Heart disease prediction using machine learning techniques

Article *in* IOP Conference Series Materials Science and Engineering · January 2021

DOI: 10.1088/1757-899X/1022/1/012046

3 authors, including:

Rizwan Khan
ABES Institute of Technology
90 PUBLICATIONS **252** CITATIONS

SEE PROFILE

Heart disease prediction using machine learning techniques

Heart disease prediction using machine learning techniques

Apurv Garg[1], Bhartendu Sharma[2] and Rijwan Khan[3]

[1,2]Scholar, ABES Institute of Technology, Ghaziabad, Uttar Pradesh – 201009, India

[3]Professor, Department of Computer Science and Engineering, ABES Institute of Technology, Ghaziabad, Uttar Pradesh – 201009, India

E-mail: 110.apurvgarg@gmail.com, bhartendu351.cs@gmail.com, rijwankhan786@gmail.com

Abstract. Machine Learning (ML), which is one of the most prominent applications of Artificial Intelligence, is doing wonders in the research field of study. In this paper machine learning is used in detecting if a person has a heart disease or not. A lot of people suffer from cardiovascular diseases (CVDs), which even cost people their lives all around the world. Machine learning can be used to detect whether a person is suffering from a cardiovascular disease by considering certain attributes like chest pain, cholesterol level, age of the person and some other attributes. Classification algorithms based on supervised learning which is a type of machine learning can make diagnoses of cardiovascular diseases easy. Algorithms like K-Nearest Neighbor (KNN), Random Forest are used to classify people who have a heart disease from people who do not. Two supervised machine learning algorithms are used in this paper which are, K-Nearest Neighbor (K-NN) and Random Forest. The prediction accuracy obtained by K-Nearest Neighbor (K-NN) is 86.885% and the prediction accuracy obtained by Random Forest algorithm is 81.967%.

1. Keywords
Heart Disease; Machine Learning; K Nearest Neighbor (K-NN); Random Forest

2. Introduction
Human body is made up of various organs, all of which have their own functions. Heart is one such organ which pumps blood throughout the body and if it does not do so, the human body can have fatal circumstances. One of the main reasons of mortality today is having a heart disease [1]. So, it becomes necessary to make sure that our cardiovascular system or any other system in the human body for that matter must remain healthy. Unfortunately, people all around the world have been facing cardiovascular diseases. Any technology that can help diagnose these diseases before much damage is done will prove as helpful in saving people's money and more importantly their lives. Data mining techniques can be useful in predicting heart diseases. Predictive models can be made by finding previously unknown patterns and trends in databases and using the obtained information [2]. Data mining means to extract knowledge from large amounts of data [3]. Machine learning is a technology which can help to achieve diagnosis of heart disease

before much damage happens to a person. As an emerging field in science and technology, machine learning can classify whether a person might be suffering from a heart disease or not.

3. Literature Review

Research has been done in this field and people have produced methods to predict cardiovascular disease using supervised machine learning algorithms. Several research papers have been written on this topic. A survey has been presented in the form of a paper which analyzes performance of various models based on machine learning algorithms and techniques [4]. In one of the papers, work has been done to create a Graphical User Interface (GUI) to predict whether a person is suffering from heart disease or not, using Weighted Association rule based Classifier [5]. In another paper, a new approach has been presented which is based on coactive neuro-fuzzy interference system (CANFIS) for the prediction of heart disease [6]. A summary of commonly used techniques for heart disease prediction and their complexities is given in one of the papers [7]. One of the papers presented a classifier approach for heart disease detection and shows how Naive Bayes can be used for classification purpose [8]. In one of the papers, a survey is done which includes different papers in which one or more algorithms of data mining have been used for heart disease prediction [9].

4. Proposed Methods

4.1. K-Nearest Neighbor (K-NN)

In K-NN algorithm a data point is taken whose classification is not available, then the number of neighbors, k is defined. After that k neighbors are selected according to the lowest Euclidian distance between the selected data points and their neighbors. The selected data point is then classified into a category, which is same as the category which has majority of neighbors among the K neighbors.

4.2. Random Forest

Random Forest works by constructing multiple decision trees of the training data. each of the trees predicts a class as an output and the class, which is the output of the greatest number of decision trees is taken as the final result, in case of classification. In this algorithm we need to define the number of trees we want to create. Random Forest is a bootstrap aggregating or bagging technique. This technique is used to decrease the variance in the results.

5. Experimental Setup

The first step for the setup is to obtain the data set containing the features of a person suffering from a heart disease and a person who is not along with the result, that whether the person is suffering from the disease or not. The data set used in this experiment is taken from a website called Kaggle (https://www.kaggle.com/ronitf/heart-disease-uci). The programming language used to do the experiment is Python. Thirteen attributes are used which are available in the data set. The information of the attributes is available on Kaggle.

The next step is to analyze the data. For this, the information of the data set is required. To gather the concise summary of the DataFrame, the info() function is used on the data set which is provided by the Pandas library.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trestbps  303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalach   303 non-null    int64
 8   exang     303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64
 11  ca        303 non-null    int64
 12  thal      303 non-null    int64
 13  target    303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Figure 1. Concise summary of the DataFrame

The describe() function provided by the Pandas library is used to retrieve some statistical information of the data set like mean of the values of the attributes used. An attribute named target is taken whose value is 1, if the patient is suffering from a heart disease, or 0, if the person is not suffering from any heart disease. Now the data set is to be checked that is it balanced or not. This is done using countplot, which is provided by the Seaborn library, on the target attribute.

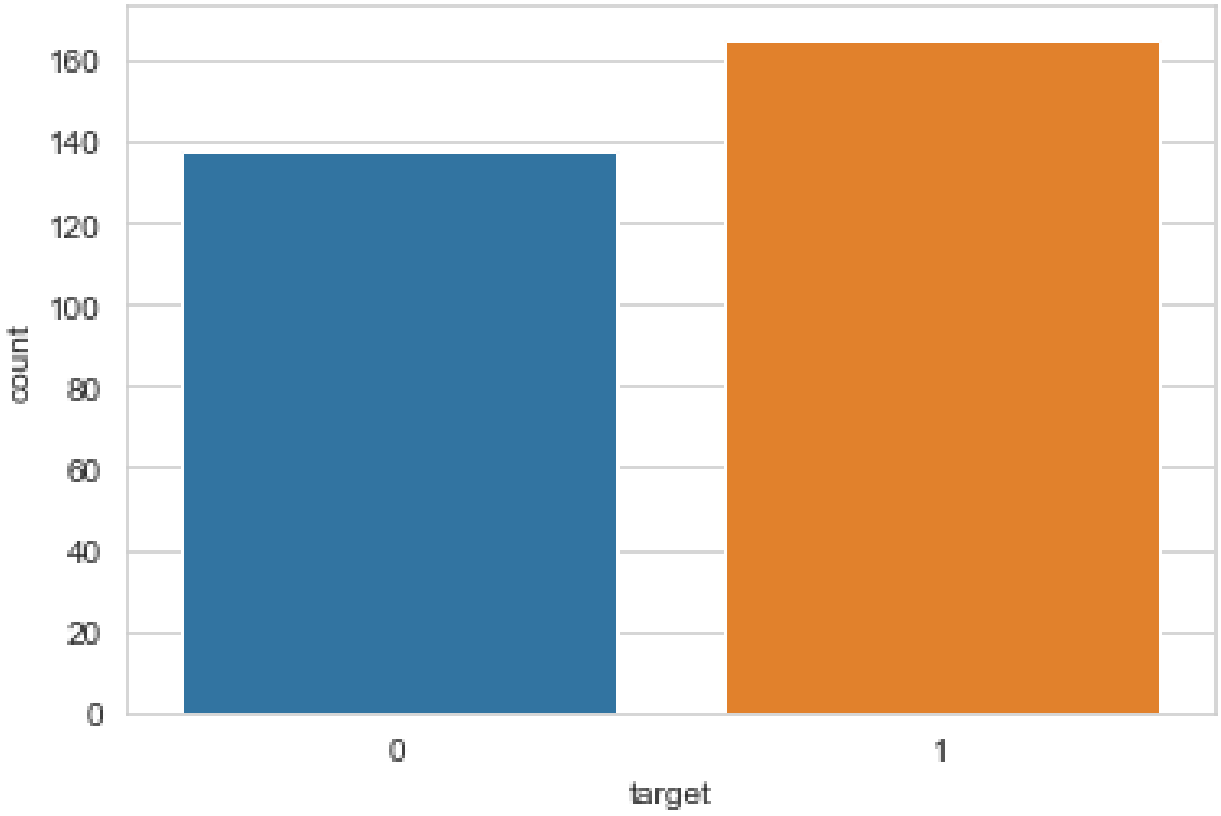

Figure 2. Count plot of target attribute

After looking at the plot, it can be concluded that the data is quite balanced. We can also use countplot with different attributes of the data set like the sex attribute which has values 1 (male) and 0 (female) and the cp (chest pain) attribute which shows the type of chest pain ranging from 0 to 3.

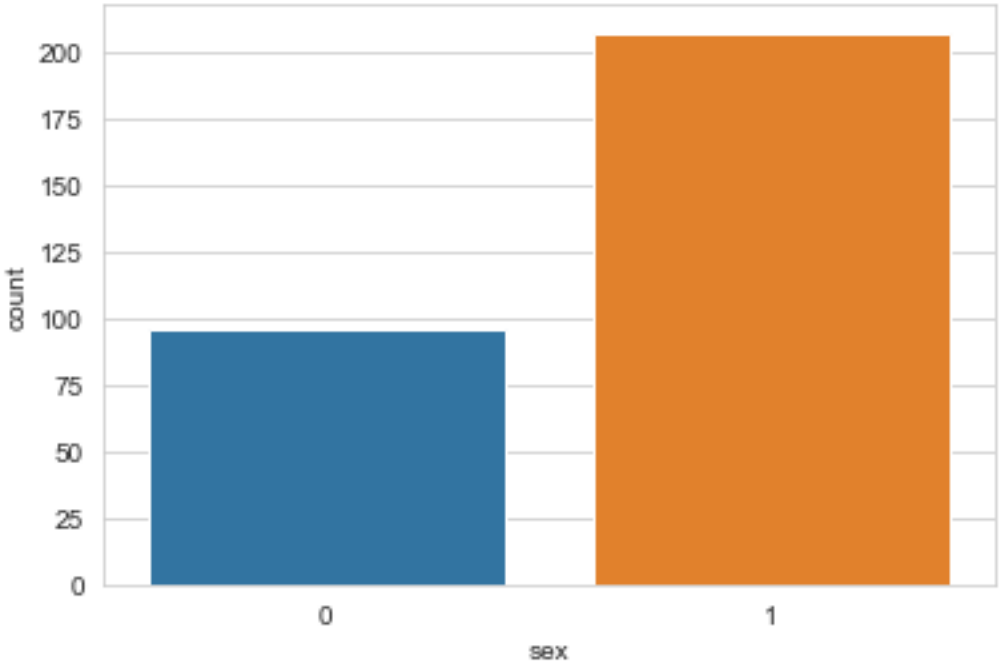

Figure 3. Count plot of sex attribute

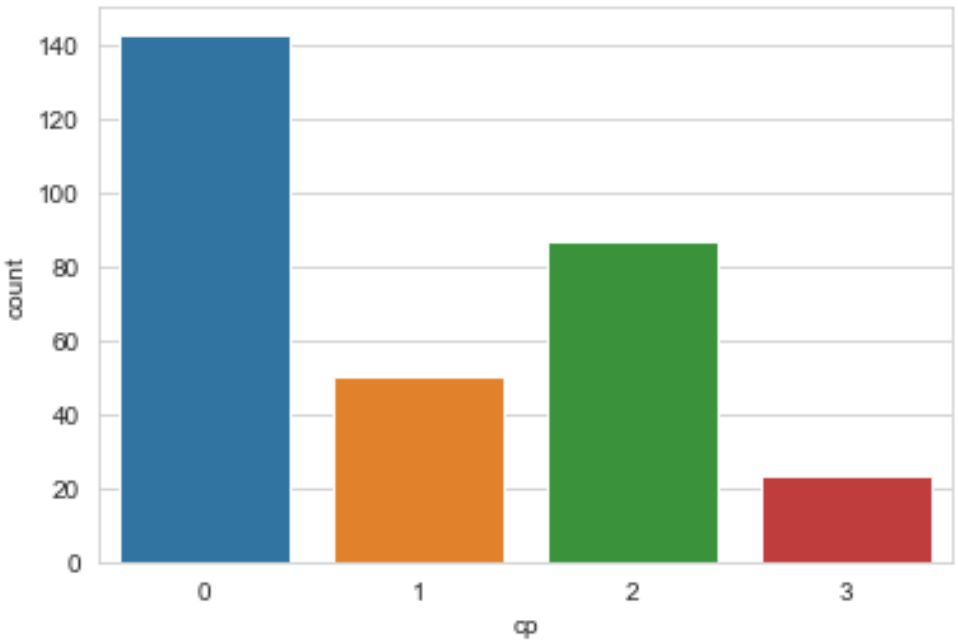

Figure 4. Count plot of cp attribute

After checking that the data is balanced, the correlation between the data is found out and is plotted as a heat map using the Seaborn library.

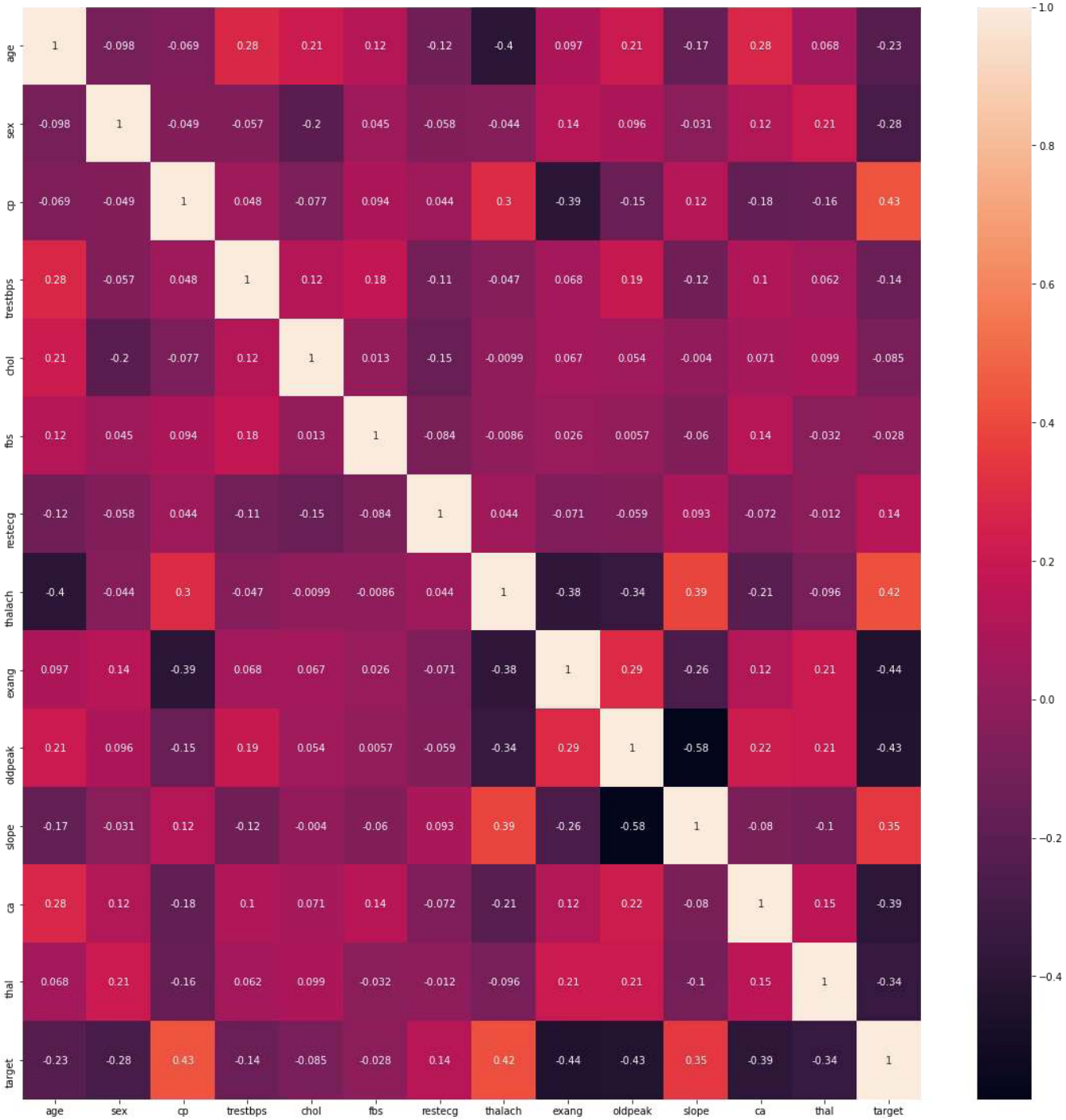

Figure 5. Correlation between variables

The heat map clearly shows that the attributes like cp (chest pain) and thalack (maximum heart rate achieved) have positive correlation with the target attribute. Now that the correlation has been checked, we need to convert categorical variables like sex, cp, fbs, restecg, exang, slope, ca and thal into dummy variables. This can be done by using get_dummies method of the Pandas library. After creating dummy variables, the data in columns like age, trestbps, chol, thalach and oldpeak needs to be standard scaled, because they have much varied quantities and units. This can be done using Scikit-learn library in Python.

The data set has been divided into two parts, training data which is 80% of the whole data set and testing data which is 20% of the whole data set. After preparing the data, the algorithms are applied and the confusion matrix has been found out. The results have been found out in term of accuracy of the algorithm. The accuracy has been found out with the use of a confusion matrix.

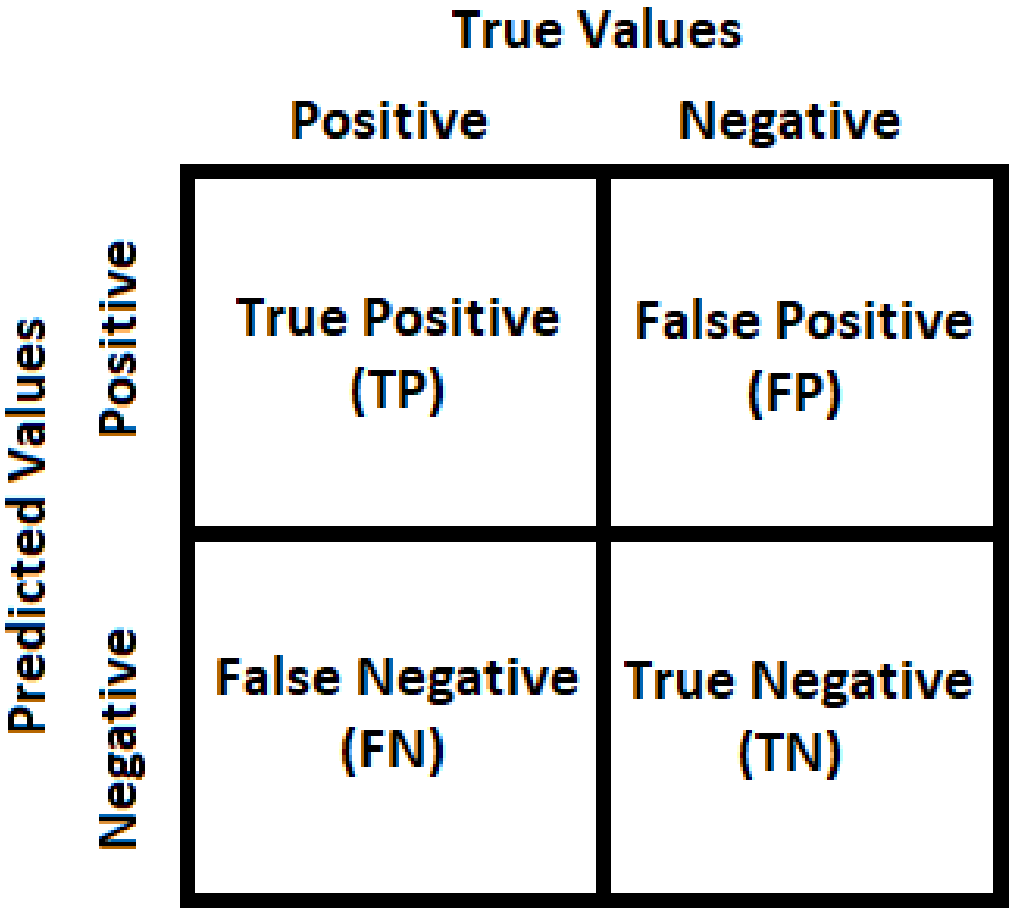

Figure 6. Confusion matrix layout

Confusion matrix can also be shown as a matrix in the following way:

[[TP FP

FN TN]]

The accuracy of the algorithm can be calculated using the formula:

$$\text{Accuracy} = \{(TP + TN) / TP + FP + TN + FN)\} * 100$$

6. Results

After applying the algorithms, the results obtained are as follows:

6.1. K-Nearest Neighbor (K-NN)

The value of k is taken as 12, as 12 was one of the values which gave the highest accuracy of the algorithm. The confusion matrix obtained was as follows:

[[23 4]

[4 30]]

From the confusion matrix, the accuracy is calculated which comes out to be 86.885%.

6.2. Random Forest

The value of number of trees is kept 10. The confusion matrix obtained was as follows:

[[22 5]

[6 28]]

From the confusion matrix, the accuracy is calculated which comes out to be 81.967%.

Table 1. Results after applying each algorithm

Algorithm Used	TP	FP	TN	FN	Accuracy
K-NN	23	4	30	4	86.885%
Random Forest	22	5	28	6	81.967%

7. Conclusion

After applying various algorithms, it can be said that machine learning is proving to be extremely valuable in predicting heart disease which is one of the most prominent problems of the society in today's world. As more and more work is being done in the field of machine learning, soon there may be new methods to make machine learning more helpful in the field of healthcare. The algorithms used in this experiment have performed really well using the available attributes. The conclusion can be finally drawn that machine learning is able to reduce the damage done to a person physically and mentally, by predicting heart disease.

8. References

[1] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542-81554.
[2] Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering*, *1*(8), 1-4.
[3] Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. *Heart Disease*, *7*(1), 129-137.
[4] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, *7*(2.8), 684-687.
[5] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Intelligent and effective heart disease prediction system using weighted associative classifiers. *International Journal on Computer Science and Engineering*, *3*(6), 2385-2392.
[6] Parthiban, L., & Subramanian, R. (2008). Intelligent heart disease prediction system using CANFIS and genetic algorithm. *International Journal of Biological, Biomedical and Medical Sciences*, *3*(3).
[7] Chitra, R., & Seenivasagam, V. (2013). Review of heart disease prediction system using data mining and hybrid intelligent techniques. *ICTACT journal on soft computing*, *3*(04), 605-09.
[8] Medhekar, D. S., Bote, M. P., & Deshmukh, S. D. (2013). Heart disease prediction system using naive Bayes. *Int. J. Enhanced Res. Sci. Technol. Eng*, *2*(3).
[9] Kaur, B., & Singh, W. (2014). Review on heart disease prediction system using data mining techniques. *International journal on recent and innovation trends in computing and communication*, *2*(10), 3003-3008.