Machine Learning for High Risk Cardiovascular Patient Identification

Muhammad Asad Arshed[*1], Fasial Riaz[2]

[1,2]*School of Systems and Technology (SST), University of Management and Technology, Lahore, Pakistan*.

Article History:

Received: 2 November 2021
Received in revised form: 14 January 2022
Accepted: 15 February 2022
Available online: 17 March 2022

Abstract

According to the WHO (World Health Organization) Cardiovascular (Heart Failure) is a fatal disease that cause estimated 17.9 million people death every year. Heart Disease risk increase due to cholesterol, overweight and hypertension in short due to harmful behavior, heart disease risk increase [1]. Further according to the American Heart Association [2] symptoms(complement) are leg swelling, sleep problem, high heart rate and cough. Diagnosis is difficult due to the common symptoms because these symptoms associate with other diseases. The collection of medical data help physician to diagnosis the disease. Machine learning playing an important role in medical field as diagnosis of disease. Machine learning is used where data is large and difficult to extract useful information from data. In this study, we have used machine learning approach with rapid miner tool for heart disease classification. The experiment results show that performance of Logistic Regression is effective with accuracy of 85.22% than other considered machine learning models.

Keywords: Heart Failure, Machine Learning, Principal Component Analysis, Classification.

I. INTRODUCTION

Heart Disease is a fatal disease that effect the heart specifically, on the other hand cardiovascular disease in which entire circulatory system disturb. In united states, heart disease is leading disease of death according to the CDC (Centers for Disease Control and Prevention). According to the world health organization the number of deaths will be increase up to 24.5 million in 2030 due to high blood pressure, smoking, diabetes and obesity. Food, Smoking, habits, diabetes and obesity can cause the cardiopathy. There are many types of heart disease e.g., Arrhythmia, Congenital heart defects, Coronary disease, Myocardial infarction and many more. Like any other disease heart disease also has some symptoms e.g., chest pain, breathing issue, fatigue and edema. Heart attack is dangers because it led to cardiac arrest and whole body stop functioning [3]. An early prediction of heart attack is necessary to save lives that is possible with symptoms and other parameters. Heart disease is common disease nowadays due to common contributing factors like diabetes, high blood pressure and many more. Doctors trying to figure out the heart disease with the analyses of characteristics and traditional ways are adopted like as electrocardiogram (ECG), blood pressure, cholesterol measuring. Many studies presented in some last year to early diagnosis of heart disease. Artificial intelligence plays an important role in medical field. The effective and reliable model presenting is a challenging task. SVM, KNN, J48 and other machine learning models prediction results and comparisons for heart disease [4]. Due to the complexity of heart disease, prioritized tests and proposed model need to be accurate, in [5] authors effectively predict heart disease with past medical history of patients and five machine learning models (Logistic Regression, AdaBoost, SVM, Naïve Bayes and Likelihood-Variation), and results were consistent with accuracy of 82%. In [6] artificial neural network-based model with back propagation (Multi-Layer Perceptron) was considered for heart disease classification. An application with machine learning model, more precisely neural network for the prediction of heart disease with basic symptoms of heart disease e.g., sex, age and pulse rate developed [7]. In [8], the accuracy of heart disease classification was improved by data mining with random forest classifier, and results shows that random forest can be used for effective results of heart disease classification task.

In this study, we have tried to find out an effective model to diagnose cardiovascular diseases from a number of machine learning classifiers list. The main contributions of this study are listed below:

1) Preprocessing the Data, like dealing with missing values with mean value, K-Nearest Neighbor and Random Forest.

2) Adaptive Synthetic Sampling Approach for Data Balancing.

3) Classification of cardiovascular disease patient with different machine learning classifiers like Logistic Regression, Decision Tree, Generalized Linear Model, Fast Large Margin, Random Forest Tree and Decision Tree etc.

4) Comparison of different machine learning models in terms of evaluation metrices.

[*] **Corresponding Author:** asad.arshed@umt.edu.pk

II. METHODOLOGY

In this study, we have presented the comparison to show the robustness of effective model for heart failure classification problem. The logistic regression is a classifier that based on probability distribution.. It is a common assumption that each feature is independent and equal. The probability of the event is found in logistic theorem and on the behalf of probability classes assigned to explanatory variables/Input.

A. Filling Missing Values

We have used open-source heart disease dataset that is available on Kaggle [9]. The dataset has 303 samples and after analysis of dataset, I have come to know that target value consist on two possible values 0 means patient is normal and 1 means patient is suffering in heart disease problem. In this dataset, number of explanatory variables are 12 and 1 response variable.

B. Filling Missing Values

Before applying the machine learning classifiers, we need to preprocess our data to achieve an effective result.

a) Data Normalization

Min Max Normalization that is also known as feature scaling is considered in this study to normalize the value of dataset features. Each value is replaced with the equation (1) after retrieving the minimum and maximum values from the data.

$$N' = \frac{O - \min(FD)}{\max(FD) - \min(FD)} \qquad (1)$$

In Equation (1) N' specify the new value and O is old value and FD stands for feature data.

C. Filling Missing Values

In this study. We have used the “MEAN Value” algorithm to overcome the problem of missing data or null data. The working of this algorithm is that it replaces the missing value with mean value of that particular feature. Due to easily implementation, this technique is used frequently in machine learning.

$$C = \sum_{k=1}^{m} \frac{Xkm}{n} \qquad (2)$$

D. Grid Search CV

Machine learning models consist on hyper parameters that is not learned by model and we need to tune it manually. Findingthe effective hyper-parameters is a challenging task. An effective approach is to find an effective pair of hyper-parameters that give an effective result and this process is called hyper-parameter optimization. There is a difference between parameters and hyperparameters as parameters are learned automatically by model and hyper-parameters need to set manually to improve the learning of model. Grid Search CV is an effective method in terms of define a grid of hyperparameters as search space and find and evaluation of grid at each position.

E. Pricipal Component Analyis (PCA)

Selection of variable in terms of magnitude of coeffcients is considered in PCA. PCA is also used for dimensionality reduction, in another words conversion of high dimension data to low dimension data. In PCA, we find PC1(Principal Component one) that described the linear combinations of features and feature with high weight(coefficients) is most important feature. We have plotted the importance of features in Figure 1.

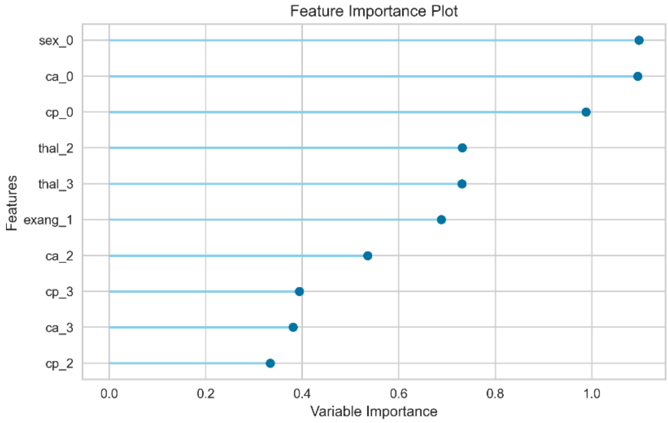

Fig. 1. Top 10 Features with Importance

F. Grid Search CV

Logistic Regression model is based on probality. It find the probability of a target variable and on behalf of probability it assign class to samples. Logistic Resgression is a classier, it intailly find probability that’s why it is so called regression but it finally predict class. Generally, normal logistic regression means binary logistic regression that means it predict class from two possible options. Our problem is also a binary classification problem. As a function of X it predict P(X=1). This classifier is used in most of the classification problems with modifactions e.g., spam detection and cancer prediction.

III. EXPERIMENTAL ANALYSIS

A. Evaluation Metrics

To evaluate the performance of machine learning models we used evaluation metrices in which we consider Accuracy, Precision, Recall, F1 and AUC.

a) Accuracy: It is the ratio of correct prediction from total number of samples.

b) Precision: The ratio of true prediction over total number of true samples in terms of dataset.

c) Recall: The ability of the model to predict true labels over trues that are expected.
d) F1: The mean (Harmonic) of F1 and Precision is called F1.
e) AUC: Aggregate performance is measured by Area Under Curve (AUC). AUC is the ability of classifier that how effectively classier distinguish between classes.

B. Confusion Matrix

We have aslo consiodered confusion matrix that is used as summary of prediction results of classfication problem. The numbers as a counts of correct and incorrect prediction in terms of classes summarized in this matrix. Figure 2 is a confusion matrix of our heart problem that we have addressed in this article.

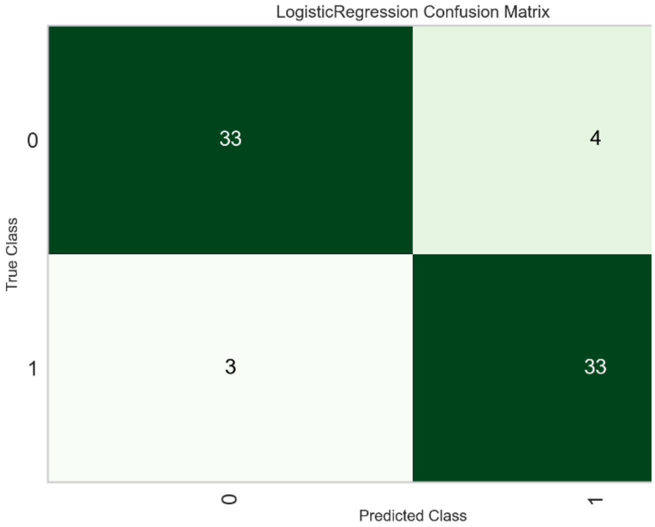

Fig. 2. Confusion Matrix of Heart Failure Problem

C. Logistic Regression Decision Boundary

Logisitic Regression blongs to linear models. Fitted Equation line with dimnension of n that denote the considered features in temrs of sigmoid transframtion is considered in logistics reression. This line denotes the class probabilities and line known as decision boundary that is shown in Figure 3.

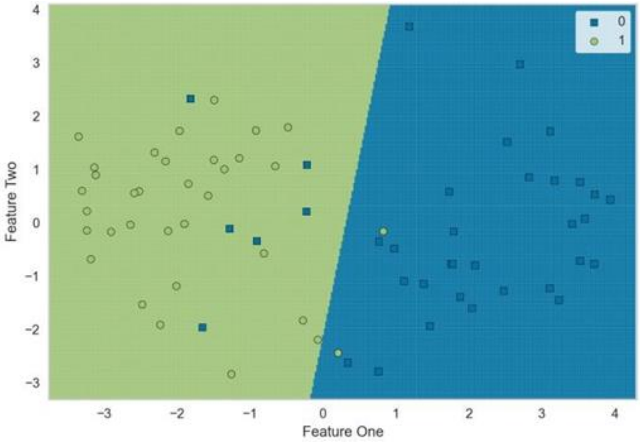

Fig. 3. Logistic Regression Decision Boundary

D. Logistic Regression Learning Curve

Learning curves plot training and validation loss of the model under training by adding the training samples in incremented manner. Learning Curve shows that where our model is suffering is it suffering in high biased (Underfitting) or high vairance (Overfitting), it aslo shows that should we need to considred more data to solve underfitting and overfiting problem. In case a model is overfitting then adding additional samples would enhance performance on unseen data. If a model is undefitted then adding training samples does not help.

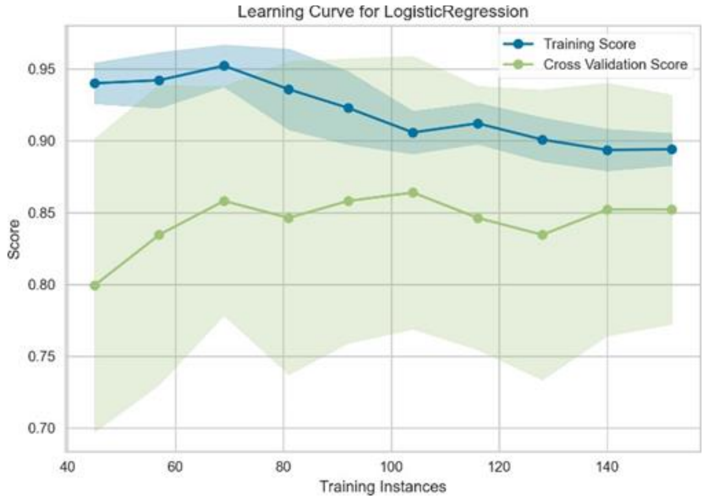

Fig. 4. Logistic Regression Learning Curve

E. Logistic Regression Lift Curve

Lift curve demosntrate the performace of considered classifers against random classifier. It shows the curves for true postive instances in terms thresholds classfiers or number of postive prediction, Figure 5 is demonstration of lift curve.

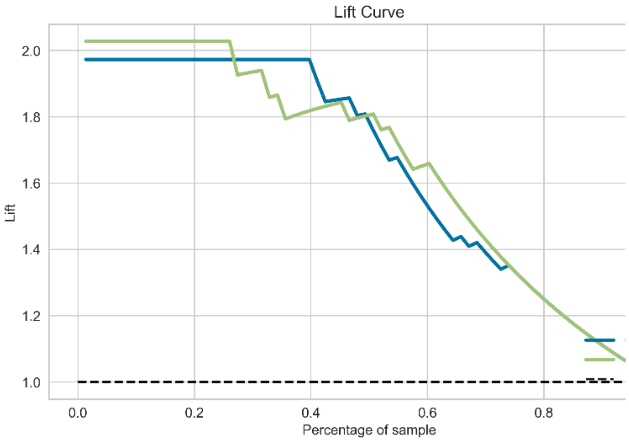

Fig. 5. Logistic Regression Lift Curve

F. Precision-Recall Curve Logistic Regression

The trade off summary using different thresholds of predicted postitve value and true postive rate. The curve consist on True postive rate (TPR) that is so called recall or sensitivity on X-axis and precison on Y-Axis. Precison(also called Positive Predicted Value - PPV) is specifically addressed that how relevant the results are that retrieved and more important. Figure 6 is PR Curve of logistisc regression for Cardiovascular problem.

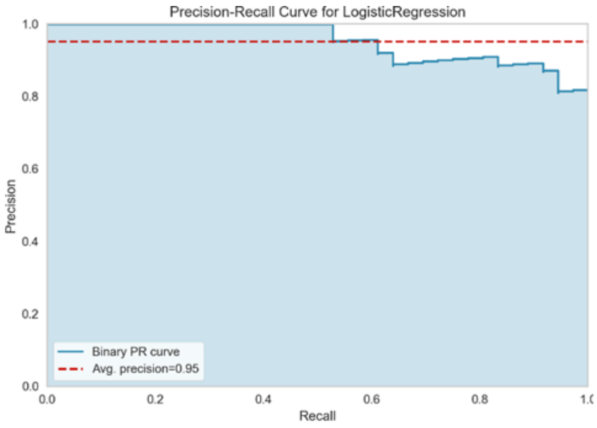

Fig. 6. PR Curve Logistic Regression

G. Threshold Curve Logistic Regression

Logistic regression is a claffier that firstly calculate probability, for example if model predict 0.95 probability for ham spam identification then it means it is likely to spam. The logistic probablity is map with binary problem with the help of threshold value and in most cases it is 0.5 but it is vary according to nature of problem. Figure 7 is of Threshold Curve of logistisc regression for Cardiovascular problem.

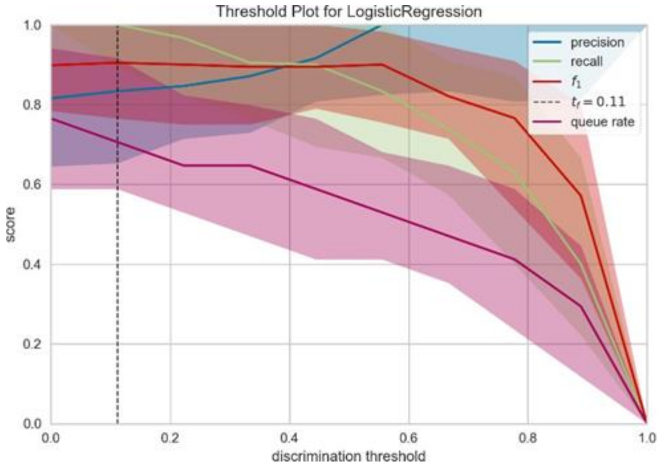

Fig. 7. Logistic Regression Thresholding

H. Cumulative Gains Curve Logistic Regression

Assess the performance of the model and compare it with random pick is considered in Cumulative Gains Curve. According to the model displaying the percentage of target reaching in terms of certain percentage of population with efective probability come into existance with culmulative gains curve. Figure 8 is of Cumulative Gains Curve of logistisc regression for Cardiovascular problem.

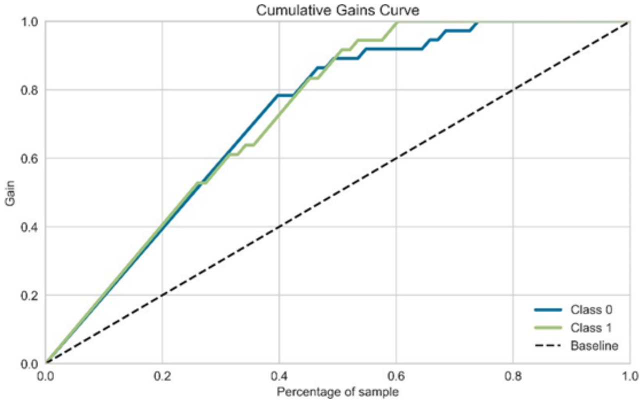

Fig. 8. Cumulative Gains Curve Logistic Regression

I. Robustness of Logistic Regression

To explore the robustness of study and model we have perform the comparison of logistic regression model with other machine learning models. Table 1 is a comparsion table in terms if basic evaluation metrices. Our model has clealy out weighed Naïve Bayes and Extra Tree Classifier on the basis of Accuracy, Precsion and better F1 score.

Table 1 is a clear justification that Logistic Regression model performance is effective than other well-known machine learning models in terms of evaluation scores. Figure 9 is graphical representation of top 5 machine learning models comparison for heart disease problem.

TABLE 1. Machine Learning Models Comparison

Model	**Accuracy**	**Precision**	**Recall**	**F1**	**AUC**
Logistic Regression	85.22 %	87.24 %	89.18 %	87.92 %	89.79 %
Naive Bayes	84.63 %	84.11 %	92.27 %	87.76 %	85.87 %
Extra Tree Classifier	84.63 %	86.61 %	89.27 %	87.53 %	88.44 %
Ridge Classifier	84.04 %	85.97 %	88.27 %	86.84 %	0.0 %
Linear Discriminant Analysis	84.04 %	85.97 %	88.27 %	86.87 %	89.29 %
Fast Large Margin	83.7 %	84.6 %	79.6 %	81.8 %	88.5 %
Random Forest Tree	82.87 %	84.45 %	88.27 %	86.12 %	88.48 %
Generalized Linear Model	84.0 %	91.3 %	72.5 %	80.3 %	90.4 %
Light Gredient Boosting	82.28 %	83.90 %	87.18 %	85.17 %	89.26 %
Gredient Boosting Classifier	79.93 %	80.45 %	88.27 %	83.87 %	87.49 %
Decision Tree Classifier	79.34 %	82.64 %	84.18 %	82.77 %	78.04 %
Ada Boost Classifier	78.12 %	79.41 %	87.27 %	82.62 %	82.29 %
Quadratic Discriminant Analysis	62.17 %	75.51 %	55.64 %	60.87 %	64.13 %
SVM - Linear Kernel	61.07 %	61.12 %	55.82 %	55.13 %	0.0000 %
K Neighbors Classifier	58.64 %	63.02 %	74.45 %	67.87 %	62.06 %

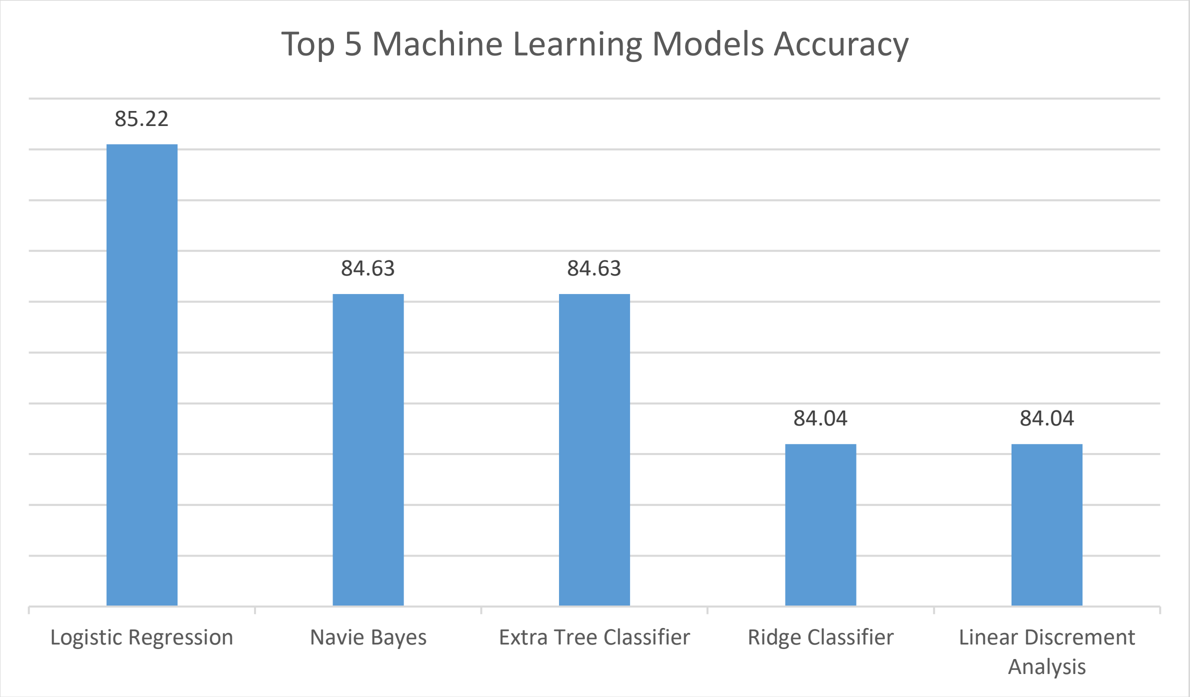

Fig. 9. Top 5 - Machine Learning Models Accuracy

IV. CONCLUSION

Machine Learning playing an important role for disease diagnosis e.g., breast cancer disease, Lungs Failure Diagnosis, Heart Failure Diagnosis and many more. In this work, Logistic Regression classifier performance is effective than other machine learning classifiers in terms of evaluation scores for Cardiovascular/Heart failure diagnosis. Further in this study, comparison of machine learning models is presented for the robustness of logistic regression model. The study has worth in terms of human life as early prediction of heart failure can save a life. The effective score of logisitc regression can help medical staff Logistic Regression Navie Bayes Extra Tree Classifier Ridge Classifier Linear Discriminant Analysis to diagnose the heart failure with symptoms of heart failure. In the future, we have decided to work with CT-Scans images of heart for heart failureidentification.

References

[1] “Breast cancer.” https://www.who.int/news-room/fact-sheets/detail/breast-cancer (accessed Jun. 11, 2021).

[2] “How Is Breast Cancer Diagnosed? | CDC.” https://www.cdc.gov/cancer/breast/basic_info/dia gnosis.htm (accessed Jun. 11, 2021).

[3] E. A. Bayrak, P. Kirci, and T. Ensari, “Comparison of machine learning methods for breast cancer diagnosis,” Apr. 2019, doi: 10.1109/EBBT.2019.8741990.

[4] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis,” in Procedia Computer Science, Jan. 2016, vol. 83, pp. 1064–1069, doi: 10.1016/j.procs.2016.04.224.

[5] H. L. Chen, B. Yang, J. Liu, and D. Y. Liu, “A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis,” Expert Syst. Appl., vol. 38, no. 7, pp. 9014–9022, Jul. 2011, doi: 10.1016/j.eswa.2011.01.120.

[6] M. W. Huang, C. W. Chen, W. C. Lin, S. W. Ke, and C. F. Tsai, “SVM and SVM ensembles in breast cancer prediction,” PLoS One, vol. 12, no. 1, p. e0161501, Jan. 2017, doi: 10.1371/journal.pone.0161501.

[7] K. Kourou, T. P. Exarchos, K. P. Exarchos,

M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” Computational and Structural Biotechnology Journal, vol. 13. Elsevier, pp. 8– 17, Jan. 01, 2015, doi: 10.1016/j.csbj.2014.11.005.

[8] “A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning.” https://machinelearningmastery.com/gentle- introduction-gradient-boosting-algorithm- machine-learning/ (accessed Jun. 16, 2021.

How to cite: M.Asad Arshed, F.Riaz. Machine Learning for High Risk Cardiovascular Patient Identification, Journal of Distributed Computing and Systems(JDCS), Vol 4, Issue 2, Pages 34-39, 2021