International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

Healthcare Diagnosis by Using Computational Intelligence Algorithms

Mrs. M.Nirmala devi,
Assistant Professor,
Department of Information Technology,
Thiagarajar College of Engineering, Madurai, India.
S.Meena,
Student, M.E Computer Science and Information Security,
Department of Information Technology,
Thiagarajar College of Engineering, Madurai, India.
Dr. S.Appavu alias Balamurugan,
Professor and Head, Department of Information Technology,
KLN College of Information Technology,Sivagangai,India.

ABSTRACT-
Healthcare domain contains a vast amount of data and to do classification and prediction is one of the foremost challenges in worldwide. Most of discoveries show that the simplest way to overcome is to forestall the risks of this is before it is occurring. With this concept we would prefer to realize how to estimate patients' risk. Data mining techniques could be used as a manner in discovering knowledge from the patient medical records and that they have shown outstanding success within the area of applying Computer Aided Diagnostic (CAD) systems. In this paper, we have applied many intelligence classifiers such as Naïve Bayesian, Decision Trees, Logistic, Random Forest, and Support Vector Machine for various datasets in healthcare. Experimental results on various datasets shows that support vector Machine has higher accuracy as compared with other different classification algorithms and it is also observed that naive bayes takes very less time to build the model but it has fourth rank in accuracy compared to other algorithms.

Keywords-Naive Bayes, Support vector Machine, Accuracy, and Execution time.

I INTRODUCTION

Today the trendy expression is "Health Care" all over the world. Early Prediction of diseases will cut back the fatal rate of human. There are vast and massive amount of data accessible in hospitals and medical related institutions. The analysis of this data is called analytics. Data that generates from healthcare will be voluminous and it gets generated from various sources and it is used to be incomplete, noisy, inconsistent so preprocessing need to be mandatory. Information technology plays a vital role in Health Care. Early prediction of diseases is quite difficult task for medical practitioners as a result of complex interdependence on varied factors. Data mining is a process to helpful information from huge database. It's interdisciplinary field of computer science which involves computational process, machine learning, statistical techniques, classification, and clustering, discovering patterns. Data mining techniques can be used for early prediction of diseases with higher accuracy in order to save lots of human life and to reduce the treatment cost. This paper explores varied Data mining techniques such as Naive Bayes, logistic regression, support vector classifier,J48,Random Forest are analyzed to predict the disease.

II DATA MINING TECHNIQUES

A. Classification:

The healthcare domain contains a huge amount of data which can't be mined so to mitigate this problem classification can be used which serves as one of the solution. Classification can be helpful to do prediction based on the given input. For doing such prediction the algorithm uses training set which has the same attribute as the original data set. Classification tries to find the relationship between attributes. By using classification the user can extract most relevant and important data. Classification involves assigning class

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

labels to the unclassified cases. Classification usually begins with the datasets where the class labels are known. Classification is being tested by comparing the predicted value with the target value. The data used for classification are: Building the model, Testing the model. Different classification algorithm uses various techniques to classify the data and to find the relationship. Classification used to be discrete which implies that it doesn't have any particular order.

B.Decision Tree:

Decision tree is a common method used in data minnig.It will predict the target variable value based on input variables. It is simple representation for classification. Decision tree is based on conditional probability. Decision trees generate rules. A rule is a conditional statement. In decision tree there won't be any backtracking mechanism. The trees are constructed using top-down recursive divide and conquer manner. It is a tree in which each internal node (non-leaf) is labeled with features edges represents the possible values for the input variable. The leaf nodes represent the target variable value. In this method a tree will be constructed for classifying the model .two steps for decision tree: Building the tree, applying the tree. The decision tree classifier generates tree as well as set of rules for the given data set. The decision tree construction is also based on training data set. The decision tree can be used to model the classification. The tree can be applied to each tuple to each database which results in the classification.

C.Naive Bayes:

Bayesian classifier is based on bayes theorem. Bayes theorem uses prior and posterior probability. It can predict whether the tuple can belong to particular class. It finds the probability based on the event which has already occurred. The conditional probability can be written as,

$$P(C\ |X) = \frac{p(C\ )p(x|c\ )}{P(x)} \text{ ------ (2.1)}$$

Where, $P(C\ |X)$-posterior probability of the class given the predictor

$p(c_k)$-class prior probability

p(x)-prior probability of predictor

$p(x|c_k)$- probability of the predictor given the class.

In other words can describe it as,

$$p \qquad = \frac{p \quad x\,li \quad ho}{e} \text{---- (2.2)}$$

Naive bayes is based on the assumption that there exist independence between every pair of feature. It is extremely fast compared to other classifiers.

D. Support Vector Machine:

SVM is based both linear, Non-Linear regression and it uses regularization property. It is automatically initialized to achieve the best average prediction across all classes. SVM utilizes priors as a weight vector that predispositions optimization and favors one class over another.

E.Logistic:

Logistic models the likelihood of the default class it is a linear method however the forecasts are changed utilizing the Logistic .The coefficients of the logistic are calculated using maximum-likelihood estimation of the training data

F.Random Forest:

It is an ensemble learning method for classification, regression and other tasks that work by developing a large number of choice trees at training time and yielding the class that is the method of the classes. It is an improvement over bagged decision trees. Random forest changes the algorithm for the way that the sub-trees are found out with the goal that the resulting predictions from the greater part of the sub trees have less correlation.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

II RELATED WORKS

Gandhi et.al [1] explains that prediction of heart disease is done using naive bayes, neural network, decision tree and each method has its own advantage and disadvantage and in decision tree there are many versions like ID3, C4.5, and C5.0. Naive Bayes has better performance if the attributes are independent, missing values are also handled easily.Tomar et al [2] portrays that the data mining techniques can be applied as hybrid because each has advantage and disadvantage. the various classification techniques which are used are: SVM, Decision tree, KNN, Neural network, Bayesian belief network Songthung et al[[6] elucidates that classification by Decision tree, Naive Bayes outperforms the traditional manual scoring. This was implemented by using rapid miner ,studio 7.0.this was tested for 12 hospital datasets and in future it was planned to explore large data sets.Masethe et al[[3] depicts that the prediction of heart disease where the patient records are collected and by using WEKA tool the results are collected it is observed that the J48,REPtree,,simpleCART. Safavian et al [4] bring out that performance of decision tree depends on number of test samples. If the test samples are large the time taken to build the design will also be large. Safavian et al[[5] interpret that naive bayes can be used for healthcare application which can serve as a best decision support system compared to back propagation neural network Safavian et.al [6] compares six most used data mining software tools based on this analysis it is said that each tool has its own advantage and disadvantage. And Weka can be suitable for classification so Weka is used in this paper for doing classification of data sets. Duggal et al[7] demonstrates that the preprocessing helps to achieve greater result than the original method. The proposed work was implemented for healthcare dataset and it was implemented in weka tool and MYSQLv6.3 for storing database. The future work which is specified that additional features and classification technique to improve the performance. Duggal et al [8] clear up that it evaluates the classification accuracy for five different algorithms: Naive bayes, SVM, Decision tree,Adaboost,Neural network. And it is found out that Random forest is an optimal algorithm by considering the evaluation metrics such as precision, recall. It was implemented in weka 3.7.Dey et al[9] clarifies the most commonly used algorithms and three algorithms were chosen namely:C4.5,Multi layer perception, Naive bayes and it is tested for different data set. The future work which is specified is that to take other technique and to evaluate in medical field and finding the optimal by comparing its own advantage and disadvantage.Palaniappan, et al[10] specifies that a prototype for intelligent heart disease prediction has been developed using 3 data mining algorithms and the effective model is to use naive bayes proceeded by neural network and decision tree.Yoo et.al [11] specifies that data mining has been used in healthcare and biomedical fields because of descriptive and predictive nature. And the classification algorithms which are used in this are: Naive bayes, SVM, Decision tree.

III METHODOLOGY

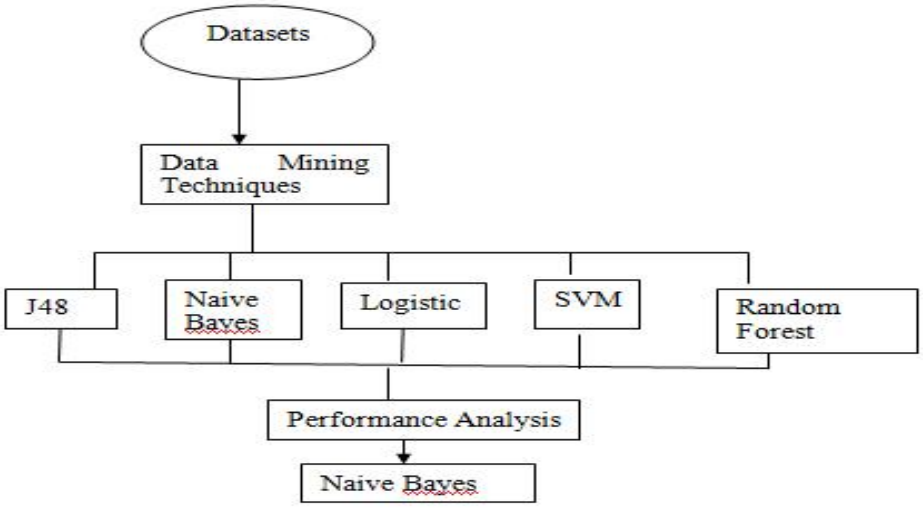

FIGURE-1 SYSTEM ARCHITECTURE

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

DATASET:

The performance of different data mining techniques has been tested for nine different data sets such as PIMA Indian Diabetes data set which is collected from- archive.ics.uci.edu/ml/datasets/pima+indians+diabetes.and it is also tested with many other data sets which are collected from UCI machine. The datasets which are taken are given in table II

IV RESULTS AND DISCUSSION

The results are filtered by using measures such as accuracy, sensitivity and specificity, Kappa mentioned in the results are calculated using WEKA (Waikato Environment for Knowledge Analysis).

A. Performance Measure

For measuring the performance of algorithms, Accuracy, Sensitivity, and Specificity are used because these three criteria are used more in the medical field.

B. Confusion Matrix

TABLE I. CONFUSION MATRIX

		Actual Class	
		C1	**C2**
Predicted Class	**C1**	**True Positive(TP)**	**False Positive(FP)**
	C2	**False Negative(FN)**	**True Negative(TN)**

For calculation of Sensitivity, Specificity and accuracy confusion matrix is required.

In confusion matrix: Actual class is the class which determined by angiography and it is existed in dataset. Predicted class is the one which is predicted by algorithms.

TP is number of samples of class C1 which has been correctly classified.

TN is number of samples of class C2 which has been correctly classified.

FN is number of samples of class C1 which has been falsely classified as C2.

FP is number of samples of class C2 which has been falsely classified as C1.

C. SENSITIVITY, SPECIFICITY AND ACCURACY

According to Confusion Matrix, Sensitivity, Specificity and Accuracy are calculated as follows:

Accuracy:

It refers to the capacity of the classifier. It predicts the class labels accurately and accuracy refers to the correctness of how well the class label is predicted for the unknown data.

$$A = \frac{T\ +T}{\mathrm{T}\ +\mathrm{T}\ +\mathrm{F}\ +\mathrm{F}} \text{ ---------(4.1)}$$

Sensitivity:

It is the ratio of true positive which is accurately identified by the classifier. It is expressed as follows:

$$\text{Sensitivity} = \frac{T}{T\ +F} \text{------- (4.2)}$$

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

Specificity:

It is the ability of the classifier to identify negative results. It is expressed as follows:

$$\text{Specificity} = \frac{T}{T + F} \text{------- (4.3)}$$

Kappa:

It is the measure of how the instances closely match the instances which is classified by the class label. It compares the observed accuracy with the expected accuracy. Kappa statics can be used as a measure to compare different classifier. Observed and expected accuracy can be computed from confusion matrix. The kappa is given mathematically as follows:

$$\text{Kappa} = \frac{\text{T a} - \text{r a}}{1 - \text{r a}} \text{----- (4.4)}$$

Where the parameters total accuracy and random accuracy is given as,

$$\text{Total accuracy} = \frac{\text{T} + \text{T}}{\text{T} + \text{T} + \text{F} + \text{F}} \text{ --- (4.5)}$$

$$\text{Random accuracy} = \frac{(\text{TN} + \text{FP}) * (\text{TN} + \text{FN}) + (\text{FN} + \text{TP}) * (\text{FP} + \text{TP})}{\text{total accuracy} * \text{total accuracy}}$$

---- (4.6)

F-Score:

It is a measure which test the accuracy. It uses both precision and recall. It is the harmonic mean of both precision and recall. It can be expressed as follows:

$$F - S = 2 * \frac{P * R}{P + R} \text{ ---- (4.7)}$$

Ten fold cross validation is applied for the betterment of results.

D.DESCRIPTION OF DATASETS TAKEN

TABLE II DATASETS

S.NO	Dataset
1	PIMA Indian diabetes
2	Breast Cancer
3	Liver
4	Thyroid-ANN
5	Thyroid-ANN train
6	Lung Cancer
7	Cancer bladder
8	Cancer ovarian
9	Dermatology

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

E RESULTS OF DIFFERENT TECHNIQUES

TABLE II.ACCURACY RESULT FOR EACH DATASET(IN %)

DATASET	Naive Bayes	logistic regression	support vector	J48	Random Forest
PIMA Indian diabetes	75.39%	75.26	73.70	73.83	72.53
Breast Cancer	95.42	94.85	95.57	94.42	94.13
Liver	58.55	60.58	61.45	57.97	59.71
Thyroid-ANN	94.92	96.12	97.05	92.71	93.38
Thyroid-ANN train	94.88	95.41	94.42	93.95	93.49
Lung Cancer	64.23	62.77	64.96	70.80	70.80
Cancer bladder	92.65	14.41	10.59	80	94.41
Cancer ovarian	65.38	38.46	42.31	65.38	50
Dermatology	97.27	95.63	96.17	94.54	86.34

TABLE III. MEAN VALUES OF ACCURACY F-MEASURE AND KAPPA FOR VARIOUS DATSETS

Algorithm Implemented	Accuracy	F-Measure	Kappa	Time taken to build the model
Naive Bayes	69.45%	0.6943	0.3722	0 seconds
logistic regression	70.39%	0.704	0.4022	14.66 seconds
support vector	70.98%	0.7043	0.4057	1.671 seconds
J48	70.15%	0.6552	0.1947	0.0155 seconds
Random Forest	59.89%	0.5717	0.2377	0.0311 seconds

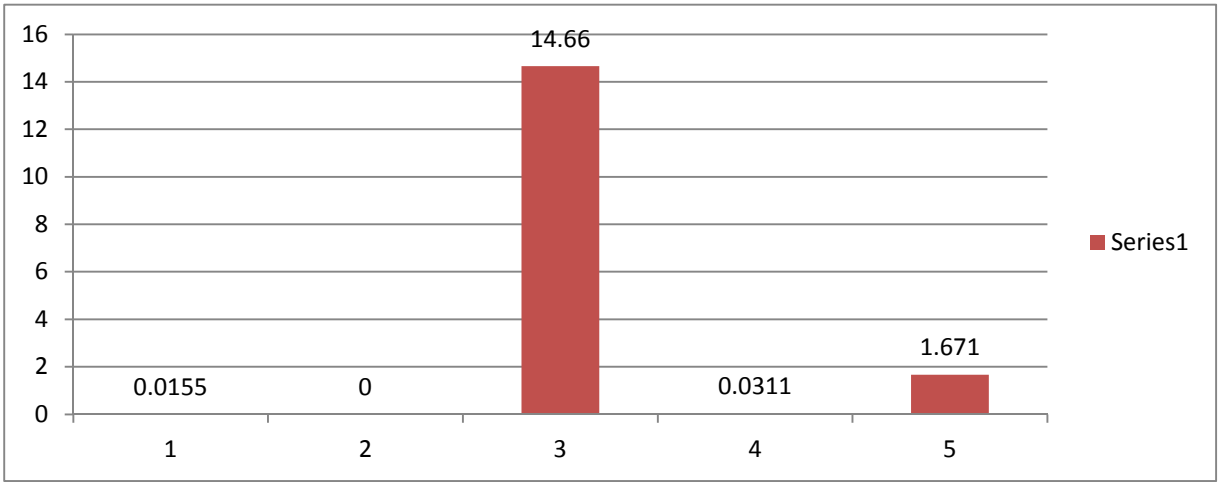

Figure2: Execution time taken for different algorithms

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 11
November 2017

where, series 1-Execution time

1- J48

2-Naive Bayes

3-Logistic

4-Random Forest

5-SVM.

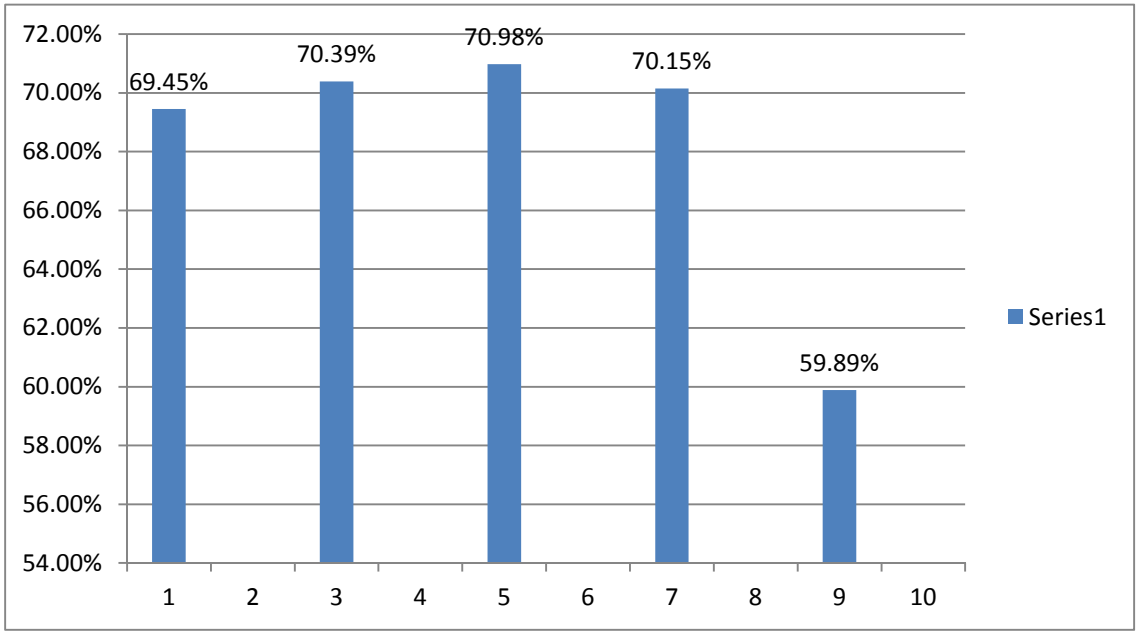

Figure3: Accuracy for different algorithms

Where, series1-accuracy

1- Naive Bayes

2-Logistic

3-SVM.

4-J48

5-Random Forest

V CONCLUSION AND FUTURE WORK

By analyzing the results from the above results for different data mining techniques for various data set it is observed that the naive bayes takes very less time to build the model but has third lower accuracy value than with other algorithm and in future the naive bayes algorithm will be improved so that it produces better accuracy than other classification techniques.

VI REFERENCES

[1] M. Gandhi and S. Narayan Singh, "Predictions in Heart Disease Using Techniques of Data Mining," 2015. [3] I. Yoo, P. Alafaireet, and M. Marinov, "Data Mining in Healthcare and Biomedicine : A Survey of the Literature," pp. 2431–2448, 2012.

[2] D. Tomar and S. Agarwal, "A survey on Data Mining approaches for Healthcare," vol. 5, no. 5, pp. 241–266, 2013.

[3] P. Songthung and K. Sripanidkulchai, "Improving Type 2 Diabetes Mellitus Risk Prediction Using Classification," 2016.

[4] H. D. Masethe and M. A. Masethe, "Prediction of Heart Disease using Classification Algorithms," vol. II, pp. 22–24, 2014.

[5] S. R. Safavian and D. Landgrebe, "A Survey of Decision Tree Classifier Methodology," *IEEE Trans. Syst. Man Cybern.*, vol. 21, no. 3, pp. 660–674, 1991.
[6] S. R. Safavian and D. Landgrebe, "A Survey of Decision Tree Classifier Methodology," *IEEE Trans. Syst. Man Cybern.*, vol. 21, no. 3, pp. 660–674, 1991.
[7] R. Duggal, S. Shukla, S. Chandra, B. Shukla, and S. K. Khatri, "Impact of selected pre-processing techniques on prediction of risk of early readmission for diabetic patients in India," *Int. J. Diabetes Dev. Ctries.*, vol. 36, no. 4, pp. 469–476, 2016.
[8] R. Duggal, S. Shukla, S. Chandra, B. Shukla, and S. K. Khatri, "Predictive risk modelling for early hospital readmission of patients with diabetes in India," *Int. J. Diabetes Dev. Ctries.*, vol. 36, no. 4, pp. 519–528, 2016.
[9] Dey,Monali,Siddarth Swarup "Study and analysis of data mining algorithms for healthcare decsision support system", *Int. J.computer science and information technology.*, vol. 5, 2014.
[10] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," *2008 IEEE/ACS Int. Conf. Comput. Syst. Appl.*, pp. 108–115, 2008.
[11] I. Yoo, P. Alafaireet, and M. Marinov, "Data Mining in Healthcare and Biomedicine : A Survey of the Literature," pp. 2431–2448, 2012.