International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 04 | April - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Using Machine Learning for Heart Disease Prediction

Siddhant Dhatrak[1], Vinayak Dhole[2], and Ruturaj Kalmegh[3]

Abstract. Research has been carried out on Heart Disease with the help of machine learning algorithms . Prediction of heart disease is a very recent field as the data is becoming available. Other researchers have used various prediction techniques as there are many ways for predicting models . We used data analytics to detect and predict disease's patients .We initiated the model with the dataset available ,where we selected the most relevant features by the correlation matrix, then we applied three data analytics techniques (Logistic Regression Decision Tree and Random Forest) on data sets of different sizes, in order to study the accuracy and stability of each of them. Found Logistic regression is easier to configure and obtain much good results (accuracy of 91.8%).

Keywords: Machine Learning · Heart Disease · prediction · detection.

1 Introduction

Of an estimated 70 million deaths from all causes worldwide in 2018, cardiovascular disease (CVD) accounted for 40%. This ratio is equal to infectious diseases, nutritional deficiencies and maternal and perinatal conditions combined [12]. It is important to recall that a significant proportion of these deaths (46%) is attributable to people under 70 years of age, in the most productive period of life. Furthermore, 79% of the burden of disease attributed to CVD is in this age group [11]. Between 2015 and 2021, deaths due to non-communicable diseases (half of which will be due to cardiovascular disease) is expected to increase by 17%, while deaths due to infectious diseases, nutritional deficiencies and maternal and perinatal conditions combined is expected to decrease by 3% [12]. Nearly half of the burden of disease in low- and middle-income countries is already due to non-communicable diseases [9].

Heart disease is the leading cause of death in the world. More than 1.6 million people die of heart disease each year. The term "heart disease" refers to several types of heart problems. The most common type is coronary artery disease, which can cause a heart attack. Other types of heart disease may involve the valves in the heart, or the heart may not pump well and cause heart failure. Some people are born with heart disease. Anyone, including children, can develop heart disease. It happens when a substance called plaque builds up in your arteries. Smoking, unhealthy eating and lack of exercise increase your risk of heart disease. High cholesterol, high blood pressure or diabetes can also increase your risk of heart disease.

To deal with this disease, there are several methods of prevention, such us natural methods, like stopping smoking, maintaining a healthy weight, adopting a healthy diet and practicing sports regularly . We also have the scientific methods such as drugs and surgeries. The prediction of this disease before being infected is part of the prevention methods, or the computer tools are the most used means in it, more precisely the Machine Learning algorithms.
Our study to this problem is part of data science applications, where we detect cardiac patients based on well-defined attributes such as (age, sex, cholesterol, blood pressure). The use of data collected from patients, is very important to train the learning algorithms, where we use a data set collected from Algerian hospitals which certain a group of people are sick, and others are not. Before starting to present our study, we present a state of the art on the most recent research work in this field. This followed by pre-processing where we select the most relevant attributes that give the best results, using the correlation matrix. Finally we apply learning algorithms on different sizes of the data set (600, 800, 1000, 1200 lines), to develop the most appropriate and stable prediction approach.

2 Related Work

The detection and prediction of heart disease an important issue. For this reason, a lot of work has been done in this area. We divide existing work into two categories: The first presents approaches that select the most relevant patient by features selection, and the second is to explore the learning algorithms that offers high accuracy.

2.1 Features selection

The choice of features is one of the major challenges to train a predictive learning algorithm. Several studies have been carried out and we present the most recent works in the literature. [6] The authors made a kind of comparison of the different Machine Learning algorithms on two classes of attributes. The first one contains 10 attributes and the second one contains 14 attributes. The problem with this study is that the authors did not mention the attributes treated. The authors of [5] developed the same approach as [6] but with only six attributes (Age, Sex,

Blood Pressure, Heart Rate, Diabetes, Hyper cholesterol). The authors of [18], in their study surveyed the most commonly used attributes in cardiac disease detection in different scientific articles, and returned three classes, mentioned in Table 1:

Table 1. The classes of the attributes used

Authors	Attributes
[14] T John Peter et al., 2012 [7] I.S Jenzi,2013 [16] S.Radhimeenakshi 2016	Number of attributes : 13 (Age, Gender, CPT, FBS, RECG, Ex-Ang, SL, Col-Ves, Thal, SC, Thalach, Old peak, RBP)
[3] Chaitrali S et al., 2012 [8] C.Kalaiselvi 2016	Number of attributes : 10 (Age, Gender, CPT, FBS, RECG, SC, Thalach, Smoking, Alchool, Obesity)
[13] Shamsher Bahadur et al., 2013 [10] Hlaudi Daniel et al., 2014	Number of attributes : 6 (CPT, Ex-Ang, Col-Ves, RBP, Num, Smoking)

2.2 Machine Learning Algorithms

An efficient Machine Learning algorithm gives more accuracy. The prediction of heart patients is very critical, because a simple mistake can lead to death of a human being. This section consists of evaluating and selecting the most frequently used algorithms with high accuracy. As in the first section, we have summarized the most recent articles. We start with [6], The authors have implemented several learning machine algorithms, Logistic Regression has given an accuracy of 93%, Random Forest 92% and Gaussian Na¨ıve Bayes 90%, we gave notice that the results are close with simple progression of Logistic Regression. Priyanka.N et al [15], made a comparison between two algorithms (Na¨ıve bayes and Random Forest), the results show that and Random Forest gives much more accuracy than na¨ıve bayes), it reaches 85% accuracy. The authors of [17], conducted a study on the prediction of heart disease, using only dataset, the algorithms used with their accuracy of Logistic Regression (99.3%), Random Forest(91.1%), and decision trees (82.3%). [18] The authors conducted a survey on the most used Machine Learning techniques that give more precision, Table 2 illustrates the algorithms cited in the articles with their accuracy.

Table 2. The most frequently cited algorithms in the papers

Algorithms	Authors	Accuracy
Logistic Regression	[14] T John Peter et al., 2012 [3] Chaitrali S et al., 2012	78% 100%
Random Forest	[14] T John Peter et al., 2012 [8] C.Kalaiselvi 2016	75% 87%

Decision Tree	[14] T John Peter et al., 2012 [3] Chaitrali S et al., 2012 [13] Shamsher Bahadur et al., 2013 [19] B.Venkata-lakshmi et al., 2014	76% 99% 99% 84%

3 Proposed Approach

Our study is based on two major parts: the pre-processing phase, where we chose the most relevant attributes, and the second one apply Machine Learning algorithms in order to select the algorithm that gives a better accuracy. Our proposal is divided into several phases, the approach is explained in detail in Fig. 2

3.1 Dataset Collection

In our case, we use a data set of people who have performed analyses and tests to detect heart disease. The data set is a matrix where the rows represent the patients and the columns represent the factors or attributes (features) to be tested.

3.2 Manual Exploration

This step is very important in the development of Machine Learning algorithms. Because we analyse the data set, where we rank or label each person as sick or not. To give the algorithms the training dataset, we formed the data set.

3.3 Data Pre-processing

Data pre-processing is an important step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we pre-process our data before feeding it into our model.

Features selection In a data set where we have a large set of features [2], we choose the most relevant ones using Pearson Correlation Method (Correlation matrix) [4]. To detect the links between attributes, we only choose those attributes that are highly dependent each other in order to apply Machine Learning algorithms and achieve better accuracy.
Splitting Data set To train the Machine Learning algorithm, we mention the

target column in the data set, then we divide the data set into two small data sets. Training-set to train the algorithm is the Test-set to test it. **Fig.1** explains how we did to apply the Machine Learning algorithm, Where we first decompose our dataset in two parts as mentioned before, then in Training Set we divide the data again, for training and validation (This second step is done automatically)

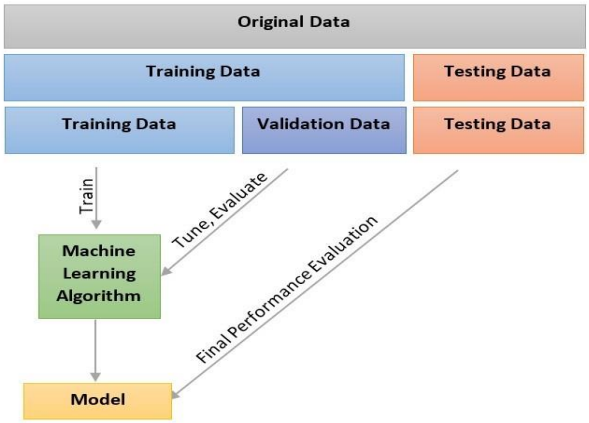

Fig.1. splitting dataset

3.4 Modelling

This is the phase where we apply and test the chosen algorithms (Logistic Regression, Random Forest, Decision Tree), to find the best between them.
Algorithms application In this step. we find that the most efficient and used

algorithms are Logistic Regression, Random Forest, Decision Tree. Our approach is based on the application of the three algorithms on data sets of several sizes to validate the accuracy of each one and above all find the one that is more stable in training and surely gives high accuracy.
Testing algorithms The accuracy ratio to test

the algorithms, on the test set versus manual exploration. A confusion matrix is a table that is often used to describe the performance of a classification model "classifier" on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing. **Choosing the best algorithm** We finish our proposal by selecting the best algorithm that gives the best accuracy, to move on to the next section and present the results obtained.

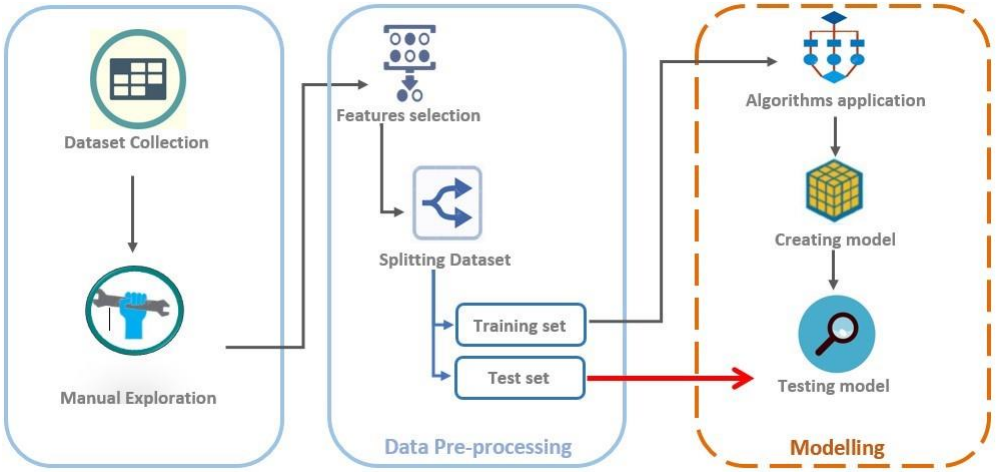

Fig.2. Approach Proposal

4 Results and discussion

In this step we follow the same steps mentioned in the approach. We apply different techniques to achieve the final results, the phases are explained as follows:

4.1 Data Collection

Data collection is defined as the procedure of collecting, measuring and analyzing accurate insights for research. A researcher can evaluate their hypothesis on the basis of collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research. In our study, we use a structured data set of Netherlands people who have done analyses

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 04 | April - 2022 Impact Factor: 7.185 ISSN: 2582-3930

at the Mohand Amokrane EHS Hospital ex CNMS located in Netherlands, with a size of 303 rows and 14 columns are presented as follows (age, sex, cp, trestbps, chol, Ex-Ang, Col-Ves, fbs, restecg, thalach, exang, oldpeak, slope, RBP, ca, thal, the data set is presented in Fig. 3

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
1	29	1	1	130	204	0	0	202	0	0.0	2	0	2
2	29	1	1	130	204	0	0	202	0	0.0	2	0	2
3	29	1	1	130	204	0	0	202	0	0.0	2	0	2
4	29	1	1	130	204	0	0	202	0	0.0	2	0	2
5	34	1	3	118	182	0	0	174	0	0.0	2	0	2
6	34	0	1	118	210	0	1	192	0	0.7	2	0	2
7	34	1	3	118	182	0	0	174	0	0.0	2	0	2
8	34	0	1	118	210	0	1	192	0	0.7	2	0	2
9	34	1	3	118	182	0	0	174	0	0.0	2	0	2
10	34	0	1	118	210	0	1	192	0	0.7	2	0	2
11	35	0	0	138	183	0	1	182	0	1.4	2	0	2
12	35	1	1	122	192	0	1	174	0	0.0	2	0	2
13	35	1	0	120	198	0	1	130	1	1.6	1	0	3

Fig.3. Data set Collection

4.2 Manual Exploration

Data exploration or Manual Explo-ration is the initial step in data analysis, where users explore alarge data set in an unstructured way to uncover initial patterns, characteristics, and points of interest. This process isn't meant to reveal every bit of information a data set holds, but rather to help create a broad picture of important trends and major points to study in greater detail. In our study, We add a column in our data set (Target) which contains zero or one (0 = is not sick, 1 = sick), to start the pre-processing. This is explained in Fig. 4

fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	1	192	0	0.7	2	0	2	1
0	0	174	0	0.0	2	0	2	1
0	1	192	0	0.7	2	0	2	1
0	1	182	0	1.4	2	0	2	1
0	1	174	0	0.0	2	0	2	1
0	1	130	1	1.6	1	0	3	0
0	0	156	1	0.0	2	0	3	0
0	1	182	0	1.4	2	0	2	1
0	1	174	0	0.0	2	0	2	1
0	1	130	1	1.6	1	0	3	0
0	0	156	1	0.0	2	0	3	0
0	1	182	0	1.4	2	0	2	1
0	1	174	0	0.0	2	0	2	1
0	1	130	1	1.6	1	0	3	0
0	0	156	1	0.0	2	0	3	0

Fig.4. Manual Exploration

4.3 Data Pre-processing

Before starting the application of Machine Learning algorithms, we prepare the data to be implemented, this phase is achieved in two steps:

Features selection This step is based on the correlation matrix. Initially we had 20 attributes mentioned before. After applying Pearson correlation matrix, we detected 13 attributes (age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal) that are related and dependent each other.The details of the selected features are explained in Table.3.

Table 3. The details of selected Features

Attribute	**Description**	**Values**
Age	Age	29 to 62 years
Sex	Sex	1 - male 2 - female
CP	Chest pain type	1- typical angina pectoris 2- atypical angina 3- non-anginal pain 4- asymptomatic
trestbps	Resting blood pressure in mm/Hg	Numeric value : example: 140mm/Hg
Chol	Serum cholesterol in mg/dl	Numeric value : example: 289mg/hg
Fbs	Fasting blood pressure>120mg/dl	Numeric value : example: 129mm/Hg
Restecg	Resting electrocardiographic results	0- normal, 1- have the ST-T 2- hypertrophy
thalach	Maximum heart rate achieved	Numeric value : Example: 140,173
Exang	Exercise induced angina	1 - Yes 2 - No
Oldpeak	ST depression induced by exercise relative to rest	Numeric Value
Slope	The slope of the peak exercise ST segment	1 - on the rise 2 - flat 3- the downhill slope
Ca	Number of major vessels coloured by fluoroscopy	0 to 3 vessels
Thal	Thalassemia	3- normal, 6- defect repaired, 7- reversible defect

This selection of Features has been proven by the correlation matrix presented in **Fig. 5.**

name	exang	old-peak	ca	sex	thal	res-tecg	cp	slope	tha-lach	tar-get	chol	trestbps	fbs
exang	1												
oldpeak	0.32	1											
ca	0.11	0.23	1										
sex	0.13	0.08	0.11	1									
thal	0.19	0.2	0.16	0.19	1								
restecg	-0.06	-0.06	-0.06	-0.04	-0.02	1							
cp	-0.4	-0.17	-0.17	-0.03	-0.17	0.04	1						
slope	-0.27	-0.57	-0.09	-0.02	-0.09	0.11	0.13	1					
thalach	-0.39	-0.35	-0.22	-0.07	-0.1	0.06	0.32	0.4	1				
target	-0.44	-0.44	-0.38	-0.26	-0.33	0.13	0.43	0.35	0.44	1			
chol	0.06	0.08	0.05	-0.22	0.11	-0.13	-0.09	-0.03	-0.03	-0.09	1		
trestbps	0.06	0.18	0.1	-0.09	0.05	-0.14	0.04	-0.12	-0.05	-0.14	0.13	1	
fbs	0.06	0.02	0.14	0.04	-0.04	-0.09	0.07	-0.07	-0.01	-0.05	0.02	0.2	1

Fig.5. The correlation matrix

Splitting Data set We divide our data set into two parts: The first part use the training data set with a size of 80%, the test part has a size of 20%. Fig. 6. illustrates how we divided our data set.

Fig.6. Splitting Data set

4.4 Algorithms Used

We apply the selected algorithms (Logistic regression, Random Forest, Decision Tree) and analyze the obtained results.

Logistic Regression:

Logistic regression is a one of the machine learning classification algorithm for analyzing a dataset in which there are one or more independent variables (IVs) that determine an outcome and also categorical dependent variable (DV)[7]. Linear regression uses output in continuous numeric whereas logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes [8].

Random Forest:

Random forest algorithm is a supervised classification algorithmic technique. In this algorithm, several trees create a forest. Each individual tree in random forest lets out a class expectation and the class with most votes turns into a model's forecast. In the random forest classifier, the more the number of trees higher is the accuracy. It is used for classification as well as regression task, but can do well with classification task, and can overcome missing values. Besides, being slow to obtain predictions as it requires large data sets and more trees, results are unaccountable.

Decision Tree:

Decision tree is a technique that is used as a decision support tool that uses a tree-like graph or model of decisions [17]. It takes as input a record or object described by a set of attributes and returns a "decision with predicted output value for the input". The input attributes can be discrete or continuous. After performing a sequence of tests decision tree reaches its decision. Each non leaf node of a decision tree corresponds to a test for the relevant attribute value, and the branches from the node are labelled with the possible outcomes of the test. Each leaf node in the tree specifies the value (decision) to be returned if that leaf is reached [10]. J48, Random Forest (RF) and Logistic Tree Model (LTM) are Decision tree implementation algorithms.

4.5 Testing algorithms

After running the 3 algorithms on the data set, we test the accuracy of each algorithm on the different data sets. Table 3 illustrates how we calculated the accuracy from the confusion matrix, on the 3 algorithms. Now we can display the accuracy of each algorithm

Choosing the best algorithm After analyzing the results found previously. we find that neural networks is the best algorithm in our study, it is always stable in its results, and gives the best accuracy.

Table 5. The accuracy of the different algorithms

	Accuracy	
Algorithm		
Logistic Regression	91.8%	92%
Random Forest	89.7%	89%
Decision Tree	85.1%	85.3%

5 Conclusion

Heart diseases have become more and more frequent among people including our country (India). Therefore, predicting the disease before becoming infected decreases the risk of death. This prediction is an area that is widely researched. Our paper is part of the research on the detection and prediction of heart disease. It is based on the application of Machine Learning algorithms, of which we

have chosen the 3 most used algorithms (Logistic Regression, Random Forest, Decision Tree), on a real data set, where we had very good results, we arrived at 91.8% of accuracy with Neural Network. The strong point of our study, we tested the stability of the algorithm on different sizes of our data set, we noticed at the end that Random Forest gives the best results.

References

1. Babu, S., Vivek, E., Famina, K., Fida, K., Aswathi, P., Shanid, M., Hena, M.: Heartdisease diagnosis using data mining technique. In: 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA). vol. 1, pp. 750–753. IEEE (2017)
2. Cai, J., Luo, J., Wang, S., Yang, S.: Feature selection in machine learning: A newperspective. Neurocomputing **300**, 70–79 (2018)
3. Dangare, C.S., Apte, S.S.: Improved study of heart disease prediction system usingdata mining classification techniques. International Journal of Computer Applications **47**(10), 44–48 (2012)
4. Fang, X., Hodge, B.M., Du, E., Zhang, N., Li, F.: Modelling wind power spatialtemporal correlation in multi-interval optimal power flow: A sparse correlation matrix approach. Applied energy **230**, 531–539 (2018)
5. Gavhane, A., Kokkula, G., Pandya, I., Devadkar, K.: Prediction of heart diseaseusing machine learning. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA). pp. 1275–1278. IEEE (2018)
6. Hasan, S., Mamun, M., Uddin, M., Hossain, M.: Comparative analysis of classification approaches for heart disease prediction. In: 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2). pp. 1–4. IEEE (2018)
7. Jenzi, I., Priyanka, P., Alli, P.: A reliable classifier model using data mining approach for heart disease prediction. International Journal of Advanced Research in Computer Science and Software Engineering **3**(3) (2013)
8. Kalaiselvi, C.: Diagnosing of heart diseases using average k-nearest neighbor algorithm of data mining. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). pp. 3099–3103. IEEE (2016)
9. Lopez, A.D., Mathers, C.D., Ezzati, M., Jamison, D.T., Murray, C.J.: Global andregional burden of disease and risk factors, 2001: systematic analysis of population health data. The Lancet **367**(9524), 1747–1757 (2016)
10. Masethe, H.D., Masethe, M.A.: Prediction of heart disease using classification algorithms. In: Proceedings of the world Congress on Engineering and computer Science. vol. 2, pp. 22– 24 (2014)
11. Organization, W.H.: The world health report 2002: reducing risks, promotinghealthy life. World Health Organization (2016)
12. Organization, W.H., of Canada, P.H.A., of Canada, C.P.H.A.: Preventing chronicdiseases: a vital investment. World Health Organization (2015)
13. Patel, S.B., Yadav, P.K., Shukla, D.: Predict the diagnosis of heart disease patientsusing classification mining techniques. IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS) **4**(2), 61–64 (2013)
14. Peter, T.J., Somasundaram, K.: An empirical study on prediction of heart diseaseusing classification data mining techniques. In: IEEE-International conference on advances in engineering, science and management (ICAESM-2012). pp. 514–518. IEEE (2012)
15. Priyanka, N., RaviKumar, P.: Usage of data mining techniques in predicting theheart diseases—na¨ıve bayes & decision tree. In: 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT). pp. 1–7. IEEE (2017)
16. Radhimeenakshi, S.: Classification and prediction of heart disease risk using datamining techniques of support vector machine and artificial neural network. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). pp. 3107–3111. IEEE (2016)
17. Raju, C., Philipsy, E., Chacko, S., Suresh, L.P., Rajan, S.D.: A survey on predicting heart disease using data mining techniques. In: 2018 Conference on Emerging Devices and Smart Systems (ICEDSS). pp. 253–255. IEEE (2018)
18. Shanmugasundaram, G., Selvam, V.M., Saravanan, R., Balaji, S.: An investigationof heart disease prediction techniques. In: 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA). pp. 1–6. IEEE (2018)
19. Venkatalakshmi, B., Shivsankar, M.: Heart disease diagnosis using predictive datamining. International Journal of Innovative Research in Science, Engineering and Technology **3**(3), 1873–7 (2014)