Correlation of Three Unique Techniques for Arrangement of Coronary Illness

[1]Shalini R
UG scholar

[2]Sasithra S
UG scholar

[3]Preethi D
Assistant professor

[1,2,3]*Department of Electronics and Communication Engineering,*
Bannari Amman Institute of Technology Erode,TamilNadu,India.

ABSTRACT

The characterization of coronary illness patients is of extraordinary significance in cardiovascular infection determination. Various information mining strategies have been utilized so far by the specialists to help medicinal services experts in the conclusion of coronary illness. For this errand, numerous calculations have been proposed in the past couple of years. This research paper considers various regulated machine learning techniques for gathering of coronary ailment data and have played out a procedural connection of these. The utilization of Logistic Regression (LR) classifier, a Naïve Bayes (NB) classifier, and a Support Vector Machine (SVM) classifier over a broad game plan of coronary ailment data. The data used in this examination is the Cleveland Clinic Foundation Heart Disease Data Set open at UCI Machine Learning Repository. It is discovered that LR beated both Naive Bayes and SVM classifier, giving the best precision rate of accurately arranging most elevated number of cases. Likewise it was found that Naïve Bayes classifier accomplished a focused exhibition however the suspicion of typicality of the information is unequivocally disregarded.

Keywords: Comparison; Heart disease data set; Logistic Regression; Naïve Bayes; Support Vector Machine

INTRODUCTION

Now a-days, the number of individuals experiencing coronary illness is expanding radically. As per the WHO reasons for death synopsis tables, the complete number of passing because of cardiovascular infection has come to nearly 17.3 million out of a year. In any case, exact finding at an underlying stage pursued by proper treatment can spare colossal measure of lives. Sadly, right finding of coronary illness at an essential stage is a significant testing assignment due to complex relationship on different components.

Exact forecast of hazard factors which are related with cardiovascular sickness is fundamentally essential for the analysis and treatment of coronary illness. Among the current strategies, regulated learning techniques are the most well known in coronary illness conclusion. Factual investigation has recognized some hazard factors related with coronary illness to be age, circulatory strain, smoking habit, total cholesterol diabetes hypertension, stoutness and absence of physical action .Different information mining procedures have been utilized by the specialists to help therapeutic experts through better precision in the finding of coronary illness.

Naive Bayes (NB), Genetic algorithm, Artificial Neural Networks, Support Vector Machine (SVM), and direct method of self-sorting out guide are a few procedures utilized so far in the classification of coronary illness. This research paper demonstrates a correlation of three discrete classifiers that might be used in machine learning techniques to be specific the Naive Bayes calculation, the Logistic regression and the SVM classifier. The investigation was roused by the need to find a mechanized strategy to locate the most appropriate machine learning system for foreseeing survivability rate of coronary illness patients.

The utilization of NB, LR, SVM keeping into record it gives the high non-typicality of our information. Initially, the results of Naive Bayes and Support vector Machine were collected and then this information was compared with the results of Logistic Regression. Also, it was observed that Support Vector Machine was effective than Naive Bayes in diagnosing the coronary illnes. Incredibly, we have discovered that Logistic regression classifier is more effective than Naïve Bayes classifier and support vector machine classifier.

CLASSIFYING TECHNIQUES

A. LOGISTIC REGRESSION

Logistic Regression is a system acquired by machine learning techniques from the field of measurements. It is the proper regression examination to be performed when the reliant variable is paired. It allows categorizing data into discrete classes by learning the relationship from a given set of labeled data. It learns a linear relationship from the given dataset and then introduces a non-linearity in the form of the Sigmoid function.It is a prescient analysis. Numerous other restorative scales used to evaluate seriousness of a patient have been created utilizing calculated relapse. Strategic relapse might be utilized to anticipate the danger of building up a given infection (for example diabetes; coronary illness), in light of watched attributes of the patient. It describes the information and clarification of connection between one ward variable and at least one ostensible or ordinal ward factors.

B. NAIVE BAYES

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.A Bayesian classifier is a quick managed characterization strategy and this is the suitable classifier for broad expectation and arrangement undertakings on composite and deficient informational collections. Naive Bayes classifiers are exceptionally versatile, requiring various parameters direct in the quantity of factors (highlights/indicators) in a learning issue. Naive Bayesian order work better when the characteristics qualities for the session are self-deciding. The Naive Bayes classifier applies to learning undertakings where each occasion 'x' is portrayed by a combination of qualities and where the objective capacity f(x) can go up against any incentive from same limited set .

C. SUPPORT VECTOR MACHINE

Support Vector Machine is a class of regulated learning calculations. Given plenty of making ready tests, every set apart to possess an area with a pair of classifications, SVM making ready calculation constructs a model that relegates new precedents into one category or the opposite, creating it a non-probabilistic parallel direct classifier. SVM demonstrate could be a portrayal of the examples as focuses in area are mapped in order that the samples of the clear categories are isolated by an affordable gap (i.e., as wide as would be prudent). The information utilized in this investigation is the Cleveland Clinic Foundation Heart Disease Data Set accessible at UCI Machine Learning Repository. This informational index has 76 crude qualities, yet all distributed examinations allude to utilizing a subset of 13 attributes.

Specifically, Cleveland informational index is the special case that has been utilized by Machine Learning scientists to this date. Therefore, to permit correlation with the writing, we limited testing to these 13 characteristics which are recorded in Table 1.The informational index comprises of 13 numeric characteristics including age, sex, chest torment type, resting pulse, cholesterol, fasting glucose, resting ECG, most extreme pulse, practice actuated angina, old peak, slant, number of vessels hued and thal. The classes include whole numbers esteemed 0 (no nearness of coronary illness) and 1 (nearness of coronary illness).

Table I: Selected heart disease attributes of Cleveland

ATTRIBUTES	TYPE	DESCRIPTION
Age	continuous	Age in years
Sex	discrete	1=male;0=female
Cp	discrete	1 = typical angina;2 = atypical angina;3 = non-anginal pain;4 = asymptomatic
Trestbps	continuous	Resting blood pressure(in mm Hg)
Chol	continuous	Serum cholesterol in mm/dl
Fbs	discrete	Fasting blood sugar>120mg/dl 1-true, 0-false
Restecg	Discrete	Resting electrocardiographic result 0 = normal; 1 = having ST-T;2 = hypertrophy
Thalach	continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina1 = yes;0 = no
Old peak	continuous	Depression induced by exercise related to rest
Slope	Discrete	The slope of the peak exercise segment 1 = upsloping;2 = flat;3 = down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that ranges between 0 and 3
Thal	Discrete	3 = normal; 6 = fixed defect;7 = reversible defect

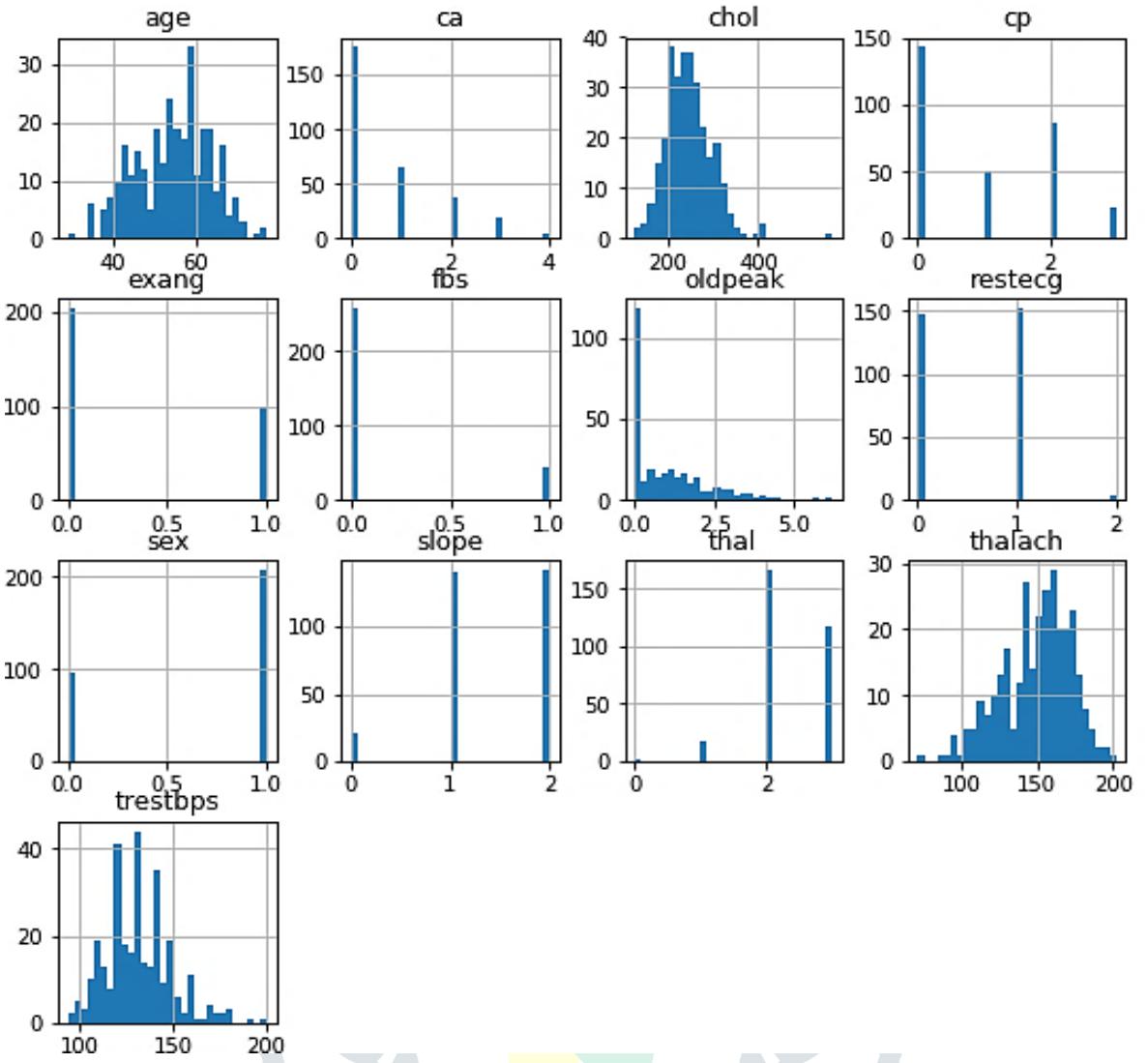

Fig 1.Histogram of 13 input attributes as per table 1

Table 2.Performance of Classifiers

S.No.	Classifiers	Classified	Misclassified
1	Support Vector Machine	(258) 85 %	(45) 15 %
2	Naïve Bayes	(256) 84.4 %	(47) 15.6 %
3	Logistic Regression	(261) 86 %	(42) 14 %

Table 3. Results of Cross-validation Score

S. No	Classifiers	Cross-Validation Score
1	Support Vector Machine	85
2	Naïve Bayes	84.4
3	Logistic Regression	86

K-FOLD CROSS VALIDATION

In K-fold cross validation, the data set 'X' is divided randomly into K equal-sized parts, X_i, I =1,….,K. To generate each pair, we keep one of the K parts out as the validation set and combine the remaining K-1 parts to form the training set. Doing this K time, each time leaving out another one of the K parts out, we get K pairs:

$V_1=X_1\ T_1=X_2\ U\ X_3\ U\ ….U\ X_K$
$V_2=X_2\ T_2=X_1\ U\ X_3\ U\ ….U\ X_K$
.
.
.
$V_K=X_K\ TK=X_1\ U\ X_2\ U\ ….U\ X_{K-1}$

There are two problems with this: First, to keep the training set large we allow validation sets that are small. Second, the training set overlap considerably, namely any two training set share K-2 parts.

K is typically 10.As K increases the percentage of training instances increases and we get more robust estimators, but the validation set becomes smaller. Furthermore, there is a cost of training the classifiers K times, which increases as K is increased. As N increases, K can be smaller; if N is small, K should be large to allow large enough training sets. One extreme case of K-fold cross validation *is leave-one-out*, where given a data set of N instances, only one instances is left out as the validation set (instances) and training uses the N-1 instances. We then get N separate pairs by leaving out the different instances at each iteration. This method helps to identify the misclassified datas. Leave-one-out does not permit stratification.

RESULTS AND DISCUSSION

The Cleveland Heart Disease Database is taken with 303 cases and 13 qualities. In this arrangement Cross validation score strategy is utilized to analyze the execution of Logistic regression, Support vector machine and Naïve Bayes classifiers. Cross valediction is a re sampling method used to assess machine learning models on a restricted information sample. The strategy has a solitary parameter considered k that alludes to the quantity of gatherings that a given information test is to be part into. In that capacity, the technique is frequently called k- cross-validation. The histogram of thirteen attributes mentioned in Table 1 has been shown in the above fig.

Support vector machine utilizes diverse sorts of calculations, for example, Kernel. The bit is likewise of various kinds where Linear part calculation is utilized in this exploration with the resilience estimation of 0.001. Naïve Bayes with defaultsmoothing variable le-9 is utilized here for the characterization. Liblinear solver which is one of the Logistic relapse calculations is utilized in this exploration.

CALCULATION

SVM using linear kernel with tolerance value 0.001-0.85081967
NB with default smoothing variable 1e-9- 0.8442622950819672
LR with liblinear solver - 0.860655737704918

As per cross valediction score, Logistic regression classifier indicates more prominent execution of about 86% than Support Vector Machine (85%) and Naïve Bayes (84.4%). Logistic Regression classifier with Liblinear solver gave the best outcomes for the coronary illness finding.

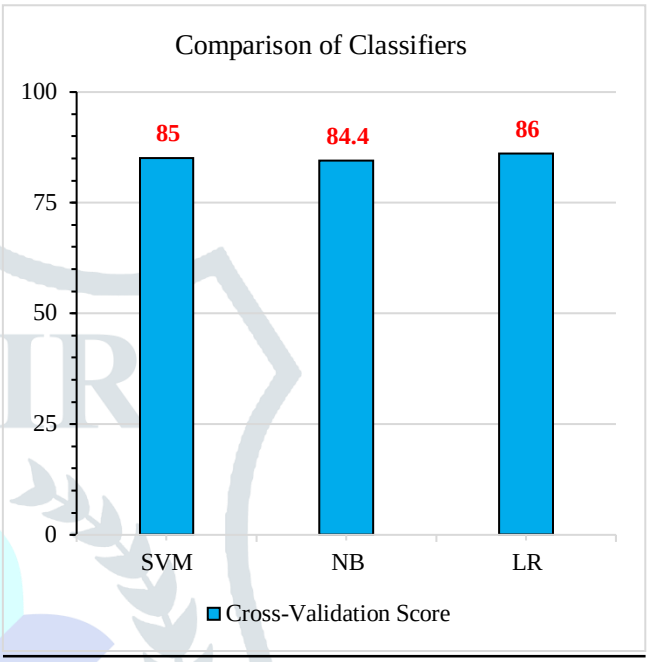

Fig 2.Comparison of Classifiers

CONCLUSION

In this paper, three different machine learning techniques are studied and have used them over a novel data set of heart disease for classification. In this paper different results are obtained for each of them. Using the whole dataset (13 attributes × 303 instances), we got the best performance from the LR classifier. In fact only 42 cases were incorrectly classified. The naïve Bayes and Support Vector Machine returned similar results but worse than the LR.

From the results, it can be stated that all classifiers achieved a reasonable performance. However, we found that, LR performed significantly better than both SVM and naïve Bayes classifier on our data set. Future research involves more intensive testing using a larger heart disease database to get more accurate results.

REFERENCE

[1] Keerthi, S.S., Shevade, S.K., Bhattacharyya, C. and Murthy, K.R.K., 2001. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural computation*, *13*(3), pp.637-649.

[2] Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, *13*(2), 415-425.

[3] Suykens, J.A. and Vandewalle, J., 1999. Least squares support vector machine classifiers. *Neural processing letters*, *9*(3), pp.293-300.

[4] Amari, Shun-ichi, and Si Wu. "Improving support vector machine classifiers by modifying kernel functions." *Neural Networks* 12, no. 6 (1999): 783-789.

[5] McCallum, Andrew, and Kamal Nigam. "A comparison of event models for naive bayes text classification." In *AAAI- 98 workshop on learning for text categorization*, vol. 752, no. 1, pp. 41-48. 1998.

[6] Patil, Tina R., and S. S. Sherekar. "Performance analysis of Naive Bayes and J48 classification algorithm for data classification." *International journal of computer science and applications* 6, no. 2 (2013): 256-261.

[7] Golub, Gene H., Michael Heath, and Grace Wahba. "Generalized cross-validation as a method for choosing a good ridge parameter." *Technometrics* 21, no. 2 (1979): 215-223.

[8] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." In *Ijcai*, vol. 14, no. 2, pp. 1137-1145. 1995.

[9] Allison, Paul D. "*Logistic regression using SAS: Theory and application*" SAS Institute, 2012

[10] Wilson, Peter WF, et al. "Prediction of coronary heart disease using risk factor categories." *Circulation* 97.18 (1998): 1837-1847.

[11] Parthiban, Latha, and R. Subramanian. "Intelligent heart disease prediction system using CANFIS and genetic algorithm." *International Journal of Biological, Biomedical and Medical Sciences* 3, no. 3 (2008).