Review of various heart disease prediction algorithms with Machine Learning (Python)

[1]Er Mohit Goyal,[2]Er Anantdeep Kaur
[1,2]Department of Computer Science and Engineering, Punjabi University Patiala

Abstract: Many diseases have found in the human body, but in the present day, heart diseases are the commonly found diseases in all age groups. Heart diseases are caused due to many other health problems such as high blood pressure, smoking, lack of exercise, poor diet, depression, and many more. This paper discusses the different methods to predict heart diseases in humans. Paper discussion different algorithms and technologies such as data mining, deep learning, and Artificial intelligence are useful in the prediction of heart disease. This includes machine learning and programming languages with automated interfaces.

KEYWORDS: **TP, FP, FN, FP, SVM**

I. INTRODUCTION

Heart Disease is also known as Cardiovascular disease (CVD) which is a whole class of the diseases that include the issues in the blood vessels and heart. This includes diseases like CAD (Coronary ArteryDiseases) such as myocardial infarction and angina which commonly known as to be the heart attack. There are many other types of CVD's also presented out there like heart failure, rheumatic heart disease, congenital heart disease, aortic aneurysms, thromboembolic disease, peripheral artery disease, valvular heart disease, and cardiomyopathy. All of these diseases and issues occur because of several other problems like smoking, high blood pressure, lack of exercise, obesity, excessive alcohol consumption, diabetes mellitus, poor diet, and many other alike issues in human's day to day life.

It has been measured that 13% of the deaths are accounted for CVD deaths because of the high blood pressure, tobacco causes 9% of deaths, 5 % of obesity, and 6% of heart issues occur because of the lack of exercise.CVD has cause 32.1% deaths in 2015, 25.8% in 1990, and this is how it increasing day by day. This is all happening because of the change in the life style of people in this generation and their dependence on artificial methods. It has found that 80% of the population who are males and 75% of females are suffering from CVD. But the risk factors of this have been found much higher in women than men. Thus, it is essential to look for certain right functionalities for better outcomes.

1.1 Prediction of Heart Disease

Doctors describe a lot of symptoms that suggest several heart-related issues and also help in determining various related problems as well. However, technology is giving new birth to the healthcare areas and making some better functionalities as well. When it comes to heart diseases, there are many more methods out there that can assist in dealing with the prediction, detection and can be dealt with the better ways. The identification of heart disease can be done by looking after several contributing factors such as high cholesterol, high blood pressure, diabetes, and many other factors. These factors help scientists to find approaches such as Machine learning and Data Mining for predicting the diseases with precise details that are needed to be known.

Machine learning is one of the processes which is automated enough and is one effective method to function in a better manner. This technology is effective enough to assist in making up the predictions and decisions that are being fetched up with the help of the data of the large quantity.

1.2 Deep Learning

Deep learning is a part of machine learning and this comprises of certain algorithms as well. This is all based on artificial neural networks and has a wide classification in it. In this fast-growing technical world, it is seen that machine learning is a part of artificial intelligence. Where machine learning is all about explicitly working on the machine and deep learning is a part of it.

It has found that deep learning does have certain algorithms which look after certain sequential rules. This follows up with the conventional machine learning that deep learning does have certain layering programs that look for some of the better aids and supports in a precise manner. It provides some of the realistic factors and also looks for certain aspects that make the automation easy by using the algorithms. These algorithms are being facilitated to look for better hands for human and to assist the human to simplify their day to day task.

It is also observed that there is a kind of difference between deep learning and machine learning and this makes some of the value differences in nature as well. It is seen that deep learning and machine learning are different in the data dependencies in a certain manner. Another difference between the two is all about the hardware dependencies as it has been observed that deep learning is cordially dependent on the high-end and high-level machines. It is seen that both having different functionalities and these are required to sort out properly. Another difference between machine learning and deep learning is all about feature engineering. It is seen that deep learning is a good domain and the technology for identifying certain

features and looking for various changes and identify the features as well.

1.3 Mining and Data Analysis

Data mining is the technique of the data which are larger in the size and mined together to analyze firmly. Whereas the data analysis is a process that inspects, transform, model, and clean the data. This always has a goal to discover the information that has some better use and also form up certain conclusions. This also helps in supporting the decision-making processes. This has multiple approaches and facts to look for the techniques under a variety of names and usually use the science, business, and alike domains. Data analysis plays an important role in the business, and alike domains.

1.4 Motivations

The evolving technology is the real motivation for this research as these technologies and techniques like data mining, data analysis, programming the factors, and integration of the system. It is seen that different constraints of heart disease should be looked for. Many factors help in predicting heart disease should be looked for.

II. LITERATURE SURVEY

Palaniappan & Awang (2008): in this paper author has worked on the data mining techniques which have been used to predict heart diseases. It has been noticed that the health care industry has a lot of data collected which is not mined, unfortunately. The current research has developed a system i.e. intelligent heart Disease prediction System (IHDPS) with the help of using the decision trees, neural networks, and the naive Bayes. The system is all based on technology and the automation for the prediction of the data related to heart diseases. Results of this study havepredicted that each of the technique is true to its strength and also assist in realizing the goals related to data mining. The study used medical profiles like age, blood pressure, and blood sugar that helped in predicting heart disease in the patients. The IHDPS systems are purely web-based, scalable, expandable, reliable, and is user friendly which has been implemented on the .net platform.

Wilson (1998) et.at: In this paper author has worked on the prediction of coronary heart disease by using the risk factor categories. This study had the objective to precisely evaluate the National Cholesterol Education Program (NCEP) and the Joint National Committee(JNC-V) association. The study has gone through several cholesterol categories along with Coronary Heart Disease (CHD)risk. This study has used the method which was based on the single-centered design which was prospective. This was designed in the proper setting of the cohort that was community-based.

This study shows that a total number of 227 women and 383 men have developed with the CHD in last 12 years. This disease was associated with significant factors like the categories of total cholesterol, blood pressure, HDL cholesterol, and LDL cholesterol. After all the processes, the accuracy of the categories was also found which was quite comparable to the CHD prediction with the help of the assistance of the precise variables which were used properly. This study has developed the algorithms which further developed the categorical variables. These variables allow several physicians to cordially predict the CHD risks of the multivariant in the patients without the overt CHD.

Chen (2011) et.al: In this study, the researchers have developed a system that will be used in the prediction of heart disease that can firmly assist medical professionals. This system will be helping in knowing the status of the disease and the symptoms that will be all based on the clinical data collected from the patient's side. Researchers have used three steps that are sorted with the help of the features and also look for certain better functionalities. The first step is to select the 13 clinical features like chest pain, age, sex, cholesterol, treetops, resting ECG, exercise-induced angina, fasting blood sugar, max heart rate, slope, old peak, thal, and several vessels colored.The second step is to develop the artificial neural networks and its algorithms which should be used for the classificationof heart diseases. The third and final step is to develop and establish the HDPS (Heart Disease Prediction System). This system will be cordially consistingof some multiple features that will also be entertaining the clinical section, a performance display section, and the ROC curve display section.

Jabbar (2013) et.al: The author in this research has worked on the HDPS (Heart Disease Prediction System) with the help of using the Associative classifications and Genetic Algorithm. This paper defines that association classification is a new technique that is rewarding and recent which helps in integrating the association mining rule along with the classification to a model that is needed for the prediction. It has been noticed that the associate classifiers are fit for the application where the accuracy is required that too at the maximum pace and the desire of the prediction is also needed to be modeled. This study has identified that heart disease causes deaths most in developing countries. It is also proved by the mortality rate of India as well that most of the death rates are caused just because of the heart disease. This includes a cause like CHD (Coronary Heart Disease) in the rural area. In this study, a genetic algorithm has also used to predict the disease with a high accuracy level of prediction and to discover some of the high

interestingness values, high predictive accuracy, and high comprehensible.

Hasan (2018) et.al: in this study author has worked on the comparative analysis of the classification approach for heart disease prediction. Study shows that it is not easy to detect heart disease because it requires proper skilled knowledge and relatable experiences as well. As it is a well-known fact that the medical datasets are usually assorted, dispersed, and widespread and dispersed. So, there is a need to sort the data, for this data mining is the technique that is best for the extraction of the data. In this study, the information gain feature selection technique is used that supports the classification techniques such as decision tree, logic regression, Random forest, KNN, and automation. This study has used the different types of performance measurement factors such as ROC curve, recall, sensitivity, accuracy, specificity, precision, and F1-score. This technique and tools are quite compatible to find out the better functionalities for the HDPS.

III. RESULTS AND DISCUSSION

It has been observed from the cordial research that the heart disease can be predicted with the help of various factors. But most of the factors like age, sex, cholesterol, heart beat rate and many more helps to predict the cause of disease. In this research the data of 303 individuals has collected and classified with the following classifiers:

1. SVM: confusion Matrix for both the training and testing set are as below:

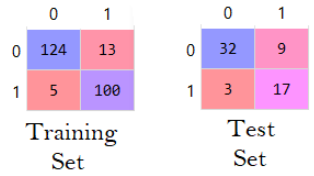

Fig: Confusion Matrix for SVM

Training Set,

Precision = TP/(TP+FP) = 124 / (124 + 13) = 0.95

Recall = TP/(TP+FN) = 124 / (124 + 5) = 0.961

Accuracy = (TP + TN)/ (TP + TN + FN +FP) = (124 + 100) / (124 + 100 + 5 + 13) = 0.926

Training Set,

Precision = TP/(TP+FP) = 32 / (32 + 9) = 0.78

Recall = TP/(TP+FN) = 32 / (32 + 3) = 0.914

Accuracy = (TP + TN)/ (TP + TN + FN +FP) = (32 + 17) / (32 + 17 + 3 + 9) = 0.803

2. NaïveBayes: Confusion Metrix for both the training and testing set are as below:

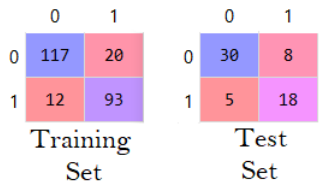

Fig: Confusion Metrix for Naive Bayes

Training Set,

Precision = TP/(TP+FP) = 117 / (117 + 20) = 0.854

Recall = TP/(TP+FN) = 117 / (117 + 12) = 0.907

Accuracy = (TP + TN)/ (TP + TN + FN +FP) = (117 + 93) / (117 + 93 + 12 + 20) = 0.868

Training Set,

Precision = TP/(TP+FP) = 30 / (30 + 8) = 0.789

Recall = TP/(TP+FN) = 30 / (30 + 5) = 0.857

Accuracy = (TP + TN)/ (TP + TN + FN +FP) = (30 + 18) / (30 + 18 + 5 + 8) = 0.787

3. logistic Regression: Confusion Metrix for both the training and testing set are below:

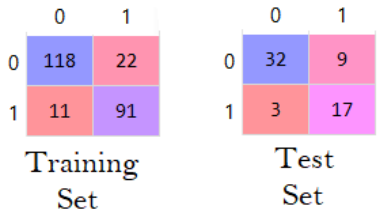

Fig: Confusion Metrix for Logistic Regression

Training Set,

Precision = TP/(TP+FP) = 118 / (118 + 22) = 0.843

Recall = TP/(TP+FN) = 118 / (118 + 11) = 0.915

Accuracy = (TP + TN)/ (TP + TN + FN +FP) = (118 + 91) / (118 + 91 + 11 + 22) = 0.864

Training Set,

Precision = TP/(TP+FP) = 32 / (32 + 9) = 0.780

Recall = TP/(TP+FN) = 32 / (32 + 3) = 0.914

Accuracy = (TP + TN)/ (TP + TN + FN +FP) = (32 + 17) / (32 + 17 + 3 + 9) = 0.803

4. Decision Tree: Confusion Metrix for training and testing set areas:

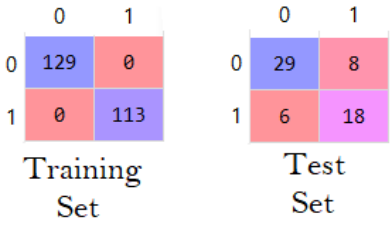

Fig: Confusion Metrix for Decision Tree

Training Set,

Precision = TP/(TP+FP) = 129 / (129 + 0) = 1

Recall = TP/(TP+FN) = 129 / (129 + 0) = 1

Accuracy = (TP + TN)/ (TP + TN + FN +FP) = (129 + 113) / (129 + 113 + 0 + 0) = 1

Training Set,

Precision = TP/(TP+FP) = 29 / (29 + 8) = 0.784

Recall = TP/(TP+FN) = 29 / (29 + 6) = 0.829

Accuracy = (TP + TN)/ (TP + TN + FN +FP) = (29 + 18) / (29 + 18 + 6 + 8) = 0.77

3.1 Comparison

3.1.1 Precision

The following shows that the precision value of all algorithms for the testing set is nearly the same. But Naïve Bayes has a little more precision than other algorithms. Therefore, based on precision, Naïve Bayer is more flexible than other algorithms.

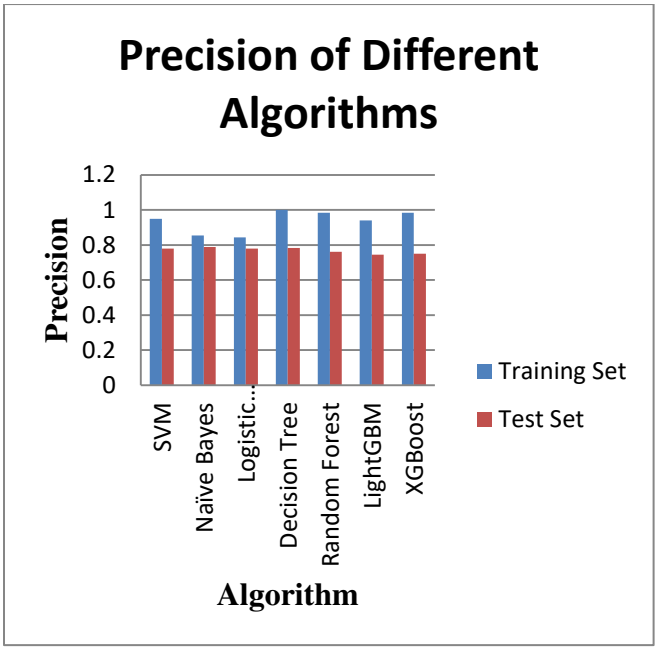

Fig: Comparison of precision of different algorithms

3.1.2 Recall

It can be noted from the graph below that the Recall value of all algorithms for the testing set is nearly the same. But SVM, Naïve Bayes, RF, and Light GBM have a little more recall than other algorithms to predict heart disease.

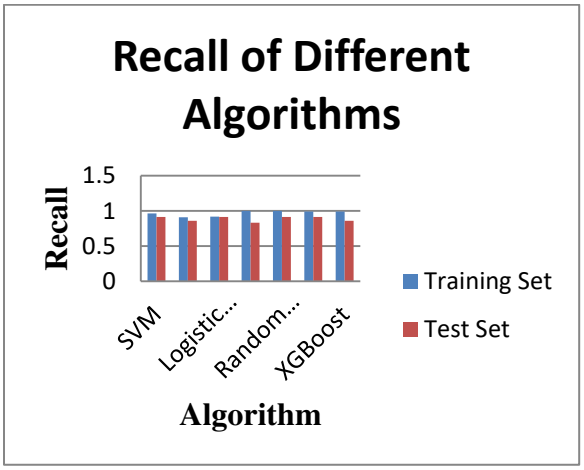

Fig: Comparison of recall value of different algorithms

3.1.3 Accuracy

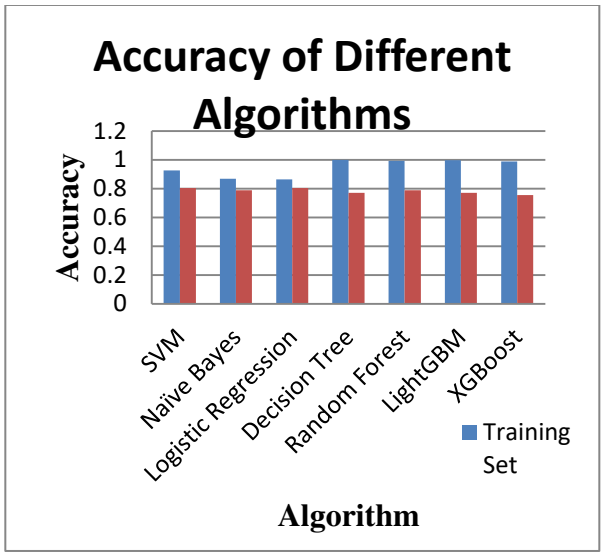

Fig: Comparison of Accuracy of different algorithms

It can be noticed from the above graph that the Accuracy value of all algorithms for the testing set is nearly the same. But SVM and Logistics Regression are more reliable than other algorithms to predict heart disease.

IV. CONCLUSION

Conclusively, it is seen that heart disease is the major issue that causes death and that too across the globe. It is necessary to have the prediction of heart disease at an early stage. The prediction can be done by analyzing the other factors associated like age, sex, cholesterol, heart diseases, thal, and many more. In this research, we have used the dataset available publicly of the 303 individuals and this has been analyzed by the help of the factors significantly essential. It has been seen that data has been classified with the help of classifier algorithms. It has found that people age 57 and more have heart disease and women age more than men who are having heart disease.

V. Future Work

The future work is that the same framework can be applied for the data for real-time scenarios. Another is that the system can be applied in small-scale hospitals and languages with automated data mining software can be used for better efficiency and the processes.

VI. References

- Cleveland Clinic. (2020). Heart Attack (Myocardial Infarction). Retrieved fromhttps://my.clevelandclinic.org/health/diseases/16818-heart-attack-myocardial-infarction
- Mendis S, Puska P, Norrving B (2011). *Global Atlas on Cardiovascular Disease Prevention and Control* (PDF). World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization. pp. 3–18. ISBN 978-92-4-156437-

3. Archived (PDF) from the original on 2014-08-17.

- Wang, H., Naghavi, M., Allen, C., Barber, M. R., Bhutta, Z. A., & Carter, A. (2015). Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, *385*(9963), 117-171. DOI:10.1016/s0140-6736(14)61682-2
- Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. *2008 IEEE/ACS International Conference on Computer Systems and Applications*. https://doi.org/10.1109/aiccsa.2008.4493524
- Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., &Kannel, W. B. (1998). Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation*, *97*(18), 1837-1847. https://doi.org/10.1161/01.cir.97.18.1837
- Chen, A. H., Hyuang, S. Y., Hong, P. S., Cheng, C. H., & Lin, E. J. (2011). HDPS: Heart disease prediction system. *IEEE*.
- Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Heart disease prediction using lazy associative classification. *2013 International Multi-Conference on Automation, Computing, Communication, Control, and Compressed Sensing (iMac4s)*. https://doi.org/10.1109/imac4s.2013.6526381
- M. Hasan, S. M., A. Mamun, M., Uddin, M. P., & A. Hossain, M. (2018). Comparative Analysis of Classification Approaches for Heart Disease Prediction. *2018 International Conference on Computer, Communication, Chemical, Material, and Electronic Engineering (IC4ME2)*. https://doi.org/10.1109/ic4me2.2018.8465594
- Cako, S., Njeguš, A., & Matić, V. (2017). Effective Diagnosis of Heart Disease Presence Using Artificial Neural Networks. *Proceedings of the International Scientific Conference - Sinteza 2017*. https://doi.org/10.15308/sinteza-2017-3-8
- Hung, C., Chen, W., Lai, P., Lin, C., & Lee, C. (2017). Comparing deep neural networks and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. https://doi.org/10.1109/embc.2017.8037515
- Saboji, R. G. (2017). A scalable solution for heart disease prediction using the classification mining technique. *2017 International Conference on Energy, Communication, Data*

Analytics, and Soft Computing (ICECDS). https://doi.org/10.1109/icecds.2017.8389755

- Raihan, M., Mandal, P. K., Islam, M. M., Hossain, T., Ghosh, P., Shaj, S. A., Anik, A., Chowdhury, M. R., Mondal, S., & More, A. (2019). Risk Prediction of Ischemic Heart Disease Using Artificial Neural Network. *2019 International Conference on Electrical, Computer, and Communication Engineering (ECCE)*. https://doi.org/10.1109/ecace.2019.8679362 [13
- Sharma, R., Singh, S. N., & Khatri, S. (2016). Medical Data Mining Using Different Classification and Clustering Techniques: A Critical Survey. *2016 Second International Conference on Computational Intelligence & Communication Technology (CICT)*. https://doi.org/10.1109/cict.2016.142
- Singh, Y. K., Sinha, N., & Singh, S. K. (2017). Heart Disease Prediction System Using Random Forest. *Communications in Computer and Information Science*, 613-623. https://doi.org/10.1007/978-981-10-5427-3_63
- Manikandan, S. (2017). Heart attack prediction system. *2017 International Conference on Energy, Communication, Data Analytics, and Soft Computing (ICECDS)*. https://doi.org/10.1109/icecds.2017.8389552
- Pandey, P. S. (2017). Machine Learning and IoT for prediction and detection of stress. *2017 17th International Conference on Computational Science and Its Applications (ICCSA)*. https://doi.org/10.1109/iccsa.2017.8000018
- Ahmed, H., Younis, E. M., Hendawi, A., & Ali, A. A. (2019). Heart disease identification from patients' social posts, machine learning solution on Spark. *Future Generation Computer Systems*. https://doi.org/10.1016/j.future.2019.09.056
- Das, S., Sanyal, M. K., & Kumar Upadhyay, S. (2020). A Comparative Study for Prediction of Heart Diseases Using Machine Learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3526776
- Patra, R., & Khuntia, B. (2019). Predictive Analysis of Rapid Spread of Heart Disease with Data Mining. *2019 IEEE International Conference on Electrical, Computer, and Communication Technologies (ICECCT)*. https://doi.org/10.1109/icecct.2019.8869194