ISSN Number: 2454-2024(Online)

International Journal of Technical Research & Science

PREDICTION OF HEART DISEASE USING MACHINE LEARNING ALGORITHMS AND NEURAL NETWORK

Deepak Upadhyay[1], Ajay Singh Dhabariya[2], Dr. Hussain Bohra[3], Bhumika Shrimali[4]
E-Mail Id: mybuzzjob@gmail.com[1], ajay.itengineer@gmail.com[2], sakanat5152@gmail.com[3]
[1,2,3]Shrinathji Institute of Technology & Engineering, Nathdwara, Rajasthan-India
[4]Department of Electrical Engineering, Aravali Institute of Technical Studies, Udaipur, Rajasthan, India

Abstract- In this paper we carried out research on heart disease from data analytics point of view. Prediction of heart disease is a very recanted as the data is becoming available. Other researchers have approached it with deferent techniques and methods. We used data analytics to detect and predict disease's patients. Starting with a pre-processing phase, where we selected the most relevant features by the correlation matrix, then we applied three data analytics techniques (neural networks, SVM and KNN) on data sets of different sizes, in order to study the accuracy
and stability of each of them. Found neural networks are easier to conure and obtain much good results (accuracy of 93%).

Keywords: Machine Learning, Heart Disease, prediction, Neural Network.

1. INTRODUCTION

An estimated 58 million deaths from all causes worldwide in 2015, cardiovascular disease (CVD) accounted for 30%. This ratio is equal to infectious diseases, nutritional deficiencies and maternal and perinatal conditions combined. It is important to recall that a Signiant 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. 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%. Nearly half of the burden of disease in low- and middle-income countries is already due to non-communicable diseases [1].

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 [2].

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 _eld. 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. 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 _rst section, we have summarized the most recent articles. We start with, the authors have implemented several learning machine algorithms, Logistic Regression has given an accuracy of 93%, Random Forest 92% and Gaussian Nave Bayes 90%, we gave notice that the results are close with simple progression of Logistic Regression [3].

The authors of tested the diagnosis of heart patients by applying two techniques: genetic algorithms and the KNN algorithm. The results gave satisfaction with the KNN algorithm. The weak point of this article is that the

authors did not mention their results. Same remark for [4], The authors have used artificial neural networks now unfortunately, they have not presented their results.
At the end of the analysis of these papers, we decided to choose the important attributes of the three classes, and the algorithms to implement are: Neural Networks (NN), KNN and SVM, to test our dataset which contains information on Algerian patients.

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 applies 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 analyze 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 preprocess our data before feeding it into our model.
Features selection in a data set where we have a large set of features [5], we choose the most relevant ones using Pearson Correlation Method (Correlation matrix) [6]. 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 (Neural Networks, SVM and KNN), to and the best between them. Algorithms application. In this step. We that the most efficient and used algorithms are: neural networks, KNN and SVM. 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 and the one that is more stable in training and surely gives high accuracy.
Testing algorithms We use the confusion matrix and 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.

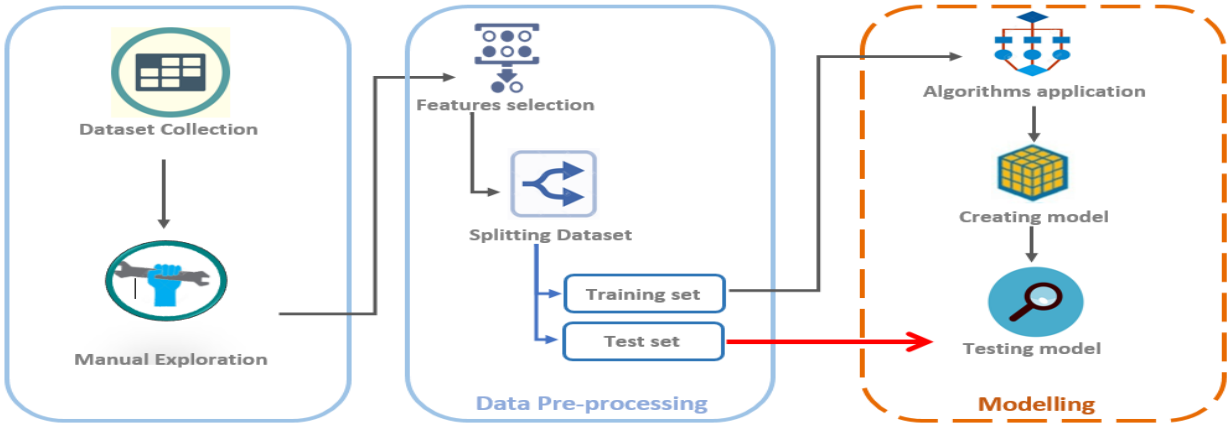

Fig. 3.1 Block diagram of Neural Network

This is the phase where we apply and test the chosen algorithms (Neural Networks, SVM and KNN), to find the best between them. Algorithms application In this step. we find that the most efficient and used algorithms are

neural networks, KNN and SVM. 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 We use the confusion matrix and 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.

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 based on collected data. In most cases, data collection is the primary and most important step for research, irrespective of the _eld of research. In our study, we use a structured data set of Algerian people who have done analyses at the Mohand Amokrane EHS Hospital ex CNMS located Algiers, Algeria, with a size of 1200 rows and 20 columns are presented as follows (age, sex, cp, trestbps, chol, Ex-Ang, Col-Ves, fbs, restecg, thalach, exang, oldpeak, slope, RBP, ca,thal, Smoking, Alchool, Obesity, OilyCough), the data set is presented in Fig. 4.1.

	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. 4.1 Data Set Collection

4.1 Manual Exploration

Data exploration or Manual Exploration is the initial step in data analysis, where users explore a large 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.

4.2 Data Pre-processing

Before starting the application of Machine Learning algorithms, we prepare thedata 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 mentionned 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.
Attribute Description Values Algorithms application The application of the three algorithms is done on our data set. We tested the algorithms on the same data set using different sizes (600, 800, 1000, 1200) in order to detect the algorithm gives more accuracy and at the same time it guarantees stability. Fig. 4.2, Fig. 4.3, Fig. 4.4 shows the graphical results of the Neural Networks algorithm, Svm Algorithm and KNN Algorithm respectively.

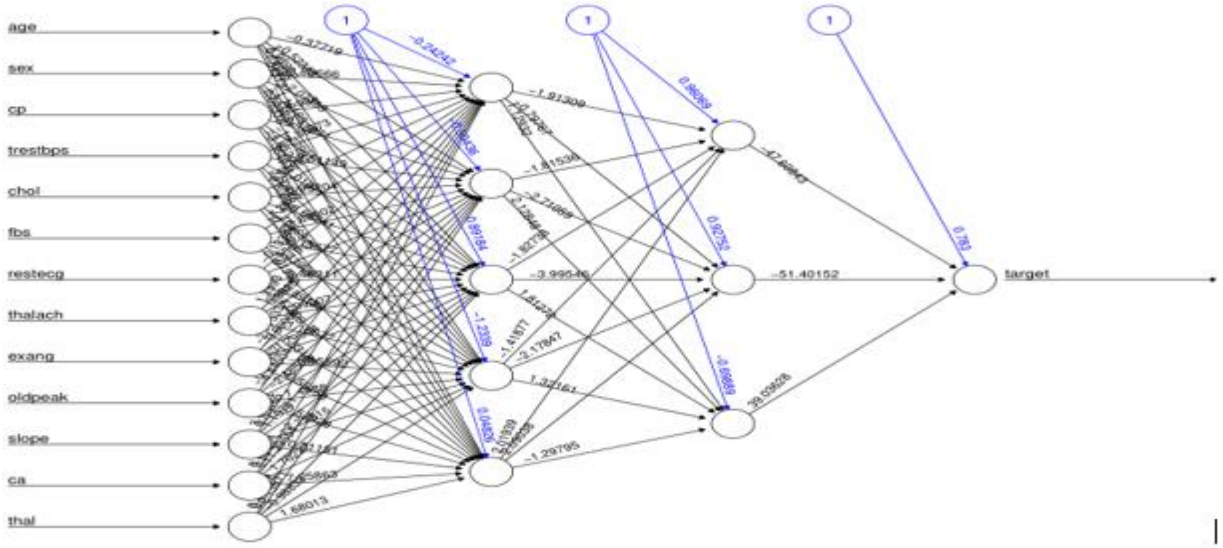

Fig. 4.4 Neural Networks Application

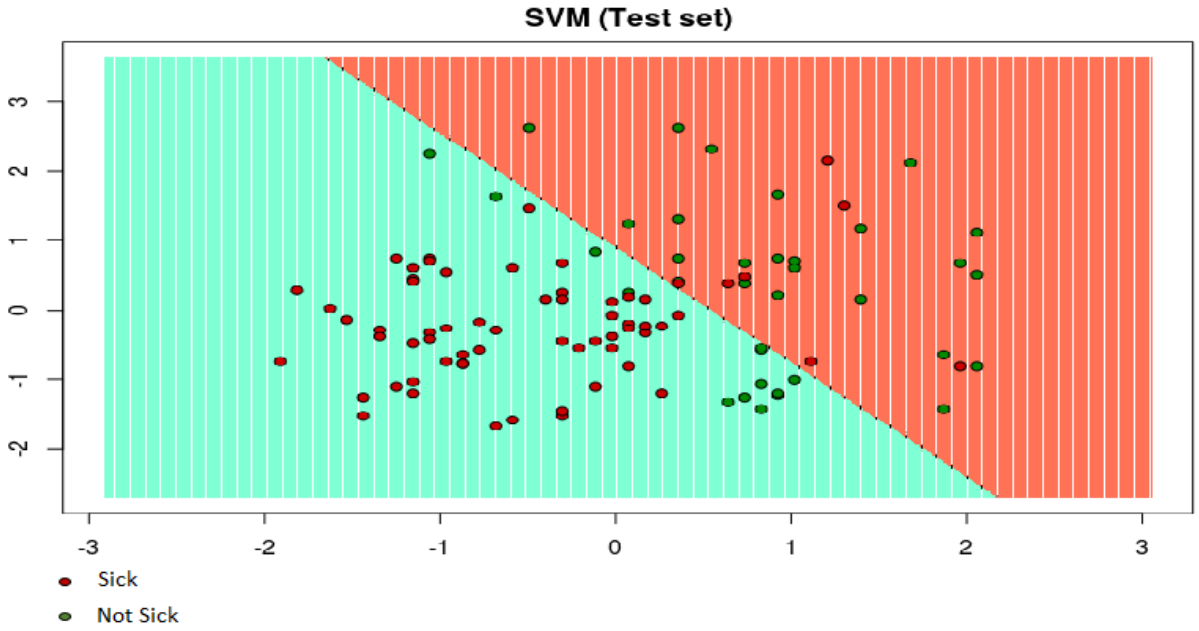

Fig. 4.5 SVM Application

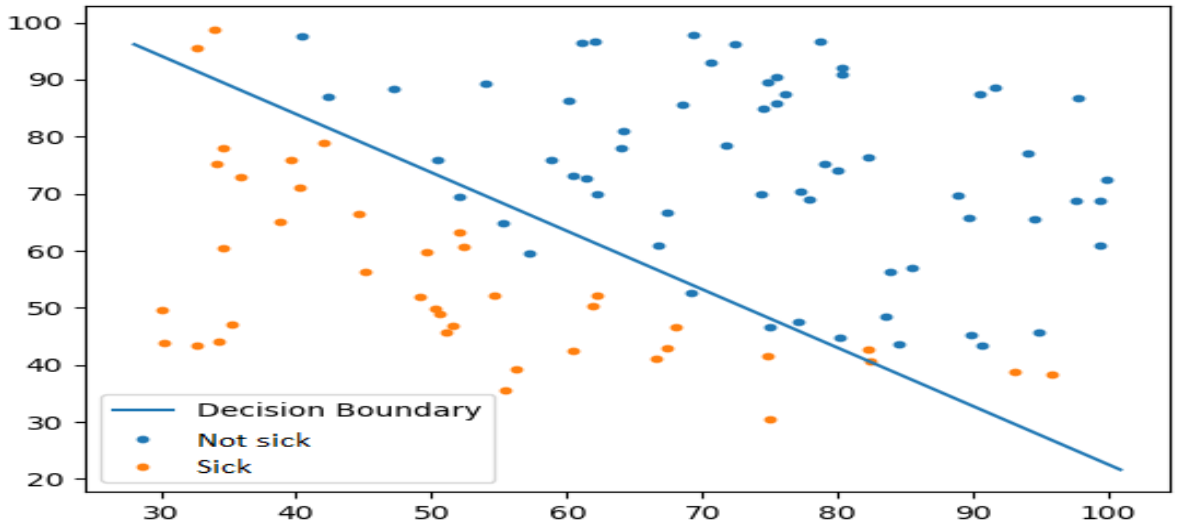

Fig. 4.6 KNN Application

4.3 Testing algorithms

After running the 3 algorithms on the 4 data sets (600, 800, 1000 and 1200 lines), we test the accuracy of each algorithm on the deferent data sets. Table 3 illustrates how we calculated the accuracy from the confusion matrix (we took the case of the last data set which contains 1200 lines), on the 3 algorithms.
Now this can display the accuracy of each algorithm on the 4 data sets, in

CONCLUSION

Heart diseases have become more and more frequent among people including our country (Algeria). 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 (Neural Network, SVM and KNN), on a real data set of Algerian people, where we had very good results, we arrived at 93% of accuracy with Neural Network. The strong point of our study, we tested the stability of the algorithm on deferent sizes of our data set, we noticed at the end that Neural Network gives the best results. Also, we made a study on the features selection, or we used the correlation matrix to detect the dependencies between the attributes. This approach can be improved in several aspects, for example applying deep Learning algorithms, using other methods for attribute selection, and even increasing the size of the data set.

REFERENCES

[1] Babu, S., Vivek, E., Famina, K., Fida, K., Aswathi, P., Shanid, M., Hena, M.: Heart disease 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 new perspective. Neurocomputing 300, 70{79 (2018).
[3] Dangare, C.S., Apte, S.S.: Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications 47(10), 44{48 (2012).
[4] N. Singh and R. Tirole, "Bumble Bees Mating Optimization Algorithm for Economic Load Dispatch with Pollution," 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), 2018, pp. 1-6, doi: 10.1109/ICACAT.2018.8933681.
[5] Gavhane, A., Kokkula, G., Pandya, I., Devadkar, K.: Prediction of heart disease using 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 and regional 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 classi_cation 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, promoting healthy life. World Health Organization (2016).
[12] Organization, W.H., of Canada, P.H.A., of Canada, C.P.H.A.: Preventing chronic diseases: a vital investment. World Health Organization (2015)
[13] Patel, S.B., Yadav, P.K., Shukla, D.: Predict the diagnosis of heart disease patients using classi_cation 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 disease using classi_cation data mining techniques. In: IEEE-International conference on advances in engineering, science and management (ICAESM-2012). pp. 514{518. IEEE (2012)