Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning

Conference Paper · November 2019

DOI: 10.1109/ICCCI48352.2020.9104210

5 authors, including:

Tarun Saxena
Indian Institute of Information Technology Nagpur
1 PUBLICATION **2** CITATIONS

SEE PROFILE

Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning

Montu Saw, Tarun Saxena, Sanjana Kaithwas, Rahul Yadav, Nidhi Lal

Dept. of Computer Science and Engineering

IIIT Nagpur, India

montu.saw@cse.iiitn.ac.in, tarun.saxena.@cse.iiitn.ac.in,
sanjana.kaithwas@cse.iiitn.ac.in, rahul.yadav@cse.iiitn.ac.in, nidhi.lal@cse.iiitn.ac.in

Abstract- In the current era deaths due to heart disease have become a major issue. Approximately one person dies per minute due to heart disease. Data is generated and has to be stored daily because of fast growth in Information Technology. The data which is collected is converted into knowledge by data analysis by using various combinations of algorithms. Medical professionals working in the field of heart disease have their own limitations, they cannot predict the chance of getting heart disease up to high accuracy. This paper aims to improve Heart Disease predict accuracy using the Logistic Regression model of machine learning considering the health care dataset which classifies the patients whether they are having heart diseases or not according to the information in the record.

Keywords:-Heart Diseases; Data Analysis; Machine Learning; Logistic Regression Algorithms.

I. INTRODUCTION

The load of cardiovascular diseases is rapidly increasing all over the world from the past few years. Even if these diseases has found as the most important source of death, it has been announced as the most manageable and avoidable disease [1]. Mainly, blockage in arteries causes heart stroke. It occurs when heart does not pump the blood around the body efficiently.

Having high blood pressure is also one of the main causes of getting a heart disease. A survey says that, in 2011 to 2014, the commonness of hypertension in the world was about 35%, which is also a cause of heart disease. Similarly, there are many more reasons for getting a heart disease such as obesity, not taking in proper nutrition, increased cholesterol and lack of physical activity. So, prevention is very necessary. For prevention, awareness of heart diseases is important. Around 47% of people dies outside the hospital and it shows that they don't act on early warning signs.

Nowadays, lifespan of a human being is reduced because of heart diseases. So, World Health Organization (WHO) developed targets for prevention of non-communicable diseases (NCDs) in 2013, in which, 25% of relative reduction is from cardiovascular diseases and it is being ensured that at least 50% of patients with cardiovascular diseases have access to relevant drugs and medical counselling by 2025 [2]. Around 17.9 million people died just because of cardiovascular diseases in 2016, which is 31% of deaths around the world.

A major challenge in heart diseases is its detection [3]. It is difficult to predict that a person has a heart disease or not. There are instruments available which can predict heart diseases but either they are expensive or are not efficient to calculate the chance of heart disease in human [4]. A survey of World Health Organization (WHO) says that medical professionals are able to predict just 67% of heart disease, so there is a vast scope of research in this field [5]. In case of India, access to good doctors and hospitals in rural areas is very low. A 2016 WHO report says that, just 58% of the doctors have medical degree in urban areas and 19% in rural areas.

In USA, someone has a heart attack every 40 seconds, that is, more than one person dies in USA due to heart attack. Apart from this, Turkmenistan have the highest rate of deaths till 2012, with 712 deaths per 100,000 people. Whereas, Kazakhstan have the second highest rate of deaths due to heart diseases. India holds 56^{th} position in this series [6]. Study also shows that, at ages 30-69 years, 1.3 million cardiovascular deaths, 0.9 million (68.4%) were caused by coronary heart disease and 0.4 million (28.0 %) by stroke

Heart diseases are a major challenge in medical science, Machine Learning could be a good choice for predicting any heart disease in humans [7]. Heart diseases can be predicted using Neural Network, Decision Tree, KNN, etc. Later in this paper, we will see that how Logistic Regression is used to find the accuracy for heart disease. It also shows that how ML will help in our future for heart disease.

II. RELATED WORK

There are many works in literature which diagnoses heart diseases using machine learning as well as data mining. A brief survey of that is presented here. A paper named 'A review of

heart disease using machine learning and analytics approach' by M. Marimuthu, M. Abinaya, K.S. Hariesh, K Mandhankumar and V. Pavithra was published on September 2018. The result shows that, through the literature survey, they concluded that, there is a need of combinational and more complex models to increase the accuracy of prediction of heart diseases.

Some papers which were published around 2 to 3 years back have a less accuracy for the prediction of heart diseases as compared to today's need. 'Efficient heart disease prediction system using decision tree' by Sharma Purshottam et al, it was published in 2015. They have used decision tree classifier as their technique and getting 86.3% accuracy. Similarly, we have, 'Prediction of heart disease using modified K-means and by using naïve bayes' by Sairabi H Mujawar et al. This paper was published in 2015. Their accuracy percentage for detection of heart disease was 93% and for undetection it was 89% [13]. This shows that the accuracy percentage depends on the technique which you are using.

Another example is of 'heart disease prediction using machine learning and data mining techniques' by Jaymin Patel, Prof. Tejpal Upadhyay and Dr. Sameer Patel from Nirma University, Gujarat. [14]

III. EXISTING SYSTEM

Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The before all existing system [6] works on sets of both Deep learning and data mining [7]. Medical diagnosis plays a vital role and yet complicated task that needs to be executed efficiently and accurately. To reduce the cost for achieving clinical tests appropriate computer-based information and decision support should be aided. Data mining is the use of software techniques for finding patterns and consistency in sets of data. Also, with the advent of data mining in the last two decades, there is a big opportunity to allow computers to directly construct and classify the different attributes or classes. Learning of the risk components connected with heart disease helps medicinal services experts to recognize patients at high risk of having Heart Disease. Statistical analysis has identified risk factors associated with heart disease to be age, blood pressure, total cholesterol, diabetes, hypertension, family history of heart disease, obesity and lack of physical exercise, fasting blood sugar, etc. but by using all the existing systems the accuracy is very less.[8]

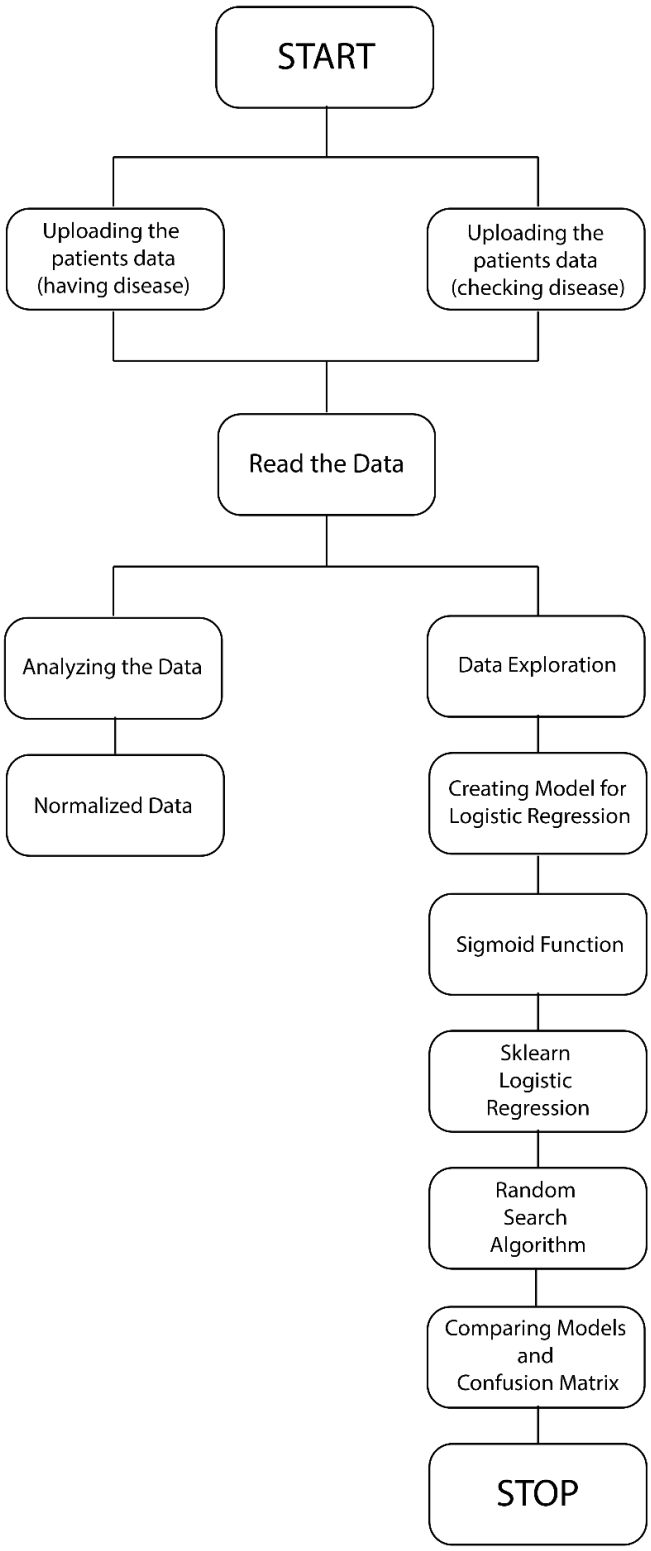

Fig.1: Flowchart of Proposed Work [10]

IV. PROPOSED SYSTEM

This proposed system has data which classifies if patients have heart disease or not according to some parameters. This proposed system can try to use this data to create a model that tries to predict (reading data and data Exploration) [9] if a patient has this disease or not. In this proposed system, using a logistic regression (classification) algorithm we use the sklearn library to calculate the score. Random search is a technique where random combinations of the hyperparameters are used to find the best solution for the built model. Finally, analyzing the results with the help of Comparing Models and Confusion Matrix. From the data we are having, it is classified into different structured data based on the features of the patient heart. From the availability of the data, we have to create a model that predicts the patient's disease using a logistic regression algorithm. First, we have to import datasets read the datasets, the data should contain different variables like age, gender, sex, chest pain, slope, target. The data should be explored so that the information is verified. Create a temporary variable and also build a model for logistic regression [10]. Here, we use a sigmoid function which helps in the graphical representation of the classified data. By using logistic regression, the accuracy is increased as compared to the previous work done in the existing system.

V.APPROACH AND METHODOLOGY

Introduction World Health Organization has estimated 12 million deaths occur worldwide; every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardio vascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression Data Preparation. Logistic Regression is a type of regression analysis in statistics used for prediction of outcome of a categorical dependent variable from a set of predictor or independent variables. In logistic regression the dependent variable is always binary. Logistic regression is mainly used to for prediction and also calculating the probability of success.[11]

	male	age	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	1	39	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
1	0	46	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
2	1	48	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
3	0	61	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
4	0	46	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0
5	0	43	0	0.0	0.0	0	1	0	228.0	180.0	110.0	30.30	77.0	99.0	0
6	0	63	0	0.0	0.0	0	0	0	205.0	138.0	71.0	33.11	60.0	85.0	1
7	0	45	1	20.0	0.0	0	0	0	313.0	100.0	71.0	21.68	79.0	78.0	0
8	1	52	0	0.0	0.0	0	1	0	260.0	141.5	89.0	26.36	76.0	79.0	0
9	1	43	1	30.0	0.0	0	1	0	225.0	162.0	107.0	23.61	93.0	88.0	0

Fig.2: Dataset Distribution

Source: The dataset as shown in Fig.2 is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). There are both demographic, behavioral and medical risk factors that we can see in Fig.3.

```
sns.countplot(x="TenYearCHD", data=heart_data)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7efd61b12908>
```

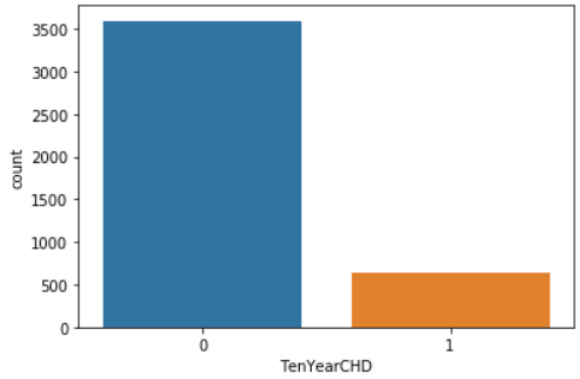

Fig.3: 10-year risk of coronary heart disease CHD

```
sns.countplot(x="prevalentHyp", data=heart_data)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7efd59adec50>
```

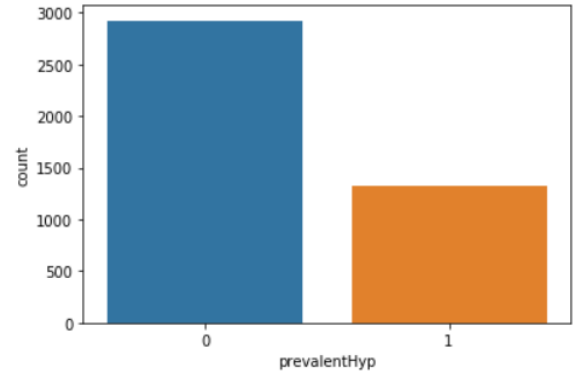

Fig.4: Patient's hypertensive nature

Hypertension was the most important single identifiable risk factor for heart failure until the last few decades. The issue has become less clear over recent years, in part, because of uncertainties in the documentation of heart failure, the lack of systematic recordings of arterial pressure prior to the onset of, and treatment for, heart failure, and the absence of systematic visualization of epicardial coronary arteries that is clearly depicted in Fig.4.[12]

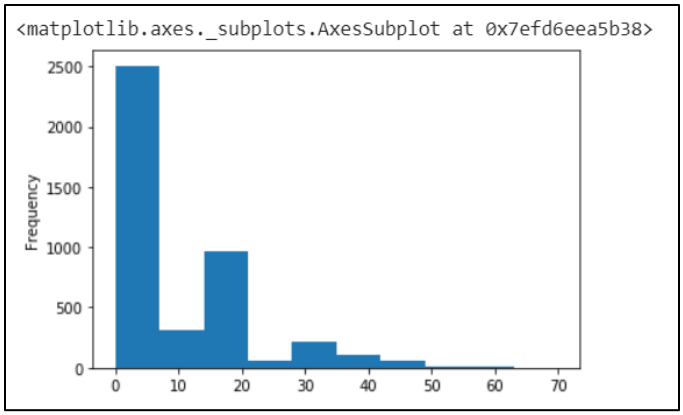

Fig.5: Cigarettes per day

Fig.5 depicts the effect of consumption of cigarettes on heart. Smoking damages the heart and blood vessels very quickly, but the damage is repaired quickly for most smokers who stop smoking. Even a few cigarettes now and then damage the heart, so the only proven strategy to keep your heart safe from the effects of smoking is to quit.

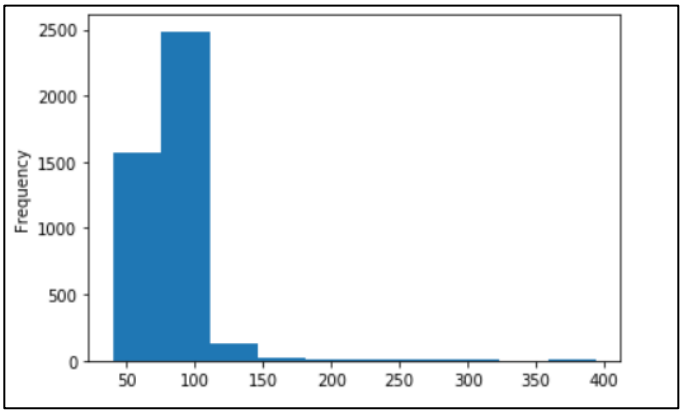

Fig.6: Glucose level

Researchers found that high blood sugar (glucose) causes stronger contraction of blood vessels and also identified a protein associated with this increased contraction. The findings could lead to new treatments to improve outcomes after heart attack or stroke that is shown in Fig.6.

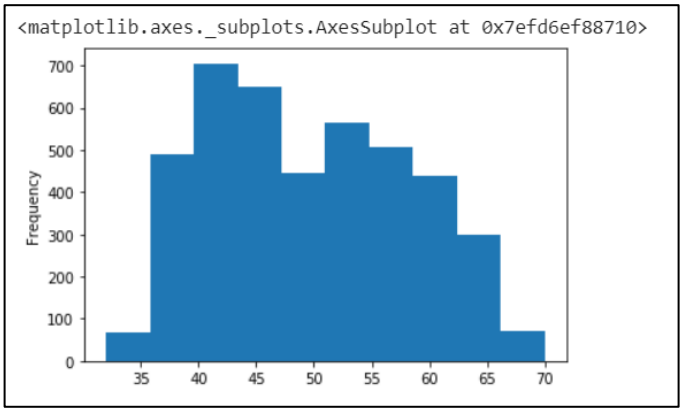

Fig.7: Age of the patient

Fig.7 depicts the effect of Age factor on cardiovascular disease. Age is the most important risk factor in developing cardiovascular or heart diseases, with approximately a tripling of risk with each decade of life. Coronary fatty streaks can begin to form in adolescence. It is estimated that 82 percent of people who die of coronary heart disease are 65 and older. Simultaneously, the risk of stroke doubles every decade after age 55.[13]

	male	age	cigsPerDay	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	1	39	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
1	0	46	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
2	1	48	20.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
3	0	61	30.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
4	0	46	23.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0
5	0	43	0.0	0	1	0	228.0	180.0	110.0	30.30	77.0	99.0	0
6	0	63	0.0	0	0	0	205.0	138.0	71.0	33.11	60.0	85.0	1
7	0	45	20.0	0	0	0	313.0	100.0	71.0	21.68	79.0	78.0	0
8	1	52	0.0	0	1	0	260.0	141.5	89.0	26.36	76.0	79.0	0
9	1	43	30.0	0	1	0	225.0	162.0	107.0	23.61	93.0	88.0	0

Fig.8: Dataset after Wrangling

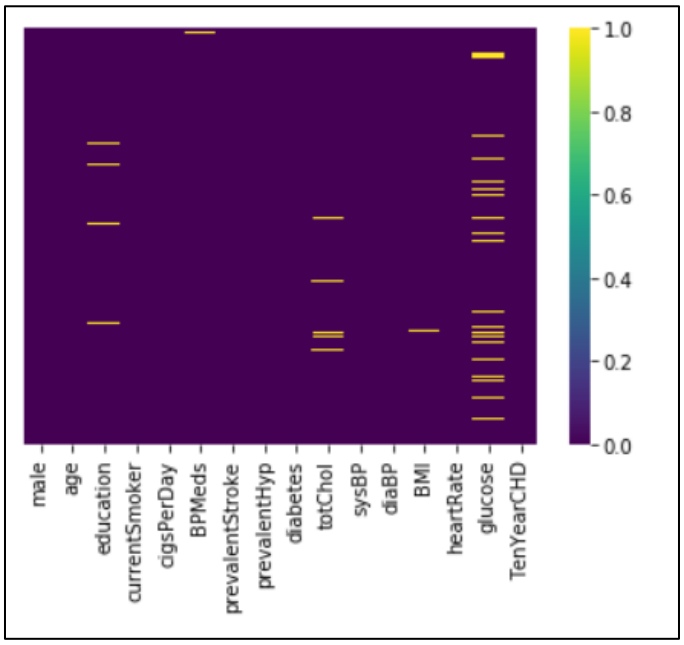

Fig.9: Before Data Wrangling

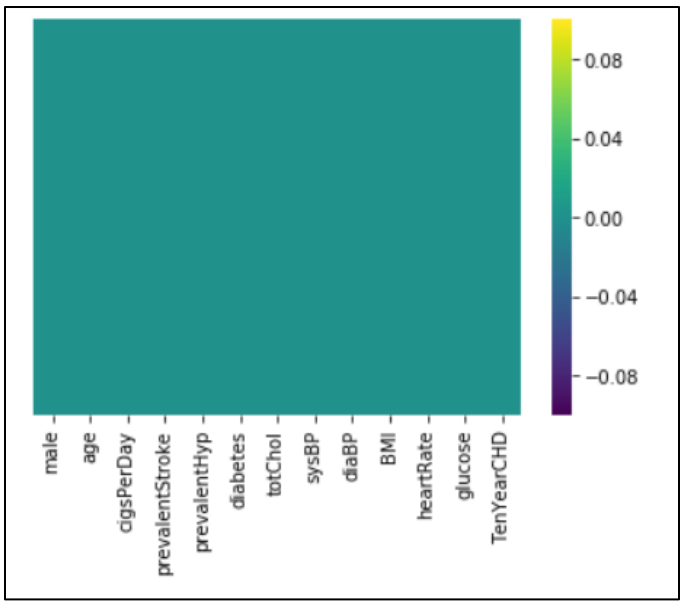

Fig.10: After Data Wrangling

```
Accuracy Score : 0.8702830188679245
Precision Score : 0.8
Recall Score : 0.10084033613445378
F1 Score : 0.17910447761194032
```

Fig.11: Accuracy Result

VI. RESULT

From the above statistics it is clear that the model is highly specific than sensitive. Men seem to be more susceptible to heart disease than women. Increase in age, number of cigarettes smoked per day and systolic Blood Pressure also show increasing odds of having heart disease. Total cholesterol shows no significant change in the odds of CHD. This could be due to the presence of good cholesterol (HDL) in the total cholesterol reading. Glucose too causes a very negligible change in odds (0.2%). The model predicted with **0.87** accuracy which can be seen in Fig.11. The model is more specific than sensitive. Overall model could be improved with more data and by using more Machine Learning models.

VII. CONCLUSION

The amount of Heart diseases can exceed the current scenario to reach the maximum point. Heart disease are complicated and each and every year lots of people are dying with this disease. It is difficult to manually determine the odds of getting heart disease based on risk factors previously shown. By using this system one of the major drawbacks of this work is that it's main focus is aimed only to the application of classifying techniques and algorithms for heart disease prediction, by studying various data cleaning and mining techniques that prepare and build a dataset appropriate for data mining so that we can use this Machine Learning in that logistic regression algorithms by predicting if patient has heart disease or not. Any non-medical employee can use this software and predict the heart disease and reduce the time complexity of the doctors. It is still an open domain waiting to get implemented in heart disease predication and increase the accuracy.

VIII. FUTURE WORK

Today's, world most of the data is computerized and everything is in the cloud which can be accessed although it is not utilized properly. By analyzing the available data, we can also use for unknown patterns. The primary motive of this research is the prediction of heart diseases with high rate of accuracy. For predicting the heart disease, we can use logistic regression algorithm, sklearn in machine learning. The future scope of the paper is the prediction of heart diseases by using advanced techniques and algorithms in less time complexity.

IX. REFERENCES

[1] Avinash Golande, Pavan Kumar T. Heart disease prediction using effective machine learning techniques.

[2] The Lancet Global Health. The changing patterns of cardiovascular diseases and their risk factors in the states of India: The global burden of disease study 1990-2016.

[3] Himanshu Sharma, M A Rizvi. Prediction of heart disease using machine learning algorithms: A survey.

[4] World health ranking.

[5] Himanshu Sharma, M A Rizvi. Prediction of heart disease using machine learning algorithms: A survey.

[6] Sana Bharti,2015. Analytical study of heart disease prediction compared with different algorithms; International conference on computing, communication, and automation (ICCA2015).

[7] Monika Gandhi,2015. Prediction in heart disease using techniques of data mining, International conference on futuristic trend in computational analysis and knowledge management (ABLAZE- 2015)

[8] Sarath Babu, 2017.Heart disease diagnosis using data mining technique, international conference on electronics, communication and aerospace technology (ICECA2017)

[9] A H Chen, 2011. HDPS: heart disease prediction system; 2011 computing in cardiology

[10] Reddy Prasad,Pidaparthi Anjali, S.Adil, N.Deepa(Feb 2019) Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning

[11] Gritsenko, Elena. "Health Care Analytics: Modeling Behavioral Risk Factors Associated With Disease." (2019).

[12] Kazzam, E., Ghurbana, B., Obineche, E. *et al.* Hypertension — still an important cause of heart failure?. *J Hum Hypertens* **19,** 267–275 (2005) doi:10.1038/sj.jhh.1001820

[13] M. Marimuthu, M. abinaya, K S Hariesh, K Madhankumar, V Pavithra. A review on heart disease prediction using machine learning and data analytics approach.

[14] Jaymin Patel, Prof. Tejpal Upadhyay and Dr. Samir Patel. Heart disease prediction using machine learning and data mining technique.