International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020
ISSN: 1475-7192

Analysis and Graphical Representation of Data Mining Techniques for Prediction of Heart Disease Using the Weka Tool

[1]K.AshaRani ,[2]Dr A.V.R Mayuri ,[3]Dr C. Sreedhar

***ABSTRACT**--In current decades, heart disease has been recognized as like the leading cause of death throughout the world. However, it is considered so the most preventable or controllable disease at the same time. According in conformity with World Health Organization (WHO), the express and timely analysis of heart disease plays a remarkable role within preventing its progress and reducing related treatment costs. Data mining methods and machine learning algorithms play a very important function within this area. The researchers accelerating their research works to boost a software with the help machine learning algorithm which perform help doctors to take a decision regarding both prediction and diagnosing of heart disease. The main objective over this research paper is predicting the heart disease regarding a patient the use of machine learning algorithms. Comparative study concerning Naïve Base Classifier, K-nearest neighbor, Support vector machines and Random Forest the precision and recall about machine learning algorithms is performed through a graphical representation concerning the results.*

***Key Words**--Data mining, Heart disease, WEKA, Naïve Base Classifier, K-nearest neighbor, Support vector machines and Random Forest*

I. INTRODUCTION

In the past decade, heart disease has been the leading cause of death in different continents and countries in the world, regardless of the income level of countries [1]. According to WHO report, heart disease is the leading cause of death across the world, accounting for 7.2 million deaths, i.e., 12.8% of all fatalities in the world [2]. Figure 1, illustrates deaths from heart disease across the world (scale: 1:100000). According to recent research predictions, cardiovascular diseases will become the leading cause of death up to 2030. Although cardiovascular diseases have been identified as the leading cause of death in the world in the past decade, they have been introduced as the most preventable and controllable diseases [3]. The complete and correct treatment of a disease depends on the timely diagnosis of that disease [4]. An accurate and systematic tool for identifying high-risk patients and extracting data for timely diagnosis of heart disease seems a critical need.

Every day, modern computer-based systems collect large amounts of data using automatic data record systems in different fields. Data mining technology is the product of the evolution of database technology, IT and

[1]Assistant Professor in, Department of Computer Science and Engineering, G.PullaReddy Engineering CollegeKurnool,Asharani.gprec@gmail.com

[2]Assistant Professor in, Department of Computer Science and Engineering, G.PullaReddy Engineering College Kurnool,mayuriavr@gmail.com

[3]Associate Professor in, Department of Computer Science and Engineering, G.PullaReddy Engineering College Kurnool, csrgprec@gmail.com

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

storage devices [5]. The current challenges is according to make data mining and knowledge discovery systems applicable in accordance with a wider range regarding domains [6]. Researchers are adopting data mining strategies to diagnose one of a kind diseases who consists of diabetes [7], stroke [8], cancer [9] and heart disease [10]. Considering the high rate about cardiovascular induced fatalities, researchers have tried in imitation of adopt data mining structures to diagnose bravery disease [11].

II. WEKA

Waikato Environment because Knowledge Analysis or WEKA is an open source software, developed into Java, issued under the GNU General Public License. Weka is basically a collection of machine learning algorithms because data mining tasks, such so data pre-processing, visualization, classification, regression and clustering.

III. Heart Disease Dataset

We performed computer simulation regarding one dataset. Dataset is a Heart dataset. The dataset is available in UCI Machine Learning Repository [12]. Dataset carries 303 samples and 14 input features as well as 1 outturn feature. The applications pencil a financial, personal, then convivial characteristic over mortgage applicants. The output feature is the decision class which has value 1 because of Good credit and 2 for Bad credit. The dataset-1 contains 700 instances shown as a Good credit while 300 instances as bad credit. The dataset contains features expressed on nominal, ordinal, or interval scales. A list regarding all those features is given in Table

Table 1: Heart Disease Features In The Dataset

Feature No.	Feature Name
1	age
2	sex
3	cp
4	trestbps
5	choi
6	fbs
7	restesg
8	thalach
9	exang
10	oldpeak
11	slop
12	ca
13	thal
14	num

IV. Precision and Recall\

Received: 27 Feb 2019 | Revised: 20 Mar 2019 | Accepted: 30 Apr 2020

Recall (**R**) and Precision (**P**) are measures that are based on confusion matrix data. Recall (R) is the portion of instances that have true positive class and are predicted as positive. On the other hand, Precision (P) is the probability of that a positive prediction is correct as shown in

$$\mathbf{R = \frac{TP}{CN}\ and\ P = \frac{TP}{RN}}$$

Classification Accuracy (**Acc**) is the most used measure that evaluates the effectiveness of a classifier by its percentage of correctly predicted instances as in

$$\mathbf{ACC = \frac{TP + TN}{N}}$$

V. Naïve Base Classifier

This classifier is a strong probabilistic representation, and its uses because classification has received considerable attention. This classifier learns from learning data the subject probability regarding every attribute AI given the class label C. Classification is after performed by using applying Bayes rule to account the probability regarding C given the particular instances concerning A1…. An and after predicting the category including the highest approximate probability. The goal about alignment is within pursuance with correctly predict the value concerning a designated distinct class variable is given a vector of predictors and attributes. In particular, the Naive Bayes classifier is a Bayesian community where the classification has no parents or each virtue has the class as much its sole parent. Although the naive Bayesian (NB) algorithm is simple, that is absolutely superb into much real-world facts units because it does entrust higher predictive accuracy than well-known methods like C4.5 or BP and is extremely efficient between that such learns regarding a linear fashion the usage of ensemble mechanisms, certain so as bagging or boosting, to combine classifier predictions. However, now attributes are redundant and not normally distributed, the predictive accuracy is reduced. Naïve Base Classifier Comparative Analysis of heart disease data set Classification Precession(0.837) and Recall(0.837) in WEKA Tool

VI. Support Vector Machine

Given availability of support vectors, Support Vector Machine (SVM) is the boundary determining the best data classification and separation. In SVM, only those data lying inside support vectors are used as the base data for machine and building a model. This means that this algorithm is not sensitive to other data. It aims to find the best data boundary with the farthest possible distance from all classes (their support vectors). SVM transfers data to a new space with respect to their predetermined classes so that data can be classified and separated linearly (using hyper planes). Then, it searches for support lines (or support planes among multi-dimensional space) and tries according to determine the equation of a straight line that maximizes the distance between each two classes. Each support vector is characterized with an equation describing the boundary line of each class. Support Vector Machine Comparative Analysis of heart disease data set Classification Precession(0.84) and Recall(0.8365) in WEKA Tool

VII. Random Forest

Random Forest consists of decision trees. Every decision tree is formed by subset of training data which randomly selected. The decision tree is a approach because displaying a series regarding laws that are leading to a category or value. The difference between the methods of decision tree is that how the distance to be measured. Decision trees that are used in accordance with predict the cluster variables called classification trees because they are located the samples within clusters or classes. Every decision tree in Random Forest provides results for classification and final results of Random Forest, is that most of the trees have announced. To build Random Forest, such can lie preserved a number about decision trees so much need to exist of the forests. One over the advantages concerning Random Forest is that it requires insignificant preprocessing. Also there is no need to choose the required variables at the beginning and Random Forest model itself chooses the useful variables [13, 14, 15]. Comparative analysis of precession and recall analyzing for heart disease data sets precession in WEKA precession (0.818) and Recall (0.819).

VIII. K-Nearest Neighbor

Nearest Neighbor algorithms are among the simplest on all machine learning algorithms. The thought is to memorize the training employ and afterwards after predict the label regarding some new instance about the basis of the labels regarding its closest neighbors into the training set.The rationale behind such a technique is based on the assumption that the features so much are used to construct the area points are relevant to theirs labeling of a course to that amount makes close by factors in all likelihood to have the same label. Furthermore, within some situations, even when the training set is immense, finding the nearest neighbor can be done extremely fast. Comparative analysis of precession and recall analyzing for heart disease data sets precession in WEKA precession 0.753) and Recall (0.752).

Table 2: Classification Algorithm Precession and Recall in WEKA Tool

Algorithm classification Average	**Precession in WEKA**	**Recall in WEKA**
Naïve base classifier	0.837	0.837
SMO or Support Vector Machine	0.84	0.8365
Random Forest	0.818	0.819
1BK or K-Nearest Neighbor	0.753	0.752

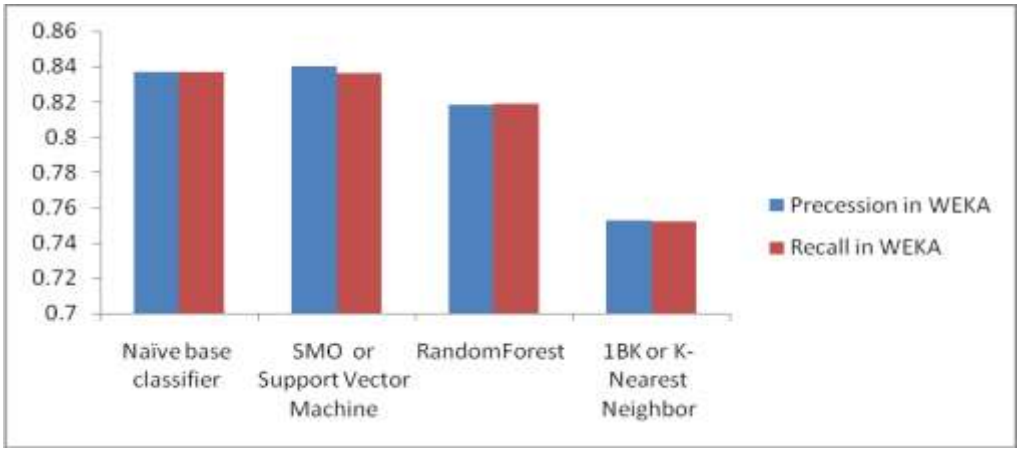

Figure1: Classification Algorithm Graph for Precession and Recall in WEKA Tool

IX. CONCLUSION

Heart disease is the lead cause concerning death in the world. It accounts for 7.2 million deaths, i.e., 12.8% regarding fatalities among the world. Although cardiovascular diseases bear been recognized so the leading cause of death in the past decade, they are the most preventable and controllable diseases at the same time. Deaths from cardiovascular diseases show an ever-increasing trend. We also choose one dataset from heart available at UCI machine learning repository. Naïve base classifier is the best in performance. In order to compare the classification performance concerning 4 machine learning algorithms, classifiers are applied in accordance with same data and results are compared on the basis regarding misclassification and accurate classification rate and in accordance to experimental results within table 1, such can lie concluded so much Naïve base classifier is the excellent as compared to Support Vector Machine, Random Forest, and K-Nearest Neighbor. This technique is expected to be implemented in future on a localized dataset with nonaggressive indices in general. This, in turn, imposes lower costs and complications on patients. Things that can be done in the future are another critical illness used in data mining in order that the role of data mining in medical science has been developed.

REFERENCES

1. World Health Organization (2011) The top ten causes of death.
2. World Health Organization (2013) Deaths from coronary heart disease.
3. Center for Disease Control and Prevention (2014) Heart Disease and Family History.
4. Paladugu S (2010) Temporal mining framework for risk reduction and early detection of chronic diseases. University of Missouri-Columbia.
5. Obenshain MK (2004) Application of data mining techniques to healthcare data. Infection Control and Hospital Epidemiology 25: 690-695.
6. A, Roddick JF (2006) Towards role based hypothesis evaluation for health data mining. Electronic. Journal of Health Informatics 1: 1-9.
7. Porter T, Green B (2009) Identifying Diabetic Patients: A Data Mining Approach.

8. Panzarasa S, Quaglini S, Sacchi L, Cavallini A, Micieli G, et al. (2010) Data mining techniques for analyzing stroke care processes. In the Proc. of the 13th World Congress on Medical Informatics.
9. Li L, Tang H, Wu Z, Gong J, Gruidl M, et al. (2004) Data mining techniques for cancer detection using serum proteomic profiling. Artificial intelligence in medicine 32: 71-83.
10. Das R, Turkoglu I, Sengur A (2009) Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications 36: 7675-7680.
11. Lakshmi K, Krishna MV, Kumar SP (2013) Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability. International Journal of Scientific and Research Publications 3: 1-10.
12. [V.A. Medical Center, Long Beach Clinic Foundation, "Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease, [Last Accessed 5 November 2015].
13. A. S. Abdullah, R. R. Rajalaxmi, "A Data mining Model for predicting the Coronary Heart Diseaseusing Random Forest Classifier", International Conference on Recent Trends in Computational Methods, Communication and Controls (ICON3C 2012), ICON3C(3), pp.22-25, April 2012.
14. K. Kalaiselvi, K. Sangeetha, S. Mogana, "Efficient Disease Classifier Using Data Mining Techniques: Refinement of Random Forest Termination Criteria", IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 14, No. 5, pp.104-111, 2013.
15. E. E. Tripoliti, D.I. Fotiadis, G. Manis, "Automated Diagnosis of Diseases Based on Classification: Dynamic Determination of the Number of Trees in Random Forests Algorithm" IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 16, NO. 4, JULY 2012.