CORRELATION OF THREE UNIQUE TECHNIQUES FOR CORONARY HEART DISEASE

L.ARULMURUGAN[1], K.JEEVITHA[2], S.NAVANEETA[3]

***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 has 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- SVM, LR, NB, Cleveland

1. INTRODUCTION

Various elements which contain in increasing the threat of heart disorder along with Family history, Smoking, Poor food regimen, High blood strain, High blood cholesterol, Obesity, Physical state of being inactive and Hyper anxiety ,these are used to analyze the Heart sickness. The heart is one of the main organs of the human body. It pumps blood through the blood vessels of the circulatory system. The circulatory system is extremely important because it transports blood, oxygen and other materials to the different organs of the body. Heart plays the most crucial role in circulatory system. Usually the heart sickness are recognized primarily based patient's check outcomes & medical doctor's enjoy. Exact forecast of chance factors which can be related with cardiovascular illness is basically critical for the analysis and remedy of coronary infection. Among the modern-day strategies, regulated mastering strategies are the maximum widely known in coronary infection conclusion. Different information mining processes had been used by the specialists to help healing experts via higher precision within the finding of coronary contamination. some of the well known data mining algorithms used for heart disease prediction. Research on data mining has led to the formulation of several data mining algorithms. These algorithms can be directly used on a dataset for creating some models or to draw vital conclusions and inferences from that dataset. Some popular data mining algorithms are Decision tree, Naïve Bayes, k-means, artificial neural network etc., This data can be used in machine learning to determine the precision among the classifiers.

1. EXISTING METHOD

Several hospitals manipulate healthcare facts using healthcare statistics system due to the fact the device incorporates good sized quantity of facts, used to extract hidden information for growing perceptive medical analysis. Research within the area of cardiovascular illnesses the use of statistics mining has been an ongoing attempt related to prediction, remedy, and risk rating evaluation with high levels of accuracy. Naive Bayes (NB), Genetic set of rules, Nearest Neighbour (NN), Artificial Neural Networks (ANN), Support Vector Machine (SVM), and direct approach of self-finding out manual are a few strategies applied so far in the classification of coronary infection.

The end result is in comparison on the premise of accuracy, sensitivity, and specificity. The technique pursuits to accomplish of two goals: the first is to carry out primary framework for heart disorder, and the second is to compare the performance of merging the effects of multiple fashions as opposed to the usage of a single version. The Healthcare industry is typically information rich, however sadly not all of the records are mined that is required for coming across unknown styles & higher cognitive method. Advanced data mining techniques are used to acquire information in scientific research.

2.1 DECISION TREES

Decision Trees are a type of Supervised Machine Learning which means that the records is continuously cut up. The tree is divided into divisions,namely decision nodes and leaves. The leaves are the very last outcomes. And the choice nodes are in which the statistics is break up. There are principal types of Decision Trees:

1.Classification trees:

The category choice trees are the unit engineered with unordered values with dependent variables.

2.Regression trees:

The regression selection trees take ordered values with non-stop values.

2.2 ARTIFICIAL NEURAL NETWORK

Neural computing refers to a sample popularity technique for machine getting to know. The resulting model from neural computing is referred to as Artificial neural network (ANN) or a neural community. Neural networks are employed in lots of business packages for pattern recognition, forecasting, prediction, and classification. Neural community computing is a key component of any information mining tool package. Neural network approach is used

for classification, clustering, characteristic mining, prediction and pattern recognition. Through training statistics mining, the neural community method step by step calculates the weights the neural network related.

2.3 RANDOM FOREST

Random forest algorithm is a supervised category set of rules. As the name suggests, the forest with some trees. In this technique within the random forest area classifier, the better the various trees in the forest area gives the highest accuracy results.

Random forest may be an indicator time period for an ensemble classifier that includes many decision trees and outputs the class that is the mode of the output of the classes via individual trees. Random forests are collections of bushes but different from each other. It randomizes the algorithm, not the data available in the training. The randomization relies upon on the algorithm don't select the best, choose randomly from the high-quality options. It commonly improves decision trees choices.

2. PROPOSED TECHNIQUES

Multiple Cardio Vascular Disease (CVD) surveys have been conducted in the area and the data set could be collected from the Cleveland Heart Clinic. The Cleveland Heart Disease Database (CHDD) has been considered.

3.1 DATA SET INFORMATION

The database contains 76 attributes, however all experiments use a subset of 10 of them. In specific, the Cleveland information is the best one that has been used by ML researchers to this date. The "goal" refers to the presence of coronary heart disease in the affected person. It is integer valued from zero (no presence) to four. The names and social security numbers of the sufferers were removed from the database and replaced with dummy variables. One file has been "processed", that one containing the Cleveland database. All additionally documents additionally exist on this directory.

The attributes used as the data for the database are

1. Sex
2. Cp
3. Exang
4. Fbs
5. Restecg
6. Ca
7. Slope
8. Thal
9. Target

Description of the attributes are shown in the Table 3.1

NAME	TYPE	DESCRIPTION
Sex	Discrete	1=male; 0=female
Cp	Discrete	1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic
Fbs	Discrete	Fasting blood sugar>120mg/dl 1-true,0-false
Restecg	Discrete	Resting electrocardiographic result 0 = normal; 1 = having ST-T; 2 = hypertrophy
Exang	Discrete	Exercise induced angina 1 = yes; 0 = no
Slope	Discrete	The slope of the peak exercise segment 1 = upsloping; 2 = flat; 3 = downslope
Ca	Discrete	Number of major vessels coloured by fluoroscopy that ranges between 0 and 3
Thal	Discrete	3= normal; 6= fixed defect; 7= reversible defect
Target	Continuous	0 or 1

Table 3.1 Selected Heart Disease Attributes of Cleveland

3.2 LOGISTIC REGRESSION

Logistic Regression is a system acquired by machine learning from the sector of measurements. It is the proper regression examination to be accomplished while the reliant variable is paired. It is a predictive analysis. Strategic relapse is probably applied to anticipate the danger of constructing up a given contamination (for instance diabetes; coronary illness), in light of watched attributes of the affected person. It describes the facts and clarification of connection between one ward variable and at the least one ostensible or ordinal ward elements.

Logistic Regression is the famous algorithm to remedy a type problem. It is likewise named as Linear Regression. The term "Logistic" is taken from the Logic feature that is used in this technique of classification.

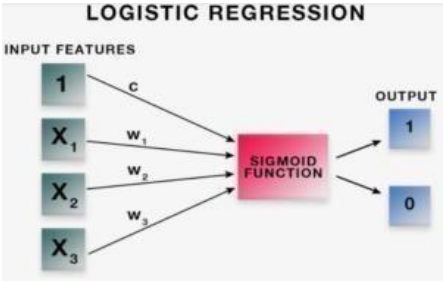

Figure 3.1 Logistic Regression

3.3 NAÏVE BAYES

Bayes theorem named after Rev. Thomas Bayes. It works on the principle of formula of conditional probability. Conditional probability is that probability that something will happen, given that something else has already occurred. Using the conditional probability, we can calculate the probability of an event using its prior knowledge. Below is the formula for calculating the conditional probability.

$$P(H|E) = P(E|H)*P(H)/P(E)$$

Where,

P(H) is the probability of hypothesis H being true. This is known as the prior probability.

P(E) is the probability of the evidence (regardless of the hypothesis).

P(E|H) is the chance given that hypothesis is true.

P(H|E) is the probability of the hypothesis given that the evidence is there.

3.4 SUPPORT VECTOR MACHINE

A support vector machine may be a sort of version used to analyze statistics and discover patters in class and regression analysis. Support vector device(SVM) is used while your data has exactly two lessons . An SVM classifies information by finding the nice hyperplane that separates all records points of one elegance from of other elegance. The large margin among the two lessons, the higher the model is a margin need to have no factors in its interior vicinity. The assist vectors are the information points that on the boundary of the margin. SVM is established on mathematical features and used to version complicated, and real international troubles. SVM plays nicely on records sets which have many attributes, together with the CHDD.

The difficult aspect is kernel choice and method selection such that your model is not over optimistic.

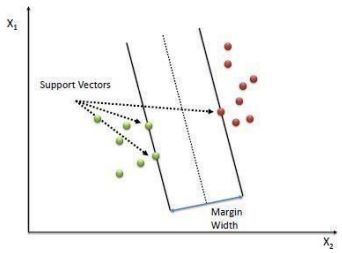

Figure 3.2 Support Vector Machine

Considering that the CHDD has a large number of instances as well as features, it is arguable whether the kernel chosen is RBF or linear. Although the relation between the attributes and sophistication are nonlinear, due to the large number of features, RBF kernel may not improve performance. It is recommended that both kernels be tested and the more efficient one be finally selected.

Support vector machines have gained popularity in the machine learning and pattern classification. The aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data . For a linearly separable dataset, a linear classification function corresponds to a separating hyperplane of (x) SVM guarentees that best such function is found by maximizing the margin between the two classes. It is a linear classifier which constructs separating hyperplane to maximize distance data.

3. PROPOSED SOLUTION

Algorithm : Cross Validation

Cross-validation is a statistical method used to estimate the ability of machine learning models. It is commonly used in applied machine learning to check and choose a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods. Cross-validation may be a sampling procedure used to evaluate machine learning models on a limited data sample.

The procedure have a single parameter also ask that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross- validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation. Cross- validation is primarily used in machine learning to estimate the ability of a machine learning model on unseen data.

It is a preferred technique because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model ability than other methods, such as a simple train/test split.

4.1 CONFIGURATION OF K

The k value should be chosen carefully for data sample. A poorly chosen value for k may result in a mis-represet idea of the skill of the model, such as a score with a high variance (that may change a lot based on the data used to fit the model), or a high bias, (such as an overestimate of the skill of the model).

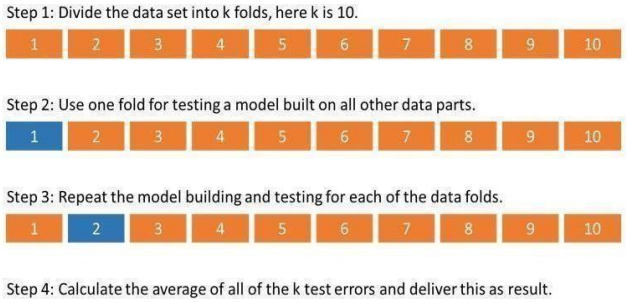

Figure 4.1 K-fold Cross Validation

4. RESULT

The main theme of this paper is to predict more accurately the presence of heart disease. In this arrangement Cross validation algorithm is used to analyze the execution of LR, SVM and NB classifiers. As per cross valediction score, Logistic regression classifier indicates more prominent execution of about (86%) than Support Vector Machine (58%) and Naïve bayes (82%). Logistic Regression classifier with Lib linear solver gave the best outcomes for the coronary illness finding.

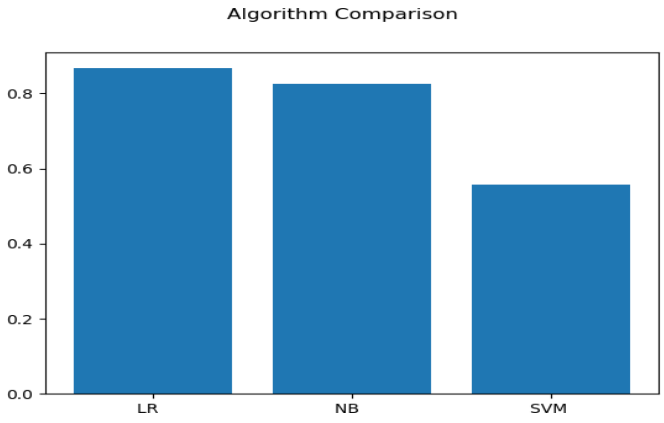

Figure 5.1 Comparison of Classifiers

5. CONCLUSION

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

REFERENCES

[1] Yanwei Xing, Jie Wang and Zhihong Zhao Yonghong Gao 2007 "Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease" Convergence Information Technology, 2007. International Conference November 2007, pp 868-872.

[2] Jianxin Chen, Guangcheng Xi, Yanwei Xing, Jing Chen, and Jie Wang 2007 "Predicting Syndrome by NEI Specifications: A Comparison of Five Data Mining Algorithms in Coronary Heart Disease" Life System Modeling and Simulation Lecture Notes in Computer Science, pp 129-135.

[3] Jyoti Soni, Ujma Ansari, Dipesh Sharma 2011 "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" International Journal of Computer Applications, doi 10.5120/2237- 2860.

[4] Mai Shouman, Tim Turner, Rob Stocker 2012 "Using Data Mining Techniques In Heart Disease Diagnoses And Treatment" Electronics, Communications and Computers (JECECC), 2012 Japan-Egypt Conference March 2012, pp 173-177.

[5] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee, Victor Froelicher 1989 "International application of a new probability algorithm for the diagnosis of coronary artery disease"The American Journal of Cardiology, pp 304-310.15.

[6] Polat, K., S. Sahan, and S. Gunes 2007 "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing" Expert Systems with Applications 2007, pp 625-631.

[7] Ozsen, S., Gunes, S. 2009 "Attribute weighting via genetic algorithms for attribute weighted artificial immune system (AWAIS) and its application to heart disease and liver disorders problems" Expert Systems with Applications, pp 386-392.

[8] Resul Das, Ibrahim Turkoglub, and Abdulkadir Sengurb 2009 "Effective diagnosis of heart disease through neural networks ensembles" Expert Systems with Applications, pp 7675–7680.