 100% accuracy, prediction is 100% practicable. The data repository should be used in hospitals to get more values in the dataset for the best prediction. We got the best accuracy from the random forest algorithm and used the pickled model in the Flask framework and deployed it on a local server using XAMPP. We used NumPy, Matplotlib, Sk-learn, and many more libraries in ML whereas in the web part we used Flask, Flask- Ext, pymysql, re, and secret types of libraries. For future attempts, we are planning to identify the particular heart disease in the output and also cover different types of disease prediction on the same platform.

REFERENCES

[1] M. K. Awang and F. Siraj, "Utilization of an articial neural network in the prediction of heart disease," Int. J. Bio-Science Bio-Technology, vol. 5, no. 4, pp. 159-165, 2013.
[2] P. Selvakumar and S. P. Rajagopalan, "A survey on neural network models for heart disease prediction," J. Theor. Appl. Inf. Technol., vol. 67, no. 2, pp. 485-497, 2014.
[3] A. Methaila, P. Kansal, H. Arya, and P. Kumar, "Early Heart Disease Prediction Using Data Mining Techniques," vol. 6956, no. October, pp. 53-59, 2014.
[4] I. S. F. Dessai, "Intelligent Heart Disease Prediction System Using Probabilistic Neural Network," no. 5, pp. 38-44, 2013.
[5] T. Karayilan and O. Kilic, "Prediction of Heart disease using neural network," in 2nd International Conference on Computer Science and Engineering, UBMK 2017, 2017, pp. 719-723.
[6] M. Mardiyono, R. Suryanita, and A. Adnan, Intelligent Monitoring System on Prediction of Building Damage Index using Neural-Network," TELKOMNIKA (Telecommunication Comput. Electron. Control., vol. 10, no. 1, p. 155, 2015.
[7] N. Guru, A. Dahiya, and N. Rajpal, "Decision support system for heart disease diagnosis using Neural Network," 2007.
[8] Kim, Jae Kwon, and Sanggil Kang. "Neural networkbased coronary heart disease risk prediction using feature correlation analysis." Journal of healthcare engineering 2017 (2017).
[9] MA.Jabbar, B.L Deekshatulu, Priti Chandra, "An evolutionary algorithm for heart disease prediction" CCIS, PP 378- 389 , Springer(2012).
[10] https://www.kaggle.com/ronitf/heart-disease-uci
[11] Jagdeep Singh, Amit Kamra, Harbhag Singh "Prediction of Heart Diseases Using Associative Classification" 2016.
[12] scikit-learn, keras, pandas and matplotlib