International Journal of Intelligence Multidisciplinary Engineering Research(IJIMER) Vol. 1 (3), Oct 2023, pp: 01-04.
ISSN: 2583-9160

Machine Learning Approach to Predict the Risk of Coronary Heart Disease

M. Akshaya[1], B. Aiswarrya[2], S. U. Muthunagai[3*], R. Anitha[4]

[1,2]Student, [3]Associate Professor, [4]Professor,
[1,2,3,4]Department of Computer Science and Engineering,
Sri Venkateswara College of Engineering, Sriperumbudur, Tamil Nadu, India
muthunagai@svce.ac.in ORCID ID: 0000-0002-9171-8784

Abstract - **Coronary Heart Disease (CHD) is the major cause of mortality in the modern world due to poor lifestyle. It is caused due to the blockage in one of the arteries, which is supplying blood to the heart, this blockage happens due to fat deposition. Though Coronary Heart Disease are lethal it can be prevented by simple lifestyle changes and early treatment. So there is a demand for a tool which can predict the risks of Coronary Heart Disease well in advance and resulting in the prevention of sudden death. Therefore, machine learning is used to predict the risks of CHD. In this proposed work, machine learning model is developed that predicts whether a person will get CHD or not with the attributes such as age, BMI, systolic and diastolic blood pressures, heart rate and blood glucose levels. Framingham dataset from Kaggle is used to train the predicting model. In this model, Boruta Feature Selection algorithm is used to select most important attributes. SMOTE is used to balance the unbalanced dataset. In this paper, algorithms like Logistic regression, KNN, Decision tree and SVM are used to train the model. Among which SVM shows best performance in the model across the metrics: Accuracy, F1 Score and Area under the ROC curve. This model can be used as a simple screening tool, through which attribute values are passed as an input to get the prediction results.**

Keywords: *Boruta Feature Selection, Coronary Heart Disease, Decision tree, Machine Learning, KNN, Logistic regression, SMOTE test, SVM.*

1. INTRODUCTION

Coronary Heart Disease is one of the major causes of mortality in global. According to the World Health Organization (WHO), an estimation of 17.9 million people has died due to heart disease in 2016, representing 31% of all global deaths. In United States it is estimated that for every 40 seconds about 8 lakh Americans have a heart disease every year (CDC 2019). In India, Coronary Heart disease has caused about 28000 deaths, for the past 5 years. Though Coronary Heart Disease (CHD) is difficult to identify, it is preventable by simple lifestyle changes. Heart is a muscle located at the centre of the circulatory system, it pumps the blood around the body and sends oxygen and nutrients through arteries and carries away carbon dioxide and waste. CHD is a direct result of muscle cells not receiving enough blood due to too much cholesterol in blood deposited on artery walls and narrow down the arteries. And thus, resulting in CHD due to decreases in the blood flow. In this paper, heart attack detection system is presented to check the person has the risk of getting heart attack in 2 years and to take preventive measures. Various attributes including daily habits is given as features to train the dataset and predicts the risk of heart attack. The sample from the patient is taken and applied into the predicting model.

The remaining section of the paper is summarized as follows: Section 2 summarizes the Existing Approaches; Section 3 summarizes about the problem statement and dataset. Section 4 summarizes about the Working model and Section 5 is about Tool development. The performance evaluation of proposed model is shown in Section 6 and Section 7 finally draws the conclusion.

2. RELATED WORKS

The authors of [1] considers minimal factors like level of cholesterol, heart rate, age and gender. Hence it has the probability of giving less accuracy. In the paper [2,3,4] continuous monitoring of heart rate has been carried out with patient only after the patient is affected by heart attack through IOT approach which leads to continuous intervention. Conventional approach of detecting the heart attack using ECG and blood test, this approach is not applicable for effective early detection of heart attack. Whereas in [5] the authors have used a logistic regression classification algorithm for heart disease detection and obtained an accuracy of 77.1%. In the paper [6] the authors have used a multi-layer perceptron (MLP) classifier for heart disease diagnosis and attained accuracy of 80%. The heart

disease classification system integrated with neural networks and artificial neural network has been addressed in the paper [7,8]. In the paper [9], Naïve Bayes (NB) and Decision Tree (DT) algorithm for the diagnosis and prediction of heart disease has been achieved with reasonable results in terms of accuracy of 82.7% with NB and 80.4% with DT.

3. METHODOLOGIES

This paper proposes the idea of Coronary Heart Disease detection system to predict the risk of heart disease in 2 years and to take preventive measures accordingly. Various attributes including daily habits is given as features to train the dataset and predicts the risk of heart attack. The sample from the patient is taken and applied into the predicting model. The dataset is loaded in the algorithms and the model is trained.

The dataset considered this model is Framingham dataset from the open-source website called Kaggle. This dataset contains 15 attributes which are as follows gender, age, current smoker, cigarettes per day, BP, prevalent stroke, prevalent hyper tension, diabetes, total cholesterol, systolic BP, diastolic BP, BMI, heart rate, glucose, CHD risk (class label). The attributes come under these four factors Demographic, Behavioral, Information on medical history and Information on current medical condition.

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
2	1	39	4	0	0	0	0	0	0	195	106	70	26.97	80	77	0
3	0	46	2	0	0	0	0	0	0	250	121	81	28.73	95	76	0
4	1	48	1	1	20	0	0	0	0	245	127.5	80	25.34	75	70	0
5	0	61	3	1	30	0	0	1	0	225	150	95	28.58	65	103	1
6	0	46	3	1	23	0	0	0	0	285	130	84	23.1	85	85	0
7	0	43	2	0	0	0	0	1	0	228	180	110	30.3	77	99	0
8	0	63	1	0	0	0	0	0	0	205	138	71	33.11	60	85	1
9	0	45	2	1	20	0	0	0	0	313	100	71	21.68	79	78	0
10	1	52	1	0	0	0	0	1	0	260	141.5	89	26.36	76	79	0
11	1	43	1	1	30	0	0	1	0	225	162	107	23.61	93	88	0
12	0	50	1	0	0	0	0	0	0	254	133	76	22.91	75	76	0
13	0	43	2	0	0	0	0	0	0	247	131	88	27.64	72	61	0
14	1	46	1	1	15	0	0	1	0	294	142	94	26.31	98	64	0
15	0	41	3	0	0	1	0	1	0	332	124	88	31.31	65	84	0
16	0	39	2	1	9	0	0	0	0	226	114	64	22.35	85	NA	0
17	0	38	2	1	20	0	0	1	0	221	140	90	21.35	95	70	1
18	1	48	3	1	10	0	0	1	0	232	138	90	22.37	64	72	0
19	0	46	2	1	20	0	0	0	0	291	112	78	23.38	80	89	1
20	0	38	2	1	5	0	0	0	0	195	122	84.5	23.24	75	78	0

Fig 1: Sample records of the dataset

As mentioned before, the proposed system uses Framingham dataset from Kaggle to train the predicting model. In this model, Boruta Feature Selection algorithm is used to select most important attributes. SMOTE (Synthetic Minority Oversampling Technique) is used to balance the unbalanced dataset. In this paper, algorithms like Logistic regression, KNN, Decision tree and SVM are used to train the model. Among which SVM was the best performing model across the metrics: Accuracy, F1 Score and Area under the ROC curve. As a result, it confirms the prediction of heart attack by giving 1 or 0 as output. Output 1 indicates risk of having CHD and Output 0 indicates no risk of CHD.

3.1 TOOL DEVELOPMENT

3.1.1 DATA CLEANING AND PREPROCESSING

It mainly checks the missing and duplicate values in the dataset. There are about 12% missing entries of the total data and therefore, it can be dropped without losing a lot of data since it is only a small percentage.

3.1.2. EXPLORATORY DATA ANALYSIS

It looks for important statistical data and detects correlation among the attributes. There are a couple of features that are highly correlated with one another and it does not makes sense using both together in building the machine learning model. These include: blood glucose and diabetes (obviously); systolic and diastolic blood pressures; cigarette smoking and the number of cigarettes smoked per day.

Fig 2: Correlation Matrix

3.1.3 FEATURE SELECTION

The results from the correlation matrix (Fig 2) prompt the need for feature selection. Boruta Feature Selection algorithm is employed for feature selection which is a wrapper method built around the random forest classification algorithm. These are the top selected features after running the algorithm for 100 iterations: Age, total cholesterol, systolic blood pressure, diastolic blood pressure, BMI, heart rate and blood glucose. Fig 3 Shows all the top selected after applying Boruta feature selection.

	CI 5%	CI 95%	Odds Ratio
age	1.011381	1.033813	1.022536
totChol	0.994963	0.999184	0.997071
sysBP	1.018236	1.031493	1.024843
diaBP	0.962258	0.984627	0.973378
BMI	0.929304	0.973798	0.951291
heartRate	0.963690	0.977730	0.970685
glucose	1.001074	1.007518	1.004291

Fig 3: Final selected attributes

3.1.4. MODEL DEVELOPMENT AND COMPARISON

The data in the training dataset is imbalanced as shown in the Fig 5, in a way that number of negative cases (people not having risk of CHD) is more than the number of positive cases (people having risk of CHD). It is not advised to train the dataset in imbalanced data, because the consequence faced will pull the accuracy leads to the issues that are more than sensitivity, as it never predicts the positive cases, since it is trained with majority negative cases.

To address this problem of balancing the data, the proposed model uses a technique called SMOTE (Synthetic Minority Oversampling Technique). SMOTE generates more synthetic positive samples and balances the dataset.

SMOTE works as follows as shown in Fig 4, SMOTE first selects a minority class x1 (a positive case record) instance at random and finds its k nearest minority class neighbors. The synthetic instance is then created by choosing one of the k nearest neighbors x2 at random and connecting x1 and x2 to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances x2 and x1.

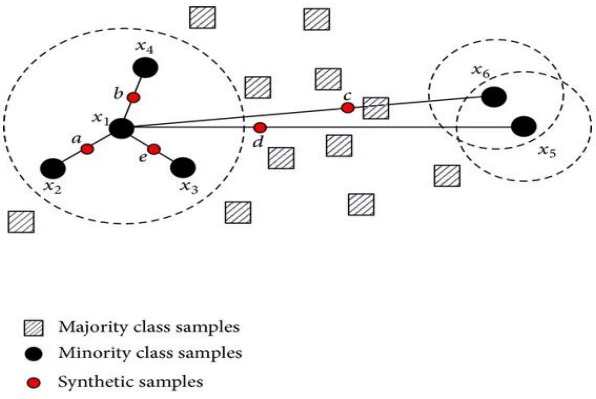

Fig 4: SMOTE

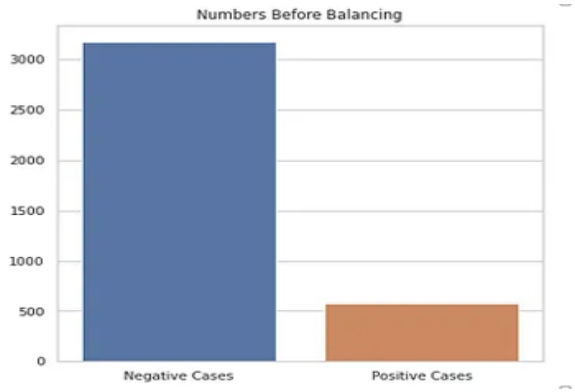

Fig 5: Imbalanced dataset

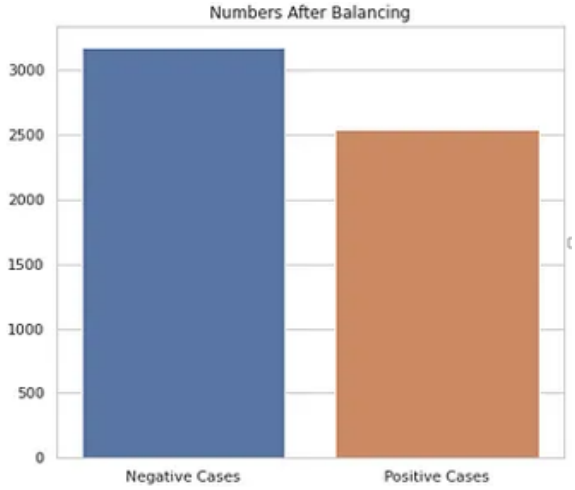

Fig 6: Balanced Dataset

Initially there was only 572 positive cases. After using this technique, the data set became more balanced with 3178 negative cases and 2543 positive cases as shown in Fig 6. After balancing the dataset, the trained model uses four algorithms namely, Logistic Regression, K Nearest Neighbor, Decision Tree and Support Vector Machine (SVM) algorithm.

4. RESULT AND DISCUSSION

In this paper, the proposed system considers three performance metrics they are accuracy, score and Area under ROC curve. Among the model trained using various algorithm, SVM algorithm performs well when compared with other three algorithms.

Fig 7, Fig 8 and Fig 9 shows the performance metrics observed in logistic regression, KNN, Decision Tree and SVM algorithms.

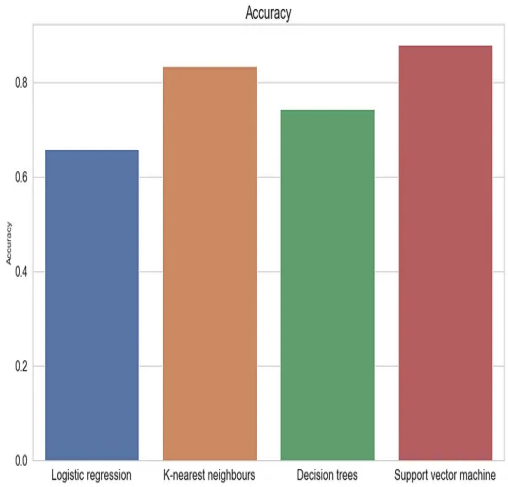

Fig 7: Accuracy

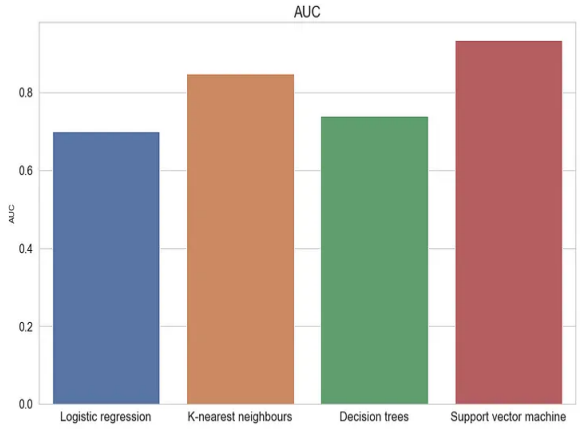

Fig 8: Area Under ROC Curve

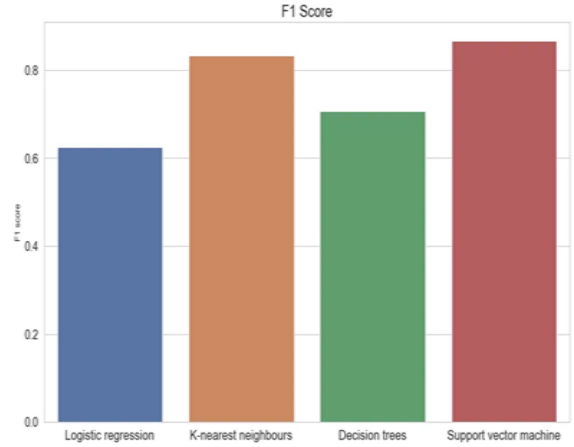

Fig 9: F1 score

5. CONCLUSION

Finally, this paper draws the conclusion with SVM algorithm performance which is chosen for tool development to detect the CHD. This model act as a simple screening tool for detecting the risk of heart attack for the following input such as age, BMI, systolic and diastolic blood pressures etc. This proposed system predicts the risk with two cases 1 and 0. 1 represents the person is under risk and 0 represents the person is not under risk stage. Thus, it takes the sample from human and predicts the risk of having CHD in prior period of time. Furthermore, the proposed system aims at predicting the various stages of risks in Coronary Heart Disease in future.

REFERENCES

[1]. Bui, A. L., Horwich, T. B. & Fonarow, G. C. Epidemiology and risk profile of heart failure. Nat. Rev. Cardiol. 8, 30 (2011).

[2]. Heidenreich, P. A. et al. Forecasting the future of cardiovascular disease in the United States: A policy statement from the American Heart Association. Circulation 123, 933–944 (2011).

[3]. Das, R., Turkoglu, I. & Sengur, A. Effective diagnosis of heart disease through neural networks ensembles. Expert Syst. Appl. 36, 7675–7680 (2012).

[4]. Allen, L. A. et al. Decision making in advanced heart failure: A scientific statement from the American Heart Association. Circulation 125, 1928–1952 (2014).

[5]. Yang, H. & Garibaldi, J. M. A hybrid model for automatic identification of risk factors for heart disease. J. Biomed. Inform. 58, S171–S182 (2015).

[6]. Samuel, O. W., Asogbon, G. M., Sangaiah, A. K., Fang, P. & Li, G. An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. Expert Syst. Appl. 68, 163–172 (2017).

[7]. Tomov, N.-S. & Tomov, S. On deep neural networks for detecting heart disease. arXiv:1808.07168 (2018).

[8]. Mohan, S., Thirumalai, C. & Srivastava, G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access 7, 81542–81554 (2019).

[9]. Ali, L. et al. An optimized stacked support vector machines based expert system for the effective prediction of heart failure. IEEE Access 7, 54007–54014 (2019).