International Journal on Emerging Technologies 11(3): 377-382(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Performance Evaluation of Supervised and Unsupervised Machine Learning Algorithms by Predicting Cardiovascular Heart Disease

K. Kavitha[1] and K. Kala[2]
[1]Assistant Professor, Department of Computer Science,
Mother Teresa Women's University, Kodaikanal, Tamilnadu, India,
[2]Associate Professor, Department of Computer Science,
Nachiappa Swamigal Arts and Science College, Koviloor, Tamilnadu, India.

(Corresponding author: K. Kavitha)
(Received 20 February 2020, Revised 14 April 2020, Accepted 16 April 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: In the recent years, Majority of population suffers with cardiovascular heart disease worldwide. Prediction of Cardiac disease at an early stage supports to avoid death rates. Early detection of Cardiac disease helps to provide proper treatments and reduce the mortality rates drastically. Medical field has focused to create a model for diagnosing the disease at an early stage to avoid critical situation. Machine Learning plays a vital role in this area nowadays. This research has proposed a model to diagnose cardiac disease in early stage with an effective manner. Many researchers have designed various Machine Learning model to predict accuracy of identifying cardio vascular disease which has considered more features. Researchers have been examined the accuracy with minimum number of training sets. The proposed method trained with 7000 records and predicts the accuracy based on the designed model. RFE and Feature Importance techniques have been applied for identifying the important parameters based on the ranking as well as Root Mean Square Error (RMSE) value in this study. The proposed model was evaluated with various machine learning algorithms like SVM, GaussianNB, KNN and K-Means. The performance of these algorithms could be evaluated using the following measures such as precision, recall, f1-score and accuracy. Based on these measures, KNN Classifier achieved better accuracy 62% for larger datasets than others and SVM achieved 64% accuracy for smaller dataset in Supervised method and K-Means Clustering algorithm achieved the better accuracy 50% in unsupervised method. Time Complexity also measured for each algorithm and compared with each other. It has clearly proven that KNN classifier algorithm achieved better accuracy with minimum computation time.

Keywords: Cardiovascular Disease [CVD], K-Nearest Neighbor [KNN], Machine Learning [ML], Naïve Bayes [NB], Root Mean Square Error [RMSE], Support Vector Machine [SVM]

I. INTRODUCTION

Heart problem is a foremost health issue that causes death worldwide severely. Cardiovascular Heart disease is one kind of heart disease that affects the functioning of heart. Many parts of the body will be affected such as kidney disorder, brain disorder etc when the heart is not functioning properly. The function of heart is to push and circulate the necessary blood to all parts of our body through arteries. If heart is not functioning well means, hearts is not able to push the necessary blood to other part which affects whole body. Due to heart diseases, WHO stated that millions of people are affects and die frequently each year. The major confront in medicine field today is that early stage diagnose of disease is too tough to recognize. Diagnosing disease globally and provides proper treatment to the affected patients which ensures the quality of life and service.

Medical records and historical medical data's maintained by experts which are larger in size with many dissimilarities. The fact is real world data forced be incomplete and occurrences of noise. It needs to be considered, focused to eliminate the dissimilar things in the dataset and replace the incomplete missing terms by some advanced techniques. An effective tool is required to diagnose the high risk diseases which recognize the patients on time based on the historical medical records. Specialist in medicine field created huge number of Patient's record in the dataset. It needs to be handled and knowledge based predictions would be necessary.

Early diagnosis of heart disease and continuous observation of suffered patients by medicine experts reduces the mortality rate. However, accurate prediction of disease and continuous monitory of patients is inappropriate in all timings. Recognising High risk heart patients are not reliable in traditional methods. Machine Learning Techniques fulfills the medicine field need through various classifier and predictive methods. In the previous study, prediction accuracy has performed with minimum training sets and maximum parameters which consumes time complexity and reduces the accuracy rate. The proposed methodology has implemented with RFE and feature importance techniques to highlights the important features and removes the un-important features which are essential for predicting the accuracy. This model has been analysed with 70000 records with parameters.

This research study considered this issue as a major thing to solve the problem immediately. The proposed model predicts the early diagnose of cardiovascular heart disease effectively with better accuracy through

machine learning algorithms. Various regression, classifier and prediction methods were applied in this proposed model to diagnose the heart disease accurately.

II. LITERATURE REVIEW

Haq *et al.*, (2018) created an intelligent model for predicting the heart disease through machine learning algorithm. The performance evaluated using various classifier metrics by machine learning algorithms. Researchers trained the data using six different classifier algorithms and predicted the better accuracy [1]. Heart disease has many varieties which includes heart attack, cardiovascular heart disease, stroke, knocks, coronory disease etc. [2].

Golande and Kumar (2019) discussed the major issues in heart diseases alongwith various machine learning techniques, algorithms and tools to predict the cardiovascular disease with better accuracy score [3].

Aslandogan and Mahajani (2004) trained the model with three machine learning classifiers such as KNN, Decision Tree and Naïve Bayes Classifier. The research study proved that the combined idea highlighted the improved accuracy score for prediction [4].

Dangare and Apte (2012) established a framework for heart diagnose with the support of 13 parameters. The performance were evaluated using decision tree, Neural network and naïve bayes classifier to analyse the heart disease and proved that neural network classifier achieved 100% accuracy than others [5]. Nashif *et al.*, (2018) designed a real time CVD monitoring system for health associate with machine learning algorithms [6]. The proposed system would be very supportive to the medicine practitioners for monitoring and visualizing the patient's health and track the health records through realtime sensor. The performance measures evaluated with various ML algorithms, finally SVM achieved 97.53% accuracy in this framework. Rajesh *et al.,* (2018) applied combination of Machine Learning algorithms for predicting heart diseases. The performance result showed that naïve bayes algorithm got better accuracy in small dataset and decision tree is better in larger dataset [7]. Wiharto *et al.*, (2016) demonstrated an intelligent framework for identifying the heart infection by learning vector NN qualtization system which predicts the nearness of coronary infirmity patients [8].

Back Propagation MLP of ANN is used to predict the heart disease. Heart disease data has been collected from UCI repository. The obtained accuracy compared with the existing models and stated to be improved [11]. Convolution Neural Network was introduced to monitor the heart cycles with various ECG signals which generate features in the testing phase of the targeters [12, 15]. Researchers have given prove that the prediction and classification of heart disease by machine learning techniques provides higher accuracy [13, 14]. Vivekanandan and Iyengar (2017) used Differential Evolution algorithm to identify optimal feature selection in the dataset for prediction of heart disease. The study examined with 300 records and 13 critical attributes and has been reduced with 9 attributes for further processing. The DE model integrated fuzzy AHP and FFNN method to predict the accuracy [16].

Dataset: This research study collected CVD dataset from Kaggle repository. The data's were updated on 2018 by Svetlana Ulianova which consists of 70,000 samples with 13 variables. This study was conducted in Spyder (Anaconda) which supports all Machine Learning features. The parameters used in this dataset are represented in below figure

id ID number
age in days
gender1 - women, 2 - men
heightcm
weightkg
ap_hiSystolic blood pressure
ap_loDiastolic blood pressure
cholesterol1: normal, 2: above normal, 3: well above normal
gluc1: normal, 2: above normal, 3: well above normal
smokewhether patient smokes or not
alcoBinary feature
activeBinary feature
cardioTarget variable

Fig. 1. Dataset Description.

III. PROPOSED METHODOLOGY

The proposed model has the following stages
– Preprocessing Stage : Dealt with Missing values and eliminate Outliers
– Feature Selection: Best features will be identified based on ML techniques
– Train & Test data : Split the dataset by 80% and 20% and train the model using 80% and test it by test data
– Model Establishment & Assessment: Model built with the support of Spyder in Python and evaluate by ML techniques such as Lasso, Ridge, Random Forest etc
– Prediction Accuracy Classifier: Trained data will be applied in supervised and unsupervised models to calculate the accuracy and measure the performance through metrics.

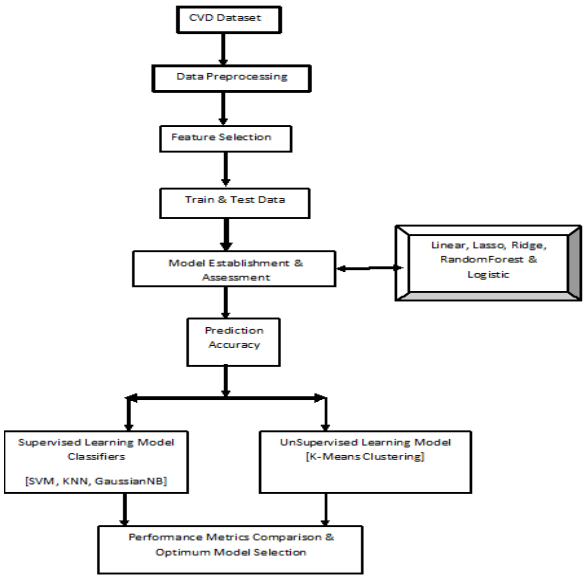


Fig. 2. Proposed Methodology.

A. Preprocessing Stage

Initially, the CVD dataset has considered that deals with the missing values and outliers in the corresponding dataframe. Because, NaN values are not considered by proposed ML model. So it has to be replaced by some conversion method. Mean function were calculated and used to replace the null values through the identified mean and replace the same in the missing places. Irrelevant data's were also has to be eliminated in the dataset for efficient processing. In this case, outliers has been noted to remove the irrelevant records in the prescribed CVD dataset.

B. Feature Selection

Machine Learning suggests various techniques to identify the importance of features in the dataset. It can be used to identify the positive, negative and zero correlation. Threshold limit has fixed for this dataset which is 0.05. Based on the threshold value (p<threshold), select the features which has given more importance for further processing and eliminate the lesser importance features (p>threshold). This research also evaluated RFE (Redundant Feature Elimination) and Feature Importance techniques for further identification of optimal feature importance using ranking for each parameters. Here, eight parameters have been selected out of 14 for further processing such as age, weight, aplo_Diastolic Blood Pressure,Cholestrol, Glucose, Smoke, active binary, CardioTarget

C. Train & Test Data

The dataset has been divided into two subgroups namely train and test dataset. Train group and test group were split by 80%, 20%. CVD dataset consists of 70,000 records in which 56000 for train and 14000 for test dataset.

D. Model Establishment and Assessment

The main intention of establishing this model was to predict the most accurate score and identify the error rates of the progression of CVD dataset. Based on the predicted root mean square error value, confidence levels of predictive features were identified easily.

To improve the training & test data and to minimize the RMSE value, model features were fixed and split data were performed with random number of samplings. The average performance metrics of the proposed model were evaluated by different machine learning techniques such as Linear, Logistic, Edge, Lasso and Random Forest. This study used Spyder (in Python) to build the proposed model and evaluate the performance metrics. The following table illustrates the training dataset (80%) and test dataset (20%) error rate of various machine learning techniques.

Table 1: Error Rate Measures.

Machine Learning Technique	MSE	RMSE	Prediction Score
Linear	0.2212	0.4	0.1157
Ridge	0.2269	0.46	0.1157
Lasso	0.2500	0.50	0.1157
ElasticNet	0.2300	0.479	0.1157
Random Forest	0.2329	0.482	0.1157

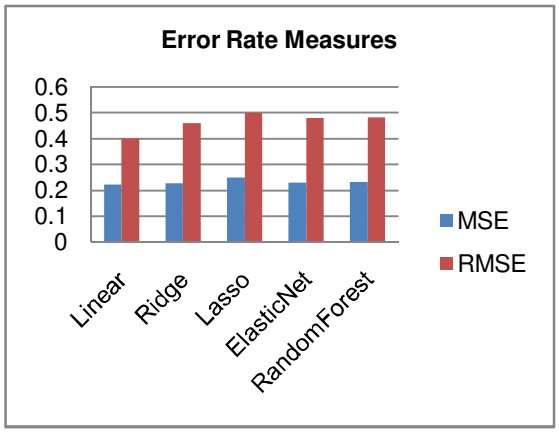


Fig. 3. Error Rate Comparison Chart.

E. Prediction Accuracy Classifiers

Prediction accuracy can be measured using various machine learning algorithms such as SVM, GaussianNB, KNN and K-Means. SVM classification method is applicable for both linear and non-linear domain values. It fixes the hyper plane with the help of support vector and the parameter-margin which is used to map with higher dimensions. GaussianNB classifier considers every parameter as independent parameter which supports to predict the accurate score. KNN classifier is the fundamental classification method in machine learning techniques. Exact prediction ie) exact similarity and location between the parameters can be decided with the support of given training dataset. By using this method, examine the similarities between parameters easily by Eucledian distance. K-Means clustering used to group the values based on the nearest classifiers ie similarity between the parameters provides better accuracy. It forms the groups based on k-variable. The proposed model was trained with 8 features and different samples. The performance metrics were evaluated using the following measures such as precision, recall, f1-score and accuracy.

Table 2: SVM Classifier Performance Metrics.

Samples	Precision	Recall	F1-score	Accuracy	Timings (in seconds)
10000	0.64	0.60	0.62	0.64	87.07
20000	0.66	0.62	0.64	0.63	254.526
30000	0.62	0.69	0.65	0.63	366.702
40000	0.64	0.61	0.62	0.63	563.3253
50000	0.64	0.64	0.64	0.64	729.7702
60000	0.65	0.59	0.62	0.64	1182.5633
70000	0.64	0.64	0.64	0.64	1209.4115

Table 3: GaussianNB Classifier- Performance Metrics (average 59%, 0.09 seconds).

Samples	Precision	Recall	F1-score	Accuracy	Timings (in seconds)
10000	0.54	0.93	0.68	0.57	0.1243
20000	0.59	0.88	0.77	0.61	0.0602
30000	0.55	0.87	0.67	0.58	0.0549
40000	0.58	0.86	0.69	0.61	0.0705
50000	0.58	0.85	0.69	0.61	0.1024
60000	0.56	0.85	0.67	0.59	0.1021
70000	0.57	0.84	0.68	0.60	0.1163

Table 4: K-Nearest Neighbor Classifier- Performance Metrics (average 0.62%, 1.81 seconds).

Samples	Precision	Recall	F1-score	Accuracy	Timings (in seconds)
10000	0.61	0.64	0.62	0.62	0.2664
20000	0.63	0.65	0.64	0.61	0.6191
30000	0.62	0.64	0.63	0.62	0.8746
40000	0.63	0.64	0.63	0.63	1.2714
50000	0.62	0.64	0.63	0.62	2.1508
60000	0.62	0.65	0.63	0.62	3.0522
70000	0.63	0.63	0.63	0.63	4.4523

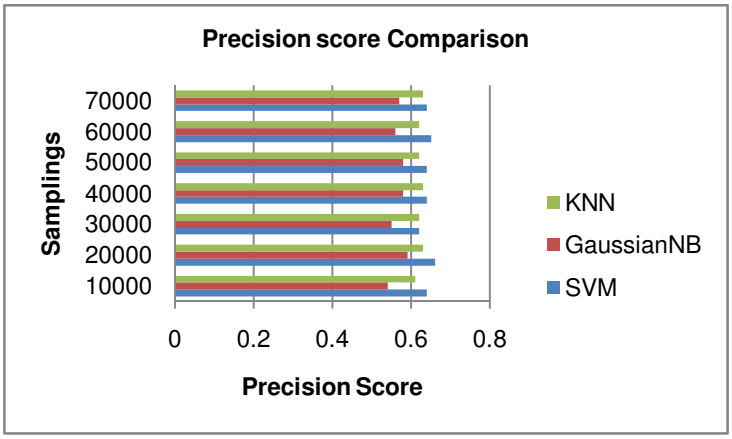
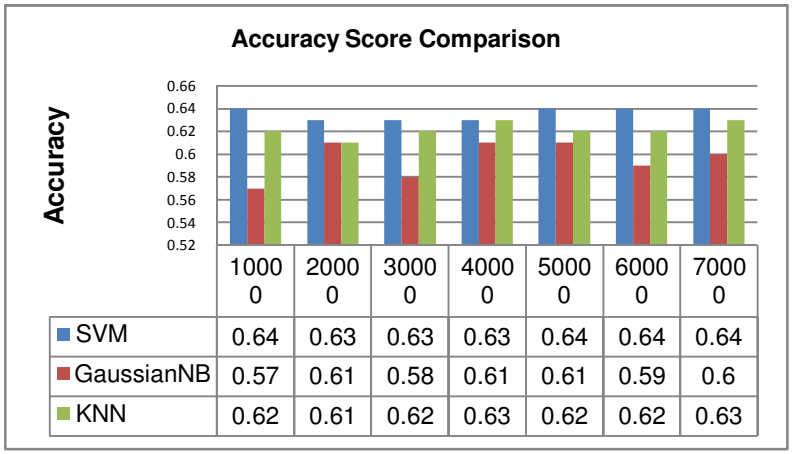


Fig. 4. Precision Score Comparison for each samplings.



	10000	20000	30000	40000	50000	60000	70000
SVM	0.64	0.63	0.63	0.63	0.64	0.64	0.64
GaussianNB	0.57	0.61	0.58	0.61	0.61	0.59	0.6
KNN	0.62	0.61	0.62	0.63	0.62	0.62	0.63

Fig. 5. Comparison of accuracy with different samplings.

The above metrics table showed the performance measures for each samplings start from 10,000 to 70,000. The above chart clearly showed that SVM achieved better accuracy than other algorithms using smaller dataset. Based on the average accuracy of supervised learning and timings, K-Nearest Neighbor classifier algorithm provides better accuracy 62% in larger dataset with 0.62 of precision, 0.64 of recall and 0.63 of f1-score with least time consuming than other algorithms which provides better accuracy than the existing model [11]. The existing model has been examined with minimum number of training sets i.e. 300 records only analysed [12].

K-Means Clustering Algorithm: K-Means clustering is the unsupervised algorithm which class labels are not known in this dataset. Based on the nearest classifiers (neighbors) ie) similarity between the parameters and values are identified. Similarity measures provides

better accuracy and the major highlight is finding the clusters in the data with number of groups which can be represented through K-variables. This algorithm tested with CVD dataset in different samplings from 10,000 to 70,000. K-means classifier achieved 50% accuracy in 70,000 samplings and 10,000 samplings and the remaining varies from 0.25 to 0.40 accordingly. After Performing cluster option, the dataset can be represented in the below graph.

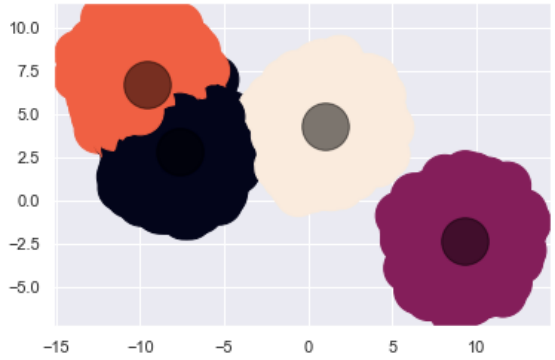


Fig. 6. K-Means Clustering with 4- Cluster for CVD dataset.

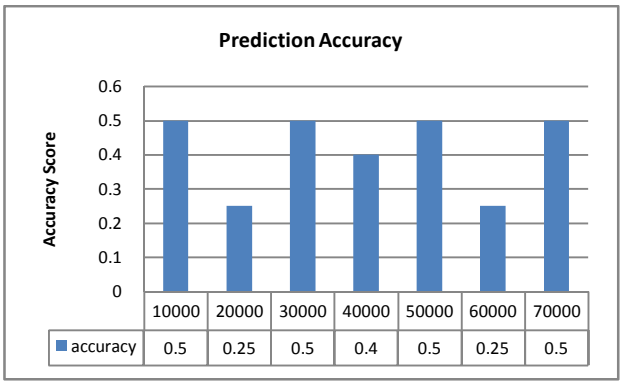


Fig. 7. Prediction Accuracy for each samplings.

IV. CONCLUSION AND FUTURE WORK

The proposed research work considered CVD dataset collected from kaggle repository and evaluated using Machine Learning techniques. The proposed model have crossed all the preprocessing methods and identified the valuable features by feature reduction techniques for diagnosing the cardiac diseases on time. This model was appraised through ML techniques such as Lasso, Ridge, Linear, ElasticNet, Random Forest which ensured the model as perfect. Finally the proposed model was reviewed by using SVM, GaussianNB and KNNeighbors and K-Means Classifier Models with different samplings. The accuracy score and time has measured for each model and has proven that KNN classifier Model could be the better model to predict the accuracy 62% in unsupervised learning and 50% achieved in K-means supervised learning method. But still accuracy rate has to be improvised and model need to be trained with many varieties of features and samplings. In future, Real time dataset will be framed and considered to train the proposed model with many varieties of features and improve the performance level in supervised as well as unsupervised learning which enables the detection of severity of the coronary disease.

Conflict of Interest. There is no conflict of interest.

ACKNOWLEDGEMENT

Dr. Kavitha. K, Assistant Professor, Mother Teresa Women's University, Kodaikanal. She is having 17 years of teaching and 10 years of research experience. She has published more than twenty papers in International Journals and presented many papers in National/International Conferences. She has published two book chapters. Her research interest is Data Mining, Cloud Computing, Data Science and Machine Learning.
Dr. Kala. K, Associate Professor, Nachiappa Swamigal arts and Science College, Koviloor. She is having 19 years of teaching and 10 years of research experience She has published more than twenty papers in International Journals and many papers in National/International Conferences. Her research interest are Data Mining, Cloud Computing, Machine Learning.

REFERENCES

[1]. Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, 1-21

[2]. Chetty, N., Vaisla, K. S., & Patil, N. (2015). An improved method for disease prediction using fuzzy approach. In *2015 Second International Conference on Advances in Computing and Communication Engineering*, 568-572.

[3]. Golande, A., & Kumar, P. (2019). Heart Disease Prediction Using Effective Machine Learning Techniques. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(1), 944-950.

[4]. Aslandogan, Y. A., Mahajani, G. A. & Taylor, S. (2004). Evidence combination in Medical Data Mining. International Conference on Information Technology: Coding and Computing(ITCC'04)Vol.2, Las Vegas, NV, 465-469.

[5]. Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. *International Journal of Computer Applications*, *47*(10), 44-48.

[6]. Nashif, S., Raihan, M. R., Islam, M. R., & Imam, M. H. (2018). Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system. *World Journal of Engineering and Technology*, *6*(4), 854-873.

[7]. Rajesh N,, Maneesha, Shaik Hafeez, & Krishna, H. (2018). Prediction of Heart Disease using Machine Learning Algorithms. *International Journal of Engineering & Technology*, *7*, 363-366.

[8]. Wiharto, W., Kusnanto, H., & Herianto, H. (2016). Intelligence system for diagnosis level of coronary heart disease with K-star algorithm. *Healthcare informatics research*, *22*(1), 30-38.

[9]. Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS international conference on computer systems and applications*, 108-115.

[10]. Vazirani, H., Kala, R., Shukla, A., & Tiwari, R. disease. *International Journal of Computer and Communication Technology*, *1*(2-4), 88-93.
[11]. Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. *Expert systems with applications*, *36*(4), 7675-7680.
[12]. Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications*, *40*(4), 1086-1093.
[13]. Ravish, D. K., Shanthi, K. J., Shenoy, N. R., & Nisargh, S. (2014). Heart function monitoring, prediction and prevention of heart attacks: Using artificial neural networks. In *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, 1-6.
(2010). Use of modular neural network for heart
[14]. Zhang, W., & Han, J. (2017). Towards heart sound classification without segmentation using convolutional neural network. In *2017 Computing in Cardiology,* 1-4.
[15]. Zaman, S., & Toufiq, R. (2017). Codon based back propagation neural network approach to classify hypertension gene sequences. In *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE),* 443-446.
[16]. Vivekanandan, T., & Iyengar, N. C. S. N. (2017). Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease. *Computers in biology and medicine*, *90*, 125-136.

How to cite this article: Kavitha, K. and Kala, K. (2020). Performance Evaluation of Supervised and Unsupervised Machine Learning Algorithms by Predicting Cardiovascular Heart Disease. *International Journal on Emerging Technologies*, *11*(3): 377–382.