As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 7(12): 109-113, 2021

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0712019
Available online at: http://www.ijmtst.com/vol7issue12.html

Heart Disease Detector Using ML

Nishu Singh | Nakul Singh

Department of Information Technology, MAIT, New Delhi, India

To Cite this Article

Nishu Singh and Nakul Singh. Heart Disease Detector Using ML. *International Journal for Modern Trends in Science and Technology* 2021, *7* pp. 109-113. https://doi.org/10.46501/IJMTST0712019

Article Info

Received: 28 October 2021; Accepted: 04 December 2021; Published: 07 December 2021

ABSTRACT

Machine Learning (ML), one of the most well-known applications of Artificial Intelligence, is revolutionising the field of research. In this work, machine learning is utilised to determine whether or not a person has cardiac disease. Many people suffer from cardiovascular diseases (CVDs), which claim the lives of people all over the world. On a monthly basis, a large amount of patient-related data is maintained. The information gathered can be used to predict the occurrence of future diseases. Machine learning may be used to determine whether a person has a cardiovascular illness based on particular characteristics such as chest discomfort, cholesterol levels, gender, and other factors. Logistic Regression (86.3% accuracy), Naïve Bayes (86.3% accuracy), Random Forest (93.6% accuracy), and K-Nearest Neighbour (87.8% accuracy) are some of the machine learning algorithms that we are using to predict cardiac disease.

KEYWORDS: Heart disease, Machine learning, cardiovascular diseases.

I.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 for 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. The techniques and algorithms can be directly used on a dataset for creating some models or to draw vital conclusions, and inferences from the dataset. Common attributes used for heart disease are Age, Sex, Fasting Blood Pressure, Chest Pain type, Resting ECG(test that measures the electrical activity of the heart), Number of major vessels colored by fluoroscopy, Threst Blood Pressure (high blood pressure), Serum Cholestrol (determine the risk for developing heart disease), Thalach (maximum heart rate achieved), ST depression (finding on an electrocardiogram, trace in the ST segment is abnormally low below the baseline), painloc (chest pain location (substernal=1, otherwise=0)), Fasting blood

sugar, Exang (exercise included angina), smoke, Hypertension, Food habits, weight, height and obesity[4].

STRUCTURE OF PAPER

The paper is organized as follows: In Section 1, the introduction of the paper is provided along with the structure, important terms, objectives and overall description. In Section 2 we discuss related work. In Section 3 we have the complete information about image processing tools. Section 4 shares information about the flexible YAML templating system created for it, its advantages and disadvantages. Section 5 tells us about the methodology and the process description. Section 6 tells us about the future scope and concludes the paper with acknowledgement and references.

OBJECTIVES

The predominant invoice processing systems are either entirely manual or they follow a rigid single template system. Whether an individual is a buyer or a seller, this leads to a lot of inefficiencies and high costs.

This project aims to address some of the problems in current systems by greatly minimizing the human intervention in the process and thus reducing costs and errors. The aim is to ease the task of both the buyer and the seller.

II. LITERATURE REVIEW

Using the UCI Machine Learning dataset, a lot of research has been done. Using different machine algorithms we got different levels of accuracy which are explained as follows.

Avinash Golande and others studied various different ML algorithms that can be used for classification of heart disease. Research was carried out to study Decision Tree, KNN and K-Means algorithms that can be used for classification and their accuracy were compared[5]. This research concludes that accuracy obtained by Decision Tree was highest, further it was inferred that it can be made efficient by combination of different techniques and parameter tuning.

Fahd Saleh Alotaibi has designed a ML model comparing five different algorithms [6]. Rapid Miner tool was used which resulted in higher accuracy compared to Matlab and Weka tool. In this research the accuracy of Decision Tree, Logistic Regression, Random forest, Naive Bayes and SVM classification algorithms were compared. Decision tree algorithm had the highest accuracy.

Anjan Nikhil Repaka, ea tl., proposed a system in [7] that uses NB (Naïve Bayesian) techniques for classification of dataset and AES (Advanced Encryption Standard) algorithm for secure data transfer for prediction of disease.

Theresa Princy. R, et al, executed a survey including different classification algorithms used for predicting heart disease. The classification techniques used were Naive Bayes, KNN (KNearest Neighbour), Decision tree, Neural network and accuracy of the classifiers was analyzed for different numbers of attributes [8].

Nagaraj M Lutimath, et al., has performed the heart disease prediction using Naive bayes classification and SVM

(Support Vector Machine). The performance measures used in analysis are Mean Absolute Error, Sum of Squared Error and Root Mean Squared Error, it is established that SVM emerged as superior algorithm in terms of accuracy over Naive Bayes [9].

III. PROPOSED MODELS

Logistic Regression

Logistic Regression is a classification algorithm mostly used for binary classification problems. In logistic regression instead of fitting a straight line or hyper plane, the logistic regression algorithm uses the logistic function to squeeze the output of a linear equation between 0 and 1. There are 13 independent variables which makes logistic regression good for classification.

Naïve Bayes

Naïve Bayes algorithm is based on the Bayes rule[]. The independence between the attributes of the dataset is the main assumption and the most important in

making a classification. It is easy and fast to predict and holds best when the assumption of independence holds. Bayes' theorem calculates the posterior probability of an event (A) given some prior probability of event B represented by P(A/B) as shown in the following equation

$$P(A|B) = (P(B|A)P(A)) / P(B)$$

Random Forest

Random Forest algorithms are used for classification as well as regression. It creates a tree for the data and makes prediction based on that. Random Forest algorithm can be used on large datasets and can produce the same result even when large sets record values are missing. The generated samples from the decision tree can be saved so that it can be used on other data. In random forest there are two stages, firstly create a random forest then make a prediction using a random forest classifier created in the first stage.

K-Nearest Neighbour

In the 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 the same as the category which has the majority of neighbors among the K neighbors.

IV. DATASET COLLECTION AND PROCESSING

The dataset used in this experiment is the heart disease dataset which is the combination of 4 databases. Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease. The dataset is already preprocessed and is available on the kaggle website.

V. EXPERIMENT SETUP

The first step was to get the dataset which contains the attributes of a patient and a result whether a patient has a heart disease or not. The dataset used in this experiment is taken from kagge website (https://www.kaggle.com/johnsmith88/heart-disease-dataset). Python programming language used in this research along with its libraries namely, pandas, numpy, seaborn, matplotlib, scikit learn etc.

The next step is to process the data and check for inconsistent data but since this dataset is already processed therefore this step is not needed
The next step is to convert the raw data into the form of tables and rows, and using the pandas library we can convert the raw dataset into the form of tables and rows as shown in figure 1

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure 1. Raw data converted into tables and rows

After that the next step is to check for any inconsistency in our dataset, by using pandas dataframe.describe() method we can check for any inconsistent data in it, but as we can see in figure 2 that there is no empty value in any row, so we are good to go

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1025 non-null   int64
 1   sex       1025 non-null   int64
 2   cp        1025 non-null   int64
 3   trestbps  1025 non-null   int64
 4   chol      1025 non-null   int64
 5   fbs       1025 non-null   int64
 6   restecg   1025 non-null   int64
 7   thalach   1025 non-null   int64
 8   exang     1025 non-null   int64
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64
 11  ca        1025 non-null   int64
 12  thal      1025 non-null   int64
 13  target    1025 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

Figure 2. Summary of the data frame

After that we do some data analysis in our dataframe to check whether the result of each patient is evenly distributed or not, as we can see in figure 3 that 45.5%

patients have no heart disease and 54.5% patients have heart disease so it is pretty evenly distributed data to work upon.

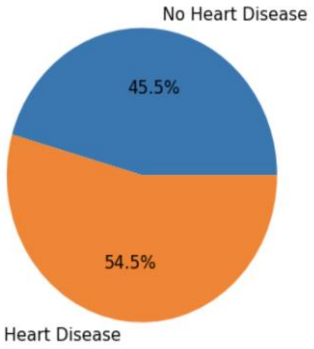

Figure 3. Percentages of patients having a heart disease and not having a heart disease

And after that we further analyze the data that what is the ratio between the two genders and as we can see in figure 3 that the count of the male patients are much higher than the female patients

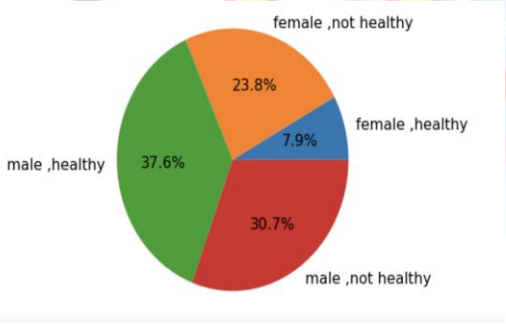

Figure 4. Division of the data of the patients (according to the result of the disease and according to their gender)

And after analyzing the data, we splitted the dataset into two parts, one part for training our model and the other part to actually test our data. We splitted our data in 8:2 ratio, and we are using SciKit Learn metrics library to check for the accuracy of our machine learning models, and for the accuracy the scikit learn library uses a confusion matrix in order to know the accuracy of machine learning model

The confusion matrix uses the following formula for checking the accuracy of a machine learning model:

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

where TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

```
[6] y = data["target"]
    X = data.drop('target',axis=1)
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state = 0)
```

Figure 5. Diving the training data and testing data in 8:2 ratio

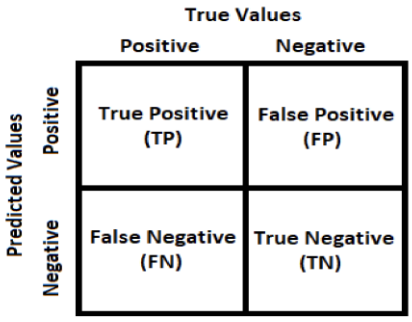

Figure 6. Layout of Confusion Matrix

VI. RESULTS

After training all the machine learning model we got the following results

Logistic Regression Model:

```
confussion matrix
[[ 77  21]
 [  7 100]]

Accuracy of Logistic Regression: 86.34146341463415
```

	precision	recall	f1-score	support
0	0.92	0.79	0.85	98
1	0.83	0.93	0.88	107
accuracy			0.86	205
macro avg	0.87	0.86	0.86	205
weighted avg	0.87	0.86	0.86	205

Naive Bayes Model:

```
confussion matrix
[[79 19]
 [11 96]]

Accuracy of Naive Bayes model: 85.36585365853658
```

	precision	recall	f1-score	support
0	0.88	0.81	0.84	98
1	0.83	0.90	0.86	107
accuracy			0.85	205
macro avg	0.86	0.85	0.85	205
weighted avg	0.86	0.85	0.85	205

Random Forest Model:

```
confussion matrix
[[ 88  10]
 [  3 104]]

Accuracy of Random Forest: 93.65853658536587

              precision    recall  f1-score   support

           0       0.97      0.90      0.93        98
           1       0.91      0.97      0.94       107

    accuracy                           0.94       205
   macro avg       0.94      0.93      0.94       205
weighted avg       0.94      0.94      0.94       205
```

K-Nearest Neighbour Model:

```
confussion matrix
[[84 14]
 [11 96]]

Accuracy of K-NeighborsClassifier: 87.8048780487805

              precision    recall  f1-score   support

           0       0.88      0.86      0.87        98
           1       0.87      0.90      0.88       107

    accuracy                           0.88       205
   macro avg       0.88      0.88      0.88       205
weighted avg       0.88      0.88      0.88       205
```

From the above results we can conclude that Random Forest is performing best among the other models at the accuracy of 93.6%, followed by K-Nearest Neighbour Model at accuracy of 87.8%.

	Model	Accuracy
0	Logistic Regression	86.341463
1	Naive Bayes	85.365854
2	Random Forest	93.658537
3	K-Nearest Neighbour	87.804878

Figure 7. Accuracy of all models

VII. CONCLUSION

Globally, heart disease is the leading cause of death. With the rising number of deaths due to heart disease, it is becoming increasingly important to build a system that can effectively and accurately forecast heart disease.The goal of the research was to discover the most effective machine learning system for detecting cardiac problems.Using the dataset present in kaggle website, this research examines the accuracy of K nearest neighbour, Logistic Regression, Random Forest, and Naive Baye's algorithms for predicting heart disease.Random forest model gives the highest accuracy among all other models which are used in this project.In the future, the study might be improved by creating a web application based on the Random Forest method and employing a larger dataset than the one used in this analysis, which would help to deliver better results and aid health professionals in successfully and efficiently forecasting cardiac disease.

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] T.Mythili, Dev Mukherji, Nikita Padaila and Abhiram Naidu, "A Heart Disease Prediction Model using SVM- Decision Trees-Logistic Regression (SDL)", International Journal of Computer Applications, vol. 68, 16 April 2013.

[5] Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.

[6] Fahd Saleh Alotaibi," Implementation of Machine Learning Model to Predict Heart Failure Disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019

[7] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, "Design And Implementation Heart Disease Prediction Using Naives Bayesian", International Conference on Trends in Electronics and Information(ICOEI 2019).

[8] Theresa Princy R,J. Thomas,'Human heart Disease Prediction System using Data Mining Techniques', International Conference on Circuit Power and Computing Technologies,Bangalore,2016.

[9] Nagaraj M Lutimath,Chethan C,Basavaraj S Pol.,'Prediction Of Heart Disease using Machine Learning', International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019.