Predicting The Risk Of Heart Disease Using Advanced Machine Learning Approach

Dr. Rakhi Waigi[1], Dr Sonali Choudhary[2], Dr Punit Fulzele[3], Dr. Gaurav Mishra[4]

[1]Asst. Professor, Computer Technology Dept. YCCE, Nagpur
[2]Prof., Dept of Community Medicine, Jawaharlal Nehru Medical College, Datta Meghe Institute of Medical Sciences, Sawangi (M), Wardha
[3]Asso. Prof, Dept of Pediatric Dentistry, Sharad Pawar Dental College, Datta Meghe Institute of Medical Sciences, Sawangi (M), Wardha.
[4]Professor Department of Radiodiagnosis, Jawaharlal Nehru Medical College, Datta Meghe Institute of Medical Sciences, Sawangi (M), Wardha.

Corresponding Author:
Dr Punit Fulzele, Asso. Prof, Dept of Pediatric Dentistry, Sharad Pawar Dental College, Datta Meghe Institute of Medical Sciences, Sawangi (M), Wardha
punitr007@gmail.com, 9890417646

Conflict of Interest: *None*
Funding: *NewGen Innovation and Entrepreneurship Development Centre, National Science and Technology Entrepreneurship Development Board (NSTEDB), Department of Science & Technology (DST), Government of India*

Abstract :

Heart diseases also called Cardiovascular Diseases (CVD) include range of conditions portraying illness of heart. These include diseases related to blood vessels, rhythm problem, chest pain, heart attack, strokes, and fluctuating blood pressure. Person suffering with CVD has fluctuating blood flow rate. CVD are the leading cause of mortality in India including both male and female. A quarter of all mortality is attributed to cardiovascular diseases. Heart diseases and strokes are the pre-dominant causes and are responsible for > 80% of CVD deaths. Therefore in this paper a machine learning model is implemented on the dataset downloaded from kaggle. This dataset contains various parameters contributing to cardiac morbidity. It contains 70000 records and contains parameters like age, cholesterol, glucose, smoking, alcoholic habit etc. The decision Tree model is used fot training and predicting the risk of heart disease. The accuracy of implemented model is 73%.

Introduction

Individuals from lower socio-economic background do not receive optimal therapy for CVD leading to major deaths and identifying individual, at risk of CVD, from high socioeconomic status need preventive measures for it. In today’s changing life style and habits of people, heart fails to supply suitable amount of blood to various organs for their normal activities. As a result heart is at risk of failure [1]. Age, sex, smoking habits, alcohol habits, improper diet, lack of exercise, obesity, high blood pressure are risk factor for possibility of heart disease. Prior symptoms of abnormal functioning of heart includes shortness in breath, fatigue, swollen feet, shoulder pain, jaw pain, neck pain etc.[2]. Identifying and diagnosing the heart disease is very complex and difficult process. There are various challenges such as advanced apparatus needed for treatment and ignorance toward symptoms of heart disease due to costly treatment [3]. Disease prediction using data mining and machine learning techniques is ongoing struggle for past decades. Most of the existing work includes applications of data mining techniques to medical profiles for prediction of diseases. Some approaches tried to predict future risk of progression of diseases but they were unable to give accurate results. Therefore, a system is needed which will continuously monitor the parameters related to blood flow and will predict risk of heart related diseases in future. Many models [4][5][6][7]were proposed in the past based on data-mining techniques but accuracy of those models are not up to the marks. Therefore recent advancement in Machine learning techniques [8 hindawi][9,10] allow researchers to use them in health care domain.

II Literature Survey: Healthcare is a major field of research since last decade. Almost all sorts of algorithms are implemented and tested positively in healthcare domain. Though medical cardiology is a critical domain, recent advancement in data mining and machine learning techniques created significant contribution in diverse domain. Large amount of medical data gets accumulated everyday and researchers have tested their algorithms on it. Developing countries are suffering with major deaths caused due heart malfunctioning[11]. Various optimization algorithms were used in the past for predicting accuracy of risk of heart disease. Neural network based on fuzzy logic was used to train genetic algorithm in paper[12] for feature extraction. It produces accuracy of 99.97%. In paper [13] heart disease was diagnosed with accuracy of 97.8% using genetic algorithm trained using neural network. Three different classifiers i.e. KNN, Decision tree, Naïve Bayes classifier were used in paper[14] for classification of data into risk and non-risk of heart disease. Due to use of Rough set and fuzzy logic techniques in[15][16],death rate due to heart disease decreases.

In this paper two classifiers are implemented for predicting risk of heart disease. Those two classifiers are Decision Tree and Naïve Bayes Classifier. Section III describes the dataset and detail method of implementation.

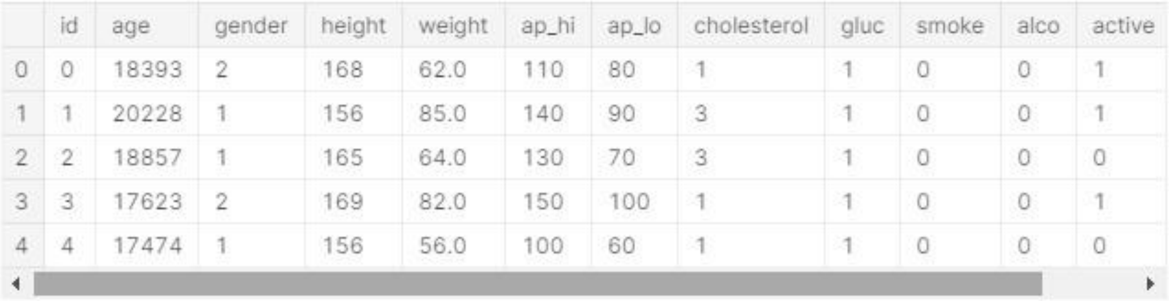
III Materials and Methods

3.1 Dataset

We have downloaded cardio dataset from kaggle website. This dataset consists of 11 parameters such as age, gender, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol level,glucose level, smoking habit, alcoholic habit, active. This dataset consists of 70000 rows and 13columns Meaning and values considered for these parameters are listed below.

Sr. No	Feature	Meaning
1.	Age	Age in Days
2.	Gender	Gender of patient 1:male, 2:female
3.	Height	Height in cm
4.	Weight	Weight in Kg
5.	Systolic Blood Pressure	Pressure exerted when blood ejected into arteries. Measured in mmHg
6.	Diastolic Blood Pressure	Pressure exerted by blood within arteries between heartbeats. Measured in mmHg
7.	Cholesterol level in blood	3 levels considered 1: normal, 2: above normal, 3: well above normal
8.	Glucose level in blood	1: normal, 2: above normal, 3: well above normal
9.	Smoking habit	Binary value 1:smoking and 0:non-smoking
10.	Alcohol habit	Binary value 1:alcoholic and 0:non-alcoholic
11.	Active	Based on physical activity 1:active,0:sluggish

Few records of the dataset are shown below in Fig.1.



	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active
0	0	18393	2	168	62.0	110	80	1	1	0	0	1
1	1	20228	1	156	85.0	140	90	3	1	0	0	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0
3	3	17623	2	169	82.0	150	100	1	1	0	0	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0

Fig1. Sample records of Cardio dataset

To do detailed analysis of the dataset we have plotted histogram of age and class parameter which is shown in Fig. 2. Blue color indicates no-risk class and red color indicates risk class for heart disease. It is observed that after 54years of age risk of cardiac disease increases.

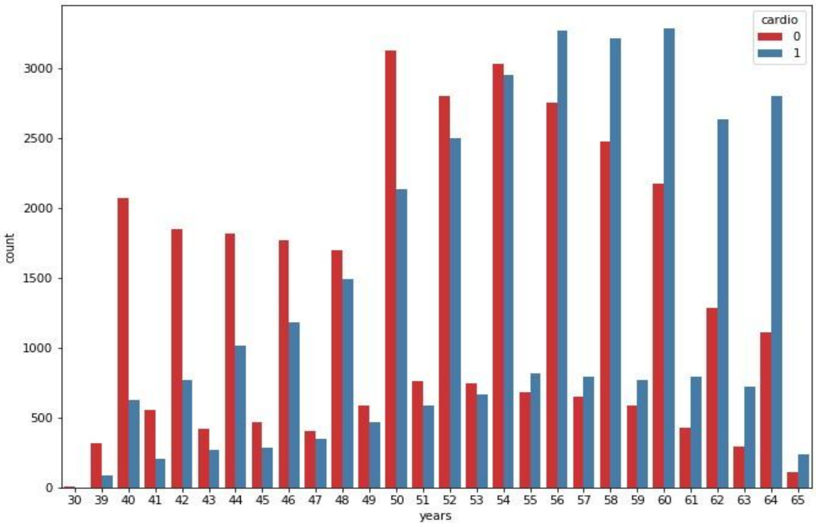


Fig.2 showing histogram of age to cardiac disease cases.

Apart from age we have also plotted histogram of discrete feature such as cholesterol, glucose, alcohol, smoking and active to know about dataset at broad level. It is shown as below in fig.3. It is observed that, in the dataset active and non-smoking and non-alcoholic cases are more with less cases of high glucose and cholesterol levels.

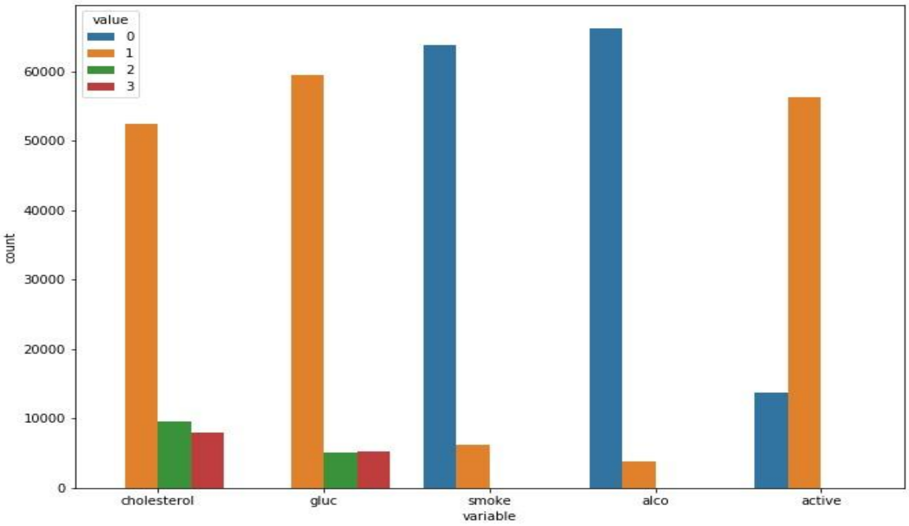


Fig.3 showing histogram of discrete valued features

In the dataset there are 45530 males and 24470 females. This we have identified from average height. Average height of gender 1 is 161.39 cm and that of gender 2 is 169.95 cm. This conclude that gender 1 is female and gender 2 is male. As part of cleaning the dataset, we have removed the irrelevant entries by considering following possibilities.

Ratio of systolic to diastolic pressure: generally blood pressure is recorded as ratio of two numbers ex.120/80. The upper value indicates systolic pressure and lower value indicates diastolic pressure. The entry with systolic value smaller than diastolic is treated as irrelevant and is deleted from the dataset. This gives us 60142 records.
Body Mass Index : BMI is calculated as ratio of weight to height. If BMI is below 9 then it indicates underweight and cholesterol level is high then it is treated irrelevant entry and is deleted.

Once data is cleaned we applied heat map on dataset which is shown in fig 4. To find close relation between attributes. From Fig. 3 it is not very clear about strong correlation between cholesterol or smoking with heart morbidity. Therefore we have implemented decision tree model on the dataset and tested its accuracy. The dataset is classified into two classes. Label 0 indicates non-risk class and label 1 indicates risk class of heart disease. There are 30779 records without risk and 29363 record with risk of heart disease. After applying decision tree model on filtered data, we got accuracy of 72.77%. Evidences from GBD studies were reported reflecting on the need of this technology [17-21]. Related articles are also reported by few of the authors [22-25].

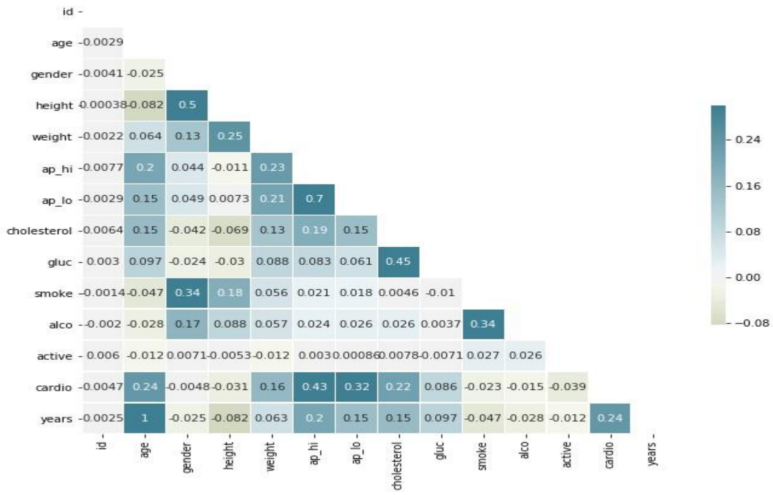


Fig.4. Heat Map showing co-relation between parameters of cardio dataset.

Conclusion and Future Scope

Thus after implementing decision tree model on cardio dataset, it is observed that the accuracy is not very good. May be Naïve Bayes classifier will be better option for this dataset for predicting risk of heart disease. Also in order to improve accuracy we will design a wrist band which will continuously monitor pulse rate, body temperature, blood flow rate. Adding these parameters in the dataset will definitely improve accuracy.

References:

[1] A. L. Bui, T. B. Horwich, and G. C. Fonarow, “Epidemiology and risk profile of heart failure,” *Nature Reviews Cardiology*, vol. 8, no. 1, pp. 30–41, 2011.

[2] M. Durairaj and N. Ramasamy, “A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate,” *International Journal of Control theory and Applications*, vol. 9, pp. 256–260, 2016.

[3] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, “Innovative artificial neural networks-based decision support system for heart diseases diagnosis,” *Journal of Intelligent Learning Systems and Applications*, vol. 5, no. 3, pp. 176–183, 2013.

[4] Ponrathi Athilingam, Bradlee Jenkins, Marcia Johansson, Miguel Labrador "A Mobile Health Intervention to Improve Self-Care in Patients With Heart Failure: Pilot Randomized Control Trial" ,*JMIR Cardio* 2017, vol. 1, issue 2, pg no:1

[5] DhafarHamed, Jwan K. Alwan, Mohamed Ibrahim, Mohammad B. Naeem "The Utilisation of Machine Learning Approaches for Med-ical Data Classification" in *Annual Conference on New Trends in Information & Communications Technology Applications* - march-2017

[6] Mai Shouman, Tim Turner, and Rob Stocker "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients", *International Journal of Information and Education Technology*, Vol. 2, No. 3, June 2012

[7] Amudhavel, J., Padmapriya, S., Nandhini, R., Kavipriya, G., Dha-vachelvan, P., Venkatachalapathy, V.S.K., "Recursive ant colony optimization routing in wireless mesh network", *Advances in Intelligent Systems and Computing*, 2016,381, pp. 341-351.

[8] Amin Ul Haq,Jian Ping Li, Muhammad Hammad Memon ,Shah Nazir,and Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", *Hindawi, Mobile Information Systems*, Vol.2018,ID:3860146,pg.1-21

[9] Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", *International Journal of Recent Technology and Engineering (IJRTE)*, Volume-8, Issue-1S4, June 2019

[10] C. Beulah Christalin Latha, S. Carolin Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", *Elsevier, Informatics in Medicine* Unlocked,2019(16),100203

[11] Vanisree K, JyothiSingaraju, "Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks" *International Journal of Computer Applications* April 2011;19(6). (0975 8887).

[12] Singh Jagwant, Kaur Rajinder, "Cardio vascular disease classification ensemble optimization using genetic algorithm and neural network" *Indian J. Sci. Technology*,2006,9(S1)

[13] KaanUyar Ahmet Ilhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks" *9th international conference on theory and application of soft computing, computing with words and perception*. Budapest, Hungary: ICSCCW; 2017. 24-25 Aug

[14] Y. Alp Aslandoganet. al., "Evidence Combination in Medical Data Mining", *Proceedings of the international conference on Information Technology: Coding and Computing* (ITCC'04) 0-7695-2108-8/04©2004 IEEE.

[15] S. Nazir, S. Shahzad, S. Mahfooz, and M. Nazir, "Fuzzy logic based decision support system for component security evaluation," *International Arab Journal of Information Technology*, vol. 15, pp. 1–9,2015.

[16] S. Nazir, S. Shahzad, and L. Septem Riza, "Birthmark-based software classification using rough sets,"*Arabian Journal for Science and Engineering*, vol. 42, no. 2, pp. 859–871, 2017.

[17] James, Spencer L, Chris D Castle, Zachary V Dingels, Jack T Fox, Erin B Hamilton, Zichen Liu, Nicholas L S Roberts, et al. "Estimating Global Injuries Morbidity and Mortality: Methods and Data Used in the Global Burden of Disease 2017 Study." *Injury Prevention* 26, no. Supp 1 (October 2020): i125–53. https://doi.org/10.1136/injuryprev-2019-043531.

[18] James, Spencer L, Chris D Castle, Zachary V Dingels, Jack T Fox, Erin B Hamilton, Zichen Liu, Nicholas L S Roberts, et al. "Global Injury Morbidity and Mortality from 1990 to 2017: Results from the Global Burden of Disease Study 2017." *Injury Prevention* 26, no. Supp 1 (October 2020): i96–114. https://doi.org/10.1136/injuryprev-2019-043494.

[19] Murray, Christopher J L, Cristiana Abbafati, Kaja M Abbas, Mohammad Abbasi, Mohsen Abbasi-Kangevari, Foad Abd-Allah, Mohammad Abdollahi, et al. "Five Insights from the Global Burden of Disease Study 2019." *The Lancet* 396, no. 10258 (October 2020): 1135–59. https://doi.org/10.1016/S0140-6736(20)31404-5.

[20] Murray, Christopher J L, Aleksandr Y Aravkin, Peng Zheng, Cristiana Abbafati, Kaja M Abbas, Mohsen Abbasi-Kangevari, Foad Abd-Allah, et al. “Global Burden of 87 Risk Factors in 204 Countries and Territories, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019.” *The Lancet* 396, no. 10258 (October 2020): 1223–49. https://doi.org/10.1016/S0140-6736(20)30752-2.

[21] Vos, Theo, Stephen S Lim, Cristiana Abbafati, Kaja M Abbas, Mohammad Abbasi, Mitra Abbasifard, Mohsen Abbasi-Kangevari, et al. “Global Burden of 369 Diseases and Injuries in 204 Countries and Territories, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019.” *The Lancet* 396, no. 10258 (October 2020): 1204–22. https://doi.org/10.1016/S0140-6736(20)30925-9.

[22] Khatib M, Sinha A, Gaidhane A, Simkhada P, Behere P, Saxena D, et al. A systematic review on effect of electronic media among children and adolescents on substance abuse. Indian Journal of Community Medicine. 2018;43(5):S66–72. https://doi.org/10.4103/ijcm.IJCM_116_18.

[23] Zodpey S, Sharma A, Zahiruddin QS, Gaidhane A, Shrikhande S. Allopathic Doctors in India: Estimates, Norms and Projections. Journal of Health Management. 2018;20(2):151–63. https://doi.org/10.1177/0972063418763651.

[24] Chaudhary, K., A. Dhatrak, B.R. Singh, and U. Gajbe. “Perimeter of the Tricuspid Valve: A Cadaveric Human Heart Study.” *International Journal of Research in Pharmaceutical Sciences* 11, no. 3 (2020): 3424–28. https://doi.org/10.26452/ijrps.v11i3.2481.

[25] Schwartz, G.G., P.G. Steg, M. Szarek, D.L. Bhatt, V.A. Bittner, R. Diaz, J.M. Edelberg, et al. “Alirocumab and Cardiovascular Outcomes after Acute Coronary Syndrome.” *New England Journal of Medicine* 379, no. 22 (2018): 2097–2107. https://doi.org/10.1056/NEJMoa1801174