Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 7, Issue 10, October -2020

Classification of Healthcare Datasets through Supervised Machine Learning Algorithms

Ravindra Singh Sapera[1], Shrwan Ram[2]

[1] Computer Science Department, M.B.M. Engineering CollegeJodhpur
[2] Associate Professor Computer Science Departments, M.B.M. Engineering CollegeJodhpur

Abstract *The work centered on methodologies based on machine learning to develop applications that are capable of recognizing and disseminating health information. In this paper, a different type of supervised machine learning approach is used for the classification. Analyzing the machine learning algorithms and finding out the most appropriate algorithms for healthcare data. In this study, designed a classification system using a Decision tree, Naïve Bayes Support Vector Machine, and KNN for medical data classification with various numbers of attributes and instances. Its include two type classification namely present or absence data distribution from the Cleveland heart disease data set. The experiment outcomes positively demonstrate that the decision tree classifier is effective in undertaking healthcare data classification tasks.*

Keywords– *Classification, Machine learning, decision tree, naïve Bayes, support vector machine algorithms, heart disease dataset.*

I. INTRODUCTION

Machine Learning is the ability of machines to adopt human behavior, in which a machine composed of different algorithms using these algorithms chooses its own choice and provides the user with the outcome or output. Machine learning is the skill of learning machines, where a machine is designed with certain algorithms from which it can make its own choices and give the user the answer.Machine Learning Algorithms are a step-by-step method for extracting information from the data set without relying on a patch programmer. Such data is useful in predicting the output of a given input. Inside the data collection, i.e. given to it.

Supervised learning algorithms aim to model relationships and dependencies between the output of the goals and the input features so that the output values for new data can be predicted based on certain relationships that have been learned from previous data sets.Healthcare data classification is a challenging task in the field of medical research. Both for a patient and the doctor, the medical record is very useful. The medical record would usually assist the doctor in classifying the illnesses, diagnosing, and treating the patient properly. In recent days, the volume of Healthcare data is huge. Therefore, the seriousness of manual diseases is difficult to identify and understand.This paper describes the classification of heart diseases through supervised machine learning.

1. **Healthcare dataset:**Cleveland's heart disease data set is a multivariate data set. It includes 76 attributes and 303 instances that range from Categorical, Integer, and True. However, the studies suggested involving the use of a subset of 13. The prediction area applies to the patient's existence of heart disease. The prediction field concentrated on simply attempting to distinguish the presence of diseased (value 2) data from non-diseased (value 1).

2. **Decision tree**: train classification decision tree to predict responses to data. Follow the decision of the root tree (beginning) node up to a leaf node to predict the answer. The leaf node contains the response. The value of one predictor (variable) is verified at each stage of a prediction. This tree predicts identifiers based on two predictors, x1 and x2. To predict, start at the top node. Check the values of the predictors in each decision to determine which branch to obey. When a leaf node is reached by the branches, the data is labeled as either form 0 or 1.

3. **Naive Bayes:** Bayes 'theorem (often called Bayes' law after Thomas Bayes), in probability theory, compares the conditional and marginal probabilities of two random events. Often used for calculating subsequent probabilities given observations. A naive classifier of Bayes is a concept that deals with a simple probabilistic classification based on applying the theorem of Bayes. Simply put, A Naïve Bayes classifier assumes that the presence (or absence) of a certain attribute of a class is irrelevant to the presence (or absence) of any other feature.

4. **Support Vector Machine:**Support Vector Machine: By finding the best hyper-plane that separates all data points of one class from those of another class, an SVM classifies information. The support vector machine algorithm performs classification by finding the hyper-plane or classifier which maximizes the margin between two classes. Two groups are divided by Hyper-plane. Classification can be seen as a task of separating classes in the space of features.

II. PROPOSED METHODOLOGY

In this work, a machine learning classifying healthcare data set using supervised machine learning. In this, a different supervised machine learning algorithm is compared using a heart disease data set to measure the performance of algorithms with different parameters. In this study Decision tree, Naive Bayes, Support Vector Machine, Algorithms are used for the classification of the healthcare dataset.

In this work, there are 2 Experimental Setup is used to measure the performance of supervised machine learning Algorithms for the healthcare dataset.

- Compare different supervised machine learning algorithms performance using heart disease data set for 8 attributes and 809 instances without 10-fold cross-validation.
- Compare different supervised machine learning algorithms performance using heart disease data set for 13 attributes and 809 instances with 10-fold cross-validation.

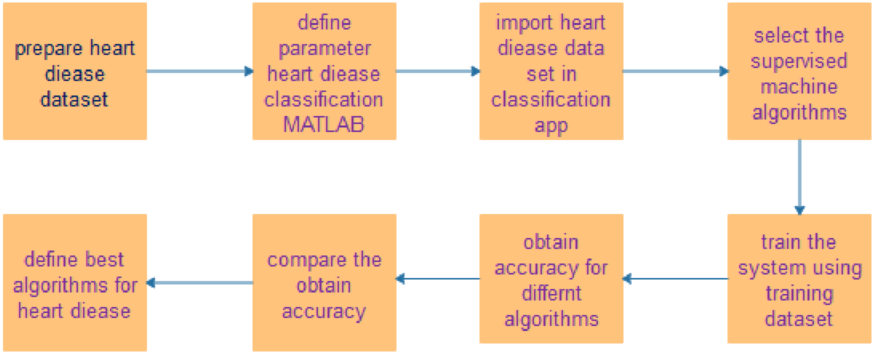
prepare heart diease dataset
define parameter heart diease classification MATLAB
import heart diease data set in classification app
select the supervised machine algorithms
train the system using training dataset
obtain accuracy for differnt algorithms
compare the obtain accuracy
define best algorithms for heart diease

Figure 1: Proposed Methodology

The flow diagram of the Proposed Methodology is shown in Figure 1, using this methodology, measure the performance of supervised machine learning algorithms, for this, obtain the accuracy of the confusion matrix, ROC curve, and compare them to get the resultant graph to show best algorithms for heart disease data set.

- In the first step heart disease data set to prepare, there is a heartdisease.xls file to include.in this dataset; we use 13 attributes and 809 instances for the heart disease dataset.

S.No.	Attributes	Description
1.	Age	Age in years
2.	Sex	Sex (1=male;0=female)
3.	Cp	Chest pain type (1=typical angina; 2=atypical angina; 3=non-anginal pain; 4= asymptomatic)
4.	Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
5.	Chol	Serum cholesterol in mg/dl
6.	Fbs	Fasting blood sugar (>120 mg/dl) (1 = true; 0 = false)
7.	Restecg	Resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV))
8.	Thalach	Maximum heart rate achieved
9.	Exang	Exercise induced angina (1 = yes; 0 = no)
10.	Oldpeak	ST depression induced by exercise relative to rest
11.	Slope	Slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
12.	Ca	Number of major vessels (0-3) colored by fluoroscopy
13.	Thal	(3 = normal; 6 = fixed defect; 7 = reversable defect)

Table 5.1Attributes ofcleveland heart disease data set

- In the second step define the parameter of heart disease data set the classification to MATLAB tool.
- In the third step importing the heart disease data to the classification learner app and after that select the supervised machine learning algorithms one by one.
- In the next step train the system using the training data set.
- At the last, obtain the accuracies for different supervised machine learning algorithms and compare these accuracies for performance measurement using confusion matrix, ROC curve, and scatter plot and then get the best accuracy of the algorithms for the heart disease dataset.

III IMPLEMENTED ALGORITHMS AND RESULT

The Supervised Machine Learning Algorithms implemented for classifying the Heart disease data set. This Experimental consists of 3 algorithms which are Decision tree, Naïve Bayes, Support Vector Machine, are Classify the heart disease dataset. Here implemented all algorithms in 2 Experimental to compare the best accuracy of algorithms.

In the first Experimental, used 8 attributes of the heart disease dataset without 10-fold cross-validation trains the data set.so implement the algorithms I give the true value of the data set and prediction value of the dataset.

confusion matrix also shows the true class or predicted class in which data is train than they give the number of observation dataset in matrix form where diagonal cells are blue, the classifier has classified observations of this true class are classified correctlyand the True Positive rate is the proportion of correctly classified observations per true class. The False Negative Rate is the proportion of incorrectly classified observations per true class.

ROC curve See the operating characteristic of the receiver (ROC) curve showing real and false-positive rates. For the currently selected trained classifier, the ROC curve indicates a true positive rate versus a false-positive rate. So we compare all classifiers and get the best accuracy. They show the values of the false positive rate (FPR) and the true positive rate (TPR) for the currently selected classifier. Now I compare all the results by graph and table and find out the best success rate of algorithms to the classification of heart disease data that present or absence.

In the Second Experimental, used 13 attributes of the heart disease dataset with 10-fold cross-validation trains the data set. Because I structure not give good results and taking much time to train the data set and implement so used 13 attributes of the heart disease dataset and applied 10-fold validation. After that implement all algorithms obtain and comparing the result for the best success rate and found a better result comparing I Experimental there is shown in the graph and table in the result section.

After the train, the system using training datasets got the following results for both the Experimental [I and II]. Shown the results of Experimental I, where accuracy values obtained by selecting algorithms. Where used 8 attributes of the heart disease dataset and no cross-validation is used.

		Predicted Class	
Classifier	**Actual Class**	**Absence**	**Presence**
Decision Tree	Absence	426	18
	Presence	19	346
Naïve Bayes	Absence	374	70
	Presence	63	302
Support Vector Machine	Absence	375	69
	Presence	68	297

Table 2: Confusion matrixfor classifying the absence and presence data of clevelandheart disease data

The above table shows the confusion matrix for classifying the absence and presence data in the actual class and predicted class in no. of observation data.

Classifier	Actual Class	TPR	FNR
Decision Tree	Absence	95.9	4.1
	Presence	94.8	5.2
Naïve Bayes	Absence	84.2	15.8
	Presence	82.7	17.3
Support Vector Machine	Absence	84.5	15.5
	Presence	81.4	18.6

Table 3: Confusion matrixfor classifying theTRP-true positive rate andFNR-falsenegative ratedata

The above table shows the result of a confusion matrix for the classification of true positive rate or false-negative rate ofthe actual class.

Classifier	ACC	TPR	TNR	Precision
Decision Tree	93.8	0.96	0.05	0.97
Naïve Bayes	83.6	0.84	0.17 .	0.90
Support Vector Machine	83.1	0.89	0.19	0.89

Table 4: Classificationrate(%) ofabsence and presence data of clevelandheart disease data

This table shows the result of the classification rate of the ROC curve they show the accuracy of all algorithms and The precision rate also.

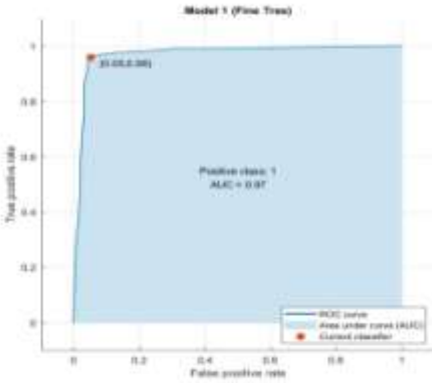

Figure 2(a): ROC curve for decision tree

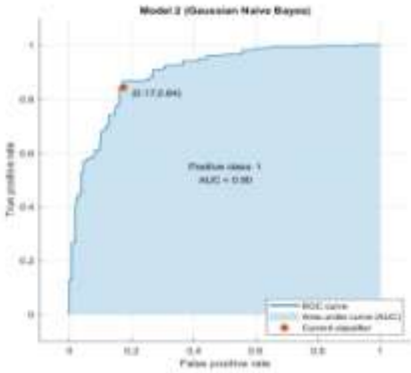

Figure 2(b): ROC curve for naïve bayes

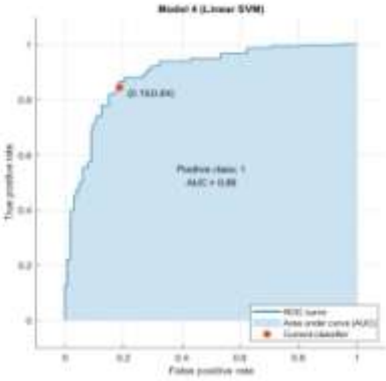

Figure 2(c): ROC curve for svm

I compare all accuracies of ROC graph a, b, and c and find out that heart disease dataset 8 attributes without cross-validation graph a Decision Tree gives better accuracy [93..80%], which is better than the accuracies of graph b Naïve Bayes[83.60%] and graph c SVM [83.10%].

Now Shown the results of Experimental II where accuracy values obtained by selecting algorithms. Where use 13 attributes of the heart disease dataset and 10-fold cross-validation is used.

Classifier	**Actual Class**	**Predicted Class**	
		Absence	**Presence**
Decision Tree	Absence	430	14
	Presence	13	352
Naïve Bayes	Absence	398	46
	Presence	57	308
Support Vector Machine	Absence	396	48
	Presence	63	296

Table 5: Confusion matrixfor classifying the absence and presence data of clevelandheart disease data

The above table shows the confusion matrix for classifying the absence and presence data in the actual class and predicted class in no. of observation data.

Classifier	**Actual Class**	**TPR**	**FNR**
Decision Tree	Absence	96.4	3.2
	Presence	96.8	3.6
Naïve Bayes	Absence	89.6	15.8
	Presence	84.4	17.3
Support Vector Machine	Absence	89.2	15.5
	Presence	81.1	18.6

Table 6: Confusion matrixfor classifying the TRP-true positive rate andFNR-false negative ratedata

The above table shows the result of a confusion matrix for the classification of true positive rate or false-negative rate
Of the actual class.

Classifier	**ACC**	**TPR**	**TNR**	**Precision**
Decision Tree	96.7	0.97	0.04	0.99
Naïve Bayes	86.4	0.90	0.16 .	0.92
Support Vector Machine	86.0	0.89	0.19	0.93

Table 7: Classificationrate(%) ofabsence and presence data of clevelandheart disease data

This table shows the result of the classification rate of the ROC curve they show the accuracy of all algorithms.

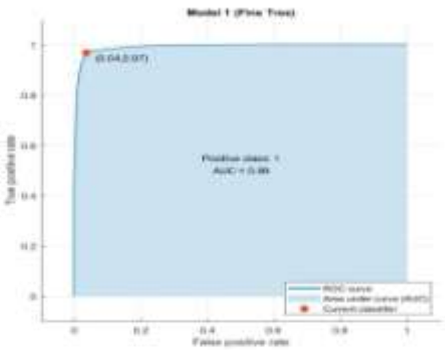

Figure 3(a): ROC curve for Decision tree

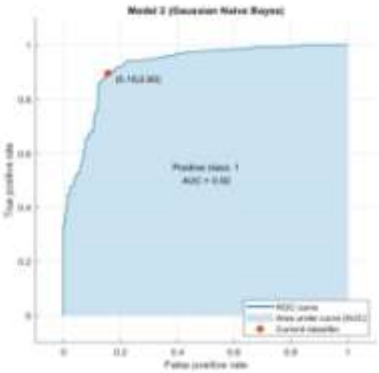

Figure 3(b): ROC curve for Naïve Bayes

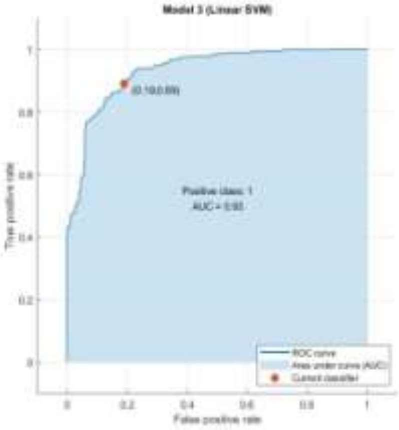

Figure 3(c): ROC curve for SVM

I compare all accuracies of ROC graph a, b, and c and find out that heart disease with 13 attributes used cross-validation graph a Decision tree gives better accuracy [96.70%], which is better than accuracies of graph b Naïve Bayes[86.40%] and graph c SVM [86.00%].

Based on both Experimental [I and II] results, the Decision Tree classifier givesprominent results in the classification of heart disease dataset where the disease is present and absences with 96.70% accuracy.

IV. CONCLUSION

In this work, Decision tree, Naïve Bayes, and Support Vector Machine, machine learning algorithms were used to measure the performance evaluation of supervised machine learning classifiers to predict operational decisions of the cesarean category. The results of the experiment show that the Decision tree achieves the highest precision rates by correctly predicting 96.7 percent of instances. The main objective of this is to evaluate the most imperative techniques for machine learning and to propose the most effective technique for the classification of healthcare data sets. However, it is not possible to assume that one algorithm is always superior to another. Instead, it could be inferred that a specific method will significantly outperform other methods for a given issue under conditions. In the future, by further increasing the number of instances and adding some more core attributes, we can improve the efficiency of Caesarian section decision making.

V. ACKNOWLEDGEMENT

I would like to thank Shrwan RamAssociate Professor, Dept. of Computer Science, MBM Engineering College, for his motivation without which I would not be able to write this paper. We are grateful to our family and friends for their support.

VI. REFERENCES

[1] DhomseKanchan B., MahaleKishorM."Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis**"** International Conference on Global Trends in Signal Processing, Information Computing and Communication 2016.

[2] MrunmayiPatil,Vivian Brian Lobo2,PranavPuranik,AditiPawaskar,AdarshPai, "A Proposed Model for Lifestyle Disease Prediction Using Support Vector Machine", IEEE - 43488

[3] R. Saravanan, PothulaSujatha (ICICCS 2018), "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification".

[4] Indu Kumar, KiranDogra , ChetnaUtreja, PremlataYadav (ICICCT 2018), "A comparative study of supervised machine learning algorithms for stock market trend prediction".

[5] MahsaMoein,Mohammad Davarpanah,M. Ali Montazeri, "Classifying Ear Disorders Using Support Vector Machines", 2012 Second International Conference on Computational Intelligence and Natural Computing (CINC).

[6] Saranya M S, Selvi M, S.Ganapathy, Muthurajkumar S, L.Sai Ramesh and A. Kannan (ICoAC 2016), "Intelligent Medical Data Storage System Using Machine Learning Approach".

[7] Daniel Vieira and JaakkoHollmen, "Resource Frequency Prediction in Healthcare: machine learning approach," 2016 IEEE 29th International Symposium on Computer-Based Medical Systems.

[8] Shu-Xia lu, JieMeng, Gui-en Cao, **"**support vector machine based on a new reduced sample method," Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 11-14 July 2010.

[9] Majali et al., "Data Mining Techniques for Diagnosis and Prognosis of Cancer", International Journal of Advanced Research in Computer and Communication Engineering, vol.3, pp. 613-616, 2015.

[10] Jalpaiguri et al.," Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", Advances in Computational Sciences and Technology, vol.10, pp. 2137-2159, 2017.