Heart-Health Status Using Machine Learning

Ebenezer Olukunle Oyebode

Computer Science Department, Ajayi Crowther University, Oyo, Nigeria

Abstract: **Heart disease is one of the killer diseases in the world. Early detection of the disease is one of the ways to salvage affected people. The use of machine learning techniques can be used to offer solution to the detection of heart diseases. In this study the accuracy of prediction of some tools of machine learning has been carried out. The performance evaluation of the three models have been carried out using precision, recall, F1-score and accuracy. The results obtained showed that Logistic regession model out performed others in terms of precision, recall, F1-score and accuracy.**

Keywords: **machine learning, recall, F1-score, precision**

I. INTRODUCTION

Heart disease is one of the killer diseases that affect humanity. World Health Organisation (WHO) confirmed that about twelve (12) million deaths occur annually through heart failure across the world. About 26 million people are having heart diseases world-wide and the number of affected people can be on the increase except right precautions/steps are taken to salvage the situation [6]. Coronary Heart Disease (CHD) and myocardial infarction (as heart attack) are variation of the heart diseases. Risk factors that can contribute to heart problem include smoking, drinking, bad weather, poor diet and dirty environment among others [4]. Features that manifest the presence of heart diseases include high level of cholesterol, high blood pressure, high pulse rates among others.

Machine learning has been applied in health sector as a tool to complement for diagnosis of diseases and yield improved results [2]. Machine learning techniques in recent times have witnessed significant progress in terms of clinical researches. Even though developing machine learning algorithms require significant amount of time and understanding of how the underlying algorithms work, machine learning algorithms can be used for detection of heart diseases.

II. RELATED WORK

[12] used A PSO Algorithm from evolutionary technique to generate the rules for heart disease. Initially, the random rules involved were encoded. Optimization was also applied on the encoded rules which showed improvement in the accuracy of results. Another approach was also used in [11] with decision support system that was cloud based in diagnosis so as to help consultant decide with existing system. A clustering technique was also used to categorize dataset in unsupervised learning. [5] used Rapid Miner, MATLAB, and Weka as software tools to develop models for detection of heart disease with feature selection method and found out that the five models of Decision Tree, Logistic Regression, Random Forest, Naive Bayes, Support Vector Machine all had improved accuracy on Rapid Miner. [10] applied a cloud based ANN for quick computation offered by Amazon as a machine learning tool for the detection of heart disease. The dataset in use was Cleveland Heart Disease Dataset. The model offered accuracy of 90.74%. [4] used Decision Tree, KNN, SVM and Random Forest with the dataset of Cleveland Heart Disease and obtained results from the machine learning models that showed KNN has the highest level of accuracy of 87% and the KNN was converted into packages with pickle library in Python to achieve a web-based framework for mobile devices. [9] used a Support Vector Machine with RBF Kernel Algorithm and applied GridSearch to achieve optimized output. Two datasets from Italian and American Datasets were used for the prediction of Cardiovascular Disease and the results attained showed an accuracy of 95.25% and 92.15% respectively in Italian Dataset and American Dataset.

III. MACHINE LEARNING TECHNIQUES

3.1.1. The logistic Regression

The logistic Regression is used to establish the certainty of event or likelihood of non-occurrence. Such can be used to detect the disease state through symptoms. Logistic regression can be applied to accommodate both categorical and continuous independent variables [5]. Logistic regression can compute the probability of independent variable as a member of the modelled category. [6] stated that It has been applied in Physical Sciences, Economics, Political Science etc. It is usually expressed using eq(1)

$$= \frac{\Pr\ of\ presence\ \ of\ feature}{\Pr\ of\ absence\ \ of\ feature} \quad (1)$$

3.1.2 Support Vector Machine (SVM)

SVM are tools useful in exploratory machine learning. It has attractive features such as geometrical explanations, generalization abilities and empirical performance [3]. SVM can be applied to machine learning problems, optimization problems especially the convex problems etc. SVM uses adopted non-parametric technique that make use of all data during training. SVM only make use of support vectors for future prediction as the support vectors define the margins of hyperplane [8]. For instance, using some points expressed in (2)

$$y_i \in [w_1, w_2] \quad (2)$$

and their distance from the hyperplane can be computed using (3)

$$\frac{|g(y)|}{||w||} \quad (3)$$

SVM solves to find w, b such that g(y) = 1 for nearest points of w_2. This can be expressed as having a marging using (4)

$$\frac{1}{||w||} + \frac{1}{||w||} = \frac{2}{||w||} \quad (4)$$

That is, using (5) and (6)

$$w^T y + b = 1, \ \forall y \in w_1 \quad (5)$$

$$w^T y + b = -1, \ \forall y \in w_2 \quad (6)$$

This can be expressed as

$$k(w) = \frac{1}{2}||w||^2 \quad (7)$$

subject to the constraint stated in (8)

$$x_i(w^T y + b) \geq 1, i = 1,2, \dots, N \quad (8)$$

3.1.3 Naïve Bayes

Naïve Bayes: It is a supervised learning model that can be used for computing the probability of independent variables. It applies the principle of Bayesian theorem for computing probabilities for the presence or absence of a particular attribute of a class of attribute with respect to another class already known. It can predict the most possible outcome with regards to the input [1]. The Naïve Bayes model can be used for computation using eq(9)

$$P(A/B) = \frac{P(B/A).P(A)}{P(B)} \quad (9)$$

IV. DATASET OVERVIEW AND EXPERIMENT SETUP

The Cleveland Heart Disease Dataset [13] used in this study was obtained from the UCI Machine Learning repository. It is an open source dataset. Some missing values were replaced in the data preprocessing stage. 54% of the dataset represented records with symptom that lead to heart disease and 46% represented the records of symptom that has no heart disease. The table below contains the dataset used in the model and the representation for each field and data types. The dataset contains fourteen attributes and 303 records of patient symptoms [7]. The work was carried out using Python alongside its libraries setup in Windows 10 with hardware configurations of 1Tb hard disk, core i3 processor speed of 2.30GHz and 5gb RAM.

Table 1: Dataset description for the heart diseases

S/N	Attribute	Description	Data type
1	Age	Years of patients	Numeric
2	Sex	Male =1 Female =0	Nominal
3	Chest pain	Angina = 1 Typica angina =2 Non-angina = 3 Asymptomatic =4	Nominal
4	Resting blood pressure	blood pressure measured in mm Hg	Numeric
5	Cholesterol level	Cholesterol level measured in mm Hg	Numeric
6	Blood sugar at empty stomach	True = 1 False = 0	boolean
7	Electrocatdiographic results at rest	-	Numeric
8	Maximum heart rate attained	-	Numeric
9	Exercise induced angina	True = 1 False = 0	boolean
10	Depression induced angina	-	Numeric
11	Peak exercise segment slope	Upsloping =1 Flat =2 Downsloping = 3	Nominal
12	Vessels coloured with fluroscopy	-	Numeric
13	Heart status	Normal = 3 Fixed defect = 6 Reversible defect =7	Numeric
14	Heart disease diagnose of narrow diameter	< than 50% >50%	Nominal

Figure 1 contains the high level architecture of the proposed system and the important modules that are contained in it. During the pre-processing stage, all data types where checked against inconsistencies and then the entire complete dataset was moved to the feature extraction module. The feature selection module can be used to select features in the dataset, in this study, all the fields represented were used. The three models were trained and tested using the datasets 80% of the dataset for training and 20% for validation.

Fig.1 : High level architecture of the proposed system

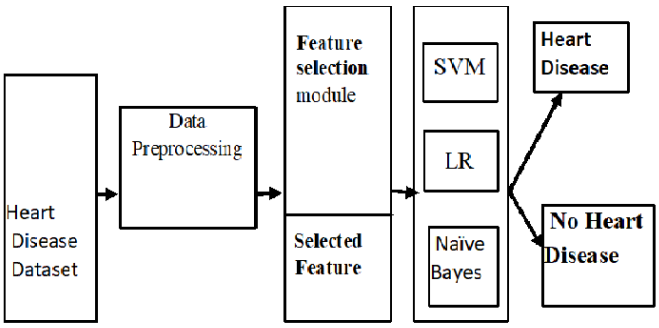

Fig.1 : High level architecture of the proposed system

V. RESULTS AND DISCUSSION

The performance of the models using precision, recall, F1-score and precision is shown in Table2. The outputs as represented in the models are 0 or 1 which means the absence or presence of heart disease. The precision gives the indication of correctly predicted positive observations. The high precision rate means low false positive rate from the model. The highest precision result obtained was given by Logistic Regression model followed by Naive Bayes and the least was given by Support Vector Machine (SVM). The

recall is a representation of sensitivity of each technique and it is the ratio of correctly predicted positive observations to all the three models. Since in all the three models, the value is above 0.5, this support the fact that all the three models are sensitive. The F1 score expresses the weighted average of precision and recall. It combines both the false positives and false negatives. The Logistic Regression model has the highest values for both presence of heart disease(1) and the absence of heart disease (0). It was followed by SVM while Naive Bayes gave the least. In all the three models, in the case of the presence of heart disease classification results, the three models gave the highest value. This is as a result of the higher presence of more records of the presence of heart disease. The accuracy is a function of the F1-score obtained in the two scenarios present in the output. In all the 3 models, logistic regression gave the highest accuracy level followed by SVM and Naive Bayes respectively.

Table 2: Results obtained for the three models

Logistic Regression				
	precision	Recall	F1-score	accuracy
0	0.97	0.78	0.87	0.88
1	0.83	0.97	0.89	
Support Vector Machine				
0	.94	.78	0.85	0.87
1	.82	.95	0.88	
Naive Bayes				
0	0.85	0.78	0.82	0.83
1	0.81	0.87	0.84	

VI. CONCLUSION

The use of machine learning techniques can be applied to detection of heart-health status but different degrees of accuracy can be obtained. The study has shown the prediction accuracy of three machine learning models to predict the presence or absence of heart diseases.

SUGGESTION FOR FUTURE WORK

The study will be extended to adapt the models to be used in mobile applications, it will also consider the identification of various categories of heart diseases.

REFERENCES

[1] Nandhini, S., Debnath, M., Sharma, A. and Pushkar 2018. Heart Disease Prediction using Machine Learning. International Journal of Recent Engineering Research and Development 3(10):39-46.

[2] Garate-Escamila, A. K., Hassani, A. H. and Andres, E. 2020. Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked 19:1-11.

[3] Sharma, S. and Parmar, M. 2020. Heart Diseases Prediction using Deep Learning Neural Network Model. International Journal of Innovative Technology and Exploring Engineering 9(3):2244-2248

[4] Srivastava, K. and Choubey, D. K. 2020. Heart Disease Prediction using Machine Learning and Data Mining. International Journal of Recent Technology and Engineering (IJRTE) 9(1):212-219.

[5] Alotaibi, F. S. 2019. Implementation of Machine Learning Model to Predict Heart Failure Disease. International Journal of Advanced Computer Science and Applications, 10(6):261:268.

[6] Kavitha, S., Baskaran, K. R. and Sathyavathi, S. 2018. Heart Disease with Risk Prediction using Machine Learning Algorithms. International Journal of Recent Technology and Engineering 7(48):314-317.

[7] Suganthi, N. Abinavi, R., Dharshini, S. D. and Haritha, V. 2020. Effective Heart Disease rediction using Distinct Machine Learning Techniques. 7(3): 3383-3388.

[8] Zriqat, I. A., Altamimi, A. M. and Azzeh, M. 2018. A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods. International Journal of Computer Science and Information Security 14(12): 868- 879.

[9] Mezzatesta, S., Torino, C., Meo, P. D., Fiumara, G. and Vilasi, A. 2019. A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. Computer Methods and Programs in Biomedicine 177: 9-15.

[10] Costa, W. L., Figueiredo, L. S. and Alves, E. T. A. 2019. Application of an Artificial Neural Network for Heart Disease Diagnosis. Brazilian Congress on Biomedical Engineering, Springer, 753-758.

[11] Maini, E., Venkateswarlu, B. and Gupta, A. 2018. Applying Machine Learning Algorithms to Develop a Universal Cardiovascular Disease Prediction System. International Conference on Intelligent Data Communication Technologies and Internet of Things 627- 632.

[12] Alkeshuosh, A. H., Moghadam, M. Z. Al Mansoori, I.and Abdar, M. 2017. Using PSO algorithm for producing best rules in diagnosis of heart disease. Int. Conf. Comput. Appl. (ICCA) 306-311. Dataset: https://archive.ics.uci.edu/ml/datasets/Heart+Disease