ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 3, March 2017

# Comparative Analysis of Medical Mining Dataset for Figure out the Heart Illness to Perform Classification Algorithm

Jeyanth Sam A[*1], Prof. B. Veniston[*2]

Research Scholar, Department of Computer Science, St.Xavier's College, Tirunelveli, Tamilnadu, India[1]

Assistant Professor, Department of Computer Application, St.Xavier's College, Tirunelveli, Tamilnadu, India[2]

**ABSTRACT**: - Data mining is a prominent research area in computer science field. It is an evaluation forward movement of determining patterns in larger data. It is a leading role in several applications such as education, Healthcare, Market Basket Analysis, Manufacturing Engineering, CRM and Fraud Detection. Etc., Medical mining is one of the important Research-Based in data mining. This involved in multiple research and development disciplines. Especially Heart illness is a major reason for morality rate in the present living style. Enormous data mining techniques for predicting disease namely classification, clustering, association rules, summarization, regression and etc. Classification is one of the important data mining techniques for classifying given set of patient health data. In this experiment grouping of heart illness data set is done with the use of Cleveland & Hungary data sets. This work has done using WEKA an open source tool. The main objective of this research work is to predict the heart illness using the classification algorithm and concentrate on finding the best classification algorithm based on the accurate categorizing and performance time.

**KEYWORDS**: Medical mining, Classification, Knowledge Data Discovery, Heart illness, disease prediction.

## I. INTRODUCTION

Data mining is a predominant technology in the medical field. The various mining tools predict the future trends; discover the knowledge from large data. Data mining also called knowledge discovery in data base (KDD). Medical mining (MLM) is one of the important research area in computer field. MLM is an application of data mining technique on health care data. The objective of MLM is to analyze such data to resolve health care research issue. The diverse tasks of the predictive and descriptive models as classification, clustering, time series analysis and regression. MLM is probable domain for extracting the hidden patterns in the data set. Predictive Analytics (PAS) uses science and technology and statistical technique to find throughout vast amount of data, analyzing it to forecast outcomes for individual patients [11]. This research work explains classification algorithms and it also evaluate the good performance of accuracy and execution time[1]. The main objective of this work is to forecast heart illness using classification algorithms. The remaining part of the paper is providing as follows. Literature reviews are discussed in section II. Section III consults with proposed methodology. Section IV holds forth the conclusion. Section V References.

## II. RELATED WORK

JieWang [2]in order to predict heart illness outcome, the work explains difference between the support vector machine (SVM), Artificial Neural Network (ANN) and Decision Tree (DT) comparison; It has been based on the sensitivity and specificity and accuracy for patients health data. Finally the best prediction is that SVM classifier is more accurate than ANN, and the accuracy level is 92.1%.

Joythisingaruju[3] analyzed Artificial Neural Network algorithm, this algorithm have been used for Congenital heart disease diagnosis for classification process; this work used in developing the tool named MATLAB. From the experimental result they absorbed that ANN back propagation is better thanalgorithm and the accuracy level is 90%.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 3, March 2017

AnkurMakwana's [4] paper is used data preprocessing, data modeling, and mining method about the difference between many of measured attributes and patient survey. Genetic algorithm, decision tree algorithm are used in this paper for improve the accuracy level to predict the disease. The approach presented in their paper reduces the cost and effort of selecting patients for health studies.

## III. METHODOLOGY

### A. *Introduction*

To evaluate risk of heart illness using the Data Mining models. This paper process the following work which include Data collection from the Database, pre-processing the data, preparing, training and testing the separate models. The proposed work is implemented in WEKA 3.6. Preprocessing data set flow chart shown in figure1.

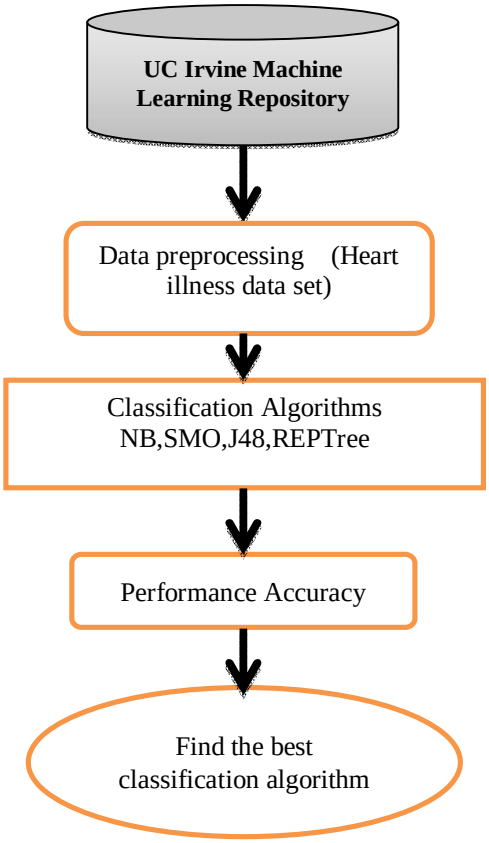


**Figure 1: Preprocessing Flow Chart ForHeart Illness Data Set**

### B. *Database*

Datasets collected from UC Irvine Machine Learning Repository, the data sets name Cleveland Clinic Foundation and Hungarian Institute of Cardiology. Each database has the same instance format. The datasets have 76 raw attributes, but only selected attributes are used; namely age, sex, chest pain location(painloc),chest pain type(cp), Resting bloodpressure(Trestbs),cholesterol(chol), and blood sugar fasting(fbs), smoke, cigarettes per day (cigs), number of years as a smoker(years), (painloc),chest pain type(cp), Resting blood pressure(Trestbs), serum cholesterol(chol), and blood sugar fasting(fbs), smoke, cigarettes per day (cigs), number of years as a smoker(years),

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 3, March 2017

history of diabetes(dm), family history of coronary artery disease (famhist), resting electro cartographic results(restecg), maximum heart rate achieved(thalach), Exercise induced angina(exang).

### *C. Pre-Processing Technique*

It is a data mining technique that involves reshaping the raw data into an intelligible format. Different types of preprocessing techniques used in this work like Data cleaning, Data Integration, data Reduction, Data Transformation[10].

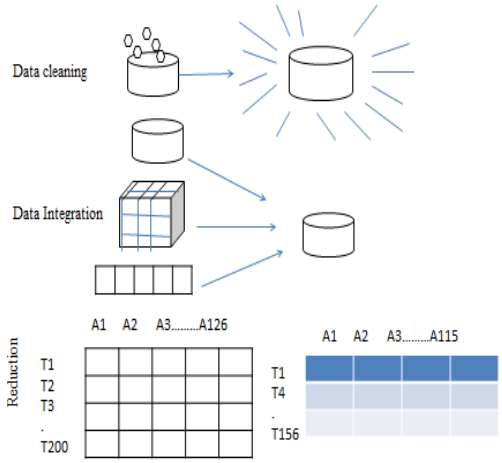


**Figure 2: Data preprocessing Technique**

Once we had details of all the heart illness patients' data sets to clean the data to fill the missing value and identify or remove outliers, and then integrate multiple database or data sets. To reduce the representation, finally transform the data understandable format. Some irrelevant attributes (patient id, social security no) are removed for predicting the future of patientheart disease risk. Totally 550 and above data's are used. Finally the "risk" attribute was added and it held the predicted results, which can be either (1, 0) 1represent risk for heart disease, 0 represent no risk for heart disease. Above all pre-processing techniques are performed WEKA tool.

### *D. Classification algorithms*

Classification involves supervised learning but in some cases unsupervised learning method also used. Different types of characteristics used in various types of classification. This may contain binary values('0' or '1'; 'male' or 'female'), Categorical data ('white', 'black', 'orange'), Ordinal('large', 'medium', 'small') integer values or real values are numbers[6]. Classification is a process of categorizing the data according to various instances. The major classification algorithms are Decision Tree, Navie Bayes method, Support Vector Machine, AdaBoost, CART, KNN, Genetic Algorithm, C4.5. [6]In this research work NB, SMO, J48, REPTree are used to classify different stages of Heart illness. The following evaluation criteria's are used for classification methods to improve accuracy, like speed, Robustness, Scalability, Interpretability, Goodness of the model, Time complexity and Flexibility [7].

Bootstrapping, Bagging and Boosting are techniques for improving accuracy of classification results. Estimating the accuracy of a supervised classification method can be difficult if only the training data is available.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 3, March 2017

### *E. Result and Discussion:*

**Table1: Performance comparison between different algorithms**

| S.No | Algorithm | Accuracy |
|---|---|---|
| 1 | Navie Bayes | 79.55% |
| 2 | SVM | 83.05% |
| 3 | J48 | 77.89% |
| 4 | REPTree | 75.00% |

Table1: above all algorithm used for find the best algorithm to heart disease prediction. The performance of data mining algorithm classification accuracy results contains true positive values and true negative values.Above all the algorithmsareimplemented by the Waikato Environment for Knowledge Analysis tool for predicting the heart illness risk.

### *F. Classification Accuracy:[9]*

The classification performance calculated by the following formula:

Accuracy $=\frac{TP+TN}{TP+FP+TN+FN}$

TP-True Positive, FB-False Positive,

TN-True Negative, FN-False Negative,

TPR$=\frac{TP}{TP+FN}$

High true–positive rate

Precision$=\frac{TP}{TP+FB}$

F-Measure$=\frac{2*Recall*precision}{Recall+precision}$

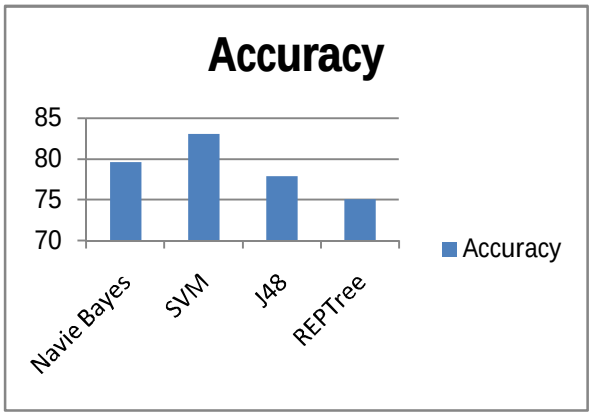


**Figure 3: Accuracy chart on Heart illness Data set**

This is detail study of data mining algorithms which are used to find the classification accuracy. The data mining algorithms are developed for different purposes. Table 1 shows the comparison of different algorithms on the base of accuracy and error. The classification accuracy is described as "percentage of exactprediction.The SVM algorithm shows the highest accuracy in heart illness classification.

## IV. CONCLUSION & FUTURE WORK

This research work is comparing the different type of classification algorithm, expeditiously predicting the heart illness from medical records of patients. The data set has been used for experimental purpose. The data mining tool WEKA has been used for predicting the data set. Thus we found that SVM is the better algorithm (accuracy -83.05%) in most of the cases. In future we planned to add the Neural Network algorithm to improve the classification accuracy and reduced the error rate.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 3, March 2017

## REFERENCES

[1] Approaches, Knowledge Oriented Applications in Data Mining,Prof. KimitoFunatsu (Ed.), ISBN:9789533071541, InTech,http://www.intechopen.com/books/
Knowledge Oriented Applications In data mining/mining Enrollment Data Using Descriptive And Predictive approaches
[2]yanwei Xing, jiewang, ZhihongzhaoYonghongGao" combination data mining methods with medical data to predicting outcome of coronary heart disease", convergence Information Technology, international conference, Nov 2007, pp868-872.
[3] vanishree K, JyothiSingaraju, "Decision Support System for Congenital Heart Disease diagnosis based on signs and symptoms using Neural Networks" International Jouranl of Computersicence applications, April 2011 Vol 19 no6.
[4] AnkurMakwana, Jaymin Patel," decision support system for heart disease prediction using data mining classification techniques" International Jouranl of Computersicence applications, may 2015 vol 117.
[5]https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names
[6] Dr. M. Safish Mary, Dr. V. Joseph Raj" A Fast Neural Classifier For Sales Optimzization",
[8]Data mining book G.K Gupta "evolution of Classification Methods"
[9]Dr.S.Vijarani, S. Dhayanand" Data Mining Algorithms For Kidney Disease Prediction" International journal on Cybernetics & Informatics (IJCI) Vol 4, August 2015.
[10]V.Boonjing,A.Kemphilam " heart disease classification and feature selection", 21st International Conference on System Engineering,2011.